



MÉTODOS ESTADÍSTICOS DESCRIPTIVOS Y DE APRENDIZAJE AUTOMÁTICO PARA LAS FINANZAS

Ciencia de datos y pronósticos

Autores

- Rolando Wilber Ordoñez Neyra
- Beto Puma Huamán
- Edgardo Martin Figueroa Donayre
- Rogger Humpiri Flores
- Hector Ito Mamani
- Cesar Eusebio Pacori Mamani

ISBN: 978-9915-9682-2-3



EST. 2021 **EMC**
EDITORIAL MAR CARIBE

ISBN: 978-9915-9682-2-3

Editorial Mar Caribe

**Métodos estadísticos descriptivos y de aprendizaje
automático para las finanzas**

2024

Métodos estadísticos descriptivos y de aprendizaje automático para las finanzas

Rolando Wilber Ordoñez Neyra, Beto Puma Huamán, Edgardo Martin Figueroa Donayre, Rogger Humpiri Flores, Hector Ito Mamani, Cesar Eusebio Pacori Mamani

© Rolando Wilber Ordoñez Neyra, Beto Puma Huamán, Edgardo Martin Figueroa Donayre. Rogger Humpiri Flores, Hector Ito Mamani, Cesar Eusebio Pacori Mamani

Primera edición: Julio, 2024

Editado por:

Editorial Mar Caribe

www.editorialmarcaribe.es

Av. General Flores 547, Colonia, Colonia-Uruguay.

RUC: 15605646601

Diseño de cubierta: Yelitza Sánchez Cáceres

Libro electrónico disponible en https://editorialmarcaribe.es/?page_id=805

Formato: electrónico

ISBN: 978-9915-9682-2-3

Hecho el Depósito Legal en la Biblioteca Nacional de Uruguay N°.: 385.374

Aviso de derechos de atribución no comercial: Los autores pueden autorizar al público en general a reutilizar sus obras únicamente con fines no lucrativos, los lectores pueden usar una obra para generar otra obra, siempre y cuando se dé el crédito de investigación y, otorgan a la editorial el derecho de publicar primero su ensayo bajo los términos de la licencia [CC BY-NC 4.0](https://creativecommons.org/licenses/by-nc/4.0/).

Índice

Introducción.....	7
Capítulo 1.....	9
El <i>Machine Learning</i> en las finanzas.....	9
La percepción socioeconómica	13
Aplicaciones en el sistema financiero	17
Las ventajas ML en el sistema financiero.....	28
Los desafíos de la ML.....	34
La competencia	36
El accionar	37
Legislación y ética.....	38
La humanidad.....	40
Capítulo 2.....	41
Análisis descriptivo en el aprendizaje automático.....	41
Los procesos en las finanzas	43
Los datos.....	47
El análisis exploratorio de los datos.....	50
Tipos de Datos	51
Los datos categóricos: cualitativos	53
Los datos numéricos.....	54
Los otros tipos de datos	56
El procesamiento de datos en la computadora	58
La estrategia en el análisis de datos con el empleo de programas de computación	61
Los malos hábitos en el empleo de la computadora.....	64
Gráficos en la estadística descriptiva	65
Gráfico de torta	66
El histograma	67
El gráfico de caja.....	68
Los gráficos de gusanos	69
Los gráficos de dispersión	69
Capítulo 3	73
El procesamiento de datos	73
Los valores duplicados.....	74
Los valores atípicos	75
Los valores faltantes	76
La estandarización.....	77

La reducción de los datos.....	78
El aprendizaje no supervisado	80
El K-means	82
El DBSCAN	85
Los otros modelos de aprendizaje no supervisado.....	88
Los coeficientes de evaluación.....	88
El aprendizaje supervisado	91
Entrenamiento y pruebas.....	92
La regresión lineal	93
La regresión logística.....	94
El árbol de decisión	95
La matriz de confusión.....	95
Capítulo 4.....	98
Redes neuronales	98
Características	98
Su estructura	99
La computación neuronal	102
Historia	104
La actualidad.....	106
ANN: aplicaciones	107
Clasificación	111
La red neuronal multicapa.....	114
Procesamiento: las entradas y las salidas	115
Las redes neuronales y los sistemas adaptativos	116
El perceptron.....	117
Los filtros adaptativos de redes neuronales	119
Las funciones adaptativas	119
El reconocimiento estadístico de patrones: redes neuronales.....	120
El reconocimiento de patrones	120
Descripción del reconocimiento	120
La regresión.....	122
Los datos de entrenamiento: validación y de testeo.....	123
El reconocimiento estadístico de los patrones.....	124
Las técnicas de reconocimiento de patrones	125
El aprendizaje y la generalización.....	127
La evaluación de la generalización	128
El entrenamiento y la generalización.....	129
La redes neuronales con mapas autoorganizados	130

Los fundamentos biológicos	134
La arquitectura	135
El algoritmo.....	136
La etapa del funcionamiento	136
La etapa de aprendizaje	137
Las fases.....	141
Las redes neuronales dinámicas.....	147
Su estructura	148
La innovación financiera	148
Inclusión	149
La innovación digital en el sector económico.....	150
Conclusiones	154
Bibliografía.....	157

Introducción

El aprendizaje automático cae bajo el paraguas de la ciencia de datos e implica la utilización de modelos estadísticos para extraer información y hacer pronósticos, y presenta la ventaja de adquirir conocimientos a través de la experiencia en lugar de depender de una programación explícita. Su función consiste en seleccionar modelos adecuados y proporcionarles datos, lo que en última instancia lleva a que el modelo ajuste sus parámetros de forma autónoma para mejorar su rendimiento.

Los especialistas en análisis de datos se dedican a la formación de modelos de aprendizaje automático utilizando conjuntos de datos existentes. Posteriormente utilizan estos modelos en escenarios prácticos. El modelo se inicia como un proceso que se ejecuta en segundo plano y tiene la capacidad de generar resultados automáticamente según su configuración. Dependiendo de los requisitos específicos de una empresa, los modelos se pueden entrenar periódicamente para garantizar que se mantengan actualizados. Algunas empresas pueden incluso actualizar sus modelos diariamente, aunque esta frecuencia puede variar en función del volumen de datos recopilados.

En el ámbito del aprendizaje automático, se entiende ampliamente que la precisión de los resultados es directamente proporcional al volumen de datos incorporados al modelo. Afortunadamente, el sector financiero cuenta con una amplia gama de datos que abarcan multitud de facetas, incluidos varios tipos de transacciones, información del cliente, detalles de facturas y mucho más. En consecuencia, es inequívoco que los datos desempeñan un papel insustituible en el ámbito del aprendizaje automático dentro de la industria financiera.

El rápido ritmo de los avances tecnológicos evoluciona constantemente y, al mismo tiempo, la cantidad de información que se genera crece exponencialmente. Estos factores combinados indican que, en el futuro cercano, las aplicaciones potenciales del aprendizaje automático en el ámbito de los servicios financieros serán cada vez más imperceptibles y aparentemente inalcanzables. Sin embargo, la mayoría de las instituciones financieras aún no están preparadas para aprovechar plenamente el inmenso potencial de esta tecnología. ¿Cuáles son las razones detrás de esta falta de preparación?

- Muchas empresas no comprenden del todo las verdaderas ventajas que puede aportar el aprendizaje automático.
- La investigación y el desarrollo de tecnología de aprendizaje automático a menudo pueden generar costes elevados.
- Actualmente, hay escasez de personas capacitadas en los campos del aprendizaje automático y la ingeniería de inteligencia artificial.
- Los administradores de instituciones financieras suelen ser personas con aversión al riesgo y cautelosas a la hora de tomar decisiones que pueden tener posibles consecuencias negativas. Además, tienden a ser lentos a la hora de implementar cambios y actualizar su infraestructura de datos, y a menudo toman una cantidad de tiempo considerable antes de adoptar nuevas tecnologías o sistemas.

Un número limitado de empresas ha adoptado la adopción de técnicas de aprendizaje automático en sus operaciones. Si bien, las empresas que han implementado con éxito esta tecnología han experimentado innumerables ventajas notables. Una de esas ventajas es la reducción significativa de los costes operativos, atribuida principalmente a la automatización de diversos procesos. Asimismo, estas empresas han observado un aumento sustancial en la generación de ingresos, que puede atribuirse a mayores niveles de productividad y una mejor experiencia de usuario. Además, la implementación del aprendizaje automático también ha reforzado las medidas de seguridad empleadas por estas empresas, fortaleciendo así sus defensas contra posibles amenazas cibernéticas.

Capítulo 1

El Machine Learning en las finanzas

Nos encontramos en una era intrigante de la tecnología, donde los constantes avances y descubrimientos descubren continuamente el inmenso potencial de cada innovación. Todo comenzó en agosto de 1981, cuando IBM presentó la primera computadora personal del mundo, inicialmente diseñada para mejorar los lanzamientos balísticos, pero que finalmente condujo al desarrollo de muchas otras aplicaciones que han transformado profundamente nuestra sociedad. A medida que se desarrollaba la tercera revolución industrial, la llegada de Internet impulsó la importancia de los teléfonos móviles como dispositivos personales para gestionar información y acceder a una gran cantidad de servicios, incluidos los financieros. Sin embargo, el futuro presenta perspectivas aún más prometedoras con la aparición de la computación cuántica. En un logro innovador anunciado por Google en octubre de 2019, realizaron con éxito un cálculo de números aleatorios en apenas tres minutos y veinte segundos, una tarea que a las computadoras tradicionales más potentes de la actualidad les habría llevado miles de años completar. Este notable progreso significa que el ámbito de la computación cuántica está avanzando rápidamente y está preparado para revolucionar el panorama tecnológico en formas que apenas podemos comenzar a comprender.

El uso generalizado de dispositivos móviles y redes sociales ha aumentado enormemente las capacidades de procesamiento y acceso a información personalizada. Esto nos ha permitido convertir datos dispersos en información valiosa, lo que nos permite identificar necesidades, preferencias y hábitos de compra individuales en tiempo real. Como resultado, el marketing directo ha mejorado enormemente. En palabras de John Naisbitt, autor de *Megatrends*, ahora tenemos una economía que depende exclusivamente de un recurso renovable y autogenerado: la información. El desafío no radica en quedarse sin información, sino en encontrar formas de navegar a través de su abrumadora cantidad. Por lo tanto, es crucial explorar más a fondo el profundo impacto de estas tecnologías que avanzan rápidamente.

El campo del análisis empresarial utiliza una metodología conocida como Machine Learning, que es un subconjunto de la Inteligencia Artificial. Esta metodología implica analizar datos e información para comprender y describir eventos pasados, determinar

las razones detrás de ellos, hacer predicciones sobre eventos futuros y proponer estrategias para lograr los resultados deseados. Al emplear técnicas de aprendizaje automático, las empresas y sus clientes pueden extraer información valiosa de las observaciones y utilizarla para tomar decisiones informadas.

El Machine Learning es un campo fascinante dentro de la inteligencia artificial que permite a las máquinas aprender y mejorar su rendimiento a través de algoritmos. Sus diversas formas de aprendizaje, como el supervisado, el no supervisado, el refuerzo y el aprendizaje profundo, permiten a las máquinas hacer predicciones, descubrir patrones y tomar decisiones informadas basadas en datos. A medida que el aprendizaje automático continúa avanzando, tiene un inmenso potencial para revolucionar numerosas industrias e impulsar la innovación en el futuro.

Así, el aprendizaje profundo consiste en una red de algoritmos que funcionan en paralelo y cada uno de los cuales contribuye al procesamiento de los datos. Como un embudo, los datos pasan por múltiples capas de algoritmos, reduciendo progresivamente la cantidad de información que se procesa. Este enfoque de múltiples capas mejora las capacidades de los sistemas de aprendizaje automático y permite un análisis y una toma de decisiones más complejos. Machine Learning (ML) es un subconjunto de la inteligencia artificial (IA) que implica el uso de algoritmos para mejorar el rendimiento de las máquinas y permitirles aprender a partir de experiencias o datos de muestra.

En términos más simples, ML permite que las máquinas aprendan sin programación explícita. Cuando estas técnicas de ML se aplican a grandes bases de datos, se denomina "minería de datos", haciendo una analogía con la extracción de materiales preciosos de las minas. Así como se obtiene una pequeña cantidad de material valioso de las minas, de la gran cantidad de datos que se analizan se extrae una pequeña cantidad de información significativa. Por otro lado, el aprendizaje no supervisado opera únicamente con datos de entrada sin ninguna variable predeterminada que predecir. Su objetivo es descubrir relaciones y similitudes dentro de los datos.

A diferencia del aprendizaje supervisado, no requiere conocimiento previo del proceso y, en cambio, se centra en agrupar puntos de datos similares e identificar patrones frecuentes como estándares de clasificación. La eficacia del aprendizaje no supervisado depende de las similitudes inherentes presentes en los datos. ML abarca

varios tipos de aprendizaje, incluido el aprendizaje supervisado, el aprendizaje no supervisado, el aprendizaje por refuerzo y el aprendizaje profundo. En el aprendizaje supervisado, el algoritmo se entrena utilizando datos de entrada y salida para replicar un proceso específico. Al comprender la relación entre entradas y salidas, el algoritmo puede predecir comportamientos futuros o tomar decisiones informadas basadas en nuevos datos.

La precisión del aprendizaje supervisado depende en gran medida de la calidad y cantidad de los datos disponibles. Los problemas de regresión y clasificación se distinguen según la naturaleza del resultado. Si el resultado es un valor numérico, se considera regresión, mientras que si implica categorizar patrones, se considera un problema de clasificación. El aprendizaje por refuerzo implica un enfoque de prueba y error, donde el programa toma decisiones y recibe recompensas o castigos en función de la corrección de sus acciones. A través de intentos repetidos y retroalimentación, el programa desarrolla un algoritmo que determina la política óptima a seguir en una situación determinada.

La inteligencia artificial, es una concepción algo difícil de explicar, fue presentada por primera vez por John McCarthy en una conferencia en Dartmouth College en 1956. McCarthy la definió como la ciencia y la tecnología detrás de la creación de máquinas inteligentes. Una definición más contemporánea, proporcionada por la Comisión Europea, caracteriza la inteligencia artificial como sistemas capaces de exhibir un comportamiento inteligente analizando su entorno y tomando acciones hacia objetivos específicos, con un cierto nivel de autonomía. El propósito de la inteligencia artificial es alcanzar una inteligencia a nivel humano.

Para comprender plenamente el concepto de inteligencia a nivel humano es necesario diferenciar entre inteligencia artificial general y específica. La inteligencia específica se refiere a la capacidad de realizar una función particular, mientras que la inteligencia general abarca capacidades cognitivas generales, similares a la inteligencia humana. McCarthy enfatizó la importancia del sentido común a la hora de replicar la inteligencia humana. Estudió cómo los individuos utilizan su conocimiento y su información para determinar el mejor curso de acción. Según McCarthy, el aprendizaje y el sentido común están estrechamente entrelazados. Imaginó programas que pudieran aprender de la experiencia con tanta eficiencia como lo hacen los humanos, con el objetivo

de mejorar el rendimiento de las máquinas mediante asesoramiento en lugar de reprogramación (Francés Monedero, 2020).

El enfoque de McCarthy, conocido como "el que toma consejos", implicó la construcción de un programa que pudiera razonar y deducir comportamientos apropiados. Concluyó que un programa posee sentido común si puede deducir de forma autónoma una amplia gama de consecuencias basándose en la información que se le proporciona y en el conocimiento existente. En 1960, él propuso un problema que destacaba la necesidad de que las máquinas pudieran aprender y comprender instrucciones. Sugirió que todos los aspectos del aprendizaje y la inteligencia podrían describirse con precisión para que una máquina pudiera simularlos. Sin embargo, a pesar de la existencia de inteligencia artificial específica, aún no se ha logrado una verdadera inteligencia artificial general. Esta idea fue cuestionada por A. M. Turing en 1950 cuando intentó determinar si las máquinas podían pensar. Turing concluyó que las definiciones de "pensar" y "máquina" eran insuficientes para responder la pregunta e introdujo la prueba del "juego de imitación" como alternativa.

La prueba tenía como objetivo determinar si una máquina podría comportarse como una persona pensante, sin centrarse en su apariencia física. Turing creía que una máquina pensante no tenía por qué parecerse a un ser humano. En el juego participaron tres participantes: una persona, una máquina y un interrogador, y el interrogador intentaba determinar cuál era la máquina en función de sus respuestas a un número ilimitado de preguntas. Al interrogador no se le permitió pedir demostraciones prácticas. Si el interrogador no pudiera distinguir entre la máquina y la persona, entonces podría decirse que la máquina piensa. Sin embargo, en la década de 1980, el filósofo John Searle criticó esta teoría presentando un escenario hipotético en el que la conversación era en chino y la persona que participaba no hablaba chino. Pese a ello, la persona tenía un libro que contenía la programación informática utilizada por la máquina para responder en chino. Searle argumentó que la persona podía simular el mismo programa que la máquina y mantener una conversación, aunque no entendiera el idioma. Usó este ejemplo para sugerir que pasar la prueba de Turing por sí sola no es suficiente para demostrar una verdadera inteligencia, ya que el comportamiento externo de una persona sería indistinguible del de una máquina.

El argumento conocido como la "Sala China" sugiere que aunque alguien pueda pasar la prueba de Turing respondiendo apropiadamente a las preguntas en chino, eso no significa necesariamente que comprenda el idioma. Este argumento plantea la cuestión de si memorizar un libro que contiene todas las respuestas posibles conduciría a una verdadera comprensión del chino. Sin embargo, Levesque sostiene que sería imposible crear un libro así debido a la gran cantidad de combinaciones que deberían incluirse. En cambio, sugiere que se podría crear un libro que explique el proceso de suma, permitiendo a la persona aprender y comprender cómo sumar. Según Levesque, esto haría válido el "Juego de la Imitación", y alcanzar una verdadera inteligencia artificial sería posible una vez superado el juego.

La percepción socioeconómica

A lo largo de la historia, la introducción de tecnologías innovadoras, como Internet, ha dado lugar a importantes cambios socioeconómicos, lo que ha provocado un sentimiento de aprensión entre las personas. De manera similar, cuando se trata de visualizar el futuro de la inteligencia artificial (IA), hay una notable falta de consenso a medida que surgen diversas perspectivas.

Existe un movimiento contemporáneo conocido como neoludismo que comparte similitudes con el movimiento ludita del siglo XIX, los cuales se oponen al progreso tecnológico. El ludismo original surgió en Gran Bretaña durante la revolución industrial y se caracterizó por un rechazo violento a la maquinaria debido a la preocupación por la pérdida de empleo. Sin embargo, el neoludismo adopta un enfoque más pasivo, rechazando el impacto positivo de los avances tecnológicos y considerándolos perjudiciales para la humanidad, la naturaleza y la sociedad en su conjunto. Ray Kurzweil, el inventor de los programas de reconocimiento óptico de caracteres (OCR), predice que este movimiento cobrará impulso a medida que la inteligencia artificial (IA) se integre más en la vida diaria. Además, Kurzweil sostiene que la evolución de la humanidad está estrechamente relacionada con el desarrollo de la tecnología (Kurzweil, 1999).

Así, según una exhaustiva investigación realizada por Timo Gnambs y Markus Appel en 2010, se reveló que un asombroso 72% de las personas encuestadas expresaron su temor de ser reemplazados por robots en sus respectivos trabajos. Esta estadística

convinciente resalta la preocupación generalizada entre la población sobre la amenaza potencial de la automatización. En este sentido, es interesante observar que sólo el 57% de los encuestados estaban abiertos a la idea de trabajar junto a un asistente robótico. Esta disparidad en las tasas de aceptación subraya aún más la compleja relación entre los humanos y la tecnología en el lugar de trabajo moderno.

Por el contrario, la Comisión Europea apoya firmemente la idea de que la inteligencia artificial (IA) no es sólo una tecnología importante de nuestro tiempo, sino más bien la más crucial y estratégica. Enfatiza que hay mucho en juego cuando se trata de IA y que nuestro enfoque hacia esta tecnología, en última instancia, dará forma al mundo en el que vivimos. De hecho, la Comisión Europea afirma que el impacto de la IA en la sociedad y en diversas industrias será tan transformador que puede compararse con el efecto revolucionario que tuvo la electricidad en numerosos sectores hace un siglo. Así como la electricidad revolucionó las industrias en aquel entonces, la IA ahora está preparada para provocar cambios profundos en las industrias a gran escala. Este sentimiento lo comparte Andrew Ng, cofundador de Coursera, quien compara la IA con la llegada de la electricidad y destaca su potencial para revolucionar y remodelar varios sectores de manera similar. Ambas perspectivas comparten el argumento común de que el impacto de la IA en nuestro mundo es inevitable, independientemente de si será ventajoso o perjudicial.

En el panorama global actual, ha habido un aumento notable en los avances tecnológicos. Esto puede atribuirse al auge de numerosas empresas de tecnología tanto en Asia como en Estados Unidos. Estas regiones han establecido la infraestructura necesaria y poseen una gran cantidad de datos, lo que contribuye al crecimiento de la industria tecnológica. Además, un examen de los cambios ocupacionales recientes en los EE. UU. revela un aumento significativo en los campos STEM, que abarcan ciencia, tecnología, ingeniería y matemáticas. Cabe señalar la aparición de diversas tecnologías nuevas, como la computación en la nube, que permite la prestación de servicios a través de Internet. Otros avances destacables incluyen el procesamiento de cantidades masivas de datos (Big Data) y la implementación de la robótica. Todas estas innovaciones desempeñan un papel fundamental a la hora de impulsar la transformación digital de las organizaciones (Banco de España, 2020).

La IA tiene el potencial de provocar la transformación tecnológica más rápida y significativa de la historia. La región de Asia y el Pacífico está liderando el camino en la adopción de la IA, y una de cada cinco empresas implementa esta tecnología. América del Norte le sigue de cerca, con una de cada diez empresas que utiliza IA. En Europa, ha habido un aumento notable en el espíritu emprendedor que rodea a las empresas relacionadas con la IA. En 2019, una de cada doce nuevas empresas centró su propuesta de valor central en la IA, en comparación con solo una de cada cincuenta empresas en 2013 (Francés Monedero, 2020).

Actualmente, Europa alberga aproximadamente 1.600 empresas de IA, con el Reino Unido a la cabeza, como el país con mayor número de estas empresas, representando un tercio del total en Europa. Francia y Alemania le siguen de cerca, mientras que España sorprendentemente ocupa la cuarta posición, superando su peso en términos de contribución. La creciente prevalencia de la IA entre los emprendedores de hoy sirve como una clara indicación de que la IA está a punto de convertirse en una fuerza omnipresente en nuestro futuro cercano (Francés Monedero, 2020).

Por lo tanto, la inteligencia artificial se ha convertido en una tendencia destacada en el mundo actual, experimentando un crecimiento rápido y sin precedentes. En consecuencia, la investigación sobre su influencia se ha convertido en un área de inmensa fascinación. En la discusión posterior, profundizaremos en el ámbito del sector financiero para explorar el impacto de la inteligencia artificial.

El sector financiero está repleto de una gran cantidad de datos, lo que allana el camino para la integración del aprendizaje automático. Sorprendentemente, la mayoría de los bancos siguen ajenos al inmenso potencial que se esconde en más del 80% de los datos que han acumulado, como revela un informe de United Consulting Group en 2018. Sin embargo, con la llegada de la digitalización y la proliferación del multicanal servicios, los datos se están volviendo cada vez más frecuentes en la industria. Esta nueva abundancia de datos presenta una oportunidad de oro para que las instituciones financieras se especialicen y se hagan un hueco ofreciendo valor añadido a sus clientes. Al aprovechar las herramientas analíticas, estas instituciones pueden anticipar con precisión las necesidades únicas de sus clientes, manteniéndose así un paso por delante en el panorama financiero en constante evolución.

Asimismo, la industria financiera ha sido reconocida como uno de los sectores líderes en el Índice de Digitalización de la Industria de MGI, lo que indica su importante adopción e inversión en inteligencia artificial. En consecuencia, no sorprende que la convergencia de las finanzas y la tecnología, conocida como Fintech, esté remodelando el panorama del sector financiero. Esta integración abarca la utilización de avances digitales y modelos de negocio inventivos habilitados por la tecnología dentro del ámbito financiero.

La inteligencia artificial, el aprendizaje automático y los Smart Data son los impulsores clave de la revolución tecnológica en este sector en particular. Entre ellas, el aprendizaje automático destaca como la disciplina de inteligencia artificial más utilizada dentro de la industria financiera (Fernández, 2019). Por lo tanto, este estudio enfatiza las aplicaciones del aprendizaje automático al tiempo que proporciona una descripción general de su proceso de implementación.

Para incorporar eficazmente el aprendizaje automático (ML) en los procesos de negocio, es crucial seguir una metodología de trabajo sistemática y desarrollar un proyecto bien definido que tenga como objetivo crear un modelo capaz de cumplir los objetivos predeterminados. Este proceso implica varias etapas, como la recopilación de datos, el preprocesamiento, la capacitación del modelo y la evaluación, que deben ejecutarse cuidadosamente para garantizar la implementación exitosa del ML en el entorno empresarial. Al adherirse a esta metodología, las empresas pueden aprovechar el poder del ML para optimizar las operaciones, mejorar la toma de decisiones, mejorar las experiencias de los clientes y, en última instancia, impulsar el crecimiento y la rentabilidad.

Para brindar una comprensión integral de las diversas etapas involucradas en el procesamiento de datos con fines comerciales, discutiremos la metodología CRISP-DM. Este modelo analítico ampliamente utilizado, conocido como Proceso Estándar Intersectorial para Minería de Datos, se originó a fines de la década de 1990 e incorpora dos elementos cruciales: una estrategia de calidad total, también conocida como mejora continua, y el concepto de tratar un proyecto como un proceso paso a paso que consta de múltiples fases. El siguiente diagrama ilustra las distintas fases de la metodología CRISP-DM.

La etapa inicial implica obtener una comprensión integral del negocio, que sirve como base para cualquier aplicación del aprendizaje automático a los datos comerciales. Esta etapa es crucial ya que nos permite evaluar la situación actual, establecer objetivos a nivel de minería de datos y desarrollar un plan de proyecto que describa claramente los resultados deseados de todo el proceso. Pasando a la segunda fase, nos centramos en los procesos de captura de datos, determinando las fuentes de las que se extraerán los datos. Esto implica buscar fuentes confiables y garantizar la calidad de los datos. Una vez identificados, decidimos un método de extracción que minimice la corrupción y cumpla con los requisitos de seguridad. Esta fase también implica la comprensión de los datos, donde realizamos tareas de exploración y gestión de calidad para identificar posibles problemas y ofrecer soluciones. A continuación, preparamos los datos estableciendo el universo de datos con el que se trabajará y realizando las tareas de limpieza necesarias. En esta fase, empleamos técnicas de preparación de datos para adaptar el conjunto de datos para su uso con varios algoritmos. La fase de modelado implica seleccionar las técnicas de aprendizaje automático más adecuadas y que produzcan los mejores resultados para nuestro conjunto de datos, teniendo en cuenta el tipo de datos y nuestros objetivos.

Adicionalmente, establecemos una estrategia para verificar la calidad del modelo. Vale la pena señalar que ningún método o algoritmo domina a los demás, ya que depende del conjunto de datos específico que se analiza. Por lo tanto, es crucial evaluar el modelo. Finalmente, diseñamos un plan de despliegue de producción de los modelos y comunicamos el conocimiento adquirido a partir de ellos dentro de nuestra organización. También llevamos a cabo una revisión integral del proyecto en su conjunto para identificar las lecciones aprendidas. Este modelo en particular evalúa sus resultados basándose en la prosperidad de la empresa en lugar de basarse en puntos de referencia estadísticos.

Aplicaciones en el sistema financiero

Las técnicas no supervisadas desempeñan un papel crucial en el análisis de grandes volúmenes de datos, particularmente en el ámbito financiero, donde se emplean para identificar posibles casos de fraude. Dado que las tarjetas de crédito se están volviendo cada vez más populares como medio de pago, la ocurrencia de actividades

fraudulentas ha mostrado una tendencia a aumentar. Esto ha llevado a las instituciones financieras a recurrir a metodologías inteligentes, ya que los enfoques convencionales de detección manual de fraude demostraron ser lentos en su capacidad para mantenerse al día con la creciente tasa de fraude.

Los bancos emplean un servicio de seguridad que se activa cuando los patrones de gasto de un cliente se desvían de sus gastos habituales. Para identificar transacciones potencialmente sospechosas (conocidas como valores atípicos), se utilizan métodos no supervisados para comparar cada transacción con las anteriores. Estos métodos implican analizar diversos factores como la ubicación del cliente, sus preferencias, sus hábitos de compra y su comportamiento típico.

Los métodos supervisados se emplean en los casos en que se construye un modelo utilizando una colección de transacciones fraudulentas y legítimas para categorizar efectivamente nuevas transacciones como fraudulentas o legítimas. En un estudio realizado por Bhattacharyya, Jha, Tharakunnel y Westland (2011), se descubrió que los bosques aleatorios, un tipo de algoritmo supervisado, superaban a otros métodos en términos de diversos criterios de rendimiento. Para identificar patrones e identificar transacciones anómalas, se utilizan datos históricos y el comportamiento del cliente.

La implementación del aprendizaje automático plantea numerosos desafíos debido a diversos obstáculos, como manejar grandes cantidades de datos, navegar a través de distribuciones cambiantes y sesgadas, manejar la variabilidad de los costos de error a lo largo del tiempo (incluidos falsos positivos y falsos negativos) y adaptarse a cambios en constante evolución. comportamientos y cuestiones sociales.

Por otro lado, la utilización de la herramienta SNA también puede ayudar en la detección de fraude. El análisis de redes sociales (SNA) es una técnica que se esfuerza por identificar y comprender las conexiones entre individuos o entidades dentro de grupos sociales. Al estudiar el impacto de las interconexiones dentro de una red, el SNA pretende obtener información sobre diversos fenómenos sociales. Una red social comprende actores interconectados, que pueden ser individuos o entidades, vinculados por lazos familiares, parentesco, membresía compartida en un grupo u organización. Esta técnica se basa en la teoría de grafos y emplea mediciones matemáticas para describir la estructura de la red y el significado de la disposición de los nodos.

En este contexto, los nodos representan a los actores dentro de la red social y las relaciones entre los participantes se representan como vínculos que conectan estos nodos. En lugar de tratar a los nodos como entidades aisladas, el SNA se centra en comprender el comportamiento humano basándose en la importancia de las relaciones sociales y sus implicaciones. La importancia del estudio de las redes sociales se remonta a la antropología y la sociología, particularmente en los modelos psicométricos. Al examinar las conexiones de red a través del SNA, los modelos se vuelven menos dependientes de la subjetividad y minimizan el riesgo de perder información crítica, ya que se pueden calcular parámetros cuantitativos de la red y se pueden interpretar gráficos.

El fraude con tarjeta no presente (CNP), que se refiere a realizar pagos con una tarjeta sin necesidad de estar físicamente presente durante la transacción, está en aumento y a menudo se asocia con redes organizadas de estafadores. Para abordar este problema, el Análisis de Redes Sociales (SNA) se ha convertido en una herramienta prometedora para detectar y comprender las conexiones entre estos estafadores. A diferencia de los métodos tradicionales de detección de fraude que se basan en el análisis de puntuaciones de riesgo, SNA se centra en las relaciones entre los actores en las redes sociales para identificar comportamientos fraudulentos. Esto significa que SNA puede analizar datos de diversas fuentes, incluidas plataformas de redes sociales, registros telefónicos y pasarelas de pago, para descubrir patrones y conexiones que puedan indicar actividad fraudulenta. Al incorporar el SNA al modelo existente de detección de fraude, los investigadores esperan obtener conocimientos más profundos sobre los mecanismos subyacentes del fraude y desarrollar estrategias de prevención más efectivas.

Los métodos supervisados, como la regresión y la clasificación, desempeñan un papel crucial en la predicción de las variables de riesgo crediticio, específicamente en la calificación crediticia. Por otro lado, los métodos no supervisados se centran en el análisis y procesamiento de datos. El objetivo principal de los métodos supervisados es identificar y detectar posibles riesgos futuros que puedan resultar en incumplimiento. Vale la pena señalar que la eficacia de los algoritmos de aprendizaje automático afecta directamente la precisión de las predicciones del sistema, incluida la determinación del momento, los datos, el contenido y el canal correctos. Esta comprensión integral permite la formulación de ofertas de pago adecuadas para motivar a las personas a cumplir con sus obligaciones de deuda.

Los cuadros de mando se utilizan con el fin de evaluar el riesgo crediticio de los clientes, particularmente después de la introducción de Basilea II en 2007 por el Comité de Basilea. Este acuerdo internacional exige que las instituciones financieras posean mecanismos efectivos para evaluar el riesgo crediticio de los clientes, lo que requiere ajustes en los sistemas de presentación de informes y métodos de análisis de la información.

Un cuadro de mando es un modelo que utiliza datos para estimar la probabilidad de que un cliente incumpla su crédito. Esta herramienta es particularmente útil para los bancos, ya que proporciona un formato estructurado y fácil de entender para evaluar la solvencia de sus clientes. El cuadro de mando consta de varios componentes, incluidas características, atributos y puntos. Las características se refieren a los datos que se analizan sobre el cliente, como su edad y situación laboral. Los atributos, por otro lado, son las diferentes respuestas o categorías dentro de cada característica. Por ejemplo, la edad podría clasificarse en rangos como 18-25 o 25-35, mientras que la situación laboral podría clasificarse como desempleado o gerente, entre otros. Se asignan puntos a cada atributo y sirven como una puntuación que indica la probabilidad de incumplimiento del cliente. Cuantos más puntos reciba un cliente, menor será su probabilidad de incumplimiento. El número específico de puntos asignados a cada atributo está determinado por los datos históricos del banco sobre el poder predictivo de la característica y la correlación entre diferentes características .

Hay dos categorías distintas de calificación crediticia conocidas como calificación de aplicación y calificación de comportamiento. La puntuación de solicitud se centra únicamente en la información proporcionada por el cliente que busca crédito, mientras que la puntuación de comportamiento incorpora tanto la información proporcionada como los datos históricos. El proceso de creación de un cuadro de mando crediticio implica tres etapas: recopilar y construir el conjunto de datos necesario, aplicar el modelo y documentar los resultados.

El paso inicial consiste en recopilar los datos necesarios y dividirlos en dos grupos distintos. Los datos de entrenamiento, que representan el 70% del conjunto de datos, se utilizarán para construir el cuadro de mando, mientras que el 30% restante se utilizará como datos de prueba para evaluar la precisión del cuadro de mando. A continuación, debemos evaluar las características de los datos y descartar aquellos que menos

contribuyen a las predicciones, simplificando el conjunto de datos. Además, es posible segmentar los datos, creando cuadros de mando separados para clientes con diferentes atributos, con el fin de mejorar las capacidades de predicción. Una vez que se ha desarrollado el modelo, se somete a pruebas utilizando los datos de prueba. Si el banco está satisfecho con la precisión del modelo a la hora de predecir resultados, se implementará para su uso.

Las instituciones financieras frecuentemente encuentran dificultades al utilizar datos históricos debido a su naturaleza desequilibrada. Este desequilibrio plantea un desafío importante para estas instituciones mientras navegan por las complejidades del análisis e interpretación de los datos. Cuando se trata de análisis de riesgo crediticio, es importante señalar que los bancos sólo tienen datos sobre los clientes a los que se les han concedido préstamos, no sobre aquellos a los que se les ha negado. En consecuencia, los datos históricos solo incluyen información sobre el número total de clientes que han recibido préstamos, no el número total de clientes que han solicitado préstamos. Los datos de las personas a quienes se les han negado préstamos no se registran, lo que genera un sesgo conocido como sesgo de rechazo.

Como resultado, los bancos tienen más casos de clientes que han pagado con éxito sus préstamos en comparación con aquellos que no lo han hecho. Equilibrar estos datos supondría un coste significativo para la empresa, ya que tendría que aprobar préstamos para personas que es poco probable que los paguen. Para abordar este problema, se pueden emplear varias técnicas de extracción de datos, como el muestreo, para mitigar el problema. La técnica más utilizada es el muestreo, que implica tres enfoques diferentes. El submuestreo implica eliminar observaciones de los datos repetidas aleatoriamente, generalmente aquellas de clientes que han pagado exitosamente sus préstamos. Por otro lado, el sobremuestreo agrega observaciones de la clase menos común, aumentando los datos existentes para esa clase en particular. Por último, se puede emplear una técnica híbrida, combinando elementos de los métodos antes mencionados para lograr un conjunto de datos equilibrado.

Un campo que está fuertemente influenciado por los algoritmos es el comercio de alta frecuencia, donde es necesario tomar decisiones en fracciones de segundo. El comercio de alta frecuencia, a menudo abreviado como HFT, es una práctica en la que los inversores emplean programas de software especializados para implementar estrategias

de inversión en múltiples valores de manera sistemática y con alta frecuencia. Esta ingeniería matemática, a menudo denominada robots financieros, permite ejecutar órdenes de acciones en cuestión de segundos, lo que ha generado controversia en el mercado. Según Martínez (2010), los sistemas automáticos representan hoy más del 60% del negocio bursátil mundial, lo que genera preocupación sobre su potencial para desestabilizar el mercado. El concepto detrás de HFT era minimizar los costos de transacción y maximizar las ganancias; así el aprendizaje automático se ha convertido en una herramienta popular para ejecutar operaciones y generar alfa.

Para mejorar la eficiencia de diversas tareas, se emplea el aprendizaje por refuerzo, ya que nos permite adquirir conocimientos sobre cómo navegar y tomar decisiones de manera efectiva dentro de un entorno determinado, en lugar de predecir únicamente los resultados deseados. Al utilizar algoritmos de aprendizaje por refuerzo, podemos determinar el curso de acción más adecuado en función de las circunstancias prevalecientes. En consecuencia, esto puede proporcionar una valiosa orientación a los inversores, ayudándoles a identificar el momento y el lugar óptimos para adquirir valores financieros, minimizando así los costos y maximizando el rendimiento de la inversión.

En el ámbito de las finanzas, el término "alfa" se refiere a la capacidad de un activo financiero para superar al mercado general y lograr una mayor rentabilidad. Para lograr alfa, los analistas e inversores emplean modelos de predicción del movimiento de precios para identificar señales ventajosas que superen los costos asociados con el comercio. Este meticuloso proceso se basa en el análisis de datos pasados para determinar tendencias futuras y tomar decisiones de inversión informadas.

Para comprender el sector financiero, es fundamental tener claro lo que implica un mercado eficiente, en donde los precios de los activos financieros reflejan con precisión toda la información disponible. Esto significa que los precios de estos activos representan su verdadero valor intrínseco y se actualizan en tiempo real a medida que hay nueva información disponible. Al adherirse a esta teoría, se desacredita cualquier noción de obtener ganancias prediciendo movimientos futuros de precios. Afirma que los productos financieros no están ni infravalorados ni sobrevalorados. Según la teoría del mercado eficiente, es posible lograr altos rendimientos, pero sólo en proporción al riesgo asociado. Sin embargo, la evidencia reciente sugiere que los principios fundamentales de la teoría financiera han quedado desacreditados. El mercado financiero no es tan eficiente

como se creía anteriormente, lo que permite a los inversores beneficiarse potencialmente de la predicción de futuras fluctuaciones de precios. Aquí es donde el aprendizaje automático juega un papel importante.

En los últimos años, la aplicación de algoritmos de aprendizaje automático en la predicción de precios del mercado de valores ha ganado fuerza. Wang y Wang (2017) realizaron un estudio en el que utilizaron técnicas de aprendizaje automático para desarrollar un modelo predictivo de los precios del mercado de valores. Su enfoque arrojó resultados prometedores, demostrando un alto nivel de precisión en la predicción de los precios de las acciones. Predecir los precios de las acciones o de cualquier activo financiero es una tarea compleja y desafiante.

Para investigadores Niederhoffer y Osborne (1966), los cambios en los precios suelen ocurrir en direcciones opuestas con más frecuencia que los cambios en la misma dirección. Esto pone de relieve la dificultad inherente a la predicción precisa de los movimientos de precios. Para abordar este desafío, comúnmente se emplean métodos supervisados. Estos métodos se basan en la disponibilidad de datos etiquetados, donde se conoce en todo momento el precio futuro de un activo financiero. A medida que pasa el tiempo, el conjunto de datos se actualiza constantemente con nuevos datos etiquetados, lo que permite una predicción más precisa.

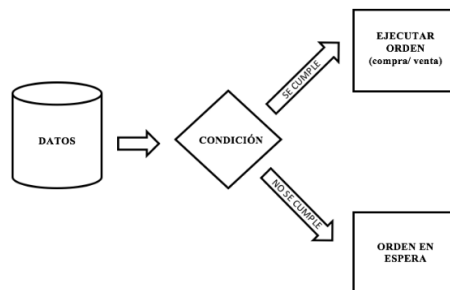
En esencia, predecir precios en los mercados financieros es una tarea compleja debido a la frecuente ocurrencia de cambios en direcciones opuestas. Para superar este desafío, se utilizan métodos supervisados, aprovechando datos etiquetados para actualizar y mejorar continuamente las predicciones. Mediante el análisis cuantitativo de precios históricos y el uso de algoritmos de aprendizaje automático, los investigadores han logrado avances significativos en el desarrollo de modelos de predicción precisos de los precios del mercado de valores. El principal objetivo de estos métodos es identificar patrones y tendencias en los precios históricos de activos financieros a través de análisis cuantitativos. Al analizar los movimientos de precios pasados, los investigadores y analistas pretenden descubrir relaciones y correlaciones significativas que puedan utilizarse para predecir precios futuros.

Otro uso del aprendizaje automático en el ámbito de los mercados financieros se conoce como algo-trading o comercio algorítmico. En este proceso, se programan

instrucciones específicas en el software para ejecutar órdenes de compra o venta. Al incorporar algoritmos en esta práctica, los comerciantes pueden potencialmente aumentar sus ganancias minimizando los gastos comerciales, al mismo tiempo que se benefician de una mayor velocidad y frecuencia de las transacciones en comparación con las operaciones manuales. Este enfoque ayuda a aliviar la influencia de la subjetividad humana en las decisiones comerciales, haciendo que el proceso sea más metódico y sistemático.

El algoritmo requiere parámetros específicos en sus instrucciones, incluido el precio de compra o venta deseado, el volumen de instrumentos a operar y el vencimiento de la orden. Si la orden es de compra, el precio indicado representa el precio máximo que el inversor está dispuesto a pagar por el instrumento financiero. El algoritmo comprará todos los instrumentos por debajo de este precio máximo, comenzando por los de menor precio, hasta alcanzar el volumen deseado. Por el contrario, si la orden es de venta, el precio indicado es el mínimo permitido, y el algoritmo venderá a precios mayores o iguales a ese precio indicado, empezando por el más alto, hasta conseguir el volumen solicitado.

Cualquier valor que no cumpla con las condiciones de precio permanecerá en el registro de órdenes pendientes hasta que expire la orden, momento en el cual la instrucción será eliminada del sistema. Esto asegura que la máquina siga condiciones específicas para ejecutar la orden y enviarla al mercado financiero a través de la plataforma del intermediario financiero. Este enfoque garantiza la rentabilidad y ayuda a limitar las pérdidas potenciales. Para proporcionar una comprensión más clara de cómo funciona el comercio algorítmico, a continuación se incluye un gráfico visual.



Una de las aplicaciones más importantes de la inteligencia artificial en el mundo actual es el uso de chatbots, que son programas diseñados para mantener conversaciones con personas en plataformas de mensajería y esencialmente actúan como reemplazos robóticos de la interacción humana. Esta interacción puede realizarse a través de texto o audio, utilizando lenguaje natural. Los chatbots han revolucionado el servicio al cliente al permitir interacciones personalizadas entre clientes y empresas, brindando a los usuarios beneficios sociales, informativos y económicos.

Asimismo, los chatbots están disponibles las 24 horas del día, los 7 días de la semana, lo que ofrece atención al cliente continua y reduce la necesidad de que los clientes viajen físicamente a los establecimientos físicos. Además, dentro de las organizaciones se implementan chatbots para ayudar a los trabajadores a resolver cualquier duda o pregunta que puedan tener. En el sector financiero, los chatbots tienen un inmenso potencial y pueden servir como asistentes de recursos humanos, asistentes de inteligencia de mercado, asistentes de flujo de trabajo, asistentes de redes sociales, asistentes de analistas financieros, asistentes de programación e incluso como embajadores de la empresa. Es de destacar que sólo el 20% del trabajo administrativo representa un asombroso 85% del costo total para los bancos. Para 2022, se estima que los chatbots generarán más de 8 mil millones de dólares en ahorros de costos anuales (United Consulting Group, 2018).

Los chatbots representan alrededor del 9,6% del suministro total de productos de IA. Sin embargo, en la industria bancaria, se utilizan en aproximadamente el 33,9% de los casos de uso de IA (Francés Monedero, 2020). Esta importante dependencia de los chatbots por parte de los bancos puede deberse a su deseo de presentarse como innovadores y tecnológicamente avanzados. Desafortunadamente, muchos bancos carecen de una estrategia clara de IA. Ian Wilson, ex director de IA de HSBC, señaló que la pregunta más común en los bancos es sobre el futuro y el retorno de la inversión, pero no hay una respuesta satisfactoria para esto, lo que plantea un problema. Además, estos programas de chatbot tienen una capacidad limitada para manejar consultas básicas de los clientes.

Debido a que el lenguaje humano es complejo, plantea un desafío para los chatbots entablar conversaciones fluidas y coherentes. Sin embargo, el programa informático Eugene logró un hito importante en 2014 al superar con éxito la prueba de Turing. Este

avance marcó un importante punto de inflexión, ya que permitió a las máquinas utilizar técnicas de aprendizaje automático para entablar un diálogo significativo con los humanos.

Inicialmente, el desarrollo de chatbots se basó en la experiencia de profesionales y siguió estándares establecidos. Si bien, a medida que pasó el tiempo, se introdujeron proyectos híbridos que combinaban técnicas de aprendizaje automático con enfoques tradicionales. Actualmente, hay un cambio hacia chatbots que utilizan únicamente el aprendizaje automático. Estos sistemas avanzados de chatbot se entrenan continuamente utilizando datos históricos, lo que les permite mejorar indefinidamente a medida que acumulan más información.

El algoritmo empleado por estos chatbots está diseñado para identificar la intención del usuario y extraer información relevante de sus mensajes. Al evaluar el problema del cliente, el sistema puede generar una respuesta adecuada. Esto se logra analizando las soluciones más comunes que se encuentran en el conjunto de datos y devolviéndolas al sistema mediante el aprendizaje supervisado. En los casos en los que hay múltiples respuestas potenciales, el chatbot emplea técnicas de aprendizaje profundo para evaluar el contexto de la conversación y selecciona la respuesta más adecuada en función de la entrada actual.

El funcionamiento de los chatbots implica la utilización de diversas técnicas. Para facilitar la interacción con los usuarios, la interfaz del software se basa en la aplicación de Procesamiento del Lenguaje Natural (NLP). La NLP, al ser una subdivisión de la inteligencia artificial, abarca la capacidad de las máquinas para comprender el lenguaje humano. Este campo se centra en procesar la información recibida de una manera que permita a las máquinas interpretarla de forma eficaz. Por otro lado, el backend, específicamente el procesador de entrada frontend, utiliza técnicas de aprendizaje supervisado y aprendizaje profundo para mejorar su funcionalidad.

España ha ostentado el título de ser el país más bancarizado del mundo, con un impresionante número de sucursales por persona, con la asombrosa cifra de 95,87 sucursales por cada 100.000 personas (Camino y de Garcillán López-Rúa, 2014). Sin embargo, esta distinción está experimentando gradualmente una transformación, todo gracias a la llegada de tecnologías innovadoras. Una famosa cita de Bill Gates de 1994

destaca este cambio de perspectiva: "Necesitamos el sistema financiero, pero no los bancos". La aparición de aplicaciones móviles y servicios automatizados en línea está revolucionando el panorama financiero, ofreciendo una experiencia altamente personalizada y adaptada a los clientes. Estas innovadoras plataformas financieras, acertadamente denominadas "roboadvisors", están allanando el camino hacia una nueva era de la banca.

Esta herramienta utiliza algoritmos de aprendizaje automático para ofrecer asesoramiento de inversión a los clientes. Al interactuar con el cliente y analizar sus ingresos y gastos pasados, la máquina puede sugerir planes de ahorro e inversión. El cliente crea un perfil y responde una serie de preguntas, lo que permite que el programa informático comprenda su perfil de riesgo, preferencias de activos y más. Utilizando esta información, el robo-advisor aplica algoritmos para brindar una recomendación personalizada. La integración de los robo-advisors en el ciclo financiero bancario permite una mejor planificación financiera que se alinea con el ciclo de vida y el perfil de riesgo del cliente, ayudándole a alcanzar sus objetivos financieros.

Al implementar este enfoque, la utilización de dicha tecnología tiene el potencial de disminuir significativamente la cantidad de oficinas operativas necesarias para atender a los clientes. Además, los robo-advisors ofrecen una oportunidad invaluable para que una parte sustancial de la población, que tal vez no posea conocimientos financieros fundamentales, participe en inversiones, ahorros e incluso planificación de la jubilación.

El campo del análisis y reconocimiento de documentos (DAR) se centra en la extracción automatizada de información de los documentos, con el objetivo de identificar las cuestiones clave que deben abordarse para la evaluación, resaltando así la información más significativa. Según Marinai y Fujisawa (2007), la integración de técnicas de aprendizaje automático (ML) en este proceso de análisis es muy ventajosa. El proceso de análisis de documentos consta de tres fases principales: recopilación y almacenamiento de información, preprocesamiento de datos y aplicación de diversas técnicas de análisis. La fase de procesamiento de la información es crucial porque los documentos normalmente contienen datos no estructurados y carecen de una organización definida. Para abordar esto, se emplean técnicas de simplificación, como reducir las palabras a sus raíces o eliminar palabras irrelevantes que no contribuyen al significado general de las

oraciones. Luego, en la fase de análisis se aplican diferentes técnicas para escudriñar el texto. Esto puede implicar un análisis cuantitativo, en el que se examina la repetición de ciertas palabras dentro del mismo documento para medir la importancia del tema en el texto presentado, o puede abarcar un análisis de sentimientos, entre otros métodos.

Esta herramienta cambia las reglas del juego en el sector financiero, ya que permite escanear y extraer información crucial de documentos extensos, brindando inmensos beneficios a través de la implementación de tecnología de aprendizaje automático. Al utilizar la plataforma COiN, JP Morgan ha revolucionado sus operaciones al recuperar sin esfuerzo datos vitales de acuerdos de crédito comerciales anuales en cuestión de segundos. Esto ha dado como resultado un asombroso ahorro de tiempo de 360.000 horas al año, que de otro modo se gastarían en tediosas revisiones manuales. La utilización de esta herramienta muestra el increíble potencial del aprendizaje automático para abordar de manera eficiente las necesidades de la industria financiera.

Las ventajas ML en el sistema financiero

En el pasado, las instituciones financieras utilizaban el precio, la velocidad y el acceso para atraer clientes. Si bien, con la llegada de la globalización, este enfoque ha cambiado. Ahora, la capacidad de analizar grandes cantidades de datos y utilizar algoritmos para identificar los productos que mejor satisfacen las necesidades de los clientes es crucial para obtener una ventaja competitiva. Las expectativas de los clientes también han evolucionado, con una demanda creciente de ofertas diversas y personalizadas. Como resultado, la estandarización ya no es una fuente importante de ingresos. En cambio, la atención se centra en ofrecer productos altamente personalizados, donde la inteligencia artificial (IA) juega un papel crucial. Este cambio en la industria ha llevado a que los grandes operadores tradicionales se conviertan en proveedores de servicios de IA.

El aprendizaje automático permite la segmentación de clientes, facilitando así el desarrollo de estrategias personalizadas basadas en sus características únicas. Es importante que los individuos dentro del mismo segmento posean similitudes entre ellos pero también muestren diferencias en comparación con otros grupos. Al implementar diversas estrategias para varios segmentos, las empresas pueden ofrecer experiencias más personalizadas a sus clientes.

Como resultado, la implementación de tecnología de aprendizaje automático otorgará una importante ventaja competitiva a las instituciones financieras que la adopten, permitiéndoles ofrecer niveles incomparables de servicio al cliente personalizado. Esta tecnología de vanguardia permite a estas entidades adaptar sus servicios a las preferencias y necesidades individuales, estableciendo así una conexión más fuerte con su clientela. Al aprovechar los algoritmos de aprendizaje automático, las empresas del sector financiero pueden obtener información valiosa sobre el comportamiento de los clientes, anticipar sus requisitos únicos y ofrecer soluciones personalizadas de forma proactiva. Con la capacidad de analizar grandes cantidades de datos de manera eficiente y precisa, el aprendizaje automático equipa a estas entidades con las herramientas para mejorar la satisfacción y lealtad del cliente. En consecuencia, las organizaciones que adopten el aprendizaje automático en la industria financiera no solo prosperarán en un mercado altamente competitivo, sino que también revolucionarán la forma en que interactúan y atienden a sus clientes.

Las instituciones financieras pueden automatizar tareas repetitivas o tareas que aportan menos valor mediante técnicas como el procesamiento del lenguaje natural o el reconocimiento de imágenes. Por ejemplo, las preguntas más frecuentes se pueden responder utilizando estas técnicas. Amazon ha implementado más de cien mil robots en todo el mundo. Según el director de operaciones, Dave Clark, el objetivo era que las máquinas realizaran tareas monótonas, permitiendo a los humanos realizar trabajos mentalmente interesantes. En lugar de humanos versus máquinas, la IA se trata de humanos mejorados por máquinas, como afirman Jubraj, Graham y Ryan. La IA tiene el potencial de mejorar las habilidades humanas y optimizar el tiempo de los empleados de la empresa liberándolos de tareas repetitivas y permitiéndoles centrarse en tareas más valiosas.

A lo largo de la historia, los economistas han sostenido la creencia de que las máquinas poseen el potencial de reemplazar a múltiples trabajadores humanos en diversas industrias. Sin embargo, su argumento final ha sido que tales avances tecnológicos conducirían a un aumento significativo en los niveles de productividad, lo que en consecuencia resultaría en un aumento sustancial tanto del ingreso como del producto interno.

Como resultado, la cadena de valor sufrirá una transformación significativa en términos de generación de beneficios. Con el tiempo, habrá un cambio notable en la distribución del valor hacia tareas que exigen experiencia especializada, mientras que los servicios que pueden ejecutarse sin la participación de profesionales experimentados serán menos valorados por la sociedad. La razón detrás de este cambio de percepción radica en la llegada del aprendizaje automático, que permite realizar estas tareas a un costo considerablemente reducido.

Dentro de la industria financiera, si se automatizaran las tareas más monótonas, se liberaría una cantidad significativa de tiempo para los equipos de gestión financiera. Según Chui, Manyika y Miremadi (2015), esto podría representar potencialmente más del 20% de su carga de trabajo total. Al delegar responsabilidades como la recopilación, verificación y consolidación de datos a las máquinas, los profesionales de este campo tendrían la oportunidad de concentrar sus esfuerzos en aspectos más cruciales como el análisis y la toma de decisiones. El nivel de automatización de tareas sería directamente proporcional a la medida en que el trabajo de un empleado comprenda actividades rutinarias.

Cuando se trata de trabajadores responsables de aprobar préstamos hipotecarios, el aprendizaje automático se hará cargo de la tarea repetitiva de recopilar y analizar datos. Esto permitirá a los empleados gestionar un mayor número de solicitudes de préstamos y brindar asesoramiento a un mayor número de clientes. Como resultado, los empleados podrán hacer un uso más eficiente de su tiempo aprovechando el análisis de datos realizado por las máquinas.

Por el contrario, los asesores financieros darán prioridad a comprender las necesidades de sus clientes y formular diversas tácticas para lograr sus objetivos, dedicando comparativamente menos tiempo a examinar el estado financiero de los clientes. Las máquinas agilizarán el procesamiento y evaluación de los datos de cada cliente, tarea que antes realizaban empleados humanos del sector.

En un estudio realizado por McKinsey en 2018, se descubrió que a pesar de la automatización de las tareas manuales, la creación de nuevos puestos de trabajo superará el número de puestos de trabajo reemplazados. El estudio sugiere que los beneficios de la inteligencia artificial irán más allá de la simple sustitución de puestos de trabajo, ya

que solo una fracción de los beneficios totales procederá de la sustitución de puestos de trabajo (Francés Monedero, 2020).

Así en la medida que las empresas adopten la IA, habrá un cambio en las habilidades necesarias para los puestos de trabajo. La demanda de habilidades físicas disminuirá, lo que resultará en un exceso de oferta, mientras que la demanda de habilidades cognitivas aumentará. En consecuencia, habrá un aumento de los salarios. El auge de las máquinas en el sector bancario, según informa Bloomberg, redefinirá los roles y funciones de los empleados bancarios. Esto ha generado una gran demanda de candidatos con experiencia en inteligencia artificial, aprendizaje automático y ciencia de datos en el sector financiero, según portales de empleo como Glassdoor y LinkedIn (Francés Monedero, 2020).

De similar forma, a medida que ciertas tareas en el sector financiero se automaticen, habrá una mayor necesidad de puestos centrados en la gestión de relaciones con los clientes (CRM) para brindar un servicio más personalizado a los clientes. Si bien las máquinas pueden generar recomendaciones, son los empleados humanos quienes mejoran la experiencia del cliente a través de interacciones personalizadas. Por tanto, la cooperación entre la inteligencia emocional humana y la eficiencia de las máquinas es crucial en esta industria.

En actividades que requieren toma de decisiones, la objetividad es crucial. El aprendizaje automático ofrece un mayor nivel de precisión, ya que puede analizar datos de diversas fuentes, en diferentes formatos, y verificar su validez. Esto reduce la influencia de la subjetividad de un individuo en la toma de decisiones. Un ejemplo de la importancia de la objetividad en las actividades empresariales es la auditoría. Cuando las empresas alcanzan un determinado nivel de capital, están obligadas por ley (en España, por ejemplo) a someterse a auditorías externas realizadas por auditores independientes. Estos auditores no deben tener ninguna asociación con las operaciones de la empresa. Esto garantiza la confiabilidad de los estados financieros y elimina los conflictos de intereses.

Esta objetividad es particularmente valiosa en sectores como el financiero, donde la toma de decisiones implica elegir entre múltiples opciones que impactan directamente el bienestar financiero de los clientes. En tales casos, es necesario disponer de información completa y seleccionar estrategias con la máxima objetividad. Los asesores humanos, por

otro lado, son subjetivos, propensos a errores y consumen mucho tiempo. Por el contrario, los algoritmos son objetivos, más precisos, más rápidos, eficientes y pueden funcionar 24 horas al día, 7 días a la semana.

El aprendizaje automático procesa datos con precisión. Por ejemplo, cuando se realiza un análisis cualitativo de la probabilidad de incumplimiento crediticio, la información de entrada utilizada para evaluar el riesgo crediticio de un cliente es a menudo subjetiva y vaga. Abordar esta cuestión es crucial para garantizar una mayor objetividad y cumplimiento de las directrices establecidas por el Banco de Pagos Internacionales (BPI). El BPI es responsable de mantener la estabilidad del sistema financiero y monetario internacional. En 2015, el BPI publicó once principios de orientación supervisora sobre el riesgo crediticio y la contabilización de las pérdidas crediticias esperadas. El principio número dos enfatiza la importancia de que los bancos adopten y se adhieran a metodologías sólidas para evaluar y calcular el riesgo crediticio en todas las exposiciones crediticias (Banco de Pagos Internacionales, 2015).

Tradicionalmente, los bancos recopilaban información de los solicitantes de préstamos a través de formularios de solicitud y otras fuentes para evaluar si el préstamo debía otorgarse y sus términos. Luego, los empleados analizarían esta información y tomarían una decisión. Sin embargo, no era raro que diferentes asesores llegaran a conclusiones diferentes sobre si se debía conceder o no un préstamo. Al utilizar el aprendizaje automático, se puede crear un modelo de calificación crediticia estandarizado, garantizando que la decisión sea universal y no dependa del análisis de un individuo.

Los datos del aprendizaje automático se caracterizan por tres factores principales: volumen, variedad y velocidad.

- El volumen de datos ha aumentado significativamente debido a la amplia cobertura y conexiones de Internet, lo que plantea el desafío de almacenar cantidades tan grandes de información.
- La variedad de fuentes de medios, diversas representaciones y disponibilidad de datos hacen que su análisis sea una tarea compleja.
- Además, la velocidad a la que se generan los datos ha aumentado exponencialmente con el avance de la difusión de datos desde diversos

dispositivos. Es fundamental analizar y almacenar estos datos de manera eficiente y rápida para evitar incurrir en costos de oportunidad, ya que puede no siempre ser factible analizar cada pieza de información.

El concepto de Internet de las Cosas, comúnmente conocido como IoT, abarca la conectividad y comunicación entre varios objetos y dispositivos a través de una red privada o basada en Internet. Esta interconexión permite una interacción perfecta y el intercambio continuo de datos entre estos dispositivos conectados. En consecuencia, el gran volumen de datos que se transmiten, que incluye información tanto cuantitativa como cualitativa, es notablemente vasto.

Los dispositivos móviles han contribuido significativamente a la expansión de las operaciones y la acumulación de una gran cantidad de datos en el sector financiero. Para aprovechar y extraer eficazmente información valiosa de esta abundancia de datos, la integración de algoritmos de aprendizaje automático se ha vuelto esencial. Estos algoritmos permiten el procesamiento de datos en tiempo real, facilitando así la recuperación inmediata de información valiosa. Esta innovadora incorporación de la inteligencia artificial en el sector financiero ha revolucionado los procesos, particularmente mediante la implementación de sistemas automáticos de reconocimiento de dígitos.

Así se tiene que, los mercados financieros funcionan las 24 horas del día y la capacidad de procesar datos comerciales rápidamente es crucial para una gestión comercial eficaz. Como resultado, la tecnología de aprendizaje automático permite a las personas tomar decisiones en tiempo real, lo que les permite posicionarse estratégicamente y prever resultados potenciales en los dinámicos mercados financieros.

En la misma línea, los cambios mencionados, que se caracterizan por su rapidez y amplitud, tienen un profundo impacto en las operaciones y funciones del Banco de España. Estos cambios no solo influyen en los procesos internos del banco sino que también tienen implicaciones significativas para sus funciones analíticas y de supervisión (Banco de España, 2020). El sector financiero se enfrenta constantemente a regulaciones y estándares de presentación de informes novedosos, lo que requiere la capacidad de adaptarse rápidamente a estos requisitos en evolución.

El aprendizaje automático no sólo acelera el procesamiento de los datos y la información de un cliente, sino que también acelera la capacitación de los empleados bancarios, lo que resulta en una mayor productividad y rentabilidad para las organizaciones. Al automatizar las tareas, se reduce el tiempo necesario para su ejecución, acelerando así la curva de aprendizaje de los nuevos empleados. Además, la utilización del aprendizaje automático infunde confianza en el trabajo de estos nuevos empleados, ya que están respaldados por la confiabilidad de los algoritmos.

En la actualidad, existe una disparidad en la accesibilidad a los servicios sanitarios y financieros, y solo determinadas clases sociales tienen fácil acceso a ellos. Sin embargo, la utilización del aprendizaje automático puede revolucionar la eficiencia de estos servicios y permitirles llegar a una gama más amplia de personas. Esto es particularmente significativo cuando se consideran préstamos pequeños, como los requeridos por las pequeñas empresas, donde el monto relativamente bajo del préstamo y el riesgo potencial involucrado para los prestamistas no justifican los gastos asociados con la evaluación de la solvencia del prestatario. En tales casos, la integración de técnicas de aprendizaje automático ofrece una solución óptima, que permite que una mayor parte de la población aproveche los servicios financieros.

Por el contrario, en los casos en que una parte de las tareas esté automatizada, las comisiones que reciben los asesores financieros se reducirían, lo que aumentaría el atractivo para una clientela más amplia. Además, si bien adoptar el aprendizaje automático (ML) tiene ventajas innegables, es importante reconocer la existencia de ciertos obstáculos y complejidades que presenta.

Los desafíos de la ML

La incorporación de la inteligencia artificial (IA) en las empresas ha creado la necesidad de contratar científicos de datos que posean la experiencia necesaria. Aun cuando, debido a la intensa competencia, encontrar personal con las habilidades adecuadas se ha convertido en una tarea desafiante. Actualmente, existe una mayor demanda de estas personas calificadas que la oferta disponible. Como resultado, las empresas que deseen implementar el aprendizaje automático en sus servicios deben hacer todo lo posible no sólo para atraer, sino también para retener a estos especialistas.

Vale la pena señalar que el 90% de las 1.600 startups de IA en Europa se centran principalmente en ofrecer servicios de IA a otras empresas. Esto indica que determinadas empresas optan por subcontratar sus servicios de aprendizaje automático a proveedores externos. Para las empresas en esta situación, los factores claves del éxito residen en integrar eficazmente estos servicios externalizados y establecer una relación de largo plazo con sus proveedores. Dado el número limitado de profesionales en este campo, si las empresas del sector financiero dependieran en gran medida del aprendizaje automático para sus operaciones, podría aumentar su riesgo operativo e incluso generar un riesgo sistémico.

Un factor crucial para la implementación exitosa de la IA es la posesión de amplios conjuntos de datos para el desarrollo de algoritmos, y garantizar el acceso a estos datos es esencial. Además, la cantidad de esfuerzo humano necesaria para recopilar estos datos es inmensa, ya que antes implicaba una cantidad significativa de tiempo para la extracción, pero ahora se centra más en la preparación. Esto es particularmente importante debido a las características específicas de los datos utilizados en el aprendizaje automático, como el volumen, la variedad y la velocidad, que requieren un proceso de captura de datos eficiente. Sin embargo, el sector financiero no enfrenta este desafío porque ya posee una vasta base de datos acumulada con el tiempo debido al sistema contable y los requisitos regulatorios. Estos registros históricos, que inicialmente no fueron plenamente reconocidos por su potencial, ahora tienen un valor significativo. También, el sector financiero ha evolucionado hacia una industria multicanal, lo que se ha traducido en un mayor volumen de datos generados desde diversos dispositivos en un período de tiempo más corto a través de transacciones monetarias. Como resultado, el sector financiero se enfrenta ahora a una cantidad tan enorme de datos que se ha hecho necesario emplear técnicas para analizarlos.

Por el contrario, el aprendizaje automático tiene el potencial de incorporar sesgos de datos históricos, incluidos los relacionados con el género y la raza. Como resultado, cuando se trata de brindar servicios financieros a grupos minoritarios, estos algoritmos pueden no ser los más adecuados. La precisión a la hora de clasificar personas con diferentes tonos de piel, ya sean hombres o mujeres, varía significativamente debido a la presencia predominante de personas de piel clara en los conjuntos de datos. En esencia, los propios algoritmos de aprendizaje automático pueden ser imparciales, pero no se

puede ignorar la naturaleza sesgada de los datos con los que se entrenan. Para garantizar la equidad y la transparencia en la toma de decisiones algorítmicas para las instituciones financieras, es crucial adoptar conjuntos de datos diversos e inclusivos, así como emplear los métodos de capacitación y enfoques de desarrollo de algoritmos más efectivos.

Un ejemplo lo constituye Amazon, que intentó desarrollar un algoritmo de aprendizaje automático con el objetivo de acelerar su proceso de selección de personal. El objetivo principal era examinar los CV de las personas y excluir a aquellos que, según los datos históricos de la empresa, era poco probable que fueran contratados. El algoritmo utilizó todos los CV de la década anterior, sin tener en cuenta el género. Sin embargo, surgió un problema notable cuando se descubrió que el modelo frecuentemente ignoraba a las mujeres como candidatas potenciales debido a sesgos inherentes en el conjunto de datos, que favorecían la contratación de hombres. Tras un examen más detenido, se descubrió que el algoritmo pasaba por alto los CV que contenían los términos "mujer" o "femenino" durante el proceso de selección. Otro ejemplo que destaca las deficiencias de los sistemas algorítmicos es la implementación por parte de Google de un algoritmo de reconocimiento facial que identificó incorrectamente a los individuos negros como gorilas. Este etiquetado erróneo fue consecuencia de observaciones insuficientes de caras negras en el conjunto de datos utilizado para entrenar el modelo.

El uso de algoritmos en el sector financiero, particularmente en actividades como ofrecer crédito o negociar activos financieros, tiene el potencial de dar lugar a casos de agregación y, con el tiempo, podría conducir a un comportamiento procíclico.

La competencia

Se espera que el surgimiento de la economía de la IA cree un panorama competitivo que favorezca a las instituciones establecidas con importantes cuotas de mercado e innovadores ágiles. En consecuencia, existe la preocupación de que esto pueda conducir a una situación monopolística en la que las grandes instituciones financieras tradicionales dominen el mercado, ya que las pequeñas y medianas empresas pueden carecer de la inversión necesaria para mantenerse al día con los avances tecnológicos. Esta preocupación está respaldada por un estudio realizado por Edelman y el Foro Económico Mundial en 2019, que encontró que el 54% del público en general y el 43% de los líderes tecnológicos creen que la IA tendrá implicaciones negativas para las personas más

vulnerables, mientras que el 67% y El 75% respectivamente prevé que beneficiará principalmente a los más ricos. Estos hallazgos resaltan la posibilidad de que surjan desequilibrios como resultado de la economía de la IA (Francés Monedero, 2020).

La brecha entre las empresas que optan por implementar el aprendizaje automático y las que no seguirán creciendo a medida que las primeras inviertan más en tecnología de aprendizaje automático y recopilen más datos. Es fundamental que las empresas se anticipen a este cambio e incorporen el aprendizaje automático en sus operaciones. De no hacerlo, estas empresas podrían verse expulsadas del mercado debido a la intensa competencia. En consecuencia, habrá un cambio en el panorama competitivo, lo que requerirá nuevos factores para lograr el éxito. El aprendizaje automático acelerará enormemente los procesos comerciales, lo que conducirá a ciclos más rápidos de innovación, adopción y consumo. Esta aceleración ha reducido históricamente el tiempo que las grandes empresas siguen siendo líderes del mercado. Es probable que el aprendizaje automático proporcione una ventaja competitiva, creando potencialmente un monopolio para las grandes empresas, dejando solo a un pequeño grupo de competidores.

El accionar

Las empresas enfrentan dificultades al intentar aplicar algoritmos que fueron desarrollados para situaciones específicas a otros casos similares pero diferentes. Este problema surge porque el uso de algoritmos en datos con los que no fueron entrenados originalmente puede crear problemas. Para superar este desafío, es necesario desarrollar algoritmos que puedan aplicarse en un contexto más amplio. Esto se puede lograr incorporando datos heterogéneos, que consisten en observaciones de varios escenarios.

En 2012, Knight Capital, una empresa especializada en negociación de acciones, desarrolló un programa de software diseñado específicamente para negociar valores de acciones en la Bolsa de Nueva York. Sin embargo, se produjo un desafortunado fallo informático que provocó que la empresa sufriera una asombrosa pérdida de 440 millones de dólares en tan solo 45 minutos. Afortunadamente, la empresa reconoció rápidamente el problema y detuvo rápidamente el funcionamiento del algoritmo. El algoritmo participaba constantemente en la compra y venta de millones de acciones, lo que provocó un aumento significativo de la demanda y, posteriormente, elevó los precios de las

acciones. En consecuencia, las acciones se sobrevaloraron, lo que resultó en pérdidas financieras sustanciales para la empresa (Harford, 2012). Aunque el algoritmo había sido programado para identificar y comprar acciones infravaloradas, no tuvo en cuenta el hecho de que comprar una gran cantidad de acciones crearía un aumento en la demanda, inflando así sus precios y anulando su estado infravalorado inicial.

Legislación y ética

La ciberseguridad es fundamental para el funcionamiento de la economía digital y tiene como objetivo salvaguardar a los usuarios, activos y recursos digitales (como redes, equipos y contenidos) del uso no autorizado en el entorno cibernético. Toda comunicación involucra un remitente, un receptor y un mensaje transmitido. Hay cuatro propiedades que garantizan la confianza o seguridad digital:

- autenticidad (verificar que el mensaje recibido fue enviado por el remitente previsto),
- integridad (garantizar que el mensaje recibido no ha sido alterado),
- confidencialidad (garantizar que el mensaje sólo puede ser leído por destinatario previsto), y
- el no repudio (evitar que el remitente niegue su autoría del mensaje).

La criptografía, un campo de las matemáticas, se ha utilizado durante mucho tiempo en contextos militares. El cifrado implica transformar los datos originales en un formato ininteligible utilizando un algoritmo de cifrado y una clave, mientras que el descifrado invierte este proceso para recuperar los datos originales utilizando un algoritmo y una clave de descifrado. En contextos históricos, la criptografía sólo proporcionaba confidencialidad a través de técnicas como la permutación de caracteres, la sustitución monoalfabética o la sustitución polialfabética. No había forma de verificar la autenticidad del mensaje ni garantizar su integridad durante la transmisión. Se depositó confianza en el mensajero y en la entrega del mensaje. Sin embargo, los protocolos criptográficos modernos se basan en principios matemáticos y herramientas o algoritmos criptográficos para garantizar las cuatro propiedades criptográficas: autenticidad, integridad, confidencialidad y no repudio.

Algoritmos o herramientas criptográficas:

- La criptografía simétrica, también conocida como algoritmo de clave secreta, implica el uso de una clave secreta compartida conocida por todas las entidades autorizadas para cifrar y descifrar datos. Este tipo de criptografía requiere un proceso computacional de moderado a intenso. Su característica inherente es la capacidad de garantizar la confidencialidad.
- La criptografía asimétrica, también conocida como algoritmo de clave pública, implica el uso de dos claves que están interconectadas pero distintas entre sí. Una clave es privada y la genera una autoridad certificadora específicamente para el ciudadano individual, mientras que la otra clave es pública y la genera la autoridad certificadora utilizando la clave privada del ciudadano. La clave privada se utiliza para cifrar datos que sólo se pueden descifrar utilizando su clave pública correspondiente. Del mismo modo, los datos cifrados con la clave pública sólo podrán descifrarse utilizando su correspondiente clave privada. A diferencia de la criptografía simétrica, la criptografía asimétrica requiere más recursos computacionales. Las características inherentes de este enfoque criptográfico incluyen garantizar la autenticidad y mantener la confidencialidad.
- La criptografía de función hash, también conocida como algoritmos criptográficos hash, es un tipo de algoritmo que opera sin necesidad de una clave y genera un resumen digital de los datos originales. Este método criptográfico implica un proceso de cálculo moderadamente complejo. Las funciones hash poseen la característica inherente de garantizar la integridad de los datos, lo que significa que brindan seguridad mediante la utilización de factores como conocimiento de algoritmos, claves y sellos.

El uso de diferentes algoritmos criptográficos en el sector financiero es crucial para garantizar la seguridad. Si bien, cuando el aprendizaje automático se incorpora a varias entidades, la cuestión de asignar responsabilidades se convierte en una consideración importante. Por ejemplo, si se descubre que un algoritmo discrimina a un segmento particular de la población y la entidad es demandada por prácticas discriminatorias, determinar quién debe rendir cuentas se convierte en una preocupación clave. Este aspecto debe abordarse antes de implementar el aprendizaje automático.

Según la ética aristotélica, la responsabilidad requiere conocimiento y control, que son difíciles de lograr cuando se trata de tecnología, particularmente de algoritmos. Como resultado, la ética en la tecnología plantea un problema desafiante. A menudo pasamos por alto la interacción entre humanos y máquinas en las tecnologías financieras, por lo que es crucial considerar este aspecto. El Foro Económico Mundial (2018) afirma que la empresa que utiliza el modelo es la responsable en última instancia.

La humanidad

Como se indicó anteriormente, existe una amplia gama de perspectivas sobre las consecuencias socioeconómicas de las tecnologías emergentes, y hay personas que expresan una fuerte oposición a la colaboración con máquinas. En consecuencia, los profesionales de los sectores bancario y comercial perciben la ausencia de una mentalidad orientada a la innovación como el obstáculo fundamental que obstaculiza la implementación generalizada de la inteligencia artificial (IA) en el ámbito bancario.

Para evitar que esta situación ocurra, es imperativo que las empresas garanticen la integración del conocimiento adquirido mediante la utilización de la IA en sus operaciones diarias y en los comportamientos de sus empleados. En consecuencia, se vuelve fundamental que la IA se arraigue profundamente en la cultura de la empresa, llevando a la transformación de ciertos procedimientos dentro de la organización.

Capítulo 2

Análisis descriptivo en el aprendizaje automático

El aprendizaje automático, considera una serie de situaciones en las que un individuo mejora sus conocimientos o habilidades para completar una tarea; implica sacar conclusiones a partir de información específica para construir una representación adecuada de algún aspecto importante de la realidad o de algún proceso. Un ejemplo común en el campo del aprendizaje automático, en inteligencia artificial, es ver la resolución de problemas como un tipo de aprendizaje que incluye, la capacidad de reconocer la situación problemática y reaccionar de acuerdo con una estrategia aprendida.

Hoy en día, la mayor diferencia que se puede hacer entre los animales y los mecanismos de resolución de problemas es que algunos animales pueden mejorar su desempeño en múltiples tareas resolviendo un problema particular. Este enfoque supone que un agente autónomo podrá realizar la misma tarea de múltiples maneras, si es posible y dependiendo de las circunstancias. Debe poder tomar decisiones sobre el curso de acción más apropiado para resolver problemas y cambiar esas decisiones según lo requieran las condiciones.

Por este motivo, una de las principales tareas de esta dirección es crear sistemas capaces de adaptarse con flexibilidad a nuevas situaciones y aprender gracias a la resolución del problema (o problemas) encontrados sin formación previa. El aprendizaje automático, también conocido como aprendizaje artificial, es un área de gran interés en el campo de la inteligencia artificial. En otros campos, como la biología, la psicología y la filosofía, la naturaleza del aprendizaje también se ha estudiado en relación con los sistemas biológicos y especialmente con los humanos.

Comprender el aprendizaje (como el aprendizaje humano) de una manera que permita a las computadoras reproducir aspectos de ese comportamiento es un objetivo muy ambicioso. Aunque algunos investigadores también han explorado esta posibilidad utilizando otros animales como modelos, los resultados han sido pobres. Como se mencionó, el aprendizaje es un término muy general que se refiere a la forma o formas

en que un animal (o máquina) aumenta sus conocimientos y mejora sus habilidades en el entorno.

El proceso de aprendizaje puede por tanto verse como un agente que produce cambios en el sistema de aprendizaje (que, por otro lado, ocurren de forma lenta y adaptativa) pueden ser reversibles o escalables. Estos cambios no sólo implican mejoras en las habilidades y el desempeño de las tareas, sino que también incluyen cambios en la presentación de los hechos conocidos. En este contexto, se hace referencia al sistema de aprendizaje automático (o alumno) como un artefacto (o conjunto de algoritmos), que, para resolver problemas, toma decisiones basadas en la experiencia acumulada, en el caso previamente resuelto, para mejorar el rendimiento. Estos sistemas deben poder manejar una amplia variedad de entradas, que pueden incluir datos incompletos, datos inciertos, ruido, inconsistencia y más.

El aprendizaje automático puede verse como un proceso de dos pasos: En el primer paso, el sistema selecciona las características más relevantes de un objeto o evento y luego las compara con otras características conocidas. La comparación se realiza mediante un proceso de emparejamiento, y cuando existen diferencias significativas, el sistema ajusta su modelo del objeto o evento en función del resultado del emparejamiento. El aprendizaje es crucial porque frecuentemente resulta en mejoras en el funcionamiento general de un sistema. Al adquirir conocimiento y comprensión, las personas y las organizaciones pueden lograr avances significativos en diversos aspectos de sus operaciones, lo que en última instancia conduce a mejores resultados y logros.

El aprendizaje en sistemas artificiales puede utilizar diversas técnicas para aprovechar el poder computacional de una computadora, sin depender necesariamente de procesos cognitivos humanos. Estas técnicas pueden implicar métodos matemáticos complejos, búsqueda en grandes bases de datos y la creación o modificación de estructuras de representación del conocimiento para facilitar la identificación de información relevante.

Una de las razones principales por las que se diseñan y construyen sistemas de aprendizaje automático es la escasez de experiencia y la naturaleza limitada e incompleta del conocimiento disponible en muchos dominios. A menudo resulta complicado proporcionar a un agente todo el conocimiento necesario, ya que es una tarea compleja

que requiere mucho tiempo y atención especializada para eliminar posibles errores. Por el contrario, a los seres humanos les lleva varios años adquirir habilidades motoras y lingüísticas básicas, e incluso más tiempo para captar conceptos complejos, aprender oficios, comprender convenciones culturales y absorber conocimientos históricos. Además, el aprendizaje humano se personaliza para cada individuo.

Un área de investigación que tiene gran importancia es la investigación sobre el nivel de conocimiento adquirido por un alumno en función de su conocimiento existente. Esto incluye examinar cómo el conocimiento previo del alumno puede contribuir al proceso de hacer inferencias, circunscribir limitaciones y, en última instancia, adquirir nuevos conocimientos. Además, existe una fuerte motivación para aprovechar el poder computacional de las computadoras como un activo valioso en la toma de decisiones. Si un sistema es capaz de aprender de experiencias pasadas, de forma similar a como lo hacen los humanos, entonces la utilidad y eficacia de dicha herramienta aumentan significativamente.

Los procesos en las finanzas

Un escenario en el que la automatización de procesos es particularmente crucial es durante eventos inconvenientes como enfermedades o accidentes. Una empresa se basa en la organización de puestos profesionales y la adecuada distribución de responsabilidades entre sus empleados. Cuando un puesto queda vacante, generalmente se espera que otros miembros del equipo o individuos puedan cubrir temporalmente las tareas necesarias hasta que se cubra el puesto. Si bien, en determinadas situaciones, como durante una pandemia, varios trabajadores pueden enfermarse y dejar numerosas tareas desatendidas. En tales casos, los procesos automatizados alivian significativamente la gravedad de este problema, haciéndolo más fácil tanto para las personas responsables como para los delegados que deban intervenir durante situaciones de emergencia.

Para garantizar que las empresas operen sin problemas y de manera eficiente, han establecido procesos para monitorear, regular y facilitar diversas actividades. Estos procesos constan de una serie de acciones o tareas que se llevan a cabo para lograr un resultado específico. Gracias a las computadoras se pueden ejecutar procesos complejos de manera eficiente, lo que lleva a un mejor desempeño en la organización. Por ejemplo,

si necesita localizar una información específica en archivos, es mucho más eficaz utilizar un programa informático que hacerlo manualmente.

De manera similar, tareas como análisis estadístico, cálculos matemáticos y evaluaciones financieras se pueden completar mucho más rápido a través de procesos automatizados en comparación con depender de un equipo de trabajadores. Como resultado, el uso de tecnología y lenguajes de programación en los procesos de las empresas es cada vez más común. Vale la pena señalar que una parte importante de las tareas realizadas por las instituciones financieras son de naturaleza regulatoria, es decir, son encomendadas por organizaciones que supervisan el buen funcionamiento de estas empresas.

En la amplia gama de empresas actuales, existe un tipo particular de empresa que se centra principalmente en el ámbito de la banca, los valores y los seguros. Estas instituciones financieras brindan productos y servicios que están legalmente clasificados como financieros, lo que, en términos más simples, significa servicios relacionados con el manejo y administración de los fondos de los clientes. El sector bancario es particularmente reconocido, sin embargo, también hay instituciones financieras no bancarias que desempeñan un papel importante, entre ellas los bancos populares de ahorro y préstamo, las compañías de seguros, las casas de bolsa, los sistemas de ahorro para el retiro, los fondos financieros, los fideicomisos y muchos otros.

Asimismo, la importancia que se otorga a la protección de los datos de los usuarios es igualmente primordial, ya que cualquier violación de esta práctica constituye una infracción de la ley y conlleva graves consecuencias para los involucrados. La protección de la información personal se considera una obligación legal, destacando la importancia de mantener la confidencialidad e integridad de los datos sensibles. Al respetar regulaciones estrictas, estas instituciones no sólo garantizan la seguridad de los activos financieros de las personas sino que también contribuyen activamente a la prevención de actividades ilícitas como la financiación del terrorismo y el lavado de dinero.

En consecuencia, las estrictas regulaciones impuestas a las instituciones financieras cuando manejan el dinero de otras personas sirven como una garantía crucial para los consumidores, garantizando la adecuada gestión y protección de sus fondos. Estas regulaciones se extienden más allá de la seguridad monetaria para abarcar la

salvaguardia de los datos de los usuarios, con sanciones severas por cualquier infracción. Al cumplir esta función, estas instituciones desempeñan un papel vital en la lucha contra las prácticas ilícitas y el mantenimiento de la integridad del sistema financiero. Así, para garantizar el manejo adecuado de los fondos de otras personas, estas instituciones están sujetas a regulaciones estrictas que tienen como objetivo proteger a los consumidores y prevenir casos de robo o actividades fraudulentas. Estas regulaciones no sólo se centran en salvaguardar los activos financieros sino que también priorizan la seguridad y privacidad de los datos de los usuarios.

Las instituciones financieras son establecimientos que acuerdan adquirir fondos y capital de fuentes externas, como individuos, corporaciones u otras instituciones financieras, a cambio de prestar servicios específicos. Uno de esos servicios lo ofrecen las empresas de corretaje, que brindan a sus clientes la oportunidad de participar en la compra y venta de acciones y otros instrumentos financieros dentro del mercado. Por el contrario, las aseguradoras ofrecen protección y cobertura financiera en caso de pérdidas específicas. Asimismo, los bancos desempeñan un papel crucial a la hora de almacenar el dinero de los clientes y facilitar el acceso a él a través de tarjetas de débito o cajeros automáticos. Los fondos adquiridos de los clientes normalmente se asignan y gestionan mediante diversos métodos de inversión, siendo el enfoque principal prestar estos fondos en forma de créditos. Estos créditos pueden servir como apoyo a inversiones o como solución a desafíos financieros imprevistos. Las inversiones pueden abarcar una amplia gama de actividades, desde utilizar una pequeña cantidad de dinero para comprar bienes con la intención de venderlos en el futuro, hasta asignar presupuestos sustanciales para proyectos inmobiliarios a gran escala. De manera similar, las dificultades financieras imprevistas pueden variar, desde la necesidad de fondos inmediatos para realizar una compra poco antes de recibir un sueldo hasta la refinanciación de una deuda hipotecaria para evitar el riesgo de perder la casa.

Cuando se trata de dinero, es fundamental manejar todos los procedimientos con extrema precaución y cumplir con los requisitos reglamentarios con la mayor seriedad. De manera similar, el enfoque para lograr rentabilidad en los negocios no debería implicar una asunción excesiva de riesgos, ya que el dinero en juego pertenece principalmente a los usuarios. Para abordar eficazmente este tema, las instituciones financieras han adoptado diversas metodologías y procedimientos a seguir, que pueden

variar según cada departamento de la empresa. Estas organizaciones pueden abarcar una amplia gama de áreas, como derecho, contabilidad y servicios informáticos, entre otras. Específicamente, este artículo se centrará en aprovechar la programación informática para automatizar, mejorar y/o simplificar la implementación de ciertos procesos de toma de decisiones y gestión de riesgos.

En el mundo actual, existe una gran cantidad de conocimientos e información valiosa oculta en grandes volúmenes de datos. Con la llegada de la revolución digital, ahora se pueden crear aplicaciones más fácilmente para darle sentido a esta información, gracias al apoyo y los avances de la tecnología. El aprendizaje automático, también conocido como machine learning en español, es un campo específico dentro de la inteligencia artificial que se enfoca en el desarrollo e implementación de algoritmos que tienen la capacidad de aprender de un conjunto de datos determinado. Este campo requiere una comprensión integral de la programación, la estadística, las matemáticas y cualquier otra rama relacionada que esté relacionada con los datos que se procesan, que puede abarcar una amplia gama de campos como los negocios y la medicina. Los algoritmos del aprendizaje automático están diseñados específicamente para sacar conclusiones basadas únicamente en el conjunto de datos proporcionado.

El aprendizaje automático abarca varios tipos, incluido el aprendizaje supervisado, no supervisado, profundo y de refuerzo. El aprendizaje supervisado, implica la tarea de clasificar o hacer una regresión de un conjunto de datos, mientras que el aprendizaje no supervisado se centra en descubrir patrones ocultos dentro de los datos. En el caso del aprendizaje por refuerzo, un modelo toma la forma de un agente que explora un espacio desconocido y determina acciones mediante prueba y error. El agente aprende de las recompensas y sanciones que recibe en función de sus acciones. Por otro lado, los algoritmos de aprendizaje profundo se basan en redes neuronales artificiales,

que constan de múltiples capas de procesamiento. Estas redes tienen la capacidad de aprender representaciones de datos en distintos niveles de abstracción.

En los últimos tiempos, las empresas han reconocido cada vez más la importancia de poseer estas habilidades, ya que han demostrado ser fundamentales para maximizar las ganancias, minimizar los riesgos, mejorar las estrategias comerciales, impulsar la productividad e incluso identificar enfermedades. Para lograr estos resultados, se debe seguir un procedimiento específico, que implica adquirir los datos, preprocesarlos y, en



última instancia, convertirlos en información valiosa. Convencionalmente, el conjunto de datos se divide en dos subconjuntos: datos de entrenamiento y datos de prueba. Los datos de entrenamiento se emplean para entrenar el algoritmo de aprendizaje y determinar los parámetros del modelo, mientras que los datos de prueba sirven para evaluar el rendimiento del modelo. Las diversas aplicaciones del aprendizaje automático se ilustran vívidamente en la figura anterior (Heros Cárdenas, 2022).

Los datos

Para considerar que un conjunto de datos es de calidad, debe ser útil para el análisis y contener valores precisos. Deben evitarse datos poco realistas, la coherencia también es importante, lo que significa que si varias fuentes proporcionan la misma información, esta debe representarse de forma coherente. Asimismo, los datos deben estar actualizados para garantizar el conocimiento más fidedigno. Deben minimizarse la redundancia y la información irrelevante, ya que pueden dificultar el análisis. Los valores faltantes son comunes al recopilar información, pero tener demasiados valores faltantes puede hacer que los datos estén incompletos y no sean aptos para el análisis.

Los datos numéricos y de texto se explican por sí solos, mientras que los datos categóricos representan diferentes categorías o grupos, que pueden tener o no un orden lógico. Por ejemplo, el género se puede representar mediante datos categóricos, asignando 0 al género femenino y 1 al género masculino. Si se necesita un orden lógico, una variable que represente la satisfacción del cliente puede usar 0 para insatisfecho, 1 para neutral y 2 para satisfecho. Los conjuntos de datos de alta calidad se someten a un proceso exhaustivo para lograr mejores resultados. Una vez que se establece la idoneidad, se realiza un análisis estadístico y gráfico para determinar el algoritmo óptimo para extraer conocimiento.

En algunos casos, es posible que sea necesario transformar los datos a otra escala o representarlos con valores diferentes sin perder la información original. Hoy en día, existen numerosos conjuntos de datos reales disponibles que contienen información valiosa relacionada con diversos campos. Los ejemplos incluyen páginas oficiales de instituciones como el Banco de instituciones financieras, así como plataformas como Kaggle y Yahoo Finance. Los datos son el ingrediente esencial para entrenar algoritmos de aprendizaje automático. Proporciona la información bruta que los algoritmos necesitan para procesar y convertir en conocimiento. Sin embargo, obtener este conocimiento no es una tarea sencilla, ya que no todos los datos pueden proporcionar la información y las respuestas esperadas. Los datos pueden ser cuantitativos o cualitativos y se transforman en variables numéricas, categóricas y textuales. Estas variables permiten manipular la evidencia recopilada para diferentes objetivos, como análisis, procesamiento o creación de modelos.

En este contexto, las técnicas de estadística descriptiva o análisis de datos exploratorios ayudan a presentar datos de manera eficaz al resaltar su estructura subyacente. Existen numerosos métodos sencillos y cautivadores para representar visualmente datos a través de gráficos, que facilitan la identificación de patrones significativos y anomalías imprevistas. Otro enfoque para describir datos implica condensarlos en unos pocos valores numéricos que capturen efectivamente su esencia, minimizando al mismo tiempo cualquier posible distorsión o pérdida de información.

El paso inicial en cualquier análisis de datos debe implicar explorar los datos. Pero ¿por qué no deberíamos simplemente analizar los datos de inmediato? Bueno, para empezar, las computadoras pueden ser rápidas, pero carecen de la capacidad de pensar

críticamente como lo hacen los humanos. Simplemente siguen las instrucciones que les damos y operan en función de los datos proporcionados. Por lo tanto, si hay errores o patrones inesperados en los datos, la computadora los procesará sin darse cuenta de que son incorrectos o inusuales. Por eso es fundamental realizar un análisis exploratorio de los datos previamente, para identificar cualquier anomalía o problema que pueda afectar la precisión y confiabilidad del análisis posterior.

De forma general, los datos suelen estar incompletos y no proporcionan una comprensión completa de una situación. Por tanto, es fundamental emplear métodos que nos permitan extraer información significativa de los datos que observamos. A pesar de la compleja teoría matemática detrás de ellas, algunas técnicas de análisis de datos son sorprendentemente fáciles de aprender y utilizar. Incluso los estadísticos enfrentan desafíos cuando manejan listas de datos. Afortunadamente, existen numerosos métodos estadísticos disponibles para ayudarnos a descubrir las características significativas e intrigantes de nuestros datos, aplicables en diversos campos del conocimiento.

Estos métodos deben utilizarse durante todo el proceso de investigación, desde su inicio hasta su conclusión. Si bien la estadística suele asociarse con el análisis de datos, es importante reconocer que también abarca aspectos cruciales relacionados con el diseño de la investigación. La elección de un método de análisis para un problema depende tanto del tipo de datos disponibles como de la forma en que fueron recopilados.

Debido al profundo impacto de los datos estadísticos y las conclusiones obtenidas a través de la metodología estadística en diversos campos de la actividad humana, especialmente en relación con la investigación en salud pública, es crucial ser cautelosos con la correcta aplicación y validez de los métodos estadísticos en trabajos científicos e informes técnicos. Esta preocupación surge porque la aplicación incorrecta de métodos estadísticos puede llevar a conclusiones incorrectas, lo que puede tener consecuencias importantes.

Asimismo, no todos los lectores tienen la experiencia necesaria para detectar errores, lo que genera un "ruido" significativo en la literatura científica. Se ha observado que los lectores sin formación metodológica tienen más probabilidades de aceptar la validez de las conclusiones publicadas en revistas de prestigio. Por lo tanto, estudiar estadísticas y adoptar un enfoque estadístico permite a las personas evaluar de manera

objetiva y efectiva la relevancia y adecuación de la información que reciben, como tablas, gráficos, porcentajes y tasas. Aunque también es necesario un conocimiento profundo del tema para una interpretación precisa, la formación básica en estadística mejora la comprensión de la información cuantitativa incluso para aquellos que no se especializan en el campo.

El análisis exploratorio de los datos

Aplicar un algoritmo de aprendizaje automático a un conjunto de datos recopilados no garantiza la obtención de conocimientos. De hecho, esto puede tener consecuencias desastrosas, incluso si los resultados inicialmente parecen prometedores. Es importante considerar el contexto y los diversos factores asociados con cada problema para poder extraer completamente el conocimiento contenido en los datos.

Dada la amplia y diversa gama de técnicas disponibles, no es aconsejable aplicarlas todas a ciegas en busca del resultado óptimo. Más bien, es esencial identificar el modelo apropiado que se alinee con las necesidades específicas y el contexto del problema en cuestión. Para afrontar este desafío, es fundamental comprender y analizar a fondo los datos disponibles. Sin embargo, no existe un enfoque único para realizar análisis de datos exploratorios, ya que depende en gran medida de la naturaleza de la información recopilada.

El objetivo principal del análisis exploratorio es examinar las características de la recopilación de datos y recopilar información sobre la misma. Esto se logra utilizando gráficos y cálculos estadísticos para identificar relaciones entre los diferentes atributos. Esta fase no solo genera posibles soluciones sino que también proporciona una perspectiva única sobre los resultados obtenidos al aplicar técnicas de aprendizaje automático.

El proceso de análisis consta de tres técnicas principales: análisis univariado, bivariado y multivariado. El análisis univariado se centra en comprender las características clave de cada variable individual, mientras que el análisis bivariado mide la relación entre pares de variables. Por último, el análisis multivariado tiene como objetivo descubrir relaciones entre un grupo de atributos. Analizar múltiples variables puede resultar más complejo debido a las diversas combinaciones que pueden surgir. Por ejemplo, al examinar dos variables, ambas pueden ser numéricas, ambas categóricas o

una puede ser numérica mientras la otra es categórica. Al analizar múltiples variables, cada par de variables se examina por separado y los resultados se representan en un diagrama o gráfico que contiene una matriz de resultados.

El análisis de datos exploratorio implica la utilización de una variedad de representaciones gráficas y medidas estadísticas para examinar los atributos de los datos. Sin embargo, es importante señalar que no todas las métricas ofrecen información valiosa. Por lo tanto, al realizar un proyecto, es crucial seleccionar sólo aquellas métricas que sean relevantes y puedan contribuir efectivamente a lograr los objetivos deseados. En este sentido, el cuadro anterior proporciona una descripción general completa de las medidas estadísticas clave utilizadas en el análisis de datos exploratorios.

Media	$\bar{\mu} = \frac{x_1 + x_2 + \dots + x_m}{m} = \frac{1}{m} \sum_{i=1}^m x_i$
Mediana	$x^{med} = x_{(m+1)/2} \text{ si es impar y } x^{med} = \frac{x_{m/2} + x_{(m+2)/2}}{2} \text{ si es par.}$
Moda	Corresponde al pico más alto de la curva de densidad empírica para un atributo. En otras palabras es el valor más repetido.
Rango	$x^{rango} = x^{max} - x^{min}$
Varianza	$\bar{\sigma}^2 = \frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{\mu})^2$
Desviación estándar	$\sqrt{\bar{\sigma}^2}$
Covarianza	$cov(a_j, a_k) = \frac{1}{m-2} \sum_{i=1}^m (x_{ij} - \bar{\mu}_j)(x_{ik} - \bar{\mu}_k)$
Correlación	$corr(a_j, a_k) = \frac{cov(a_j, a_k)}{\bar{\sigma}_j \bar{\sigma}_k}$

Así, se tiene que:

- m es el número total de observaciones en el atributo;
- x_i representa la i -ésima observación;
- x^{max} es el valor observado más grande (sólo aplica en variables numéricas);
- x^{min} es el valor observado más pequeño (sólo aplica en variables numéricas);
- a_j, a_k son un par de variables a comparar.

Tipos de Datos

Características:

- La unidad de análisis u observación, se refiere al tema o entidad que se está estudiando. Este tema puede abarcar diversas entidades como individuos, familias, países, regiones, instituciones o cualquier otro objeto de interés.
- La variable, se refiere a cualquier aspecto o rasgo que queramos medir o registrar sobre un objeto o sujeto. Es algo que puede cuantificarse o representarse mediante un valor numérico cuando se mide u observa.
- El valor de una variable, observación o medida es la representación numérica que describe una característica específica de interés dentro de una unidad de observación determinada.
- Un caso o registro se refiere a la recopilación de mediciones realizadas en una unidad de observación específica. Esta unidad podría ser un individuo, un grupo, una organización o cualquier otra entidad que se esté estudiando.

Caso	Sexo	Lugar nacimiento	Edad	PAS
1	F	J1	35	110
2	M	J2	28	120
3	M	J2	59	136

↑
VARIABLE

↘
OBSERVACIÓN

← **REGISTRO**

El sexo, el lugar de nacimiento, la edad y la presión arterial sistólica son factores que contribuyen a describir a un individuo. Estas variables abarcan el sexo del individuo, el lugar donde nació, su edad actual y su lectura de presión arterial. Cada una de estas variables tiene valores específicos únicos para la persona en cuestión, lo que mejora aún más su descripción general.

Al realizar una investigación, el objetivo es analizar el impacto de una o varias variables (conocidas como variables independientes) sobre una o más variables de interés (conocidas como variables dependientes). Un excelente ejemplo de esto se observa en los experimentos, donde el investigador manipula deliberadamente las condiciones experimentadas por los sujetos (variable independiente) y posteriormente observa y analiza los efectos resultantes sobre características o condiciones específicas dentro de los propios sujetos (como la aparición o alteración de ciertos rasgos, características o circunstancias).

Al comenzar a gestionar un conjunto de datos, es fundamental determinar la cantidad de variables registradas y el método de registro para cada variable. Esta información ayudará a diseñar la estrategia de análisis adecuada. En el ejemplo mencionado anteriormente, las variables se pueden clasificar en tipos numéricos y categóricos, con algunas variables representadas por números y otras por letras que indican categorías. Es importante mencionar que varios autores pueden utilizar diferentes criterios para clasificar los datos.

Los datos categóricos: cualitativos

Las variables categóricas son el resultado de documentar la presencia de un determinado atributo. Al diseñar un estudio de investigación, es fundamental definir claramente las categorías de una variable cualitativa. Estas categorías deben ser exclusivas y cubrir todas las posibilidades. Esto significa que cada observación debe clasificarse sin ambigüedades en una categoría y debe haber una categoría para cada individuo.

Es importante considerar todos los escenarios potenciales al crear variables categóricas, incluidas opciones como "No sabe/No responde", "No registrado" u "Otro". Estas categorías adicionales aseguran que todos los individuos observados puedan clasificarse adecuadamente según los criterios de la variable. Los datos categóricos se pueden clasificar además como dicotómicos, nominales u ordinales:

- Los datos dicotómicos con dos categorías: La unidad de observación se puede clasificar en dos grupos distintos. Normalmente, esta categorización representa la presencia o ausencia de un atributo específico, con el código 0 asignado a la ausencia y el código 1 asignado a la presencia. Por ejemplo, se puede clasificar a las personas como hombres o mujeres (a), embarazadas o no (b), fumadores o no fumadores (c) e hipertensos o normotensos (d). Es importante señalar que los ejemplos (a) y (b) abarcan todas las categorías posibles, mientras que los ejemplos (c) y (d) son versiones simplificadas de categorías más complejas. En el ejemplo (c), la asignación de exfumadores no está clara, y en el ejemplo (d), se tuvo que establecer un criterio de corte para convertir una variable numérica en categórica.
- Más de dos categorías:

- Categorías Nominales se refieren a categorías que no tienen un orden o jerarquía clara. Ejemplos de categorías nominales incluyen país de origen, estado civil y diagnóstico.
- Por otro lado, las categorías ordinales tienen un orden o jerarquía natural entre las categorías. Por ejemplo, al considerar los hábitos de fumar, las categorías pueden variar desde no fumar hasta ser exfumador, fumar menos o igual a 10 cigarrillos por día y fumar más de 10 cigarrillos por día. De manera similar, al evaluar la gravedad de una patología, las categorías pueden variar desde ausente hasta leve, moderada y grave.

Aunque los datos ordinales se pueden representar numéricamente, como en los estadios I a IV del cáncer de mama, es importante señalar que los valores numéricos no reflejan una diferencia proporcional en la variable subyacente. Por ejemplo, un paciente en estadio IV no tiene un pronóstico dos veces peor que un paciente en estadio II, ni la diferencia entre el estadio I y II es la misma que entre el estadio III y IV. Esto contrasta con variables cuantitativas como la edad, donde 40 años es el doble que 20 años y una diferencia de 1 año es consistentemente significativa en todo el rango de valores.

- Debido a las diferencias inherentes entre variables cualitativas y cuantitativas, es crucial manejar las variables cualitativas con cuidado, especialmente cuando han sido codificadas numéricamente. Tratarlos como números y calcular promedios u otras estadísticas numéricas puede dar lugar a interpretaciones incorrectas. En la práctica clínica, las escalas se utilizan a menudo para definir grados de síntomas o enfermedades, como el uso de 0, +, ++, +++ para indicar una gravedad creciente. Es importante establecer definiciones operativas claras para este tipo de variables y evaluar su confiabilidad para garantizar que diferentes observadores clasifiquen al mismo paciente en la misma categoría.

Los datos numéricos

Se considera numérica una variable, si el resultado de la observación o medición corresponde a un valor numérico. Estas variables se pueden clasificar en diferentes tipos según sus características y propiedades específicas:

- La variable es discreta, lo que significa que sólo puede tener un conjunto específico de valores. Normalmente, estos valores se determinan mediante conteo. Por ejemplo, esto podría incluir variables como la cantidad de personas en un hogar, la cantidad de procedimientos quirúrgicos realizados o la cantidad de casos reportados de una condición médica particular.
- Variable continua, suele ser medidas y expresadas en unidades. En teoría, una medición puede adoptar un conjunto infinito de valores posibles dentro de un rango determinado. En la práctica, los valores posibles de una variable están limitados por la precisión del método de medición o modo de registro. Por ejemplo: altura, peso, pH, niveles de colesterol en sangre.

Comprender la distinción entre datos discretos y continuos es crucial a la hora de determinar qué método de análisis estadístico emplear, ya que ciertos métodos suponen que los datos son continuos. Tomemos como ejemplo la variable edad. La edad suele considerarse una variable continua, pero si se mide en años, se vuelve discreta. En estudios con adultos con un rango de edad de 20 a 70 años, tratar la edad como continua no plantea problemas debido a la gran cantidad de valores potenciales. Si bien, cuando se trata de niños en edad preescolar, registrar la edad en años requeriría tratarla como discreta.

Por el contrario, si la edad se registra en meses, se puede tratar como continua. De manera similar, la variable del número de pulsos por minuto es técnicamente discreta, pero a menudo se trata como continua debido a la multitud de valores posibles. Vale la pena señalar que los datos numéricos, ya sean discretos o continuos, pueden transformarse en datos categóricos y analizarse en consecuencia. Si bien este enfoque es correcto, puede que no siempre sea eficiente, por lo que es preferible registrar el valor numérico de la medición. Esta práctica permite un análisis e interpretación más precisos:

- Cuando analiza la variable como numérica, permite un análisis estadístico más sencillo y efectivo.
- Generar clasificaciones adicionales basadas en estándares alternativos.

Sólo hay determinadas situaciones en las que es más ventajoso categorizar datos numéricos en lugar de registrarlos como valores cuantitativos. Esto ocurre principalmente cuando se reconoce que la medición tiene una imprecisión inherente,

como la cantidad de cigarrillos consumidos diariamente o la cantidad de tazas de café consumidas en una semana.

Los otros tipos de datos

- Los porcentajes:

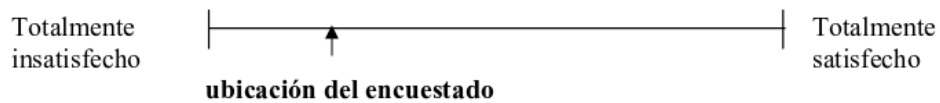
Los porcentajes se calculan dividiendo dos cantidades. Por ejemplo, puede calcular el porcentaje de reducción de la presión arterial después de usar un medicamento o el peso corporal relativo dividiendo el peso observado por el peso deseable. En el primer ejemplo, ambas cantidades se miden al mismo tiempo, mientras que en el segundo ejemplo, el denominador es un valor estándar predeterminado.

Aunque los porcentajes pueden verse como variables continuas, pueden plantear desafíos durante el análisis, particularmente cuando pueden exceder el 100% o estar por debajo del 0% (como en el caso del peso corporal relativo) o cuando pueden resultar en valores negativos (como en el caso del peso corporal relativo). caso de reducción porcentual de la presión arterial). Por ejemplo, si un paciente tiene una presión arterial sistólica (PAS) de 150 mm Hg y experimenta un aumento del 20% en la PAS, alcanzará los 180 mm Hg. Sin embargo, una disminución posterior del 20% lo reducirá a 144 mm Hg. Por lo tanto, es necesario tener precaución al analizar dichos datos.

- Las escalas analógicas visuales:

Cuando se requiere que un individuo exprese el alcance o la intensidad de un atributo no cuantificable, como satisfacción, malestar, salud general, disfrute, consenso, etc., la escala visual analógica se presenta como una herramienta valiosa. Esta técnica permite la adquisición de categorías ordinales, ya que implica presentar al encuestado una línea recta, que generalmente mide 10 centímetros, donde cada extremo de la línea representa los extremos del atributo que se está midiendo. Luego se le pide al encuestado que marque un punto en la línea que mejor represente su percepción personal de su propio estado en relación con el atributo en cuestión. Por ejemplo, si se desea evaluar el nivel de satisfacción

experimentado con un tratamiento en particular, la utilización de la siguiente escala puede resultar beneficiosa.



Estas escalas brindan una valiosa ayuda para evaluar las variaciones dentro de un individuo. Si bien una puntuación única de 3,7 puede no tener un significado significativo por sí sola, una disminución de 2 puntos en la puntuación de un paciente sí ofrece información significativa. Sin embargo, se debe tener precaución al manejar este tipo de datos ya que, a diferencia de los datos numéricos, incluso cuando se expresan como números, la escala de medición subyacente puede diferir entre dos individuos distintos.

- Los scores:

Sirven como medio para evaluar el estado de un individuo teniendo en cuenta diversas variables, normalmente de naturaleza categórica. En entornos clínicos, estas puntuaciones se elaboran considerando los síntomas y signos que presenta un paciente, asignándoles puntuaciones respectivas y posteriormente agregándolas para obtener una puntuación acumulativa que proporcione información sobre la condición general del individuo. En el análisis, los scores deben considerarse y manejarse de la misma manera que se utilizan habitualmente en la práctica, es decir, estableciendo categorías ordinales en lugar de tratarlas como variables numéricas.

- Lo datos censurados:

Una observación censurada se refiere a una situación en la que no se puede obtener la medida exacta, pero somos conscientes de que está por encima o por debajo de un umbral específico. Es decir, tenemos información sobre el valor mínimo o máximo que pueden tomar los datos. Exploremos algunos ejemplos para ilustrar mejor este concepto:

- Al realizar mediciones de oligoelementos, es posible que el nivel del elemento en la muestra sea inferior al que puede detectarse con la técnica elegida. En

tales casos, se dice que los datos han salido de la censura porque se desconoce el valor real del elemento, pero sí tenemos conocimiento de un límite superior para él.

- Se han realizado más investigaciones que se centran en la duración de la supervivencia. En los casos en que los pacientes continúan viviendo más allá de la duración del estudio, se desconoce la duración exacta de su supervivencia. Sin embargo, se establece que su tiempo de supervivencia supera la duración del estudio. Este tiempo de supervivencia se clasifica como censurado por la derecha, ya que sólo disponemos de una estimación mínima del mismo.
- Una investigación posterior tiene como objetivo examinar la duración entre la aparición de una condición médica y su posterior reaparición. En los casos en que los participantes ya no sean parte del estudio por diversas razones como abandono, muerte por causas no relacionadas o cualquier otro factor, pero se confirmó que estaban libres de la condición hasta su último examen, la información relativa al tiempo entre la ocurrencia inicial y la recurrencia posterior se consideran datos censurados correctamente.

La determinación del método de análisis apropiado y válido depende en gran medida del tipo de datos que se analizan, ya que cada método de análisis estadístico se adapta específicamente a un determinado tipo de datos. La diferenciación más significativa radica en la categorización de los datos en formas numéricas y categóricas.

El procesamiento de datos en la computadora

Las computadoras desempeñan un papel crucial en la simplificación de los laboriosos aspectos del análisis estadístico y son capaces de generar cálculos precisos. Sin embargo, es importante reconocer que su uso no garantiza automáticamente la validez y corrección de los resultados obtenidos. En esta discusión, exploraremos las ventajas y desventajas de utilizar computadoras para el procesamiento de datos y también examinaremos varios enfoques para compilar archivos de datos.

Las ventajas:

- La combinación de precisión y rapidez es fundamental cuando se trata de software de alta calidad, ya que garantiza la adquisición oportuna de resultados precisos.
- Una de las ventajas notables de la estadística es su versatilidad, ya que ofrece una amplia gama de técnicas estadísticas que van más allá del alcance de cualquier curso de estadística para abarcarlas por completo.
- Se pueden crear gráficos para representar visualmente los datos originales o los resultados obtenidos, mejorando la capacidad de comprender e interpretar la información.
- Otra ventaja de utilizar una base de datos es su flexibilidad, ya que permite modificaciones sencillas y repetición de análisis. Después de construir la base de datos, es posible realizar pequeños ajustes y volver a analizar los datos. Por ejemplo, se pueden excluir ciertos casos y realizar análisis en subgrupos o estratos específicos. Esta flexibilidad brinda a los investigadores la oportunidad de refinar su análisis y explorar diferentes perspectivas dentro de los datos.
- Crear nuevas variables es una tarea sencilla que se puede realizar de varias maneras. Por ejemplo, se puede calcular la diferencia entre las mediciones tomadas antes y después de un tratamiento, determinar la edad restando fechas de nacimiento, convertir variables numéricas en categóricas, reclasificar variables cualitativas, aplicar transformaciones, etc.
- Una ventaja de ciertos programas es su capacidad para manejar una gran cantidad de datos, sin limitaciones en la cantidad de registros o variables que pueden procesarse.

Las desventajas:

- Hay varios problemas relacionados con errores de software que pueden ocurrir al utilizar paquetes estadísticos. Es importante ser consciente de estos errores, ya que pueden afectar la precisión y confiabilidad de los resultados obtenidos. Entre los paquetes estadísticos más utilizados, hay algunos que se consideran más seguros que otros. Estos incluyen SAS, S-PLUS, STATA y SPSS. Sin embargo, incluso con estos paquetes seguros, aún pueden surgir errores en algunos procedimientos. Para garantizar la calidad del software utilizado, es recomendable verificar la

exactitud de los resultados comparándolos con ejemplos proporcionados en libros o utilizando otro software de alto nivel. De esta manera, se pueden identificar y abordar cualquier discrepancia o inconsistencia, mejorando así la confiabilidad de los hallazgos. En conclusión, los errores de software pueden plantear desafíos al realizar análisis estadísticos. Es crucial seleccionar un paquete estadístico confiable y seguro y validar los resultados obtenidos cotejándolos con fuentes confiables. Al tomar estas precauciones, los investigadores pueden mitigar los riesgos asociados con los errores de software y garantizar la validez de sus análisis.

- Uno de los beneficios de tener una amplia gama de métodos estadísticos para elegir es la capacidad de ser versátiles en nuestro análisis. Si bien, esta ventaja puede convertirse rápidamente en desventaja si no tenemos cuidado. La abundancia de métodos estadísticos puede hacer que resulte tentador utilizar uno inadecuado para nuestro análisis. Por lo tanto, es fundamental que los usuarios comprendan claramente sus propias limitaciones en el conocimiento estadístico y utilicen únicamente métodos que comprendan plenamente. En los casos en que el problema en cuestión parezca requerir el uso de métodos desconocidos, se recomienda encarecidamente buscar la orientación de un estadístico profesional. Al hacerlo, podemos garantizar que nuestro análisis sea preciso y confiable.
- El concepto de caja negra hace referencia a la posibilidad de perder contacto con los datos. Cuando el análisis se realiza automáticamente, existe el peligro potencial de pasar por alto los aspectos más importantes de los datos o de no capturar información sobre personas que exhiben un comportamiento inusual. Esto resalta la importancia de monitorear e interpretar activamente los datos para evitar cualquier posible pérdida o supervisión. La precisión y validez de los resultados obtenidos de un método de análisis estadístico dependen de la calidad del archivo de datos. En el caso de que los datos estén registrados de manera inadecuada o contengan discrepancias y el investigador no pueda identificar estos problemas, las conclusiones extraídas del análisis serán erróneas, independientemente de la complejidad y el refinamiento de las técnicas estadísticas empleadas. Es crucial que los investigadores evalúen diligentemente y garanticen la integridad de los datos con los que están trabajando, ya que cualquier deficiencia o inexactitud en

los datos puede afectar significativamente la confiabilidad de los hallazgos del estudio.

- La precisión de los resultados depende del calibre del archivo de datos. En el caso de que los datos estén documentados de manera inexacta o contengan discrepancias que pasen desapercibidas para el investigador, los hallazgos inevitablemente serán defectuosos, independientemente de la complejidad y delicadeza de la técnica de análisis estadístico empleada.

La estrategia en el análisis de datos con el empleo de programas de computación

- La definición de variables:

Codificar todas las variables categóricas con números puede facilitar la carga de datos, haciéndola más rápida y precisa. Además, asignar etiquetas a cada categoría ayuda a identificarlas fácilmente y mejora la facilidad de uso de los resultados estadísticos. Cuando se trata de fechas, es fundamental determinar el formato que se utilizará para la variable, como día/mes/año, mes/día/año o día-mes-año. Sin embargo, es importante tener en cuenta que es posible que algunos paquetes de software no reconozcan formatos de fecha específicos y, en su lugar, traten los valores de fecha como caracteres alfanuméricos (texto). En tales casos, estas fechas no se pueden utilizar en operaciones algebraicas, ya que no se consideran valores numéricos.

Al registrar variables numéricas, es importante mantener el mismo nivel de precisión que cuando se obtuvieron los datos originalmente, sin redondear ni categorizar los valores. Si se observa al mismo individuo varias veces, como durante el seguimiento del embarazo o en un ensayo, se deben recopilar mediciones repetidas para esa persona específica. Cada visita o medición no debe tratarse como un registro separado, ya que sería incorrecto tratar estas observaciones como si pertenecieran a individuos diferentes.

El análisis de este tipo de datos requiere técnicas estadísticas especializadas conocidas como técnicas de medidas repetidas. Para simplificar el proceso de carga de datos y garantizar la precisión, es recomendable asignar un nombre de no más de 10 letras a cada variable. Si es necesario, se puede asignar un nombre

completo a la variable mediante una etiqueta. Vale la pena señalar que algunos paquetes de software solo aceptan nombres de variables con un máximo de 8 letras, truncando los caracteres adicionales. Ciertos caracteres, como los puntos, no están permitidos en los nombres de variables y no se deben dejar espacios dentro de los nombres.

- La consistencia de los datos:

Pueden surgir errores en varias etapas al tratar con mediciones y datos. Estos errores pueden ocurrir durante el proceso de medición inicial, al registrar los datos en fuentes como registros médicos, durante la transcripción a una hoja de cálculo o durante el armado de una base de datos. A menudo es difícil determinar si los datos son totalmente exactos, pero es importante asegurarse de que sean al menos plausibles. Aquí es donde entra en juego el concepto de coherencia de los datos. El objetivo no es necesariamente corregir todos los errores, sino más bien identificar y abordar los más evidentes. La coherencia de los datos tiene como objetivo identificar y, si es posible, rectificar estos errores dentro de los datos. El primer paso en este proceso es verificar si hay errores tipográficos. En el caso de archivos más pequeños, es una práctica común imprimirlos y revisarlos detenidamente. Sin embargo, para archivos más grandes, se recomienda escribir los datos dos veces y comparar ambas versiones. Cierta software, como EpiInfo, ofrece un procedimiento de "VALIDAR" que facilita esta comparación y genera una lista de las discrepancias encontradas entre las dos versiones.

- Los datos categóricos: En este escenario, es sencillo verificar la validez de todos los valores de las variables porque existe un rango predeterminado de valores posibles para la variable. Por ejemplo, consideremos la variable "Grupo sanguíneo" que puede tener valores de 0, A, B o AB. Para garantizar la precisión, podemos crear una tabla de frecuencia para cada variable categórica y verificar que las categorías se alineen con las categorías predefinidas. Vale la pena señalar que ciertos paquetes de software distinguen entre letras mayúsculas y minúsculas, por lo que pueden tratar "a" como una categoría distinta de "A" en el caso de los grupos sanguíneos. Es aconsejable compilar una lista completa de tablas de frecuencia para todas las variables categóricas antes de comenzar el análisis estadístico de los datos.

- Los datos numéricos: Para garantizar la precisión, es importante proponer el rango de valores esperado o posible para cada variable. Por ejemplo, la edad materna al momento del parto podría oscilar entre 12 y 50 años, mientras que la presión arterial sistólica podría oscilar entre 70 y 250 mg Hg. Un error común es colocar mal la coma o el punto decimal, lo que puede generar datos inexactos. Vale la pena señalar que los valores fuera del rango esperado no son necesariamente incorrectos, sino más bien improbables o incluso imposibles. Desafortunadamente, determinar el límite exacto entre valores improbables e imposibles puede resultar un desafío. Sin embargo, si hay evidencia de error, es importante corregir valores improbables pero posibles. Además, al importar una base de datos desde un programa de software diferente, es fundamental verificar que se haya mantenido el tipo de variable. En concreto, las variables numéricas no deben transformarse en texto si no se reconoce el indicador del símbolo decimal (coma o punto). Es importante tener en cuenta que no se pueden realizar operaciones algebraicas con variables de texto.
- El chequeo lógico: Hay cierta información que sólo se releva en ciertos casos. Un caso muy simbólico es, número de embarazos, sólo es relevante si sexo = femenino, pero para sexo = masculino, esta variable debería ser '.' o "no corresponde". Los datos deben satisfacer los criterios de inclusión y exclusión del estudio. Ejemplo: Estudio de agentes anti-hipertensivos, los pacientes que entran en el estudio deben tener valores de la presión arterial dentro de un cierto rango al ingreso. Evaluar la consistencia de los datos es algo más complicado cuando existen valores de algunas variables que dependen de valores de otras variables. Existen combinaciones de valores de ciertas variables que son inaceptables, aun cuando cada una de ellas se encuentre dentro de límites razonables. El investigador debe proponer chequeos lógicos que permitan detectar aberraciones en los datos. Ejemplos: es poco probable que un sujeto se ubique en el percentil 5 de presión diastólica y en el percentil 95 de presión sistólica, o es poco probable que un niño nacido con 30 semanas de gestación pese 3800 g. Cuando una variable se mide varias veces en la misma unidad de observación puede graficarse a lo largo del tiempo para ver si el comportamiento es acorde a lo esperado.

- Las fechas: Los intervalos de tiempo entre eventos se determinan utilizándolos como punto de referencia. Por ejemplo, con estos se puede calcular la edad de un paciente en el momento de la consulta o la duración de la supervivencia. Para garantizar la precisión, es fundamental examinar si las fechas se encuentran dentro de plazos razonables. Esto incluye evaluar si las fechas de las evaluaciones se alinean con el período de desarrollo de la investigación o si las fechas de nacimiento cumplen con los criterios de edad requeridos para la inclusión y exclusión. Asimismo, es de suma importancia secuenciar correctamente las fechas de cada individuo, como nacimiento, hospitalización y muerte.
- Los datos faltantes: Otra cuestión que es necesario abordar es cómo se manejan los datos faltantes. Cuando existe un espacio en blanco en la información cargada, es importante considerar que ciertos paquetes estadísticos pueden asignar un valor de cero a ese espacio en blanco. A veces, a los datos faltantes se les asignan valores poco realistas como 99999 o valores negativos para datos que solo pueden ser positivos. El problema surge cuando estos valores atípicos no se excluyen durante el análisis, ya que los resultados serán inexactos ya que cualquier programa aceptará el valor cero o 99999 como válido. Sin embargo, EpiInfo aborda este problema representando los datos faltantes con un punto, lo que ayuda a evitar este problema. EpiInfo también ofrece una función llamada CHEK, que garantiza la coherencia de los datos durante el proceso de carga.
- El análisis exploratorio de los datos:

Para analizar adecuadamente los datos, es fundamental crear representaciones visuales como gráficos y tablas. Estas ayudas visuales desempeñan un papel importante en la identificación de patrones de datos inusuales o anormales. El siguiente capítulo estará dedicado exclusivamente a explorar y abordar este aspecto en particular.

Los malos hábitos en el empleo de la computadora

Además de los inconvenientes antes mencionados de depender de computadoras para la gestión de datos, es fundamental abordar ciertos usos indebidos y abusos que deben evitarse:

- Al realizar investigaciones con objetivos vagos y recopilar datos basados en intereses potenciales, los investigadores suelen realizar numerosos análisis estadísticos para identificar diferencias entre grupos o correlaciones entre variables. Si bien, es importante reconocer que en tales análisis existe una alta probabilidad de encontrar relaciones significativas puramente por casualidad, sin que en realidad reflejen ninguna relación verdadera dentro de la población. Los análisis exploratorios sirven como herramientas valiosas para generar nuevas hipótesis, que luego deberían probarse en estudios separados. No es apropiado utilizar el mismo estudio tanto para la generación como para la verificación de hipótesis.
- No es aconsejable someter los datos a análisis estadísticos complejos únicamente porque estén disponibles en software, ya que esto puede no ser lo mejor para el estudio. Más bien, el análisis debe limitarse al mínimo necesario para abordar las preguntas de investigación en cuestión. Realizar análisis más simples ofrece varios beneficios, uno de los cuales es la facilidad de interpretar y comunicar eficazmente las conclusiones resultantes.
- Un problema en el análisis estadístico es la presencia de precisión espuria en los resultados generados por los programas estadísticos. Estos resultados suelen incluir un número significativo de decimales, pero es importante comunicar los hallazgos con la precisión adecuada. Por ejemplo, considere un cálculo en el que el porcentaje se determina dividiendo 17 entre 45 y multiplicando por 100, lo que da como resultado 37,778%. En este caso, sería más apropiado informar el porcentaje como 38% porque la adición de solo un caso más, lo que daría como resultado 18 de 45, modificaría el porcentaje al 40%. Este ejemplo resalta la necesidad de evitar decimales innecesarios y comunicar con precisión los hallazgos en el análisis estadístico.

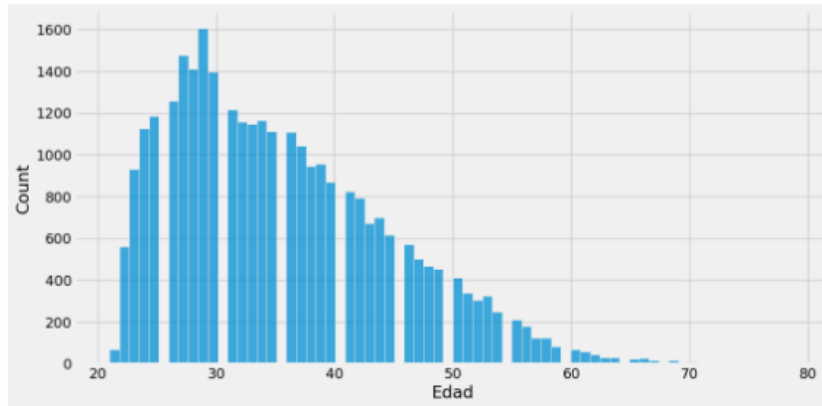
Gráficos en la estadística descriptiva

representa la frecuencia relativa correspondiente. Como cualquier otro gráfico, es fundamental incluir el número total de sujetos para proporcionar una comprensión completa. Este gráfico en particular se conoce como histograma.

El histograma

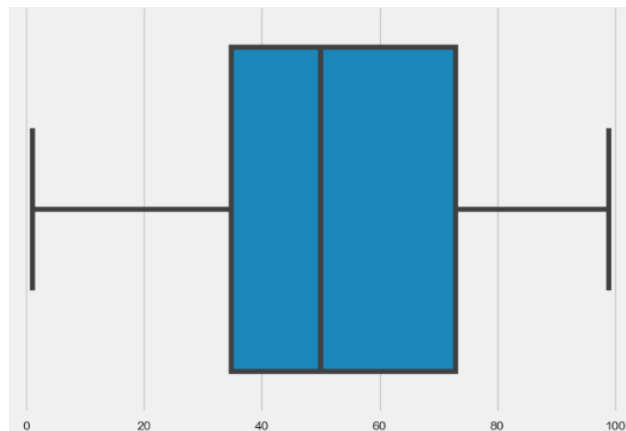
El histograma es ampliamente reconocido como el gráfico más popular para resumir conjuntos de datos numéricos y cumple el mismo propósito que un diagrama de tallo y hoja. Si bien un diagrama de tallo y hoja es ventajoso para preservar los valores de observación individuales, resulta menos práctico para conjuntos de datos más grandes. Aunque la creación manual de un histograma lleva más tiempo en comparación con un gráfico de tallo y hojas, el software estadístico suele ofrecer la opción de generar histogramas. El paso inicial para construir un histograma implica construir una tabla de frecuencias.

El histograma representa visualmente la distribución de datos. Utiliza un eje vertical para mostrar la frecuencia de los valores que aparecen en el eje horizontal. Esto nos permite observar fácilmente cuántas veces aparece un número particular o un rango de números en un conjunto de datos en comparación con otros. Por ejemplo, en la figura que sigue (Heros Cárdenas, 2022), podemos examinar un histograma que muestra el número de individuos de una edad específica en una muestra. Si bien el examen de los datos en su conjunto puede no llevar a conclusiones concluyentes, el histograma nos permite determinar de forma rápida y segura que hay un número significativo de individuos entre 25 y 30 años en nuestra población, mientras que el grupo más pequeño está formado por aquellos personas cercanas a los 70 años.



El gráfico de caja

El diagrama de caja es una representación gráfica que proporciona información sobre un conjunto de datos resaltando sus tres cuartiles. La sección inferior del cuadro, o el lado izquierdo dependiendo de la orientación del gráfico, representa el primer cuartil, que es el punto donde se acumula el 25% de los datos recopilados. La línea media dentro del cuadro indica el segundo cuartil o mediana, donde se acumula el 50% de los datos. Por otro lado, la parte superior o derecha del cuadro representa el tercer cuartil.

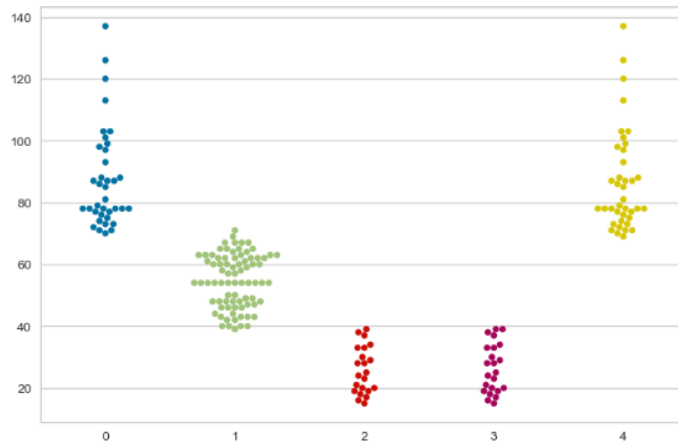


La altura del cuadro se conoce como rango intercuartil, lo que proporciona información sobre la dispersión de los datos. Este tipo de representación es particularmente útil para identificar valores atípicos, que son valores que se encuentran a 1,5 rangos intercuartiles del primer y tercer cuartil hacia los extremos del conjunto de

datos. En la anterior figura podemos observar la variable de gastos, que no presenta valores atípicos ya que no hay observaciones fuera del cuadro. Vale la pena señalar que aproximadamente la mitad de los datos se agrupan alrededor del valor 50 en términos de gasto, con el primer cuartil apareciendo antes de 40 y el tercer cuartil antes de 80.

Los gráficos de gusanos

El diagrama de gusano es una herramienta extremadamente valiosa cuando se trata de comprender visualmente la distribución de datos y cómo se compara con otras variables. Sirve como un medio beneficioso para hacer comparaciones. Al examinar la siguiente figura, se puede observar claramente cómo los datos correspondientes a los gusanos azules y amarillos se distribuyen entre valores más altos, mientras que los gusanos rojos y morados constan de valores más pequeños.



Los gráficos de dispersión

El diagrama de dispersión, también conocido como gráfico de dispersión, sirve como una valiosa herramienta de control y apoyo para evaluar la presencia de una correlación o relación entre dos variables cuantitativas. Su propósito radica en examinar la posible conexión de causa y efecto entre estas variables y evaluar las hipótesis disponibles. Es una herramienta que presenta visualmente una comparación entre dos conjuntos de valores en un gráfico, siendo particularmente útil para representar datos como resultados de encuestas, puntajes de exámenes e información demográfica.

El diagrama de dispersión se emplea en situaciones donde hay una gran cantidad de puntos de datos diversos y el objetivo es enfatizar las similitudes dentro del conjunto de datos. Esta técnica resulta ventajosa cuando se intenta identificar puntos de datos excepcionales o inusuales, así como para obtener información sobre los patrones de distribución generales que exhiben los datos.

Cuando los puntos de datos crean una banda que se extiende diagonalmente desde la parte inferior izquierda hasta la parte superior derecha, es muy probable que exista una correlación positiva entre las dos variables que se analizan. Por el contrario, si la banda se extiende desde la parte superior izquierda hasta la parte inferior derecha, es más probable que se observe una correlación negativa. En situaciones en las que resulta difícil discernir cualquier patrón discernible dentro de los datos, es probable que no exista correlación entre las variables examinadas.

El diagrama de dispersión representa gráficamente la relación entre dos variables. Se utiliza comúnmente para identificar patrones o tendencias en puntos de datos. El diagrama de dispersión consta de un eje X horizontal y un eje Y vertical, donde cada punto de datos se traza según sus valores correspondientes para las dos variables. El gráfico resultante muestra la distribución de los puntos de datos y puede proporcionar información sobre la correlación, si la hay, entre las variables que se analizan. La forma o patrón formado por los puntos de datos en el diagrama de dispersión puede revelar información sobre la fuerza y dirección de la relación entre las variables.

Asimismo, los diagramas de dispersión se pueden mejorar con elementos visuales adicionales, como el color o el tamaño, para representar dimensiones adicionales de los datos. Estas características hacen de los diagramas de dispersión una herramienta valiosa para el análisis y visualización de datos en diversos campos, como la estadística, la economía y las ciencias sociales. Permiten a investigadores y analistas identificar posibles valores atípicos, grupos o tendencias dentro de los datos, facilitando la interpretación y comprensión de relaciones complejas. Al representar visualmente puntos de datos, los diagramas de dispersión brindan una forma clara e intuitiva de explorar y comunicar datos, lo que los convierte en una herramienta esencial para investigadores, educadores y tomadores de decisiones. También, se considera una herramienta valiosa utilizada en el campo de la estadística para examinar y confirmar la presencia de una correlación o asociación entre dos variables cuantitativas. Su propósito es evaluar e investigar la posible

relación causa-efecto entre estas variables y, en última instancia, proporcionar evidencia para apoyar o refutar las hipótesis.

Es particularmente adecuado para analizar e interpretar varios tipos de datos, incluidos resultados de encuestas, puntajes de exámenes académicos e información demográfica. Al utilizar este componente, los usuarios pueden identificar fácilmente patrones, tendencias y correlaciones entre diferentes variables, lo que les permite obtener información valiosa y tomar decisiones informadas basadas en los datos presentados. Ya sea examinando la relación entre la satisfacción del cliente y las características del producto o estudiando el impacto de la edad y los ingresos en el comportamiento de compra, el componente del gráfico de dispersión proporciona un medio completo y visualmente atractivo para comprender conjuntos de datos complejos.

Cuando los puntos de datos se organizan en una banda diagonal que se extiende desde la esquina inferior izquierda hasta la esquina superior derecha, es muy probable que exista una correlación positiva entre las dos variables que se analizan. Por el contrario, si la banda de puntos de datos va desde la esquina superior izquierda hasta la esquina inferior derecha, es muy probable que exista una correlación negativa. En los casos en los que resulta difícil discernir algún patrón discernible entre los puntos de datos, se puede inferir que no existe correlación entre las variables consideradas. En general, los diagramas de dispersión son una herramienta poderosa para visualizar y analizar la relación entre dos variables. Proporcionan una representación clara y concisa de los datos, lo que permite a los investigadores y analistas sacar conclusiones significativas y tomar decisiones informadas basadas en los patrones y tendencias observados en el gráfico (Ortega, 2021).



Los gráficos de tiempo muestran información en diferentes intervalos de tiempo. Los datos representados en estos gráficos consisten en valores numéricos y se supone que ocurren durante un período de tiempo uniforme. Al examinar los gráficos de tiempo, se puede realizar un análisis inicial de las características de los datos de series de tiempo en pruebas y estadísticas básicas. Este análisis permite extraer información valiosa de los datos antes de construir modelos. Los gráficos de tiempo emplean varias técnicas analíticas, incluida la descomposición, Dickey-Fuller aumentado (ADF), correlaciones (ACF/PACF) y análisis espectral.

Capítulo 3

El procesamiento de datos

Los datos son un recurso valioso que puede proporcionar una gran cantidad de conocimientos a quienes los poseen, pero su verdadero potencial depende de la diligencia y la experiencia de los especialistas a la hora de extraer conocimientos de ellos. A menudo, los datos se presentan en su forma cruda, tal como fueron recopilados, lo que puede plantear desafíos para los analistas. A pesar de no ser ampliamente discutido en el contexto del aprendizaje automático, el preprocesamiento de datos es una etapa crucial por la que todo conjunto de datos debe pasar y cada especialista debe reconocer su importancia.



Es fundamental enfatizar la importancia de asegurar la calidad de los datos que servirán de base a cualquier modelo y evaluar su utilidad y confiabilidad. La preparación de datos para el análisis es un tema complejo, pero existen varias técnicas bien respaldadas que pueden mejorar significativamente la calidad de los datos. La implementación de estas técnicas puede mejorar enormemente los datos y tener un impacto sustancial en los resultados resultantes. Contrariamente a la creencia popular, las personas que trabajan con datos, ya sea para análisis o extracción de datos, dedican la mayor parte de su tiempo a la etapa de preprocesamiento de datos. En los últimos años, ha habido un aumento de roles especializados dedicados únicamente al preprocesamiento de datos para su posterior análisis por parte de otros departamentos o equipos.

Índice de Cobertura (B)				
	Banco 1	Banco 2	Banco 3	Banco 4
Cartera total	200.974544	201.946929	213.102568	216.411742
Cartera total	200.974544	201.946929	213.102568	216.411742
Cartera total de créditos comerciales	211.402936	219.34811	227.067873	233.710117
Cartera de empresas	188.631933	198.420656	205.229844	212.130982
Cartera de entidades financieras	4156.14642	4263.00498	4486.71109	4597.94261
Cartera de entidades gubernamentales	n. a.	n. a.	n. a.	n. a.
Cartera de tarjeta de crédito	182.13607	186.104835	194.264094	987.749303

La tabla anterior (Heros Cárdenas, 2022), muestra un pequeño conjunto de datos que presenta varios problemas de datos, incluidos valores atípicos, valores faltantes y entradas duplicadas. Un valor duplicado notable es la variable cartera total, que aparece dos veces. Mantener ambas filas podría tener un impacto negativo en el modelo, por lo que sería innecesario incluir una de ellas. Además, la variable de Cartera de entidades gubernamentales muestra cuadros con n.a., lo que indica información faltante que podría ser valiosa para el modelo. Es importante señalar que identificar datos atípicos requiere una comprensión y un conocimiento profundos del tema que se está estudiando. El índice de cobertura, que indica la adecuación de las reservas bancarias para una cartera específica, es otro factor importante a considerar. Por ejemplo, un índice de cobertura del 200% significa que la institución ha reservado el doble de la cantidad de dinero perdida. Por lo tanto, la inclusión de valores de índice de cobertura del 4000% para la cartera de entidades financieras en la tabla anterior es claramente errónea.

Los valores duplicados

En diversos escenarios, como cuando las empresas recién establecidas necesitan realizar análisis exploratorios de sus datos, o cuando hay presupuestos limitados para la recopilación de información, o incluso cuando obtener información es un desafío, nos encontramos con conjuntos de datos con información limitada. Estos tipos de conjuntos de datos, comúnmente denominados conjuntos de datos pequeños, plantean un mayor riesgo cuando contienen valores duplicados. Esto es particularmente significativo en situaciones donde el cálculo preciso de la probabilidad de que ocurra un evento es esencial, ya que la presencia de valores duplicados puede afectar significativamente los resultados de manera más severa.

Por el contrario, cuando se trata de grandes cantidades de datos, es común encontrar múltiples instancias de información duplicada. En consecuencia, eliminar estos duplicados no sólo mejoraría la calidad general del conjunto de datos sino que también haría que su manejo y análisis fuera más conveniente. Como resultado, rectificar el problema de los datos duplicados puede generar ahorros significativos en los gastos de almacenamiento, lo que en última instancia mejora la eficiencia y la velocidad del procesamiento de la información y facilita la ejecución de varios algoritmos.

Del mismo modo, incluso si asumimos que tener valores duplicados no afecta la precisión de nuestros modelos y aún produce resultados favorables, todavía surgen problemas al utilizar estos valores duplicados. Imaginemos un escenario en el que planeamos vender un nuevo producto a nuestros clientes existentes y nuestra estrategia consiste en enviar folletos junto con obsequios personalizados a compradores potenciales. Ahora digamos que obtenemos una lista de direcciones de un modelo que hemos desarrollado para enviar esta información y los obsequios acordados. Resulta evidente que tener varios valores duplicados en esta lista generaría gastos innecesarios para la empresa, lo que a su vez provocaría pérdidas financieras. Afortunadamente, en la era actual, es una tarea relativamente sencilla abordar este problema mediante la implementación de instrucciones que puedan identificar con rapidez estos casos y posteriormente eliminarlos.

Los valores atípicos

Los resultados de cualquier modelo de aprendizaje automático pueden verse significativamente influenciados por el método utilizado para recopilar datos. A pesar de los esfuerzos por garantizar una recopilación de datos de calidad, siempre habrá algunos valores poco fiables. Los datos se pueden recopilar mediante entrevistas, encuestas, observaciones y otros métodos, todos los cuales son susceptibles de errores. Por ejemplo, la precisión de los datos obtenidos a través de entrevistas y encuestas depende de la honestidad y la calidad de las respuestas de los participantes, así como de otros factores que pueden parecer no relacionados, como el clima o la hora del día. De manera similar, las observaciones deben cumplir ciertas condiciones para que sean útiles en el modelo. Para abordar esta cuestión, se emplea un tratamiento atípico. Si, por ejemplo, alguien proporciona información falsa, como afirmar que tiene 20 hermanos en una encuesta, ese

dato se eliminará del conjunto de datos porque es probable que sea erróneo. Incluso si la información fuera cierta, todavía se consideraría un valor atípico y no se tendría en cuenta en el modelo.

El proceso de manejo de valores atípicos comienza con la detección de estas irregularidades no deseadas. Para lograrlo, se puede confiar en un concepto estadístico conocido como dispersión, que ofrece varias medidas. Sin embargo, a los efectos de esta explicación, nos concentraremos en los cuartiles y el rango intercuartil. Un enfoque implica considerar cualquier valor que quede fuera del rango especificado en la ecuación como valor atípico. En donde C1 y C3 representan el primer y el tercer cuartil, respectivamente, y RI es el rango intercuartil.

$$(C1 - 1,5 * RI, C3 + 1,5 * RI)$$

Los valores faltantes

Con base en lo que se mencionó anteriormente, es evidente que los datos a menudo existen en forma cruda, lo que genera varios problemas. Un obstáculo común que se encuentra es la presencia de valores faltantes, que ocurre cuando se omite información vital de las variables, lo que resulta en una pérdida de datos potencialmente significativos. Por lo general, no es aconsejable ignorar o intentar analizar un conjunto de datos que contiene valores faltantes, por lo que es necesario utilizar diferentes técnicas para abordar esta complejidad.

Existen numerosos métodos disponibles para manejar los valores faltantes, pero uno de los enfoques más simples es eliminar cualquier observación o característica que tenga valores faltantes. Sin embargo, esta estrategia presenta ciertas desventajas, particularmente para conjuntos de datos más pequeños, ya que se puede perder información valiosa en el proceso. Además, también puede ocurrir la eliminación de características importantes que contribuyen a la precisión y confiabilidad de los resultados. Normalmente, es más común eliminar observaciones que características, y es aconsejable eliminar características sólo si se consideran irrelevantes.

Una técnica que puede resultar más eficaz para abordar los datos faltantes es incorporar información coherente obtenida de los datos existentes. Al hacerlo, podemos

retener una cantidad significativa de información recopilada, incluidas características y observaciones, lo que en última instancia facilita la aplicación de varios modelos. Sin embargo, surge una pregunta crucial: ¿cómo podemos determinar qué información es adecuada para llenar los vacíos que faltan? La respuesta a esta pregunta depende en gran medida del tipo de datos que se consideren. Por ejemplo, en el caso de una variable numérica continua, es aconsejable asignar el valor medio o promedio derivado de todos los resultados disponibles de esa variable en particular. Por otro lado, si estamos ante una variable numérica discreta que sólo toma valores enteros, lo mejor sería optar por la moda.

Existen enfoques más complejos, como el empleo de regresiones o interpolaciones para sustituir valores. Estas técnicas, aunque más avanzadas, pueden producir datos más precisos. Sin embargo, su eficiencia puede variar según los atributos específicos del conjunto de datos que se analiza. Estos métodos se emplean normalmente cuando los valores faltantes tienen una importancia significativa.

La estandarización

La estandarización o normalización, abarca una variedad de técnicas empleadas para mejorar el rendimiento de los modelos de aprendizaje automático mediante la transformación de datos. Entre estas técnicas, los métodos Min-max y Z-index son ampliamente reconocidos (Vercellis, 2011). El objetivo principal de la estandarización es simplificar la comparación y el análisis de variables dentro de un conjunto de datos, mitigando efectivamente las diferencias de escala. Al hacerlo, pretende evitar que los modelos favorezcan excesivamente los atributos con valores más altos, garantizando así una consideración equilibrada de todos los atributos.

El método Min-max es una técnica utilizada para estandarizar valores dentro de un rango específico, generalmente entre -1 y 1 o 0 y 1. En este método, el valor más alto de un atributo se escala a 1, mientras que el valor más bajo se escala a ya sea -1 o 0. Los otros valores luego se ajustan proporcionalmente para ajustarse al intervalo elegido. Este ajuste se logra mediante el uso de una ecuación matemática.

$$x'_{i,j} = \frac{x_{i,j} - x_{\min,j}}{x_{\max,j} - x_{\min,j}} (x'_{\max,j} - x'_{\min,j}) \uparrow x'_{\min,j}$$

Siendo:

$x_{\min,j} = \min_i x_{i,j}$ es decir, el valor más pequeño correspondiente a la determinada columna j ,

$x_{\max,j} = \max_i x_{i,j}$ es decir, el valor más grande correspondiente a la determinada columna j ,

$x'_{\min,j} = -1$ y $x'_{\max,j} = 1$, o bien, $x'_{\min,j} = 0$ y $x'_{\max,j} = 1$

siendo $x'_{i,j}$ los valores ya transformados y $x_{i,j}$ los valores originales.

Entretanto, el método Z-index emplea la transformación:

$$x'_{i,j} = \frac{x_{i,j} - \bar{\mu}_j}{\bar{\sigma}_j}$$

En este método, los valores de μ_j y σ_j representan la media muestral y la desviación estándar de la columna j , respectivamente. Cuando los datos siguen una distribución normal, es muy probable que este enfoque produzca valores dentro del intervalo de (-3,3).

La reducción de los datos

Se considera que la reducción de datos es una técnica valiosa empleada para manejar y procesar eficazmente grandes cantidades de datos, garantizando que los algoritmos de aprendizaje automático funcionen de manera eficiente y que la calidad de la información permanezca intacta. Existen indicadores específicos que señalan la conveniencia de reducir el conjunto de datos, uno de los cuales es el tiempo computacional requerido para trabajar con la información. Minimizar este tiempo es

crucial, ya que ciertos algoritmos poseen complejidades intrincadas y tener un conjunto de datos más pequeño puede marcar una diferencia significativa en el logro de la eficiencia. Además, seleccionar las características apropiadas para usar en los modelos contribuye a obtener resultados más precisos. Por tanto, la simplicidad juega un papel vital, ya que permite a los analistas comprender los criterios implementados. De hecho, algunos expertos incluso están dispuestos a sacrificar cierto grado de precisión en favor de la simplicidad.

La reducción de datos ofrece importantes beneficios en la implementación del modelo. Estas técnicas de optimización son valiosas no sólo para aprender algoritmos sino también para empresas que manejan grandes volúmenes de datos. El objetivo principal de estas empresas es guardar la mayor cantidad de información posible en la forma más compacta. Si bien algunos servicios se especializan en almacenar datos en la nube, normalmente cobran según el uso del almacenamiento. Por lo tanto, reducir el tamaño del conjunto de datos se vuelve crucial para evitar gastos innecesarios en recursos.

El PCA, consiste en una metodología ampliamente utilizado para el análisis de datos con el fin de reducir atributos. Su principal objetivo es sustituir el conjunto original de atributos por un número menor, conseguido mediante combinaciones lineales. Para comprender el concepto y la implementación de PCA, es necesaria una comprensión sólida del álgebra lineal. Se ha demostrado que esta técnica produce resultados más precisos y confiables; sin embargo, es importante tener en cuenta que se recomienda encarecidamente estandarizar los datos antes de aplicar PCA.

En el estudio realizado por Hull (2012), se destaca que el procedimiento PCA juega un papel crucial en la captura de la estructura subyacente de los datos al reducir el número de variables correlacionadas a un conjunto más pequeño de variables no correlacionadas. Para iniciar el proceso de PCA, es necesario calcular una matriz de varianza y covarianza a partir de los datos proporcionados. Esta matriz sirve como base para cálculos posteriores. El paso siguiente consiste en calcular los valores propios y los vectores propios de la matriz de varianza y covarianza.

Es importante señalar que los vectores propios elegidos se normalizan para que tengan una longitud de 1. El vector propio correspondiente al valor propio más alto se

identifica como el primer componente principal, seguido por el vector propio asociado con el segundo valor propio más alto como el segundo componente principal, y pronto. En consecuencia, el valor propio de cada componente principal, expresado como porcentaje de la suma total de todos los valores propios, representa la proporción de la varianza general explicada por ese componente en particular. Además, la raíz cuadrada de cada valor propio indica la desviación estándar de la puntuación del componente correspondiente.

En términos más simples, el propósito de este proceso es crear nuevas variables combinando las variables originales de manera lineal. Esto implica multiplicar los valores de cada variable por un escalar y luego normalizar la ecuación para que la suma de todos los escalares sea igual a 1. El objetivo es identificar las nuevas variables que tienen la varianza más alta y no están correlacionadas entre sí. La nueva variable con la varianza más alta se denomina primer componente principal, seguida del segundo componente principal con la siguiente varianza más alta, y así sucesivamente.

Cuando se trata de variaciones, este método se ve muy afectado por los valores atípicos, por lo que es recomendable abordarlos de antemano. De manera similar, las variaciones se evalúan en función de las escalas de las variables, lo que significa que todas las variables deben estandarizarse para garantizar resultados consistentes. Es importante señalar que, si bien diferentes programas informáticos pueden producir los mismos resultados de PCA, los signos de estos resultados pueden variar, lo que en última instancia no tiene ningún impacto en la interpretación general.

Hay varios métodos adicionales disponibles para reducir la dimensionalidad, incluido kPCA, que es una extensión de PCA que utiliza métodos del kernel. Otra técnica es la descomposición de valores singulares, que permite descomponer una matriz en otras matrices. El análisis de componentes independientes es otro método que se puede utilizar. Todas estas técnicas se pueden encontrar en la biblioteca Scikitlearn para Python. De manera similar, existen herramientas como LASSO, una técnica de análisis de regresión que selecciona variables específicas para mejorar la precisión de un modelo estadístico.

El aprendizaje no supervisado

El aprendizaje no supervisado, es un campo dentro del aprendizaje automático que se centra en descubrir patrones ocultos dentro de un conjunto de datos. Estos patrones suelen ser difíciles de discernir mediante el análisis manual, incluso para personas con amplia experiencia y conocimientos. El objetivo principal del aprendizaje no supervisado es construir un modelo robusto capaz de identificar y comprender con precisión estos patrones intrincados dentro de los datos. A diferencia de otras categorías de aprendizaje, el aprendizaje no supervisado normalmente implica trabajar con información sin etiquetar, lo que significa que no hay etiquetas o clasificaciones preexistentes adjuntas a los datos.

En el campo del aprendizaje no supervisado, existe una categoría de modelos conocidos como modelos de agrupación. Estos modelos tienen como objetivo identificar grupos de puntos de datos que comparten una mayor similitud dentro de su propio grupo en comparación con otros grupos. Si bien la determinación de la similitud generalmente se basa en medir distancias entre puntos de datos, también es factible agrupar datos según variables categóricas.

El aprendizaje no supervisado tiene una importancia significativa en lo que respecta a la exploración e identificación de patrones dentro de grandes cantidades de información. Un ejemplo ilustrativo de esto se encuentra en el contexto de la clientela de una empresa, donde ya se conocen ciertos atributos de los clientes. Al emplear técnicas de aprendizaje no supervisadas, resulta factible agrupar y categorizar a estos clientes en función de sus similitudes, lo que permite a la empresa dirigirse y comercializar eficazmente sus productos a grupos específicos. Además, a través del análisis de los comportamientos de compra colectivos dentro de un grupo particular, la empresa también puede hacer recomendaciones informadas a las personas dentro del grupo, basadas en los productos que sus pares han comprado con frecuencia.

Para completar la tarea de agrupación, existen varios algoritmos y enfoques, cada uno con sus propias ventajas únicas. Al seleccionar el modelo apropiado en función de las características de los datos que se analizan, se pueden minimizar las desventajas de estos algoritmos. Los enfoques principales incluyen algoritmos jerárquicos, particionales y basados en densidad. Los algoritmos jerárquicos implican minimizar la distancia o maximizar las medidas de similitud y pueden clasificarse además como aglomerativos o disociativos.

Por otro lado, los algoritmos particionales requieren un conocimiento previo del número deseado de grupos y tienen como objetivo optimizar criterios específicos o funciones objetivo. Por último, los métodos basados en densidad utilizan diversas técnicas, como gráficos, histogramas y núcleos, para determinar los grupos.

Hay varios algoritmos de agrupamiento disponibles para su uso, incluidos K-Means, DBSCAN, propagación de afinidad, cambio medio, agrupamiento espectral, agrupamiento jerárquico y otros (como se muestra en la figura 2.10). Para este estudio en particular, se emplearon los algoritmos K-Means y DBSCAN, ya que se incluyen en las categorías de métodos particionales y de densidad, respectivamente. Sin embargo, vale la pena señalar que la biblioteca Scikit-learn para Python ofrece una variedad de otros algoritmos de agrupación, con documentación completa y ejemplos prácticos sobre cómo implementarlos de manera efectiva.

El K-means

El aprendizaje no supervisado o las técnicas de agrupamiento implica invariablemente mencionar el algoritmo K-means (también K-medias), introducido por MacQueen en 1967. Sin duda, este algoritmo ha ganado una inmensa popularidad en estos dominios, principalmente debido a su sencilla implementación. y requisitos computacionales mínimos. Estos dos atributos esenciales lo convierten en el método preferido tanto para principiantes como para profesionales experimentados.

El K-means, se considera un método dentro de la familia de métodos de agrupación que tiene como objetivo categorizar las observaciones en función de sus características compartidas y distinguirlas de otros grupos con peculiaridades distintas. Para lograr este objetivo, es fundamental que los datos no estén etiquetados de antemano, ya que K-means se basa en identificar puntos en común entre las observaciones. Este método ha demostrado ser muy preciso y exitoso, particularmente cuando se aplica a grandes bases de datos. Su versatilidad es evidente en su uso generalizado en diversos campos de interés. Por ejemplo, las empresas suelen utilizar K-means para segmentar su base de clientes, lo que les permite diseñar estrategias y mejorar sus operaciones comerciales de manera efectiva.

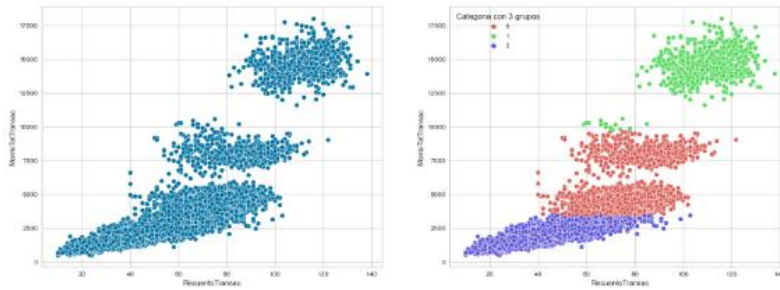
Por lo tanto, K-means es un tipo de algoritmo de agrupamiento que se incluye en la categoría de agrupamiento basado en prototipos. En español se le conoce como

clustering basado en prototipos, indicando que un prototipo, que típicamente es un punto dentro del conjunto de datos, representa a cada grupo similar. Este prototipo es comúnmente el centroide o centro del grupo, de ahí el nombre "K-medias", donde K se refiere al número de centros y las medias representan el promedio o media de los puntos de datos dentro de cada grupo.

Existe un inconveniente importante asociado con las K-medias, que es el requisito de especificar el número exacto de grupos que el método debe identificar desde el principio. Este desafío se puede aliviar mapeando visualmente y analizando los datos para tener una idea inicial de cuál podría ser un número adecuado de conglomerados. Si bien, cuando se trata de conjuntos de datos grandes y complejos que constan de numerosas variables, pueden surgir problemas de visualización, lo que hace extremadamente difícil determinar visualmente con precisión el número óptimo de grupos. Para abordar este problema, se emplean varias estrategias junto con K-means, como el método del codo, la puntuación de silueta y el análisis de componentes principales (PCA), que se explicó en la sección anterior. Es importante señalar que, si bien existen numerosas técnicas disponibles, ninguna ha sido universalmente reconocida como superior a las demás.

Estos pasos se repiten iterativamente hasta que se cumple una condición de parada. Hay varias formas de determinar cuándo detener el proceso. Un método consiste en comprobar si los centroides ya no se pueden reubicar, ya que siempre se ubican en la misma posición. Otro enfoque es establecer un número máximo de iteraciones, después de las cuales el algoritmo debe terminar.

Dadas las características inherentes de K-means, es importante reconocer que los resultados de esta implementación pueden no ser consistentes. Los grupos finales formados pueden diferir dependiendo de la posición inicial de los centroides, por lo que es aconsejable establecer una estrategia para determinar qué resultados retener. En la siguiente figura, el lado izquierdo muestra una colección de puntos de datos sin ningún agrupamiento, mientras que el lado derecho muestra el mismo conjunto de datos después de haber sido agrupado usando el algoritmo K-means.



El modelo

K-medias se puede definir de una manera más formal y matemática como el proceso de selección de centroides que minimicen la suma de errores cuadrados entre estos centroides y cada observación presente en el grupo respectivo.

En este contexto, la variable x_i representa las observaciones individuales, mientras que μ_j representa el valor promedio de las observaciones dentro de un grupo específico, que también puede verse como el punto central o centroide de ese grupo. La letra C , por su parte, denota el grupo que se ha formado en base a ciertos criterios o características.

$$\sum_{i=0}^n \min_{\mu_j \in C} (\|x_i - \mu_j\|^2)$$

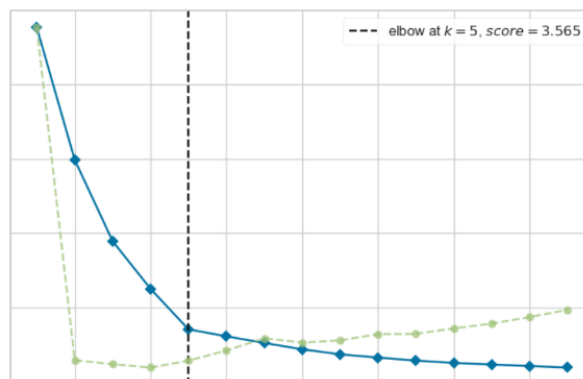
El método del codo

Como se indicó anteriormente, un inconveniente del algoritmo K-means es el requisito de predefinir el número de clústeres para la partición de datos. Para abordar este problema, el método del codo ofrece una solución al determinar visualmente el número ideal de segmentos para la separación de datos.

La técnica implica ejecutar el método K-means varias veces, utilizando cada vez un número consecutivo diferente de grupos para la segmentación. Luego, se calcula la suma de las distancias al cuadrado entre los centros del grupo y sus respectivos puntos

de datos. Es importante señalar que a medida que aumenta el número de grupos, esta métrica disminuye. Esto se debe a que al haber más grupos, los centros están más cerca de cada observación. Por lo tanto, seleccionar el algoritmo con la menor distancia o error no es la solución deseada. En cambio, el objetivo es encontrar el número de grupos que muestran el cambio más significativo en las distancias, lo que lleva a disminuciones más pequeñas.

Cuando los valores que representan la distancia se trazan en un gráfico, la forma resultante se puede comparar con un brazo. La sección que se parece visualmente al codo se considera el número óptimo de agrupaciones para el modelo. Sin embargo, hay casos en los que esta interpretación puede no ser evidente de inmediato, por lo que se puede emplear un enfoque analítico. En la próxima figura, se puede observar como el punto de interés se ubica en un valor de 5, lo que indica que el número ideal de grupos en este escenario también sería 5.



El DBSCAN

El algoritmo DBSCAN, también conocido como agrupación espacial de aplicaciones con ruido basada en densidad, ha ganado una popularidad significativa en el campo de los algoritmos de agrupación. Inicialmente fue presentado por Martin. E., Kriegel, H., Sander, J., Xu, X. en 1996. La característica distintiva de DBSCAN radica en

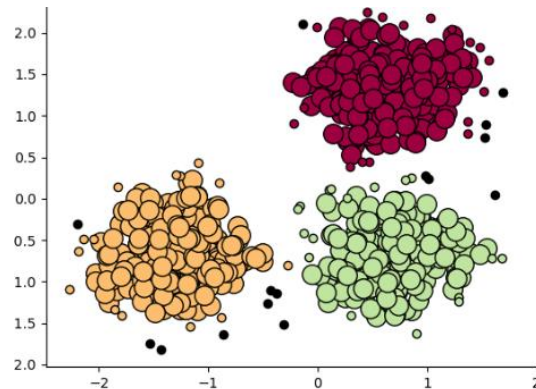
su capacidad para agrupar puntos de datos según su densidad. En otras palabras, si se encuentra que un conjunto de puntos están muy próximos entre sí, se asignarán al mismo grupo. Esta separación se produce debido a la escasez de puntos entre los grupos, lo que da como resultado grupos distintos.

Este método posee varias características distintas, siendo la principal que no impone ninguna expectativa sobre las formas o estructuras que deben tener los grupos. Si bien algunos modelos pueden dar como resultado grupos con forma esférica, el algoritmo DBSCAN permite infinitas posibilidades de formas de grupo siempre que mantengan una densidad mínima requerida. A diferencia de K-means, otra diferencia notable es que no es necesario especificar la cantidad de grupos que se generarán al final. Sin embargo, para garantizar la ejecución adecuada del algoritmo DBSCAN y lograr resultados precisos, se deben especificar otros dos parámetros, aunque no es necesario indicar explícitamente el número exacto de grupos.

El requisito inicial es el número mínimo de puntos necesarios para que se forme un grupo. Este criterio determina cuántos puntos deben estar lo suficientemente cerca entre sí para ser considerados un grupo. El número específico de puntos mínimos requeridos puede variar según el conjunto de datos, el problema en cuestión o las preferencias de la empresa que solicita la agrupación. Es posible determinar el número óptimo de puntos mínimos para el mejor rendimiento del modelo, pero los requisitos específicos pueden influir en esta decisión. Por ejemplo, una empresa que atiende a varios tipos de clientes puede querer ejecutar diferentes estrategias publicitarias para cada grupo. En tal caso, si la empresa tiene una gran cantidad de clientes, puede decidir lanzar campañas únicamente para grupos que consten de más de 10.000 clientes. En consecuencia, el número mínimo de puntos necesarios para formar un grupo en este escenario sería 10.000.

El siguiente parámetro, conocido como ϵ , juega un papel crucial a la hora de determinar la distancia máxima entre dos puntos para que se clasifiquen dentro del mismo grupo. Como se mencionó anteriormente, DBSCAN opera según la densidad de puntos, pero no determina cuándo los puntos están lo suficientemente cerca como para ser considerados parte del mismo grupo. Este parámetro requiere un análisis cuidadoso o un enfoque estratégico para definirlo, ya que establecerlo demasiado pequeño puede dar como resultado que ningún par de puntos cumpla la condición, mientras que

establecerlo demasiado grande puede hacer que todo el conjunto de datos pertenezca a un solo grupo. La figura a continuación proporciona un ejemplo de un conjunto de datos que se ha agrupado utilizando DBSCAN.



El modelo

La densidad se puede describir como la medida de qué tan estrechamente empaquetados o concentrados están los puntos dentro de un radio determinado, denotado como ϵ . De acuerdo con los principios del algoritmo DBSCAN, el etiquetado de cada punto individual está determinado por un conjunto de condiciones específicas:

- Un punto central se define por la presencia de un número mínimo de puntos vecinos dentro de un radio dado ϵ .
- Un punto límite se refiere a un punto que está muy cerca de un punto central dentro de una cierta distancia (ϵ), pero que tiene un número menor de puntos cercanos (MinPoints) dentro de esa distancia.
- Todos los puntos que no forman parte del conjunto principal se clasifican como puntos de ruido.

Después de nombrar los puntos, DBSCAN se puede resumir en dos pasos:

- Crea grupos distintos para cada punto central individual o grupos de puntos centrales que estén vinculados entre sí (los grupos se consideran conectados si están ubicados a una distancia mayor de ϵ).
- El proceso consiste en asignar cada punto fronterizo al grupo que corresponde a su punto central.

Los otros modelos de aprendizaje no supervisado

- El Affinity Propagation, un modelo de aprendizaje automático, consiste en una técnica de aprendizaje no supervisado que facilita la creación de grupos o clústeres basados en dos matrices. La matriz de responsabilidad evalúa el nivel de responsabilidad o influencia de cada observación dentro del conjunto de datos, mientras que la matriz de disponibilidad determina el número de puntos vecinos asociados con cada observación. En particular, el modelo no requiere especificación previa del número de grupos resultantes, lo que permite flexibilidad en el proceso de agrupación.
- Mean Shift, un algoritmo de agrupamiento que opera buscando regiones de alta densidad en los datos. Este enfoque se basa en el concepto de centroides, con el objetivo de determinar el punto central óptimo para cada grupo. Los centroides se actualizan iterativamente calculando el promedio de las observaciones dentro de la ventana evaluada. Una ventaja de Mean Shift es que no requiere especificación previa del número de clústeres.

Los coeficientes de evaluación

Los coeficientes de evaluación tienen un impacto significativo en el avance de los modelos de aprendizaje automático, ya que brindan información valiosa sobre el rendimiento y la eficacia de los modelos implementados. Estos coeficientes nos permiten determinar la utilidad de los modelos y evaluar su efectividad. En el ámbito del aprendizaje no supervisado, existen numerosas medidas de evaluación que se utilizan para calificar los algoritmos de aprendizaje automático. Entre estas medidas, el coeficiente de silueta, el coeficiente de Calinski y el coeficiente de Davies son ampliamente reconocidos como los coeficientes más populares y ampliamente utilizados.

El coeficiente de silueta:

Este coeficiente toma un valor en el rango de -1 a 1, donde -1 significa que las agrupaciones son incorrectas y 1 significa que son correctas. La forma de determinar el coeficiente de silueta es de la siguiente manera:

$$s = \frac{\bar{d} - a}{\max(a, b)}$$

En este contexto:

- "a" representa la distancia promedio entre una observación particular y todas las demás observaciones dentro del mismo grupo.
- Por otro lado, "b" representa la distancia promedio entre una observación específica y todas las demás observaciones en el grupo vecino más cercano.

El coeficiente de Calinski

El coeficiente de Calinski nos dice que cuanto mayor es el coeficiente, mejor calidad de agrupación tiene el modelo. Cómo obtener el coeficiente de Calinski:

$$s = \frac{\text{tr}(B_k)}{\text{tr}(W_k)} \times \frac{n_E - k}{k - 1}$$

En donde:

$$W_k = \sum_{q=1}^k \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$

$$B_k = \sum_{q=1}^k n_q (c_q - c_E)(c_q - c_E)^T$$

- E representa el conjunto de datos,

- mientras que n_E representa el número de muestras en el conjunto de datos,
- k se refiere al número de grupos en los que se categorizaron las observaciones,
- B_k representa la matriz de dispersión entre estos grupos,
- mientras que W_k representa la matriz de dispersión dentro de un grupo específico,
- $\text{tr}()$ calcula la traza de una matriz, que es esencialmente la suma de sus elementos diagonales,
- q denota un grupo particular,
- mientras que C_q representa el conjunto de puntos dentro de ese grupo,
- c_q significa el centro de un grupo específico,
- mientras que c_E representa el centro de todo el conjunto de datos E ,
- por último, n_q se refiere al número de puntos dentro de un grupo determinado.

El coeficiente de Davies

El número óptimo de grupos se determina encontrando el valor del coeficiente de Davies que se minimiza. El coeficiente de Davies representa la similitud promedio entre cada grupo C_i (donde i es un número entre 1 y k) y su grupo más similar C_j . En este contexto, la similitud se mide por el valor de R_{ij} . Para calcular el coeficiente de Davies-Bouldin utilizamos esta información:

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} R_{ij}$$

En donde:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \quad (2.10)$$

La distancia promedio entre cada punto del grupo i y su centro correspondiente se denota como s_i . Además, la distancia entre los centros de los grupos i y j se representa como d_{ij} .

El aprendizaje supervisado

El aprendizaje supervisado, es un área específica dentro del campo del aprendizaje automático que se centra en utilizar la información disponible sobre un conjunto de datos determinado para crear un modelo. Este modelo está diseñado para tomar un conjunto de características y predecir con precisión una clasificación o regresión específica para ese conjunto en particular. En el aprendizaje supervisado, el conjunto de datos normalmente contiene una variable objetivo que ya ha sido etiquetada y el objetivo es asignar la misma etiqueta a datos nuevos sin etiquetar. Esta asignación se logra mediante un proceso llamado entrenamiento, donde el modelo se expone a la mayoría o a la totalidad del conjunto de datos, según el tamaño de la muestra.

El aprendizaje supervisado se considera un arsenal muy completo de técnicas que permiten obtener conocimientos y perspectivas a partir de la información disponible. Implica estudiar las relaciones entre varias variables y una variable objetivo específica. El aspecto crucial de este enfoque de aprendizaje es la utilización de este conocimiento adquirido para predecir el comportamiento de la variable de interés ante diferentes escenarios que involucran a las otras variables. Esta capacidad predictiva se vuelve particularmente valiosa en aplicaciones como las predicciones de pagos predeterminados, ya que puede generar ganancias financieras sustanciales para la institución que implementa esta metodología de aprendizaje.

La computadora utiliza los datos para adquirir conocimiento y comprender patrones, lo que le permite hacer predicciones sobre la variable que nos interesa. Esta variable puede tomar diferentes formas, incluyendo nominal (categórica), binaria (sí/no), numérica e incluso texto. La elección del modelo depende del tipo de variable que intentamos predecir y nuestro objetivo es encontrar el modelo más adecuado para nuestras necesidades específicas. Entre los diversos tipos de aprendizaje supervisado, este estudio se centrará principalmente en dos técnicas altamente efectivas y de uso común: regresión logística y árboles de decisión. Sin embargo, vale la pena mencionar que existen otros algoritmos de aprendizaje supervisado disponibles.

Es evidente que para poder identificar con precisión el comportamiento potencial de nuevos datos es fundamental haber recibido una formación previa. Además, la calidad de los datos juega un papel importante en este proceso, ya que impacta directamente en la precisión y confiabilidad del etiquetado. Además, a medida que pasa el tiempo, es imperativo monitorear consistentemente los resultados y alimentar continuamente al modelo con nueva información y conocimientos. Este conocimiento y capacitación continuos permiten que el modelo mejore su rendimiento con el tiempo, lo que en última instancia conduce a resultados más favorables y precisos.

Entrenamiento y pruebas

Después de completar un preprocesamiento exhaustivo de los datos y realizar un análisis exploratorio integral, el siguiente paso es elegir cuidadosamente las variables independientes que se utilizarán para explicar los patrones y tendencias en la variable objetivo. Es muy recomendable, y de hecho casi necesario, dividir nuestro conjunto de datos en dos grupos distintos: un conjunto de entrenamiento y un conjunto de prueba. Esta división debe basarse en la valoración y juicio cualitativo del analista, ya que depende de las características específicas y la naturaleza de los datos.

Un posible enfoque, cuando se trata de un gran conjunto de datos, es dividirlo en dos subconjuntos utilizando una proporción de 90/10 o incluso 99/1. En este escenario, alrededor del 90% o el 99% de los datos se asignarían a fines de capacitación, mientras que el porcentaje restante se utilizaría para evaluación. La razón detrás de esta división es que con una cantidad sustancial de datos, habrá suficientes instancias para la evaluación incluso con un porcentaje pequeño, lo que permitirá una capacitación más efectiva. Es importante señalar que el término "enorme" es subjetivo y puede variar según el contexto. Sin embargo, este enfoque resulta beneficioso cuando se trata de conjuntos de datos que contienen cientos de miles o incluso millones de observaciones. Al utilizar una parte importante de los datos para el entrenamiento, el modelo puede aprender a reconocer y manejar incluso casos raros o poco comunes que de otro modo se pasarían por alto si el conjunto de pruebas fuera más grande.

Si bien, es importante tener en cuenta que si el conjunto de datos es muy pequeño, puede ser más apropiado dividir el conjunto en 100/0, ya que el enfoque más lógico sería utilizar todos los datos disponibles con fines de capacitación. En tal escenario, dividir los

datos de manera más uniforme podría resultar en evaluaciones engañosas, ya sean positivas o negativas, dependiendo de los datos específicos seleccionados para la capacitación. En consecuencia, estas evaluaciones no serían particularmente confiables. Además, cabe mencionar que la determinación de lo que constituye un pequeño conjunto de datos es subjetiva y varía entre los analistas. Esto es especialmente común cuando se trata de información que las empresas recopilan mensual, trimestral o anualmente, particularmente cuando el período de recopilación de datos es relativamente corto.

Las divisiones más utilizadas para el análisis de datos suelen dividirse en proporciones 80/20 o 70/30. Aunque, es importante señalar que la decisión óptima para las proporciones de división depende en gran medida del conjunto de datos específico y del análisis cualitativo realizado por la persona a cargo. Además de las divisiones comúnmente utilizadas, también existe una tercera división conocida como datos de validación. Aunque esta división en particular no es directamente relevante para este trabajo específico o sus aplicaciones, vale la pena reconocer su existencia.

Sobreajustar un modelo es un concepto crucial de entender, ya que se refiere a una situación en la que el modelo está tan bien ajustado a los ejemplos de entrada etiquetados que no logra predecir con precisión los resultados de los ejemplos de datos no etiquetados que no se incluyeron en la fase de entrenamiento. Este fenómeno puede ser problemático ya que indica que el modelo esencialmente ha memorizado los datos de entrenamiento en lugar de generalizar patrones y relaciones que pueden aplicarse a datos nuevos e invisibles. En otras palabras, el sobreajuste da como resultado un modelo excesivamente complejo y demasiado específico para los datos de entrenamiento, lo que conduce a un rendimiento deficiente cuando se enfrenta a ejemplos invisibles. Por tanto, es vital abordar el sobreajuste para garantizar la fiabilidad y eficacia de los modelos de aprendizaje automático.

La regresión lineal

La regresión lineal es un método ampliamente reconocido y ampliamente utilizado en el aprendizaje automático supervisado. Establece una conexión lineal entre varios atributos y un resultado o etiqueta específica. Durante la fase de entrenamiento de un modelo se adquieren los valores óptimos de sus parámetros. En el aprendizaje supervisado, lograr esto implica emplear un algoritmo que analiza numerosas instancias

etiquetadas y tiene como objetivo identificar los valores de estos parámetros del modelo que minimizan el componente de error.

Los algoritmos de regresión se utilizan para modelar la conexión entre varias variables de entrada mediante el empleo de una medida de error, que pretendemos reducir mediante un proceso iterativo para mejorar la precisión de las predicciones. En esta discusión, profundizaremos en dos tipos específicos de algoritmos de regresión: regresión lineal y regresión logística. Vale la pena señalar que la principal distinción entre estos dos radica en la naturaleza de su producción. La regresión logística se emplea cuando la salida es discreta, mientras que la regresión lineal se utiliza cuando la salida es continua.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$$

- La variable y representa la etiqueta o el resultado que nos interesa.
- Las características x_i describen la etiqueta o las variables independientes.
- La pendiente de la línea, denotada como β_i con $i = 1, 2, \dots, k$, se conoce comúnmente como peso. Es uno de los dos parámetros que deben aprenderse del modelo durante el proceso de capacitación para poder utilizar el modelo en futuras inferencias.
- El símbolo ϵ representa el punto de intersección de la línea en el eje, también conocido como error.

La regresión logística

La regresión logística es un tipo de algoritmo de clasificación que se incluye en el ámbito de las técnicas de aprendizaje supervisado. A pesar de su aparente diferencia, es esencialmente un modelo lineal empleado para segregar un problema en dos categorías distintas utilizando variables predictivas como puntos de referencia. Estas variables predictivas se integran en una ecuación que se asemeja a la regresión lineal, con el objetivo de determinar el ajuste óptimo. En realidad, la regresión logística también se puede emplear para abordar problemas que abarcan múltiples clases, ampliando así su aplicabilidad.

Crear una implementación exitosa de la regresión logística es sin duda una tarea compleja que va más allá de los límites de este proyecto en particular. No obstante, es crucial abordar ciertos factores que es esencial considerar para lograr resultados satisfactorios. Para comprender plenamente estos aspectos, es imperativo poseer una sólida comprensión de los principios matemáticos y estadísticos.

El modelo logit, también conocido como modelo de regresión logística, opera evaluando la probabilidad de que cada observación pertenezca a una categoría específica. Esta probabilidad está influenciada por los valores de las variables predictoras. Posteriormente, las observaciones se categorizan en función de la probabilidad calculada.

$$f(z) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)}}$$

El árbol de decisión

La fama y popularidad de los modelos de árboles de decisión, específicamente los árboles de clasificación y los árboles de regresión, son ampliamente reconocidas. Además, también han adquirido importante renombre la ampliación y modificación de estos modelos, conocidos como Bosques de Clasificación y Bosques de Regresión. Una de las principales razones del alto índice de aprobación de estos algoritmos es su simplicidad visual, que permite a los analistas comprender e interpretar fácilmente los datos de forma intuitiva. Estos modelos son particularmente fáciles de usar cuando se presentan gráficamente.

A diferencia de otros algoritmos, los árboles de decisión no requieren un preprocesamiento extenso del conjunto de datos. Sin embargo, es crucial realizar dicho preprocesamiento debido a su susceptibilidad a manejar grandes cantidades de datos. Además, existe una amplia gama de enfoques para implementar este método, como la técnica de inducción de arriba hacia abajo de árboles de decisión. Este proceso se inicia designando cada observación como el nodo raíz del árbol, y estos nodos luego se incorporan a la lista de nodos activos, conocida como L.

La matriz de confusión

La matriz de confusión es una herramienta valiosa en el campo del aprendizaje automático que nos permite evaluar la eficacia y precisión de los modelos bajo evaluación. Esencialmente, esta matriz proporciona una visión integral de la comparación entre los resultados predichos generados por el modelo y los valores reales. Al organizar estas predicciones en columnas y los valores reales en filas, obtenemos una comprensión clara de cómo el modelo se desempeñó en sus predicciones. Esta matriz sirve como una ayuda crucial para evaluar y ajustar los modelos de aprendizaje automático para un rendimiento óptimo.

Esta herramienta abarca cuatro categorías: verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Los verdaderos positivos se refieren a casos en los que el modelo predijo con precisión un resultado positivo, que resultó ser verdadero. Los verdaderos negativos son casos en los que el modelo predijo correctamente un resultado negativo. Por otro lado, los falsos positivos ocurren cuando el modelo predice incorrectamente un resultado positivo, mientras que los falsos negativos ocurren cuando el modelo predice incorrectamente un resultado negativo. Para ilustrar, consideremos un escenario en el que un banco emplea un modelo predictivo para determinar qué clientes realizarán sus pagos mensuales. En este caso, un verdadero positivo sería cuando el modelo predice correctamente que un cliente pagará y realmente lo hace. Un verdadero negativo ocurriría cuando el modelo predice con precisión que un cliente no pagará. Un falso positivo surgiría si el modelo predice incorrectamente que un cliente pagará, pero en realidad no lo hace. Por último, un falso negativo ocurriría si el modelo predice erróneamente que un cliente no pagará, pero en realidad sí lo hace.

- La métrica de exactitud evalúa la cantidad de pronósticos correctos sumando los valores verdaderos positivos y verdaderos negativos del número total de pronósticos realizados.
- La precisión es una métrica que evalúa la exactitud de las predicciones positivas determinando la proporción de verdaderos positivos con respecto a la suma total de verdaderos positivos y falsos positivos. En términos más simples, mide qué tan bien un modelo o sistema identifica correctamente instancias positivas entre todas las instancias que etiqueta como positivas.

- La sensibilidad, también conocida como tasa de verdaderos positivos, se refiere a la capacidad de una prueba o modelo de diagnóstico para detectar con precisión casos positivos. Se calcula dividiendo el número de verdaderos positivos (casos positivos correctamente identificados) por la suma de verdaderos positivos y falsos negativos (casos positivos clasificados erróneamente como negativos). En términos más simples, la sensibilidad indica la eficacia de una prueba para identificar correctamente a las personas que realmente padecen la afección que se está examinando. Además, la sensibilidad también se puede interpretar como la probabilidad de que una persona con un resultado positivo en la prueba realmente tenga la afección.
- El concepto de especificidad es similar al de sensibilidad, pero se centra en la identificación precisa de los casos negativos. En otras palabras, la especificidad mide la proporción de verdaderos negativos entre todos los casos negativos, incluidos tanto los verdaderos negativos como los falsos positivos. También puede interpretarse como la probabilidad de que un caso negativo sea correctamente identificado como negativo.
- El F1 score es una métrica que combina los conceptos de precisión y sensibilidad en una sola medida. Se calcula mediante la fórmula $2 * (\text{Sensibilidad} * \text{Precisión}) / (\text{Sensibilidad} + \text{Precisión})$. Para que un modelo se considere aceptable, debe tener una puntuación F1 de al menos 80. Esto significa que el modelo es capaz de lograr un buen equilibrio entre identificar correctamente las instancias positivas (sensibilidad) y clasificarlas con precisión (precisión). Al utilizar F1 score, podemos evaluar el rendimiento general de un modelo de una manera más completa, teniendo en cuenta tanto los falsos positivos como los falsos negativos.

Al tener en cuenta estas métricas, un analista obtiene información valiosa sobre el rendimiento de su modelo y puede determinar sus objetivos principales. Al aprovechar estas medidas, los analistas pueden evaluar eficazmente si están progresando y logrando los resultados deseados, garantizando así que están en la trayectoria correcta.

Capítulo 4

Redes neuronales

El cerebro humano es ampliamente considerado como el sistema informático más complejo que existe. Si bien tanto las computadoras como los humanos se destacan en diversas tareas, hay ciertas actividades que son naturales para los humanos pero que resultan desafiantes para las computadoras, y viceversa. Por ejemplo, reconocer el rostro de una persona es una tarea relativamente sencilla para los humanos, pero presenta dificultades para las computadoras. Por otro lado, gestionar la contabilidad de una empresa es una tarea compleja y costosa para un experto en contabilidad, pero un programa informático básico puede realizarla con facilidad.

Las notables capacidades del cerebro humano, incluida su capacidad para pensar, retener información y resolver problemas, han despertado el interés de innumerables científicos que buscan replicar su funcionamiento en una computadora. Esta búsqueda interdisciplinaria ha reunido a profesionales de campos como la ingeniería, la filosofía, la fisiología y la psicología, todos impulsados por el inmenso potencial que encierra esta tecnología. En colaboración, están explorando diversas aplicaciones en sus respectivos dominios. Un grupo concreto de investigadores ha dedicado sus esfuerzos a crear un modelo informático que emule las funciones fundamentales del cerebro humano. Este logro revolucionario ha dado lugar a una tecnología innovadora conocida como Computación Neural o Redes Neuronales Artificiales.

La renovada fascinación por este novedoso método de realizar cálculos, que se había pasado por alto durante los últimos veinte años, puede atribuirse a los notables avances y logros tanto en el ámbito teórico como en el práctico que se han logrado en los últimos tiempos.

Características

Las redes neuronales artificiales, comúnmente conocidas como ANN (por sus siglas en inglés), se inspiran en la intrincada red de neuronas biológicas que se encuentran en el cerebro humano. Los componentes fundamentales imitan el comportamiento de las neuronas biológicas, emulando así sus funciones primarias. Estos componentes están

meticulosamente organizados dentro de la estructura de ANN, reflejando la organización observada en el cerebro humano.

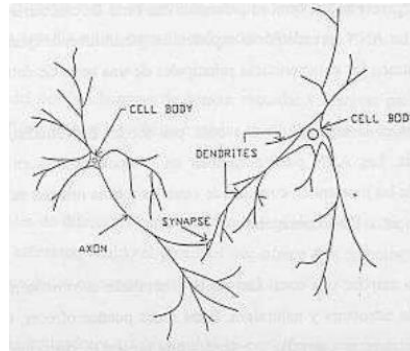
Las ANN no poseen únicamente un parecido visual con el cerebro, también exhiben varias características similares a el cerebro. Una característica notable es su capacidad para adquirir conocimientos a través de la experiencia, lo que les permite asimilar información de instancias pasadas y aplicarla a escenarios novedosos. Además, las ANN poseen la capacidad de extraer características fundamentales de un conjunto de datos determinado, lo que les permite identificar y comprender los atributos esenciales que definen los datos.

- Aprender significa adquirir conocimientos sobre algo participando en actividades como estudiar, practicar o adquirir experiencia de primera mano. Las (ANN) tienen la capacidad de modificar su comportamiento en función de las condiciones de su entorno. Cuando se les presenta una colección de entradas, las ANN tienen la capacidad de adaptarse y ajustarse para generar resultados confiables y consistentes.
- La generalización se refiere al proceso de extender o expandir algo. Las ANN también tienen la capacidad de generalizarse automáticamente, ésto es producto de su estructura y características inherentes. Estas redes tienen la capacidad de proporcionar respuestas precisas a entradas que pueden presentar ligeras variaciones causadas por factores como ruido o distorsión, dentro de un cierto rango.
- El proceso de abstracción implica separar o analizar mentalmente las cualidades de un objeto por sí solas. En el contexto de las redes neuronales artificiales, ciertas ANN poseen la capacidad de abstraer la naturaleza fundamental o los elementos centrales de un grupo de entradas, incluso cuando estas entradas parecen carecer de características obvias o compartidas.

Su estructura

La neurona, que es la piedra angular del sistema nervioso y específicamente del cerebro, sirve como un componente crucial en el procesamiento de la información. Cada neurona opera como una unidad de procesamiento básica, responsable de recibir e

integrar señales de otras neuronas. Cuando la entrada colectiva supera un cierto umbral, la neurona se activa y produce una salida. Para comprender mejor la estructura y función de una neurona, observe la siguiente figura que ilustra los diversos componentes que constituyen una neurona (Basogain Olabe, s.f.).

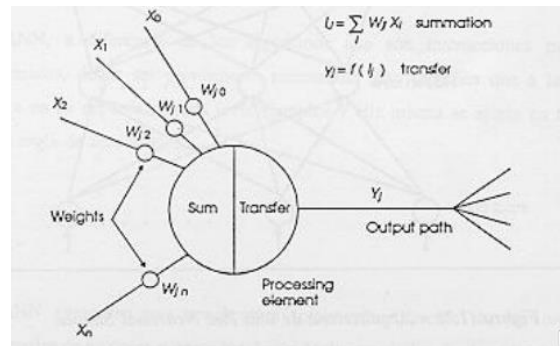


El sistema nervioso central se compone por miles de millones, o incluso billones, de neuronas que se encuentran intrincadamente interconectadas. Estas neuronas se comunican entre sí a través de conexiones especializadas llamadas sinapsis, donde el axón de una neurona se ramifica y se conecta con las dendritas de otras neuronas. A través de esta intrincada red de sinapsis, la información se transmite y procesa dentro del cerebro. Curiosamente, la eficacia de estas sinapsis puede modificarse y ajustarse a medida que el cerebro aprende y se adapta. Esta flexibilidad en la función sináptica juega un papel crucial en la configuración de la capacidad del cerebro para aprender, recordar y procesar información.

En el ámbito de las redes neuronales artificiales, el elemento de procesamiento (PE) actúa como el equivalente de una neurona biológica. Un elemento de procesamiento posee la capacidad de recibir múltiples entradas, que después se combinan, generalmente mediante un proceso de suma sencillo. La suma resultante de las entradas sufre modificación mediante una función de transferencia, y el valor resultante de esta modificación se transmite directamente a la salida del elemento de procesamiento.

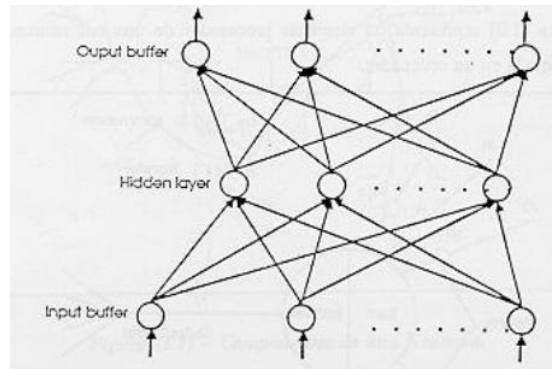
La salida de una neurona artificial, también conocida como elemento de procesamiento, se puede vincular a la entrada de otras neuronas artificiales, formando una red de PE interconectados. Estas conexiones están ponderadas, lo que significa que

la fuerza de la conexión está determinada por la eficiencia de la sinapsis entre las neuronas. El siguiente diagrama muestra un componente de procesamiento de una red neuronal artificial basada en computadora (Basogain Olabe, s.f.).



Una red neuronal se compone de un conjunto de unidades elementales, conocidas como elementos de procesamiento, que están interconectadas de una manera específica. La importancia de las redes neuronales artificiales no reside sólo en el modelo PE individual, sino también en las intrincadas conexiones entre estos elementos de procesamiento. Normalmente, los elementos de PE se agrupan en niveles o capas, formando una estructura jerárquica. Una red neuronal estándar consta de una serie de estas capas, con conexiones establecidas entre cada capa adyacente consecutiva.

La red neuronal artificial consta de varias capas, incluida una capa de búfer de entrada responsable de recibir datos del entorno externo y una capa de búfer de salida que almacena la respuesta de la red a la entrada. Estas dos capas sirven como interfaz entre la red y el mundo exterior. Aparte de estas capas, existen capas ocultas adicionales dentro de la red. La figura siguiente representa visualmente la estructura de una Red Neuronal Artificial (Basogain Olabe, s.f.).



La computación neuronal

Para resolver un problema utilizando técnicas de programación tradicionales, es necesario desarrollar un algoritmo, definido como un conjunto de instrucciones que describen los pasos que debe seguir un sistema informático para alcanzar con éxito el resultado deseado, que es la solución al problema en cuestión.

Crear una secuencia de instrucciones para abordar un problema contable es una tarea relativamente sencilla, pero existen numerosos escenarios del mundo real en los que diseñar un algoritmo para resolver problemas complejos puede resultar sumamente desafiante. Un ejemplo de ello es el desarrollo de un programa de reconocimiento de imágenes, especialmente cuando se trata de identificar el rostro de una persona. En este caso, el algoritmo debe tener en cuenta varios factores, como diferentes expresiones faciales, incluidas caras serias o felices, así como otras variaciones generales que pueden existir dentro de la imagen de una persona.

Las ANN se distinguen de los algoritmos preprogramados en su necesidad de un entrenamiento previo. En otras palabras, la red está expuesta a un conjunto de ejemplos a través de su capa de entrada y luego se adapta de acuerdo con una regla de aprendizaje específica.

Así, estas redes neuronales artificiales poseen una estructura distintiva que las diferencia de las computadoras convencionales de un solo procesador. A diferencia de las máquinas tradicionales que siguen el modelo de Von Neuman, las ANN constan de múltiples elementos de procesamiento en lugar de una CPU (Unidad de proceso de control) solitaria responsable de ejecutar todos los cálculos de acuerdo con la secuencia algorítmica programada. A diferencia de las CPU, que están equipadas para ejecutar más

de cien comandos elementales que abarcan operaciones como suma, resta y desplazamiento, las ANN cuentan con un marco completamente diferente.

Los comandos o instrucciones se ejecutan en un orden específico y coordinados con el reloj del sistema. Sin embargo, en los sistemas de computación neuronal, cada elemento de procesamiento (PE) se limita a realizar solo uno o unos pocos cálculos. La eficacia de las redes neuronales artificiales está determinada principalmente por la frecuencia con la que se actualizan las interconexiones durante el entrenamiento o el aprendizaje. Por otro lado, el rendimiento de las máquinas Von Neumann se mide en función del número de instrucciones ejecutadas por segundo por la unidad central de procesamiento (CPU).

La estructura de las ANN se deriva de la disposición de sistemas de procesamiento paralelos, donde varios procesadores están conectados entre sí. No obstante, estos procesadores son unidades básicas para la computación, diseñadas para sumar numerosas entradas y al mismo tiempo tener la capacidad de adaptar y modificar la fuerza de las conexiones entre ellas.

Por lo tanto, los sistemas expertos y la programación tradicional tienen una distinción notable en la forma en que manejan el procesamiento del conocimiento. En los sistemas expertos, la base de conocimientos está separada del motor de inferencia, que se encarga de procesar ese conocimiento. Esta separación permite que el sistema incorpore nuevos conocimientos sin necesidad de una reprogramación completa de todo el sistema. Esencialmente, permite la expansión y mejora continua de la base de conocimientos del sistema. Sin embargo, para que esta técnica sea efectiva, es imperativo contar con un experto en el campo relevante que pueda aportar su experiencia y establecer reglas que puedan codificar efectivamente ese conocimiento dentro del sistema.

Al crear una red neuronal, no es necesario programar manualmente el conocimiento ni los métodos para procesar ese conocimiento. En cambio, la red neuronal adquiere la capacidad de procesar conocimiento modificando la fuerza de las conexiones entre las neuronas en varias capas de la red. En los Sistemas Expertos el conocimiento se representa explícitamente a través de reglas. Por otro lado, en la computación neuronal, las Redes Neuronales Artificiales tienen la capacidad de generar sus propias reglas aprendiendo de los ejemplos que se les presentan durante la fase de entrenamiento. Este

proceso de aprendizaje se ve facilitado por una regla de aprendizaje, que ajusta los pesos de las conexiones dentro de la red en función de los ejemplos de entrada proporcionados y, potencialmente, también en función de los resultados deseados. Esta característica única de las ANN les permite adquirir conocimientos a través de la experiencia.

Entonces, un aspecto clave de las redes neuronales artificiales es cómo almacenan información. A diferencia de los sistemas informáticos tradicionales, las ANN distribuyen su memoria o conocimiento entre todas las conexiones ponderadas dentro de la red. Además, ciertas ANN poseen el atributo de ser "asociativas", lo que significa que cuando se les presenta una entrada parcial, la red puede identificar la entrada más similar almacenada en su memoria y generar una salida que se alinea con la entrada completa. Esta capacidad asociativa permite a las ANN procesar y responder eficientemente a datos incompletos o fragmentados. Así, las ANN tienen una condición única denominada memoria, que les facilita la eficaz adaptación y generación de respuestas adecuadas, aun cuando se enfrentan a entradas imperfectas o distorsionadas. Este atributo invaluable se describe comúnmente como la capacidad de la red para "generalizar" su comprensión y proceso de toma de decisiones.

Otra característica importante de las redes neuronales artificiales es su capacidad para tolerar fallos. La tolerancia a fallas se refiere a la capacidad de las ANN de continuar funcionando incluso si algunos de los elementos de procesamiento (PE) o conexiones dentro de la red están dañados o alterados. En tales casos, el comportamiento global de la red puede sufrir ligeras modificaciones, pero el sistema en su conjunto no colapsa ni deja de funcionar. Esta característica única de tolerancia a fallos en las ANN se puede atribuir a la forma en que se distribuye y almacena la información en toda la red, en lugar de concentrarse en una única ubicación. Esto garantiza que incluso si ciertas partes de la red se ven afectadas, el funcionamiento general y el rendimiento de la ANN permanezcan intactos.

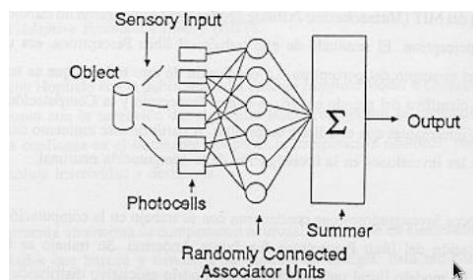
Historia

En 1956, se produjo un hito importante en la Inteligencia Artificial cuando pioneros como Minsky, McCarthy, Rochester y Shanon organizaron la primera conferencia sobre Inteligencia Artificial. Esta conferencia, patrocinada por la Fundación Rochester, tuvo lugar durante el verano de 1956 en la localidad inglesa de Dartmouth.

Muchos libros se refieren a este verano como el encuentro inicial significativo con las redes neuronales artificiales. Durante esta conferencia, Nathaural Rochester, del equipo de investigación de IBM, presentó su propio modelo de red neuronal, que puede considerarse el primer software de simulación de redes neuronales artificiales.

Un año más tarde, en 1957, Frank Rosenblatt hizo una importante contribución al campo de la computación neuronal con la publicación de su extenso trabajo de investigación. Este trabajo se centró en el desarrollo de un elemento conocido como "Perceptron". El Perceptron es un sistema para clasificar patrones y tiene la capacidad de identificar patrones tanto geométricos como abstractos. El primer Perceptron demostró la capacidad de aprender y mantener una funcionalidad sólida, y solo se veía afectado si los componentes del sistema resultaban dañados. Además, mostró flexibilidad y continuó comportándose correctamente incluso después de que algunas células fueron destruidas. Diseñado originalmente para el reconocimiento de patrones ópticos, el Perceptron incorporó una rejilla de 400 fotocélulas, que representan neuronas retinianas sensibles a la luz, para recibir estímulos ópticos. Estas fotocélulas estaban conectadas a elementos asociativos que recogían los impulsos eléctricos emitidos por las fotocélulas.

Las conexiones entre los elementos asociativos y las fotocélulas se establecieron de forma aleatoria. Si las celdas recibieran un valor de entrada que superara un umbral predeterminado, el elemento asociativo produciría una salida. La figura muestra la estructura de la red Perceptron (Basogain Olabe, s.f.).



En 1982, la publicación de John Hopfield del modelo Hopfield o red asociativa Crossbar, junto con la invención del algoritmo de retropropagación, reavivó el interés y la confianza en el campo de la computación neuronal después de dos décadas de inactividad y desinterés. Hopfield presenta un sistema informático neuronal que comprende elementos de procesamiento interconectados. Estos elementos trabajan juntos

para minimizar la energía dentro de la red. Este sistema específico, con su función energética y mecanismo de respuesta, cae dentro de la categoría más amplia de redes que Grossberg exploró previamente.

La actualidad

Numerosos grupos de investigación, repartidos por varias universidades de todo el mundo, participan activamente en la realización de amplios estudios relacionados con las redes neuronales artificiales. Estos grupos, que comprenden una amplia gama de profesionales, incluidos neurólogos, psicólogos cognitivos, físicos, programadores y matemáticos, aportan distintos enfoques y motivaciones dentro de este ámbito. En consecuencia, aportan perspectivas novedosas y conocimientos intuitivos al campo de la tecnología en constante evolución.

Grossberg colabora actualmente con Carpenter en la Universidad de Boston, mientras que Teuvo Kohonen trabaja en la Universidad de Helsinki. En los últimos años, uno de los grupos de investigación más destacados ha sido el grupo PDP (Parallel Distributed Processing), establecido por Rumelhart, McClelland y Hinton. Rumelhart, un distinguido académico de la Universidad de Stanford, ha desempeñado un papel fundamental en la popularización de la red neuronal de retropropagación, que se utiliza ampliamente en diversas aplicaciones en la actualidad.

El grupo de investigación de McClelland en la Universidad Carnegie-Mellon es particularmente digno de mención por su exploración de posibles aplicaciones de la retropropagación. Por otro lado, Hinton y Sejnowski de la Universidad de Toronto han desarrollado una máquina denominada Boltzman, que presenta la red de Hopfield con dos modificaciones importantes. Además, Bart Kosko ha ideado una red llamada BAM (Memoria Asociada Bidireccional) que se basa en la red de Grossberg.

Asimismo, es importante reconocer la presencia de importantes grupos de investigación en instituciones estimadas como el Instituto de Tecnología de California, el Instituto de Tecnología de Massachusetts, la Universidad de California Berkeley y la Universidad de California en San Diego. No pasemos por alto las importantes inversiones económicas y técnicas que están realizando empresas privadas en los Estados Unidos, el Japón y la Comunidad Económica Europea. Para ilustrar la magnitud de estas

inversiones, basta señalar que sólo Estados Unidos gasta más de 100 millones de dólares al año.

ANN: aplicaciones

La industria del automóvil se ha beneficiado enormemente de la aplicación de redes neuronales artificiales. Con el auge de los vehículos autónomos, las ANN desempeñan un papel crucial al permitir el reconocimiento de objetos, peatones y otros vehículos, facilitando una navegación segura y eficiente. Además, los sistemas de computación neuronal se utilizan para optimizar el rendimiento del motor, mejorar la eficiencia del combustible y predecir fallas de componentes, lo que mejora la confiabilidad general del vehículo.

Asimismo, las redes neuronales artificiales encuentran un amplio uso en el ámbito de las finanzas y la economía. Los modelos ANN se emplean para pronosticar precios de acciones, predecir tendencias del mercado y optimizar carteras de inversión. Al utilizar datos históricos y algoritmos complejos, estos sistemas pueden identificar patrones y fluctuaciones ocultos en los mercados financieros, ayudando a los inversores a tomar decisiones informadas.

Las aplicaciones de las redes neuronales artificiales son amplias y abarcan diversos dominios. Una de esas áreas en las que ANN sobresale es en las tareas de reconocimiento de patrones. Ya sea que se trate de identificar patrones complejos en imágenes, voz o texto, los sistemas de computación neuronal han demostrado ser invaluable para extraer información significativa de estas entradas. Las ANN ha demostrado su eficacia en áreas como la minería de datos, donde puede descubrir patrones y correlaciones ocultos dentro de conjuntos de datos vastos y complejos. Otra aplicación importante de las redes neuronales artificiales se encuentra en el campo de la bioinformática y la genómica.

Con la capacidad de analizar datos biológicos a gran escala, incluidas secuencias de ADN y estructuras de proteínas, ANN ayuda a comprender códigos genéticos, predecir funciones de proteínas e identificar posibles objetivos farmacológicos. Esto tiene inmensas implicaciones para los avances en el campo médico, incluida la medicina personalizada y el descubrimiento de fármacos.

En esencia, la versatilidad y el enfoque humano de los sistemas de computación neuronal han abierto una amplia gama de aplicaciones en numerosas industrias. Desde el reconocimiento de patrones hasta la bioinformática, las finanzas y la automoción, las redes neuronales artificiales continúan revolucionando la forma en que procesamos y analizamos los datos y, en última instancia, mejoran nuestra comprensión y capacidad de toma de decisiones. Las características únicas y excepcionales de los sistemas de computación neuronal los hacen muy versátiles y aplicables en una amplia gama de campos. Al imitar la capacidad del cerebro humano para reconocer y percibir información, la computación neuronal ofrece un enfoque más holístico de los cálculos en comparación con los métodos convencionales. Esto significa que las redes neuronales artificiales pueden producir resultados más precisos y confiables, particularmente en escenarios donde los datos de entrada son propensos a tener ruido o carecen de integridad.

- Terrence Sejnowski es un destacado defensor de la computación neuronal en el campo de la conversión de texto a voz. Este proceso implica transformar el texto escrito en lenguaje hablado alterando los símbolos visuales del texto. Sejnowski y Rosemberg introdujeron un sistema de computación neuronal llamado NetTalk, que convierte texto en fonemas y utiliza un sintetizador de voz llamado Dectalk para generar voz. Una ventaja importante de la computación neuronal en la conversión de texto a voz es su capacidad de eliminar la necesidad de programar en la computadora reglas de pronunciación complejas. Si bien el sistema NetTalk demuestra un rendimiento impresionante, la aplicación de la computación neuronal en esta área abre numerosas oportunidades para la investigación y el desarrollo comercial.
- El procesamiento del lenguaje natural implica el examen y análisis de cómo se formulan y estructuran las reglas del lenguaje. En el ámbito de la ciencia del conocimiento, los investigadores Rumelhart y McClelland han incorporado con éxito una red neuronal al campo del procesamiento del lenguaje natural. Este sistema avanzado ha logrado dominio en el aprendizaje de las conjugaciones verbales en tiempo pasado de los verbos en inglés. Las características únicas de la computación neuronal, como la capacidad de extrapolar a partir de datos incompletos y la capacidad de conceptualizar conceptos abstractos, permiten que

el sistema produzca predicciones precisas para verbos desconocidos o inexplorados.

- La compresión de imágenes es el proceso de convertir datos de imágenes en una forma diferente que ocupe menos espacio de almacenamiento o que pueda reconstruirse como una imagen visualmente indistinguible. Investigadores de la Universidad de San Diego y Pittsburgh, concretamente Cottrel, Munro y Zisper, han desarrollado un sistema de compresión de imágenes que utiliza una red neuronal.
- El reconocimiento de caracteres implica el análisis visual y la categorización de símbolos. Nestor, Inc. ha logrado avances significativos en este campo al crear un sistema de computación neuronal que, luego de un período de entrenamiento con una variedad de variaciones de caracteres de letras, posee la capacidad de descifrar e interpretar caracteres o letras desconocidos.
- El reconocimiento de patrones de imágenes se utiliza comúnmente en diversas aplicaciones, como clasificar objetivos identificados por sonar y realizar inspecciones industriales. En el campo del sonar, se han desarrollado redes neuronales artificiales (ANN) que utilizan el algoritmo de retropropagación para imitar el comportamiento de los operadores humanos. Estas ANN han demostrado capacidades similares a las de los humanos para clasificar objetivos con precisión. Además, el reconocimiento de patrones de imágenes se emplea ampliamente en entornos industriales con fines de inspección.
- Los problemas de combinatoria son un tipo específico de problema en el que el método de cálculo tradicional requiere una cantidad significativa de tiempo de procesamiento que aumenta exponencialmente con el número de entradas. Un ejemplo de tal problema es el problema del vendedor, que implica encontrar la ruta más corta para que un vendedor viaje a través de un número limitado de ciudades en un área geográfica específica. Hopfield ha logrado una solución a este problema, quien ha desarrollado una Red Neuronal Artificial que proporciona resultados efectivos para resolver este tipo de problemas combinatorios complejos.
- El procesamiento de señales implica la utilización de redes neuronales artificiales en diversas aplicaciones. Dentro de este dominio, las ANN se han utilizado

ampliamente para tres tipos distintos de tareas de procesamiento de señales: predicción, modelado de sistemas y filtrado de ruido.

- En el ámbito del mundo físico existen numerosos fenómenos que pueden entenderse y predecirse en función de los patrones y tendencias observados en una serie de datos o valores registrados. Una extensa investigación realizada por Lapedes y Farber en el prestigioso Laboratorio de Investigación de Los Álamos ha revelado que la implementación de la retropropagación de la red supera las técnicas tradicionales de predicción lineal y polinomial cuando se trata de pronosticar con precisión series temporales caóticas, lo que demuestra una sorprendente mejora diez veces mayor en la precisión predictiva.
- El modelado de sistemas implica la caracterización de sistemas lineales utilizando una función de transferencia. Esta función de transferencia representa la relación entre la variable de salida y una variable independiente, así como sus derivadas. De manera similar, las redes neuronales artificiales tienen la capacidad de aprender e imitar el comportamiento de una función de transferencia, emulando efectivamente el sistema lineal que se está modelando.
- Las redes neuronales artificiales también se pueden emplear de forma eficaz para eliminar el ruido de una señal, lo que demuestra su capacidad para preservar las estructuras y valores fundamentales de los filtros convencionales. En el ámbito de los modelos económicos y financieros, la previsión juega un papel fundamental en la predicción de diversos aspectos, como los precios de las acciones, el rendimiento de los cultivos, las tasas de interés, los volúmenes de ventas y más. En este ámbito, las redes neuronales han demostrado superar a los métodos tradicionales y ofrecer resultados superiores en la previsión financiera.
- El control de un sistema de servomecanismo complejo plantea una tarea desafiante debido a la necesidad de un método de cálculo computacional adecuado para tener en cuenta las variaciones físicas que surgen dentro del sistema. Uno de los principales obstáculos es la dificultad para medir con precisión estas variaciones y la importante cantidad de tiempo necesaria para calcular la solución matemática. Para abordar este problema se han desarrollado y entrenado diversas redes neuronales para replicar o pronosticar el error que se produce en la posición final

del robot. Al integrar este error con la posición deseada, se logra una corrección de posición adaptativa, mejorando en última instancia la precisión de la posición final.

Clasificación

Hay varias formas de categorizar las redes neuronales según su propósito y uso. Una forma es clasificarlos como clasificadores o regresores.

- Los clasificadores son redes neuronales diseñadas para clasificar datos en diferentes categorías o clases. Asignan una clase discreta a un vector de entradas, lo que significa que determinan a qué categoría pertenece una entrada determinada.
- Por otro lado, los regresores son redes neuronales que se utilizan para tareas de regresión. En lugar de asignar clases discretas, asignan un vector de salida continuo o analógico a un vector de entrada continuo. Esto significa que pueden predecir o estimar valores numéricos basándose en las entradas dadas.

Basadas en su arquitectura:

- Monocapa: Perceptron simple
- Multicapa

La red neuronal monocapa se caracteriza por su simplicidad ya que consta de una sola capa de neuronas. En este tipo de red, las entradas se utilizan directamente para obtener las salidas. Por otro lado, las redes multicapa son más complejas ya que incorporan múltiples capas ocultas entre la capa de entrada y la capa de salida. Estas capas ocultas desempeñan un papel crucial en el procesamiento de la información sensorial y la generación del resultado final de la red.

Basadas en el método de aprendizaje:

- Con método de aprendizaje supervisado
- Con método de aprendizaje no supervisado
- Con método de aprendizaje reforzado

En el aprendizaje supervisado, la red neuronal recibe patrones de entrada y los correspondientes patrones de salida deseados durante el proceso de entrenamiento. Utiliza esta información para ajustar los parámetros internos de cada neurona. Por el contrario, en los enfoques de aprendizaje no supervisados, se desconocen los resultados deseados. En cambio, la red neuronal adapta sus parámetros internos identificando patrones comunes o similitudes dentro de los datos de entrada. Esto podría implicar agrupar puntos de datos y determinar las distancias entre ellos para formar clases o grupos. El aprendizaje por refuerzo, por otro lado, se centra únicamente en evaluar la exactitud del resultado, sin un conocimiento explícito de los patrones deseados.

Se utilizan diferentes técnicas para el aprendizaje no supervisado:

- Por agrupamiento
- Según el Análisis de Componentes Principales (PCA)
- El Aprendizaje competitivo.
- Los Mapas autoorganizados (SOM)

En el proceso de aprendizaje competitivo, las neuronas participan en una batalla, esforzándose por superarse unas a otras. Las neuronas victoriosas se determinan en función de qué tan cerca se alinean sus pesos con el patrón de entrada dado. Como resultado, el aprendizaje se produce a medida que las conexiones de la neurona triunfante se refuerzan, mientras que las conexiones de las otras neuronas experimentan un efecto de debilitamiento.

En los mapas autoorganizados (SOM), el proceso de agrupar datos en función de la similitud se emplea para proyectarlos de manera efectiva en un mapa, facilitando la creación de distintos grupos o clases. Por otro lado, el Análisis de Componentes Principales (PCA) sirve para reducir la dimensionalidad de los datos describiéndolos en un nuevo conjunto de variables que no están correlacionadas entre sí. Esta reducción de dimensionalidad permite que la red neuronal maneje los datos de manera más eficiente al simplificar su complejidad. Además, PCA se puede utilizar en redes neuronales para agrupar los datos de manera competitiva, proporcionando más información y comprensión.

Basada en su dependencia temporal:

- Las redes neuronales estáticas: dependencia estática
- Las redes neuronales dinámicas: dependencia temporal

Las redes neuronales estáticas no poseen la capacidad de retener información ni recordar experiencias pasadas. Inicialmente se les enseña a través de un proceso de capacitación y posteriormente se les aplica una serie de insumos para generar los resultados correspondientes. Una vez completada la capacitación, los resultados dependen únicamente de los insumos proporcionados. Estas redes se emplean ampliamente y tienen una gran prevalencia en diversos ámbitos.

Por el contrario, las redes neuronales dinámicas ofrecen la capacidad de establecer conexiones entre las entradas y/o salidas actuales y anteriores, dando como resultado la incorporación de memoria. Esto significa que estas redes están diseñadas con la utilización de ecuaciones diferenciales o ecuaciones en diferencias para minimizar cualquier discrepancia o inexactitud en los resultados generados.

Las redes dinámicas se pueden clasificar en dos tipos según su proceso de evolución: redes recurrentes, que experimentan una evolución recurrente, y redes diferenciales, que experimentan una evolución continua.

Basada en su conexión:

- Las redes neuronales prealimentadas
- Las redes neuronales recurrentes

Una red neuronal feedforward (prealimentada) es un tipo de red que no tiene bucles ni ciclos en sus conexiones entre neuronas. Esto lo distingue de las redes neuronales recurrentes. En una red feedforward, la información fluye de manera unidireccional, comenzando desde la capa de entrada, pasando por la capa oculta y finalmente llegando a la capa de salida. Esta falta de bucles garantiza que no haya retroalimentación ni flujo de información hacia atrás en estas redes.

Las redes simples:

- El Perceptron imple
- El percepción multicapa

- Adaline y Madaline
- Las redes neuronales de base radial (RBF)
 - Las redes neuronales dinámicas recurrentes (RNN)
 - Las redes neuronales profundas o convolucionales (CNN)
 - Las redes neuronales derivativas

En su investigación, Widrow y su alumno Hoff realizaron un examen detallado de la red Adaline de una capa y su regla de aprendizaje asociada. Introdujeron el concepto de utilizar el algoritmo LMS, también conocido como mínimo cuadrado medio, para entrenar esta red. La red Adaline, que es un tipo de red lineal, comparte similitudes con la red perceptron. Sin embargo, la distinción clave radica en la función de transferencia utilizada por Adaline, que es de naturaleza lineal en lugar de limitante.

Esta característica única permite a Adaline producir valores analógicos como salidas, en contraste con la red de perceptrones, que está limitada a generar salidas de 0 o 1. En consecuencia, las redes de perceptrones solo son capaces de resolver problemas linealmente separables. Vale la pena señalar que la red Madaline, por otro lado, es una red Adaline multicapa.

La red neuronal multicapa

Las redes neuronales suelen constar de varias capas, y cada capa comprende una matriz de peso W , un vector de polarización b y un vector de salida a . La notación de superíndice se emplea para diferenciar las variables que pertenecen a distintas capas en las matrices de ponderación, sesgos y vectores de salida. La siguiente figura demuestra visualmente la utilización de esta notación en una red neuronal de tres capas, mientras que las ecuaciones correspondientes a cada capa se representan en la parte inferior de la figura (Vorobioff, 2022).

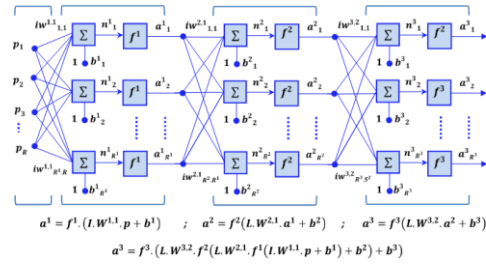
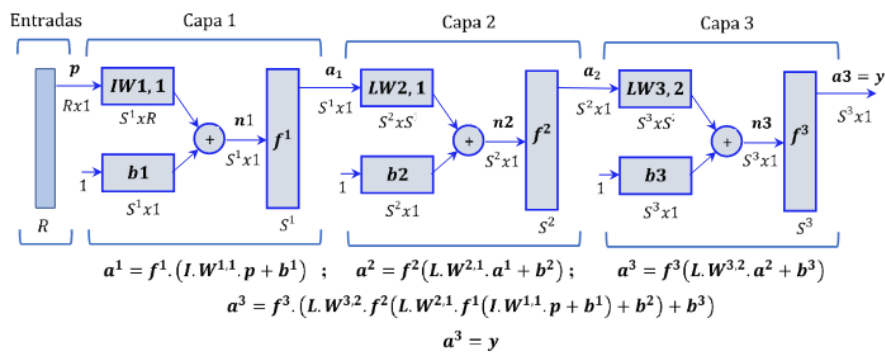


Fig. 9 – Red Neuronal Multicapa

De igual manera la red neuronal de tres capas se puede representar de forma abreviada (Vorobioff, 2022).



Al utilizar una red neuronal sencilla de dos capas que incorpora funciones de activación sigmoidea en la primera capa y funciones lineales en la segunda capa, es posible aproximar cualquier función, aunque con un número limitado de discontinuidades. Esta configuración particular de una red neuronal de dos capas se emplea ampliamente en diversas aplicaciones de redes neuronales, junto con la utilización del algoritmo de retropropagación. En la segunda ilustración se muestra la salida de la red que corresponde a la salida de la capa final, específicamente denominada capa 3, denotada como $yy = aa3$. Vale la pena señalar que yy significa la salida de la red.

Procesamiento: las entradas y las salidas

Las entradas de la red pueden ir acompañadas de funciones de procesamiento que modifican los datos de entrada para hacerlos más convenientes o eficientes para la red. Un ejemplo de dicha función es la función `mapminmax` en Matlab®, que transforma los datos de entrada para que todos los valores estén dentro del rango de -1 a 1. Esto puede

mejorar el proceso de aprendizaje para muchas redes. Otra función comúnmente utilizada en Matlab® es `removeconstantrows`, que elimina las filas del vector de entrada que corresponden a elementos de entrada que siempre tienen el mismo valor. Esto se debe a que estos elementos no aportan ninguna información útil a la red.

La función `fixunknowns` es otra función de procesamiento que desempeña un papel en la recodificación de datos desconocidos (representados por valores NaN en Matlab) en un formato numérico adecuado para la red. Esta función también mantiene la información sobre qué valores se conocen y cuáles se desconocen. De manera similar, las salidas de la red también pueden estar sujetas a funciones de procesamiento. Estas funciones transforman los vectores de salida para alinearlos con las salidas esperadas, a menudo reescalándolos. Al hacerlo, los datos de salida conservan las mismas características que los objetivos originales proporcionados por el usuario.

Las redes neuronales y los sistemas adaptativos

En el ámbito de los sistemas adaptativos, nuestro enfoque implica la utilización de redes neuronales que poseen la capacidad de aprender y adaptarse en función de su entorno. Este proceso de aprendizaje se lleva a cabo de forma supervisada, es decir, proporcionamos a la red el resultado deseado, permitiéndole ajustar sus parámetros a través de un proceso de aprendizaje iterativo. Este ajuste se logra modificando sistemáticamente una colección de parámetros libres conocidos como pesos sinápticos. Estos pesos sinápticos juegan un papel crucial en el almacenamiento y retención de información obtenida de los datos de entrada.

Cuando se trata de aplicaciones de procesamiento de señales adaptativas, las redes neuronales ofrecen varias ventajas en comparación con los sistemas adaptativos lineales:

- La no linealidad se refiere a la característica de numerosos sistemas físicos donde su comportamiento no sigue un patrón lineal. Este concepto es esencial ya que permite la inclusión y consideración de una amplia gama de sistemas que exhiben un comportamiento no lineal. Al reconocer la prevalencia de la no linealidad en varios sistemas físicos, podemos comprender y analizar mejor sus complejas dinámicas e interacciones.

- La capacidad de aprendizaje de un sistema se puede determinar en función de la información que recibe como entradas y los resultados que produce como salidas.
- La generalización se refiere a la capacidad de manejar eficazmente sistemas desconocidos mediante el procesamiento de entradas que no se han encontrado previamente dentro de la red. Esto implica que la red posee la capacidad de adaptarse y funcionar con precisión incluso en situaciones en las que carece de conocimientos o experiencia previa.
- El concepto de tolerancia a errores en las redes neuronales reconoce que incluso si ciertas neuronas dentro de la red no funcionan de manera óptima, el sistema en general aún puede producir resultados satisfactorios.
- Integración a gran escala y capacidades de procesamiento paralelo

El perceptron

En 1943, W. McCulloch y W. Pitts introdujeron el concepto de neuronas artificiales, lo que marcó un hito importante en el desarrollo de la inteligencia artificial. Estos primeros modelos operaban calculando una suma ponderada de señales de entrada y comparándola con un umbral predeterminado. Si la suma excediera o igualara el umbral, la salida de la neurona sería 1; de lo contrario, sería 0.

El aspecto destacable de estas redes fue su capacidad para realizar diversas funciones aritméticas y lógicas, lo que las hacía muy versátiles y capaces de resolver problemas complejos. A diferencia de sus homólogas biológicas, los parámetros de estas neuronas artificiales debían diseñarse manualmente, ya que en aquel momento no existían métodos de entrenamiento establecidos. Sin embargo, esto no detuvo a los científicos e investigadores que quedaron fascinados por el potencial de estos ordenadores digitales, inspirados en el intrincado funcionamiento de las neuronas biológicas.

En 1957, un grupo de investigadores dirigido por F. Rosenblatt logró un avance significativo en el campo de las redes neuronales al desarrollar los primeros perceptrones. Estos perceptrones, similares a las redes de McCulloch y Pitts, estaban equipados con una regla de aprendizaje que les permitía entrenarse en la resolución de problemas de reconocimiento de patrones. Esta introducción de una regla de aprendizaje fue un paso

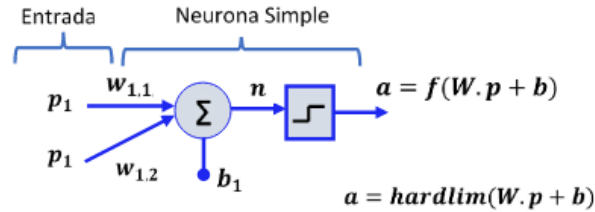
crucial en el avance de las redes neuronales, ya que aseguró que los perceptrones siempre convergerían a los pesos correctos de la red, siempre que hubiera pesos disponibles para resolver el problema en cuestión. Los investigadores lograron esto presentando a la red ejemplos de comportamiento deseado y permitiéndole aprender de sus errores. Incluso cuando se inicializó con valores aleatorios para ponderaciones y sesgos, el perceptron pudo aprender y mejorar mediante el uso de la regla de aprendizaje. Sin embargo, cabe señalar que estas primeras redes de perceptrones tenían sus limitaciones. No pudieron implementar ciertas funciones elementales, lo que obstaculizó sus capacidades generales.

No fue hasta la década de 1980 que estas limitaciones se superaron con el desarrollo de redes de perceptrones multicapa y sus reglas de aprendizaje asociadas. Estas redes mejoradas pudieron abordar tareas más complejas y proporcionaron un avance significativo en el campo de las redes neuronales. A pesar de estos avances, la red de perceptrones sigue teniendo importancia en la actualidad. Sirve como una red rápida y confiable para resolver aplicaciones simples. Además, el estudio de la red de perceptrones proporciona una base sólida para comprender redes más complejas y profundizar en el campo de las redes neuronales. En las siguientes secciones, profundizaremos en el funcionamiento de la red de perceptrones y exploraremos su regla de aprendizaje con mayor detalle.

En el ámbito de la biología, se sabe que una neurona libera una señal como salida sólo cuando la suma acumulada de las señales entrantes supera un umbral específico. Este fascinante comportamiento se replica en el modelo de perceptron al calcular la suma ponderada de todas las señales de entrada, reflejando el funcionamiento de las redes neuronales biológicas. Además, la salida generada por el perceptron se transmite posteriormente a otras redes de perceptrones interconectadas, como destacó Kundella en su investigación realizada en 2020.

La red neuronal Perceptron se basa en una función de transferencia conocida como función "hard lim". Esta función se utiliza para determinar la salida de la red comparando la suma ponderada de las entradas con un valor umbral (Vorobioff, 2022).

$$a = \text{hardlim}(W \cdot p + b)$$



Los filtros adaptativos de redes neuronales

Las redes adalinas, también conocidas como neuronas lineales adaptativas, comparten similitudes con los perceptrones pero difieren en el uso de una función de transferencia lineal en lugar de una función limitante. Esta distinción permite a las redes Adaline generar resultados con una gama más amplia de valores, mientras que los perceptrones están restringidos a producir 0 o 1 como resultado. Tanto las redes Adaline como las de perceptrones son capaces de resolver problemas que son linealmente separables. Si bien, la regla de aprendizaje empleada en las redes Adaline, conocida como LMS o mínimos cuadrados medios, es significativamente más poderosa que la regla de aprendizaje utilizada en los perceptrones. El LMS, también conocido como regla de aprendizaje de Widrow-Hoff, tiene como objetivo minimizar el error cuadrático medio y, en consecuencia, desplaza los límites de decisión lo más lejos posible de los patrones de entrenamiento.

Un sistema lineal adaptativo equipado con una red neuronal tiene la capacidad de adaptarse y responder a las variaciones de su entorno en tiempo real. Estas redes lineales son capaces de ajustar sus pesos y sesgos en cada paso del tiempo teniendo en cuenta nuevos vectores de entrada y salida. El objetivo es encontrar los pesos y sesgos óptimos que minimicen la suma del error cuadrático medio de la red para los vectores de entrada y objetivo más recientes. Estas redes han encontrado amplias aplicaciones en sistemas de control, procesamiento de señales y sistemas de cancelación de errores. Los pioneros en este ámbito, Widrow y Hoff, acuñaron el término Adaline para describir estos elementos lineales adaptativos.

Las funciones adaptativas

La función de adaptación es responsable de modificar gradualmente los pesos y sesgos de una red mientras se entrena. Cabe señalar que la regla de Widrow-Hoff, que es

un método utilizado para entrenar redes lineales de una sola capa, posee una limitación en su aplicabilidad. Sin embargo, esta limitación no es un inconveniente importante porque las redes lineales de una sola capa son tan poderosas como sus contrapartes multicapa. En otras palabras, para cada red lineal multicapa, existe una red lineal equivalente de una sola capa que puede alcanzar el mismo nivel de rendimiento.

El reconocimiento estadístico de patrones: redes neuronales

El reconocimiento de patrones implica la investigación, examen y manipulación de datos provenientes de procedimientos científicos y tecnológicos que pertenecen a entidades tangibles e intangibles. El objetivo final es discernir y extraer conocimientos valiosos de varios grupos o categorías de estas entidades.

El reconocimiento de patrones

El aprendizaje automático a menudo se describe como un proceso que imita el funcionamiento del cerebro humano. Sin embargo, es importante señalar que, si bien los algoritmos imitan la inteligencia humana hasta cierto punto, operan de una manera distinta. El aprendizaje automático ha demostrado ser eficaz para resolver problemas complejos y generar mejores resultados en diversos campos. Su implementación puede resultar desafiante en determinadas aplicaciones, especialmente cuando faltan datos de entrenamiento suficientes o cuando las variables son difíciles de medir.

Los humanos poseen la capacidad de detectar e interpretar fácilmente diversas variables o percepciones, como hacer diagnósticos médicos, interpretar radiografías o analizar el comportamiento social. No obstante, con programas adecuadamente capacitados, el aprendizaje automático puede lograr resultados superiores en comparación con los humanos. Aun cuando, es fundamental tener precaución al utilizar estos resultados, ya que los sistemas de aprendizaje automático no son infalibles.

Descripción del reconocimiento

Un patrón se refiere a una ocurrencia o disposición consistente dentro de una colección de información o en aspectos conceptuales específicos. El proceso de reconocimiento de patrones implica identificar y descubrir consistencias y semejanzas entre los datos mediante el empleo de diversas técnicas, como la medición o el

aprendizaje a partir de datos. Estas semejanzas se pueden detectar mediante análisis estadístico, examinando datos pasados o empleando algoritmos.

En el campo del reconocimiento de patrones, el paso inicial consiste en recopilar datos. Luego, estos datos se someten a filtrado y preprocesamiento para permitir que el sistema identifique y extraiga características relevantes. La selección del algoritmo apropiado para el reconocimiento de patrones, ya sea Clasificación, Agrupación o Regresión, depende del tipo particular de sistema de datos.

Patrones:

- La clasificación supervisada se refiere a un conjunto de algoritmos que tienen la capacidad de categorizar nuevos objetos utilizando información obtenida de muestras previamente categorizadas. Para lograr esto, estos algoritmos reciben datos de entrenamiento que consisten en entradas junto con las respuestas correspondientes o grupos de membresía. Este tipo de clasificación implica que el algoritmo asigne etiquetas específicas a los datos en función de atributos predeterminados. Cabe mencionar que la clasificación supervisada es un componente fundamental del aprendizaje supervisado.
- La clasificación no supervisada implica el proceso de identificar la clasificación de una muestra no clasificada sin ningún conocimiento previo de las respuestas correctas. Durante el proceso de formación las respuestas no se dan ni se conocen.
- La agrupación, también conocida como clustering, es un proceso algorítmico que implica dividir datos en múltiples grupos según la similitud de sus características. Este tipo de aprendizaje se considera no supervisado porque no se basa en etiquetas o clasificaciones predeterminadas. En ciertos casos, los datos que necesitamos analizar pueden ser extremadamente complejos, lo que dificulta categorizarlos en distintos grupos. Sin embargo, las redes neuronales tienen la capacidad de reconocer patrones y características únicos dentro de estos datos complejos, lo que les permite clasificarlos en diferentes grupos sin ningún conocimiento o información previa sobre los datos. Esta técnica es particularmente valiosa en los campos de la minería de datos, tanto con fines comerciales como científicos, ya que permite la extracción eficiente de conocimientos y patrones significativos a partir de conjuntos de datos complejos.

- Los algoritmos de regresión están diseñados para establecer conexiones entre variables de entrada y salida, permitiendo la predicción de variables dependientes desconocidas utilizando las relaciones identificadas. Estos algoritmos operan bajo el marco del aprendizaje supervisado.
- Una aplicación de las redes que funcionan correctamente es la capacidad de predecir valores futuros basándose en una secuencia conocida de valores. Este concepto se puede ver en diversos campos, como el mercado de valores, donde predecir el comportamiento futuro de las acciones resulta de gran interés.
- A las redes neuronales se les puede enseñar a reconocer y recordar una secuencia de patrones. Esto significa que cuando se muestra una versión ligeramente alterada de un patrón específico, la red es capaz de vincularla al patrón más similar que se le haya enseñado y recuperar la versión original de ese patrón específico.

La clasificación, en un contexto matemático, implica la partición de un espacio multidimensional en múltiples regiones. Su finalidad es determinar la región a la que pertenece un punto determinado del espacio. Este concepto encuentra aplicación en numerosos escenarios de la vida real, como en diversos programas de reconocimiento de patrones.

En estos programas, cada patrón se convierte en un punto multidimensional y luego se clasifica en un grupo específico, cada uno de los cuales representa un patrón conocido. La selección de variables apropiadas se vuelve crucial, ya que es necesario identificar las características o variables más adecuadas para describir y analizar los objetos bajo estudio. El reconocimiento de patrones se puede lograr mediante el uso de redes neuronales o empleando métodos estadísticos (Reconocimiento estadístico de patrones: REP).

La regresión

Cuando se trata de problemas de ajuste de datos, el objetivo principal de una red neuronal es identificar las salidas numéricas asociadas con un conjunto determinado de entradas numéricas. Para realizar esta tarea, nftool, que es una herramienta de ajuste de datos, emplea una red neuronal feedforward de dos capas. Esta red neuronal se entrena mediante algoritmos como Levenberg Marquardt, gradiente conjugado o métodos

bayesianos. Al utilizar esta herramienta, los usuarios tienen la flexibilidad de cargar sus propios datos o importar un conjunto de datos desde Matlab®.

En el ámbito de los problemas de reconocimiento de patrones que implican clasificación, la red neuronal está diseñada para categorizar las entradas en un conjunto predeterminado de categorías de salida. En otras palabras, el objetivo es determinar la categoría específica que corresponde al insumo analizado. Para realizar esta tarea en Matlab®, tiene la opción de utilizar la herramienta `net = patternnet(hiddenSizes, trainFcn, performFcn)`.

Cuando se trata de problemas de agrupamiento, el objetivo es emplear una red neuronal que pueda categorizar datos según sus similitudes. Un enfoque eficaz es utilizar redes equipadas con mapas autoorganizados (SOM). Estas redes están compuestas por una capa competitiva que posee la capacidad de clasificar un conjunto de datos que comprende vectores de diferentes dimensiones. El número de clasificaciones que puede realizar está determinado por el número de neuronas dentro de la capa. Estas neuronas están dispuestas en una estructura bidimensional, lo que permite que la capa cree una representación de la distribución del conjunto de datos y genere una aproximación bidimensional de su topología.

Los datos de entrenamiento: validación y de testeo

Los datos de entrenamiento se utilizan para entrenar el modelo ajustando sesgos y ponderaciones, particularmente en el caso de una red neuronal. En otras palabras, el modelo aprende de los datos de entrenamiento. Por otro lado, el conjunto de datos de validación sirve para evaluar la generalización del modelo y determinar cuándo detener el entrenamiento si el rendimiento del modelo ya no mejora. Este conjunto de datos permite la evaluación del rendimiento del modelo mientras se entrena con los datos de entrenamiento. Por tanto, los datos de validación influyen indirectamente en la etapa de desarrollo del modelo.

Por el contrario, el conjunto de datos de prueba no afecta el proceso de capacitación y se utiliza únicamente para medir el rendimiento de la red de forma independiente después de haber sido entrenada. Antes de utilizar los datos de prueba, la red debe entrenarse utilizando tanto los datos de entrenamiento como los de validación. Ocasionalmente, el conjunto de validación se puede utilizar como sustituto del conjunto

de prueba, pero no se recomienda esta práctica. Idealmente, el conjunto de prueba debería abarcar datos que representen todas las clases posibles, lo que permitirá que la red funcione correctamente en escenarios del mundo real. En la siguiente ilustración se puede ver un ejemplo de cómo asignar porcentajes para datos de entrenamiento, datos de validación y datos de prueba. Estos porcentajes pueden variar y no existe un criterio universalmente aplicable; se pueden encontrar diferentes recomendaciones en la literatura (Vorobioff, 2022).



El reconocimiento estadístico de los patrones

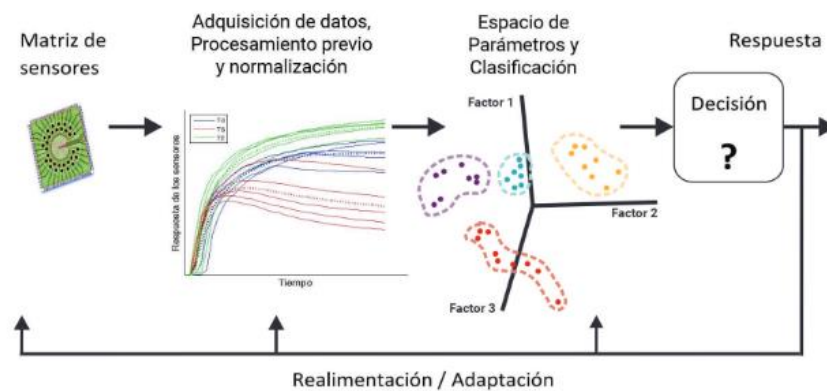
El reconocimiento de patrones estadísticos (REP) es un método que se basa en la teoría de la probabilidad y la estadística para analizar datos. Opera bajo el supuesto de que las mediciones que se analizan siguen distribuciones de probabilidad conocidas. El proceso de reconocimiento implica utilizar estas distribuciones para hacer inferencias y decisiones. En cuanto al aspecto de reconocer los patrones, se considera que es un proceso integral que abarca todos los aspectos de la investigación y la resolución de problemas. Implica analizar datos mediante discriminación y clasificación, con el objetivo de comprender y evaluar los resultados obtenidos.

El sistema de reconocimiento de patrones dentro de un sistema de medición, que se puede dividir en distintas etapas:

- el sistema de adquisición de datos, responsable de recolectar las mediciones;
- el sistema de extracción de parámetros, que identifica características relevantes de los datos;
- los clasificadores, que categorizan los datos en función de sus patrones; y

- la estrategia de toma de decisiones, que determina los resultados finales con base en los resultados de la clasificación.

Cuando se trata de mediciones que se dividen en varios grupos, los clasificadores no siempre proporcionan una indicación clara del grupo específico de la medición que se está realizando. En cambio, es posible que solo proporcionen un valor analógico. En tales casos, se hace necesario establecer límites y técnicas para determinar el resultado apropiado. La próxima ilustración representa un sistema básico de reconocimiento de patrones, en el que las diferentes etapas pueden incorporar retroalimentación de salida (Vorobioff, 2022).



Las técnicas de reconocimiento de patrones

Existen diversas formas que se pueden clasificar en:

- Reconocimiento de patrones estadísticos (REP)
- Técnicas que utilizan inteligencia artificial (IA), incluido el uso de redes neuronales y lógica difusa.

Los datos se pueden dividir en dos categorías principales: variables parcialmente independientes, también conocidas como variables de medición, y variables dependientes, que pueden denominarse clases o grupos.

Los algoritmos de reconocimiento de patrones se pueden clasificar en enfoques supervisados y no supervisados. Los métodos no supervisados implican la exploración

de datos, mientras que los métodos supervisados implican entrenar los algoritmos con resultados conocidos para fines de clasificación.

- Los algoritmos supervisados se utilizan para asignar un descriptor o salida a un vector de datos de entrada determinado en función de sus mediciones. Este descriptor, también conocido como salida o respuesta del sistema, es un vector que se determina durante el proceso de entrenamiento. La capacitación implica organizar un conjunto de mediciones en diferentes grupos o clases en función de sus similitudes, y el vector descriptor ayuda a identificar el grupo específico al que pertenece cada medición. Para que los algoritmos clasifiquen con precisión nuevos datos, es necesario entrenarlos y establecer sus parámetros internos. Una vez finalizada la fase de aprendizaje o entrenamiento, el algoritmo puede analizar datos de una medición que pertenece a un grupo desconocido y clasificarlos en función del entrenamiento al que ha sido sometido.
- Los algoritmos no supervisados se diferencian de los supervisados en que no se basan en descriptores preasignados. En cambio, estos algoritmos se entrenan y modifican sus parámetros identificando y aprovechando patrones o similitudes dentro de los datos.

Los métodos estadísticos se clasifican como paramétricos porque operan bajo el supuesto de que los datos pueden caracterizarse mediante funciones de densidad de probabilidad. Dentro de esta categoría, hay varias técnicas disponibles, como el análisis de componentes principales (PCA), el análisis de factores discriminantes (DFA), el análisis de la función de densidad de probabilidad utilizando el teorema de Bayes, el método de regresión de mínimos cuadrados parciales (PLS) y los algoritmos de separación de grupos como la agrupación jerárquica y k-means.

Las técnicas de inteligencia artificial (IA), tal como las definen, abarcan una amplia gama de métodos que se inspiran en modelos biológicos. Estas técnicas se pueden clasificar en tres subgrupos distintos, cada uno de los cuales ofrece su propio enfoque único y un conjunto de principios intuitivos. Al explorar y aprovechar el poder de estas técnicas de IA, los investigadores y desarrolladores han podido profundizar en el ámbito de los sistemas inteligentes y crear soluciones innovadoras que imitan las capacidades de los organismos vivos.

- Las redes neuronales artificiales (ANN) constan de varios componentes, como la propagación hacia atrás de errores (BP), el perceptron multicapa (MLP) y las redes de funciones de base radial (RBF). Además, existen mapas autoorganizados (SOM), cuantificación de vectores de aprendizaje (LVQ), redes dinámicas recurrentes y sistemas adaptativos. Estos diferentes elementos y técnicas contribuyen a la funcionalidad y eficacia general de las redes neuronales artificiales.
- Los algoritmos de lógica difusa, también conocidos como lógica difusa, y la aplicación de reglas o razonamientos difusos, desempeñan un papel importante en diversos campos e industrias. Estos algoritmos y reglas están diseñados para manejar la incertidumbre y la vaguedad en los procesos de toma de decisiones, proporcionando así un enfoque más flexible y matizado para la resolución de problemas. Al incorporar lógica difusa, los sistemas pueden capturar y procesar con precisión información imprecisa o incompleta, lo que genera mejores resultados y capacidades de toma de decisiones. Los algoritmos de lógica difusa se utilizan ampliamente en áreas como la inteligencia artificial, sistemas de control, reconocimiento de patrones y análisis de datos, entre otras, destacando su versatilidad y eficacia para abordar problemas complejos del mundo real. La esencia de la lógica difusa radica en su capacidad de imitar el razonamiento humano al considerar múltiples grados de verdad, lo que permite modelar con mayor precisión fenómenos inciertos o subjetivos. Como resultado, los algoritmos de lógica difusa han ganado prominencia y continúan mejorando las capacidades de diversas aplicaciones, permitiendo sistemas más inteligentes y adaptables en un mundo cada vez más incierto y complejo.
- Se han empleado algoritmos genéticos para seleccionar parámetros.

El aprendizaje y la generalización

La generalización de una red neuronal tiene como objetivo lograr un buen desempeño cuando se le presentan nuevos inputs que no fueron utilizados durante el entrenamiento. Estas nuevas entradas, conocidas como entradas de prueba, no se utilizan para ajustar los parámetros internos de la red. Las redes neuronales pueden funcionar como clasificadores, asignando clases discretas a vectores de entrada, o como regresores,

asignando vectores de salida continuos a vectores de entrada continuos. Inicialmente no se conocen las funciones de clasificación y regresión inferidas por las redes neuronales. En cambio, se utiliza un conjunto de entrenamiento para proporcionar ejemplos de entrada y salida de la función.

A través del entrenamiento, la red neuronal identifica esta "función desconocida" basándose únicamente en los datos de entrenamiento proporcionados. Los parámetros de la función, como los pesos y los sesgos de las neuronas, se estiman para replicar la relación entre las entradas y salidas del entrenamiento con la mayor precisión posible. Además, se espera que la red tenga un buen desempeño con nuevos datos, lo que indica que se ha generalizado con éxito. Sin embargo, lograr un rendimiento de generalización óptimo en datos nuevos no significa necesariamente replicar los datos de entrenamiento a la perfección.

Por ejemplo, si solo hay unos pocos patrones de entrenamiento pero una red neuronal grande, puede ser fácil encontrar pesos que reproduzcan el conjunto de entrenamiento, pero es poco probable que la red resultante haya aprendido efectivamente a manejar nuevos datos. Por el contrario, si existen numerosos patrones de entrenamiento y la red está entrenada para replicarlos, es más probable que responda correctamente a nuevos datos. Estas intuiciones deben perfeccionarse utilizando métodos que mejoren las capacidades de generalización de las redes neuronales.

La evaluación de la generalización

La generalización de una red se puede determinar probando su rendimiento utilizando nuevos datos. Sin embargo, es importante tener cuidado con el tipo de datos utilizados. Si utilizamos constantemente el mismo conjunto de datos de entrenamiento, incluso si el algoritmo de entrenamiento ya no lo utiliza, esencialmente nos estamos esforzando por lograr el mejor rendimiento únicamente en ese conjunto específico. Normalmente, trabajamos con tres conjuntos de datos: el conjunto de entrenamiento, que se utiliza para ajustar los pesos y sesgos de la red; el conjunto de validación o desarrollo, que se utiliza durante la capacitación para evaluar el desempeño actual de la red y guiar el proceso de capacitación; y el conjunto de prueba, que consta de datos desconocidos para los que queremos encontrar respuestas una vez que la red ha sido entrenada. Es

importante tener en cuenta que el conjunto de pruebas no influye en el proceso de formación.

El entrenamiento y la generalización

Durante la fase de entrenamiento, una red neuronal puede realizar diversas funciones ajustando sus parámetros en función de una arquitectura específica. Sin embargo, no siempre es posible que la red aprenda con precisión todos los datos de entrenamiento en las tareas de reconocimiento de patrones. De hecho, hay casos en los que ni siquiera es deseable que la red lo haga. Para evaluar el desempeño de la red en problemas del mundo real después del entrenamiento, se utiliza una métrica llamada generalización. Esta medida evalúa qué tan bien se desempeña la red en datos de evaluación que no se utilizaron durante el proceso de capacitación. Aunque la red compleja puede responder bien a los datos de entrenamiento, es posible que no se generalice bien a datos nuevos e invisibles.

Por otro lado, la red simple con una representación en línea recta puede ser más adecuada para representar con precisión los nuevos datos de prueba. En el proceso de entrenamiento, la red se entrena de la forma habitual minimizando la función de error con respecto al conjunto de datos de entrenamiento. Sin embargo, el rendimiento en un conjunto de entrenamiento se estima utilizando el conjunto de validación. Por ejemplo, si comparamos el rendimiento de dos redes en un pequeño conjunto de datos de entrenamiento, siendo una red simple que implementa una línea recta (que representa una red de una sola capa) y la otra es una red multicapa más compleja con numerosas redes ocultas. unidades. La red compleja puede aproximarse perfectamente a la función, mientras que la red de una sola capa ajusta los datos razonablemente bien con una línea recta, pero con una mayor tasa de error.

Generalmente se puede decir:

- Lograr buenos resultados con los datos de entrenamiento no siempre garantiza buenos resultados con los datos generalizados.
- Las soluciones más simples tienden a tener una mayor probabilidad de generalizarse de manera efectiva en comparación con las soluciones complejas, a menos que haya una cantidad sustancial de datos de entrenamiento que sugieran

lo contrario. Esto significa que, en la mayoría de los casos, mantener las soluciones simples y directas conduce a un mejor rendimiento y aplicabilidad generales. Si bien, es crucial considerar la influencia de un volumen significativo de datos de entrenamiento, ya que puede revelar excepciones en las que las soluciones complejas superan a las más simples.

- Las redes complejas requieren una mayor cantidad de datos de entrenamiento. Esto significa que para que las redes complejas aprendan y se adapten eficazmente, necesitan introducir una cantidad significativamente mayor de datos en su sistema. Sin una cantidad suficiente de datos de entrenamiento, las redes complejas pueden tener dificultades para procesar y comprender con precisión patrones y relaciones complejos dentro de los datos que reciben. Por lo tanto, es crucial proporcionar una cantidad sustancial de datos de entrenamiento para garantizar el rendimiento y la eficiencia óptimos de redes complejas.
- Un modelo que carece de complejidad es rígido y propenso a sesgos importantes.
- Un modelo complejo posee un mayor nivel de adaptabilidad cuando se trata de datos de entrenamiento y muestra sesgos mínimos.

La redes neuronales con mapas autoorganizados

Las redes neuronales autoorganizadas, comúnmente conocidas como SOM, son técnicas no supervisadas sofisticadas e intrincadas que se utilizan en el análisis de datos. Estas redes tienen la capacidad de transformar y proyectar datos en un espacio novedoso, generando así mapas que exhiben representaciones discretas. Las neuronas constituyentes de la red poseen la capacidad de organizarse de forma autónoma y participar en interacciones competitivas entre sí.

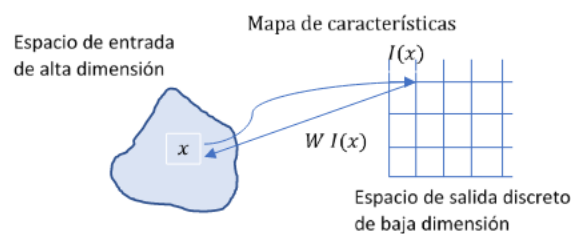
Esta autoorganización se facilita mediante la utilización de funciones de proximidad o vecindad. Una de las ventajas clave de los SOM es su eficacia para reducir la dimensionalidad de los datos de entrada, lo que a menudo da como resultado la visualización de resultados a través de mapas bidimensionales. Vale la pena señalar que los SOM también se denominan mapas de Kohonen, en homenaje al profesor Teuvo Kohonen, quien inicialmente conceptualizó y describió este modelo.

Un tipo fascinante de sistema no supervisado se centra en el aprendizaje competitivo, donde las neuronas de salida participan en una batalla por la activación, lo que da como resultado que solo se activen las neuronas victoriosas. Para facilitar esta competencia, se establecen vías de retroalimentación entre las neuronas, obligándolas a organizarse en consecuencia. Esta red se conoce como mapa autoorganizado (SOM).

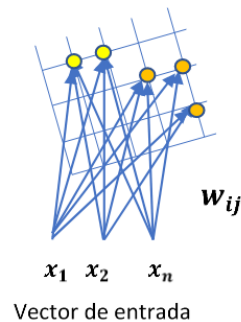
El objetivo principal de un mapa autoorganizado (SOM) es convertir un patrón de una señal de entrada con cualquier número de dimensiones en un nuevo mapa discreto con una o dos dimensiones. Esta transformación debe realizarse de manera que sea adaptable y mantenga un orden específico. Para lograr esto, configuramos nuestro SOM colocando neuronas en los nodos de una red que puede ser unidimensional o bidimensional. Si bien es posible utilizar mapas con dimensiones mayores, normalmente no se utilizan.

Las neuronas poseen la capacidad de adaptarse específicamente a diferentes patrones de entrada o categorías de patrones de entrada a través del aprendizaje competitivo. Esto implica organizar las neuronas según su ubicación y establecer un nuevo sistema de coordenadas para las entradas en la red neuronal. Este proceso puede verse como una expansión del PCA, que es un método lineal para analizar componentes principales.

En la siguiente figura, los datos se representan en un mapa autoorganizado mediante un proceso de mapeo de las entradas x del espacio de entrada al espacio de salida, lo que da como resultado los puntos $I(x)$. Luego, cada punto I en el espacio de salida se asigna nuevamente a su punto correspondiente $w(I)$ en el espacio de entrada. Este proceso de mapeo permite una representación visual de la relación entre los espacios de entrada y salida en el mapa autoorganizado (Vorobioff, 2022).



En este análisis, nos centraremos en la Red Kohonen, que es un tipo específico de Mapa Autoorganizado (SOM). La estructura de este SOM se caracteriza por una disposición feedforward, que consta de una única capa computacional organizada en filas y columnas. Dentro de esta red, cada neurona está conectada a todos los nodos fuente presentes en la capa de entrada. La proyección de datos para estas redes neuronales se puede observar en la próxima ilustración. Cabe señalar que si el mapa se representara de manera unidimensional, solo constaría de una sola fila o columna dentro de la capa computacional (Vorobioff, 2022).



El proceso de autoorganización consta de cuatro componentes principales:

- En la etapa de inicialización, los pesos del sistema se establecen asignándoles pequeños valores aleatorios.
- En un entorno competitivo, las neuronas de la red realizan cálculos para determinar los valores de una función discriminante. Esta función sirve como base para la competencia, ya que la neurona con el valor más bajo es reconocida como la ganadora.
- La cooperación juega un papel crucial en el funcionamiento de las redes neuronales, ya que implica la interacción y coordinación entre neuronas vecinas. Este proceso es facilitado por la neurona ganadora, que ayuda a determinar la ubicación espacial del vecindario topológico que contiene las neuronas excitadas.

En otras palabras, la cooperación asegura que las neuronas próximas trabajen juntas en armonía para lograr el resultado deseado.

- El proceso de adaptación implica que las neuronas excitadas disminuyan los valores de su función discriminante en relación con el patrón de entrada ajustando los pesos de conexión en consecuencia. Este ajuste conduce a una mejora en la respuesta de la neurona ganadora cuando se aplica nuevamente un patrón de entrada similar.

Durante el proceso de competencia, tenemos la capacidad de establecer nuestra función discriminante. Esta función se define como la distancia euclidiana al cuadrado entre el vector de entrada x y el vector de peso w_{ji} para cada neurona j . Esta ecuación nos permite determinar el nivel de competencia entre las neuronas.

$$d_j(x) = \sum_{i=1}^D (x_i - w_{ji})^2$$

En concreto, la neurona ganadora se determina identificando el vector de peso que es más similar al vector de entrada. Al hacerlo, establecemos una conexión entre el espacio de entrada continuo y el espacio de salida discreto de las neuronas a través de un proceso sencillo de competencia neuronal.

En 1982, un investigador Kohonen introdujo el modelo de red conocido como mapas autoorganizados (SOM), que se inspiró en hallazgos fascinantes en el campo de la neurociencia. Este concepto innovador mostró una tremenda promesa para aplicaciones del mundo real, presentando una oportunidad emocionante para una mayor exploración y desarrollo.

La red se distingue porque emplea aprendizaje competitivo no supervisado. Así, a diferencia del aprendizaje supervisado, en el que un profesor externo proporciona retroalimentación sobre el rendimiento de la red, el aprendizaje no supervisado funciona sin dicha orientación. En consecuencia, la red autoorganizada debe identificar de forma autónoma características, patrones, correlaciones o categorías comunes dentro de los datos de entrada e integrarlos en su estructura interna de conexiones.

Lo anterior implica que las neuronas dentro de la red deben autoorganizarse en respuesta a estímulos (datos) recibidos de fuentes externas. Dentro del ámbito del aprendizaje no supervisado, existe un subconjunto de modelos de red que emplean el aprendizaje competitivo. Para el aprendizaje competitivo, las neuronas entablan una competencia entre sí para realizar una tarea determinada. El objetivo de este enfoque de aprendizaje es activar sólo una neurona de salida (o un grupo de neuronas vecinas) cuando se le presenta un patrón de entrada. Por lo tanto, las neuronas compiten entre sí, lo que finalmente resulta en que una neurona emerja como ganadora mientras que las neuronas restantes son suprimidas y forzadas a sus valores mínimos de respuesta.

El objetivo principal de este método de aprendizaje es categorizar o agrupar los datos de entrada que se introducen en la red. Esto implica clasificar información similar como perteneciente a la misma categoría, activando así la misma neurona de salida. La propia red debe generar estas clases o categorías, ya que opera de manera no supervisada, basándose en correlaciones entre los datos de entrada.

Los fundamentos biológicos

En el córtex de los animales superiores existen regiones específicas donde las neuronas que detectan características están dispuestas de manera sistemática (Kohonen, 1989, 1990). Esto significa que la información obtenida del entorno circundante a través de nuestros sentidos se representa internamente como mapas bidimensionales. Por ejemplo, en el área somatosensorial, las neuronas que reciben señales de sensores ubicados cerca de la piel también se encuentran cerca de la corteza.

Como resultado, estas neuronas esencialmente crean un mapa que se asemeja a la superficie de la piel dentro de un área específica de la corteza cerebral. De manera similar, en el sistema visual se han identificado mapas del espacio visual en diferentes regiones del cerebro. Además, en lo que respecta al sentido del oído, hay regiones específicas del cerebro que representan mapas tonotópicos. Estos mapas están organizados de manera que las neuronas que detectan ciertas características relacionadas con el tono de un sonido se organizan en un patrón bidimensional.

Se cree que una parte considerable de la organización neuronal está determinada por la genética, pero también hay evidencia que sugiere que un cierto grado de ella puede estar influenciado por el aprendizaje. Esto implica que el cerebro podría poseer una

capacidad innata para crear mapas topológicos de la información que recibe del entorno externo.

Por el contrario, se ha observado que el impacto que tiene una sola neurona sobre sus neuronas vecinas depende de la distancia entre ellas. Cuando las neuronas están muy separadas, esta influencia es mínima. Las investigaciones han demostrado que ciertos primates experimentan interacciones laterales entre sus neuronas. Estas interacciones pueden ser excitadoras o inhibitoras, dependiendo de la proximidad de las neuronas. Las interacciones excitadoras ocurren dentro de un radio de 50 a 100 micrones, mientras que las interacciones inhibitoras tienen lugar en un anillo circular que varía de 150 a 400 micrones de ancho alrededor del círculo anterior. Asimismo, se producen interacciones excitadoras muy débiles que prácticamente no tienen efecto a partir de ese punto hasta una distancia de varios centímetros. Este patrón distintivo de interacción se asemeja a la forma de un sombrero mexicano, como exploraremos con más detalle más adelante.

El modelo de red autoorganizada propuesto por Kohonen está diseñado para replicar, de manera simplificada, la capacidad del cerebro para crear mapas topológicos basados en señales entrantes del entorno externo.

La arquitectura

Un modelo de mapa autoorganizado (SOM) consta de dos capas de neuronas. La primera capa, conocida como capa de entrada, está compuesta por N neuronas, y cada neurona corresponde a una variable de entrada. Su función principal es recibir y transmitir información desde el entorno externo a la segunda capa, conocida como capa de salida. La capa de salida, por otro lado, es responsable de procesar la información recibida y crear un mapa de características. Normalmente, las neuronas de la capa de salida están dispuestas en un mapa bidimensional, sin embargo, en algunos casos, también se utilizan capas unidimensionales (cadena lineal) o tridimensionales (paralelepípedo).

La red se compone de dos capas y las conexiones entre ellas son siempre en dirección directa. Esto significa que la información fluye desde la capa de entrada a la capa de salida. Cada neurona de entrada está conectada a cada neurona de salida mediante un peso. Estos pesos están representados por un vector de pesos llamado vector

de referencia o libro de códigos. El vector de referencia sirve como prototipo o promedio de la categoría que representa la neurona de salida.

En la capa de salida existen conexiones laterales que pueden describirse como excitación e inhibición implícitas. Aunque estas neuronas no están directamente vinculadas, todavía ejercen una influencia sobre las neuronas vecinas. Esta influencia se establece mediante un proceso competitivo entre las neuronas y la utilización de una función conocida como vecindad, que se analizará más adelante.

El algoritmo

El algoritmo utilizado en el modelo de mapa autoorganizado (SOM) consta de dos etapas principales. En primer lugar, está la etapa operativa donde la red entrenada recibe un patrón de entrada y lo relaciona con la neurona o categoría que tiene el vector de referencia más similar. En segundo lugar, está la etapa de entrenamiento o aprendizaje donde las categorías que componen el mapa se ordenan de manera no supervisada, en base a las relaciones identificadas en el conjunto de datos de entrenamiento.

La etapa del funcionamiento

Cuando se presenta un patrón de entrada p X_p : $x_{p1}, \dots, x_{pi}, \dots, x_{pN}$, se transmite directamente desde la capa de entrada a la capa de salida. En la capa de salida, cada neurona realiza cálculos para determinar la similitud entre el vector de entrada X_p y su propio vector de peso W_j o vector de referencia en función de una medida de distancia específica o un criterio de similitud establecido. Este proceso competitivo implica seleccionar como ganadora la neurona cuyo vector de peso sea más similar al de entrada. La siguiente expresión matemática representa la activación de las neuronas M cuando se presenta el patrón de entrada X_p .

$$y_{pj} = \begin{cases} 1 & \min \|X_p - W_j\| \\ 0 & \text{resto} \end{cases}$$

La variable "ypj" se utiliza para representar la salida o el nivel de activación de las neuronas de salida, y esto está determinado por el resultado de la competición. En esta competición, a una neurona ganadora se le asigna un valor de 1, mientras que a una neurona no ganadora se le asigna un valor de 0. La expresión " $\|X_p - W_j\|$ " se utiliza para medir la similitud entre el vector de entrada (X_p) y el vector de peso (W_j) de las conexiones entre las neuronas de entrada y la neurona de salida j . Esta medida de similitud es crucial para determinar la neurona ganadora.

Durante esta etapa particular de operación, el objetivo principal es identificar el vector de referencia que se parezca mucho al vector de entrada. Esto nos permite determinar qué neurona es la ganadora y, lo que es más importante, determinar la ubicación específica de esta neurona dentro del espacio de salida bidimensional, teniendo en cuenta las interacciones excitadoras e inhibitoras entre las neuronas. Por tanto, la red Self-Organizing Map (SOM) funciona como una herramienta de clasificación, ya que la neurona de salida activada corresponde a la clase a la que pertenece la información de entrada. Además, como entradas similares activan las mismas neuronas de salida o las vecinas, debido a la similitud entre clases, se garantiza que las neuronas topológicamente próximas respondan a entradas físicamente comparables. Como resultado, esta red resulta particularmente valiosa para establecer conexiones no identificadas previamente entre conjuntos de datos.

La etapa de aprendizaje

En primer lugar, es importante reconocer que no existe un algoritmo de aprendizaje universalmente aceptado para la red SOM. A pesar de esto, el procedimiento en sí es conocido por su resiliencia, ya que el resultado final no se ve significativamente influenciado por los detalles específicos de la implementación. Como resultado, nos esforzaremos en esbozar el algoritmo más frecuente típicamente vinculado a este modelo, como lo describe Kohonen en sus trabajos publicados en 1982a, 1982b, 1989 y 1995.

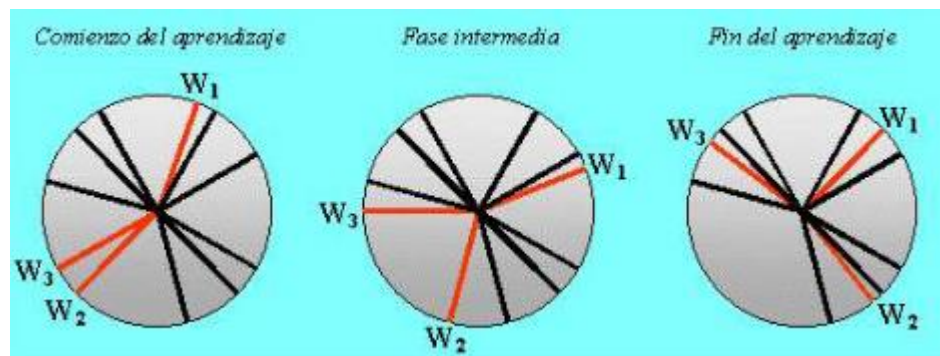
El objetivo principal del algoritmo de aprendizaje es determinar distintas categorías, representadas por las neuronas de salida, mediante la presentación de un conjunto de patrones de entrenamiento. Estas categorías luego se utilizarán durante la fase operativa para clasificar nuevos patrones de entrada.

Así, el proceso de aprendizaje se puede simplificar y entender de la siguiente manera. Cuando se introduce y procesa un vector de entrada, la neurona ganadora se determina comparando su vector de peso con el vector de entrada. La neurona con el vector de peso más similar se considera ganadora. Posteriormente, el vector de peso de la neurona ganadora se ajusta para que se parezca más al vector de entrada. Este ajuste garantiza que cuando se presente el mismo patrón de entrada en el futuro, la neurona ganadora responderá aún con más fuerza. Este proceso se repite para un conjunto de patrones de entrada que se presentan repetidamente a la red. Finalmente, los diferentes vectores de peso se alinean con uno o más patrones de entrada, creando dominios específicos dentro del espacio de entrada. Si estos dominios se agrupan, cada neurona se especializa en uno de ellos. Esta interpretación nos permite ver la función principal de la red como un análisis de conglomerados.

Una forma interesante de comprender el funcionamiento de la red SOM es a través de una interpretación geométrica propuesta por Masters en 1993 (Palmer et al., 2002). Esta interpretación arroja luz sobre el proceso de aprendizaje de la red. Esencialmente, la regla de aprendizaje empleada en la red SOM tiene como objetivo acercar repetidamente el vector de peso de la neurona con mayor actividad (también conocida como ganadora) al vector de entrada. En términos más simples, la regla de aprendizaje garantiza que la neurona con mayor actividad ajuste continuamente su vector de peso para alinearse más estrechamente con el vector de entrada. Este proceso iterativo de rotación y enfoque facilita que la red aprenda y se adapte a los datos de entrada. En cada iteración del proceso de aprendizaje, el vector de peso de la neurona ganadora sufre una rotación hacia el vector de entrada. Esta rotación va acompañada de un acercamiento hacia el vector de entrada, y el alcance de este acercamiento está determinado por la magnitud de la tasa de aprendizaje.

Inicialmente, durante las primeras etapas del entrenamiento, los vectores de peso de las tres neuronas (representados por vectores rojos) se distribuyen aleatoriamente alrededor de la circunferencia. Sin embargo, a medida que avanza el proceso de aprendizaje, estos vectores de peso se acercan gradualmente a las muestras del espacio de entrada. Con el tiempo, se estabilizan y sirven como centroides de los tres grupos. En general, la próxima ilustración proporciona una representación visual de cómo opera la regla de aprendizaje en el contexto de un espacio de entrada bidimensional, mostrando

la convergencia de los vectores de peso hacia los patrones en el espacio de entrada, lo que finalmente resulta en el establecimiento de centroides para los grupos (Palmer et al., 2002).



Cuando finaliza el proceso de aprendizaje, el vector de referencia de cada neurona de salida se alinear\u00e1 con el vector de entrada que activa con \u00e9xito esa neurona espec\u00edfica. En escenarios donde la cantidad de patrones de entrenamiento excede la cantidad de neuronas de salida, es necesario asignar m\u00faltiples patrones a la misma neurona, formando as\u00ed una clase. Para lograr esto, los pesos que componen el vector de referencia se derivan calculando el promedio (centroide) de estos patrones.

Adem\u00e1s del esquema de aprendizaje competitivo mencionado anteriormente, el modelo de mapa autoorganizado (SOM) introduce un concepto importante al incorporar las relaciones entre neuronas vecinas en el mapa. Esto se logra mediante la implementaci\u00f3n de una funci\u00f3n de zona de vecindad, que define un entorno que rodea a la neurona ganadora. Esta funci\u00f3n juega un papel crucial en el proceso de aprendizaje ya que permite la actualizaci\u00f3n simult\u00e1nea tanto de los pesos de la neurona ganadora como de las neuronas vecinas.

En general, el modelo SOM no s\u00f3lo incorpora el aprendizaje competitivo sino que tambi\u00e9n considera las relaciones entre neuronas vecinas. Esto le permite capturar la estructura topol\u00f3gica del espacio de entrada y proporcionar una poderosa herramienta

para la representación y el análisis de datos. Al actualizar los pesos de las neuronas cercanas, el modelo SOM garantiza que estas neuronas se adapten a patrones similares. Esto da como resultado un orden topológico que se refleja en el mapa, representando la estructura subyacente del espacio de entrada. Este mecanismo permite que el modelo SOM capture las relaciones espaciales entre diferentes puntos de datos, mejorando su capacidad para representar conjuntos de datos complejos.

Para comprender el proceso de aprendizaje del modelo SOM de una manera más matemática, es importante reconocer la identificación de la neurona de salida ganadora cuando se presenta un patrón de entrenamiento. Esto implica encontrar la neurona cuyo vector de peso sea más similar al patrón de entrada. Un criterio de similitud comúnmente utilizado es la distancia euclidiana, que se puede calcular mediante una expresión específica.

$$\min \|X_p - W_j\| = \min \sum_{i=1}^N (x_{pi} - w_{ji})^2$$

Según este criterio, la similitud entre dos vectores aumenta a medida que disminuye su distancia. Otra medida de similitud, más sencilla en comparación con el método euclidiano, es la correlación o producto escalar:

$$\min \|X_p - W_j\| = \max \sum_{i=1}^N x_{pi} \cdot w_{ji}$$

Cuanto más correlacionados estén dos vectores, más similares serán. Una vez que hemos determinado la neurona ganadora en función de su similitud, podemos ajustar su vector de peso y el de sus neuronas vecinas usando la regla de aprendizaje.

$$\Delta w_{ji}(n+1) = \alpha(n) (x_{pi} - w_{ji}(n)) \quad \text{para } j \in \text{Zona } j^*(n)$$

La variable

"n" representa el número de ciclos o iteraciones en el proceso de aprendizaje, indicando cuántas veces se ha presentado y procesado todo el conjunto de patrones de entrenamiento. La tasa de aprendizaje, denominada " $\alpha(n)$ ", comienza con un valor inicial

entre 0 y 1 y disminuye a medida que aumenta el número de iteraciones. La zona de vecindad, denominada $Zonaj^*(n)$, abarca el área circundante de la neurona ganadora j^* donde se encuentran las neuronas con pesos actualizados. De manera similar a la tasa de aprendizaje, el tamaño de esta zona disminuye gradualmente en cada iteración, lo que da como resultado un conjunto más pequeño de neuronas vecinas.

En el enfoque convencional, los pesos de una red neuronal se ajustan después de presentar cada patrón de entrenamiento. Esto está en línea con la regla de aprendizaje que se ha utilizado ampliamente. Sin embargo, algunos investigadores, como Masters (1993), sugieren un enfoque diferente donde se acumulan los incrementos calculados para cada patrón de entrenamiento. Una vez presentados todos los patrones, los pesos se actualizan en función del promedio de estos incrementos acumulados. Este método alternativo tiene como objetivo evitar que el vector de peso cambie entre diferentes patrones, acelerando así el proceso de convergencia de la red.

El proceso general de aprendizaje se puede dividir en dos fases. La primera fase se centra en la organización de los vectores de peso en el mapa. Esto se logra utilizando inicialmente una tasa de aprendizaje y un tamaño de vecindario altos, que luego se reducen gradualmente a medida que avanza el aprendizaje. En la segunda fase, el objetivo es ajustar el mapa alineando los vectores de peso con mayor precisión con los vectores de entrenamiento. Esta fase suele ser más larga e implica mantener la tasa de aprendizaje constante en un valor pequeño, como 0,01, y mantener un radio de vecindad fijo de 1.

No existe una regla establecida para determinar el número exacto de iteraciones necesarias para entrenar un modelo de forma eficaz. Sin embargo, el número de iteraciones debe verse influenciado por el número de neuronas en el mapa; más neuronas generalmente requieren más iteraciones. Por otro lado, el número de variables de entrada no tiene un impacto significativo en las iteraciones necesarias. Si bien se considera adecuado un número recomendado de 500 iteraciones por neurona, normalmente de 50 a 100 iteraciones son suficientes para la mayoría de los problemas, como sugirió Kohonen en 1990.

Las fases

La inicialización de los pesos

Al crear un mapa autoorganizado por primera vez, es necesario asignar valores a los pesos para comenzar el proceso de capacitación. Normalmente, hay poco debate sobre este asunto y los pesos se inicializan con pequeños valores aleatorios. Estos valores suelen elegirse dentro de un rango, como entre -1 y 1 o 0 y 1, como sugirió Kohonen en 1990. Sin embargo, también es posible inicializar los pesos con valores nulos, o incluso mediante una selección aleatoria de patrones de entrenamiento.

El entrenamiento de la red

Para modificar los vectores de peso de las neuronas en el conjunto de entrenamiento, proporcionaremos una serie de recomendaciones prácticas sobre tres parámetros de aprendizaje. Estos parámetros son cruciales para lograr resultados óptimos en el proceso de aprendizaje, pero sus valores ideales no pueden determinarse de antemano debido a la singularidad de cada problema.

La medida de similitud

Anteriormente se analizaron los dos métodos más comúnmente empleados para determinar la neurona ganadora durante la presentación de un patrón de entrada tanto en la etapa operativa como en la de aprendizaje de la red. Sin embargo, es importante considerar que la medida de similitud y la regla de aprendizaje utilizadas en el algoritmo deben ser compatibles en términos de métricas. Si no son compatibles, estaríamos empleando diferentes métricas para identificar la neurona ganadora y ajustar el vector de peso asociado, lo que podría generar problemas en el desarrollo del mapa.

La distancia euclidiana y la regla de aprendizaje mencionada en el pasaje son compatibles en términos de sus métricas, por lo que no hay ningún problema al respecto. Sin embargo, cuando se trata de la correlación o producto escalar y la regla de aprendizaje, no son compatibles porque la regla de aprendizaje se deriva de la métrica euclidiana. La correlación solo se alinea con esta métrica cuando se utilizan vectores normalizados, lo que da como resultado una coincidencia entre la distancia euclidiana y la correlación. En consecuencia, si optamos por emplear la correlación como criterio de similitud, es necesario utilizar vectores normalizados. En cambio, si optamos por la distancia euclidiana, no se requiere normalización. Para garantizar la coherencia, es aconsejable tener el mismo rango de valores potenciales para las variables de entrada, como -1 a 1 o 0 a 1.

La tasa del aprendizaje

Como se indicó anteriormente, la tasa de aprendizaje (α) juega un papel crucial a la hora de determinar cuánto se ajustan las ponderaciones cuando se presenta un patrón de entrada. La tasa de aprendizaje se establece inicialmente entre 0 y 1, como 0,6, y disminuye gradualmente a medida que aumenta el número de iteraciones (n). Esto significa que a medida que el algoritmo de aprendizaje presenta repetidamente todo el conjunto de patrones de aprendizaje, la tasa de aprendizaje disminuye a un valor prácticamente insignificante, lo que resulta en modificaciones insignificantes en los pesos. Normalmente, el ajuste de este parámetro se logra mediante la utilización de una de varias funciones.

$$\alpha(n) = \frac{1}{n} \quad \alpha(n) = \alpha_1 \left(1 - \frac{n}{\alpha_2} \right)$$

Cuando 1 es igual a un valor de 0,1 o 0,2, y 2 representa un valor cercano al número total de iteraciones de aprendizaje, que muchas veces se considera 10000. Cabe mencionar que la elección entre utilizar una función u otra no tendrá un impacto significativo en el resultado final.

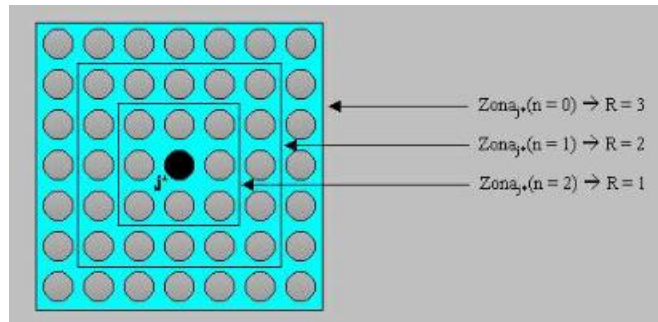
La zona de vecindad

También denominada $Zonaj^*(n)$, es una función que determina si una neurona de salida es parte o no de la vecindad que rodea a la neurona ganadora j^* durante cada iteración n . Esta vecindad es simétrica y su centro es j^* . Puede tener varias formas, como circular, cuadrada, hexagonal o cualquier otro polígono regular.

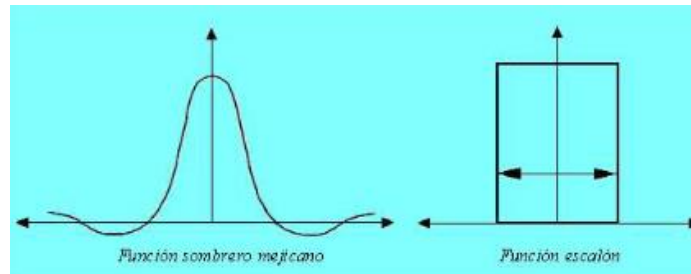
Generalmente, a medida que avanza el aprendizaje, el valor de $Zonaj^*(n)$ tiende a disminuir. Esta disminución está influenciada por un parámetro conocido como radio de vecindad $R(n)$, que indica la extensión o tamaño de la vecindad actual.

El tipo de paso es el tipo de función de vecindad más simple y más utilizado. En este escenario, una neurona j se considera parte de la vecindad de la neurona ganadora j^* sólo si la distancia entre ellas es menor o igual al valor de $R(n)$. Este tipo de función da forma a los vecindarios con aristas distintas, como cuadrados, círculos, hexágonos, etc., centrados alrededor del ganador, como se ilustra en la próxima figura (Palmer et al.,

2002). En consecuencia, en cada iteración, solo se actualizan las neuronas que están dentro del rango de $R(n)$ del ganador.



Ocasionalmente también se emplean funciones gaussianas o en forma de sombrero mexicano, como se muestra en la siguiente figura (Palmer et al., 2002). Estas funciones son continuas y diferenciables en cada punto, y en lugar de definir límites distintos, crean distintos niveles de membresía al definir vecindades decrecientes en el dominio espacial.



La función estilo sombrero mexicano se basa en la forma en que interactúan ciertas neuronas en la corteza, que se discutió anteriormente en el documento. Esta función implica que una neurona central envíe señales que excitan un área pequeña a su alrededor. A medida que aumenta la distancia desde la neurona central, el nivel de excitación disminuye hasta volverse realmente inhibitorio. Cuando la distancia es bastante grande, la neurona central emite una señal excitadora débil. Por otro lado, la función paso es una versión simplificada de la función en forma de sombrero mexicano y define discretamente el grupo de neuronas involucradas en el aprendizaje.

La zona vecinal tiene una forma específica, pero su tamaño cambia con el tiempo. Inicialmente, el radio se establece en un valor grande, por ejemplo, igual al diámetro total del mapa. Esto se hace para garantizar que el mapa esté ordenado globalmente. A medida que avanza el tiempo, el radio, denominado $R(n)$, disminuye continuamente hasta alcanzar un valor final de 1. En este punto, sólo se actualizan los pesos de la neurona ganadora y sus neuronas vecinas.

$$R(n) = R_0 + (R_f - R_0) \frac{n}{n_R}$$

Aquí n simboliza la iteración y n_R representa la cantidad de iteraciones para alcanzar R_f .

Evaluación del ajuste del mapa

En los mapas autoorganizados, los vectores de peso finales están influenciados por varios factores, incluidos los pesos aleatorios iniciales, la tasa de aprendizaje, el tipo de función de vecindad y la tasa de reducción de estos parámetros. Es importante encontrar un mapa óptimo que represente con precisión las relaciones entre los patrones de entrenamiento. El mapa ideal es aquel en el que los vectores de peso coinciden estrechamente con los vectores de entrenamiento. Esto se puede determinar calculando el error de cuantificación promedio, que mide la diferencia entre cada vector de entrenamiento y el vector de peso de su neurona ganadora. En nuestras simulaciones, utilizamos la expresión del error de cuantificación promedio como la media de la distancia euclidiana entre el vector de entrenamiento y su vector de peso asociado.

$$\text{Error}_{\text{medio}} = \frac{\sum_{p=1}^P \sum_{i=1}^M (x_{pi} - w_{ji})^2}{P}$$

La visualización y funcionamiento del mapa

Después de elegir el mapa más adecuado, podemos pasar a la etapa de visualización examinando las coordenadas en el mapa donde se encuentra cada neurona del patrón de entrenamiento. Esto nos permite transformar el espacio multidimensional de entrada en un mapa bidimensional y, debido a la similitud entre neuronas vecinas,

identificar grupos o categorías de datos ordenados por la red. Esto hace que el modelo de mapa autoorganizado sea particularmente valioso para descubrir conexiones previamente no identificadas entre conjuntos de datos.

Durante la fase operativa, la red posee la capacidad de funcionar como un clasificador de patrones. Esto es evidente ya que la neurona de salida que se activa por una entrada recién introducida simboliza la clase particular a la que pertenece la información de entrada. Además, ante otra entrada que se parece a una anterior, se activa la misma neurona de salida o una que se encuentra muy próxima a ella. Este hecho se puede atribuir a las similitudes entre las distintas clases, lo que garantiza que las neuronas topológicamente adyacentes respondan a entradas que comparten similitudes físicas.

El análisis de la sensibilidad

Una de las principales críticas dirigidas a la utilización de redes neuronales artificiales gira en torno a la dificultad inherente a la comprensión de las representaciones internas generadas por la red en respuesta a un patrón de entrada determinado. A diferencia de los modelos estadísticos tradicionales, no es inmediatamente evidente cómo cada variable de entrada contribuye a la salida del modelo dentro de una red. Sin embargo, es fundamental señalar que la percepción de las ANN como "cajas negras" inescrutables no es del todo exacta.

El análisis de sensibilidad implica evaluar el impacto de cambiar una variable de entrada sobre otra. En el caso de un modelo SOM, este análisis se realizó para determinar cómo los pequeños cambios en las variables de entrada afectan la salida del modelo. Para este análisis se utilizó el conjunto de datos Iris, un conjunto de datos ampliamente utilizado en el reconocimiento de patrones, junto con otros conjuntos de datos bien conocidos en el campo, como los utilizados para la discriminación del cáncer. Si bien el análisis de sensibilidad se ha aplicado a las redes de retropropagación en estudios anteriores, existe una investigación limitada sobre su aplicación a los modelos SOM.

De manera similar a la red de retropropagación, Hollmén y Simula (1996) realizaron un estudio en el que hicieron ajustes menores a una de las variables de entrada mientras mantenían las otras variables en un valor promedio. Luego observaron cómo, como resultado, la posición de la neurona ganadora en el mapa cambiaba. Este enfoque

les permitió determinar el nivel de correlación o significancia que cada variable de entrada tenía en la salida de la red.

Las redes neuronales dinámicas

Las redes neuronales se pueden clasificar en estáticas o dinámicas. Las redes estáticas calculan la salida directamente a partir de la entrada mediante conexiones feedforward, mientras que las redes dinámicas consideran no solo la entrada actual sino también las entradas y salidas o estados anteriores. Las redes dinámicas, como aquellas con filtros adaptativos o redes Hopfield, tienen un elemento de memoria ya que su salida está influenciada por entradas pasadas.

El entrenamiento de redes neuronales implica el uso de algoritmos que se basan en gradientes, como el algoritmo de gradiente conjugado y de descenso más pronunciado, o jacobianos, como los algoritmos de Gauss-Newton y Levenberg-Marquardt. El proceso de entrenamiento de redes estáticas y dinámicas difiere en cómo se calcula la matriz jacobiana o de gradiente. Las redes dinámicas incorporan bloques de retardo que procesan entradas secuenciales y el orden de las entradas es significativo. Pueden tener conexiones de retroalimentación como filtros adaptativos o incluir conexiones de retroalimentación recurrentes conocidas como redes neuronales recurrentes (RNN). Debido a sus capacidades de memoria, las redes dinámicas se pueden entrenar para aprender patrones secuenciales o variables en el tiempo, lo que las hace adecuadas para aplicaciones en diversos campos como sistemas de control, predicción de mercados financieros, ecualización de canales de comunicación, clasificación, detección de fallas y reconocimiento de voz.

Es posible entrenar redes dinámicas utilizando métodos estándar de optimización de redes estáticas, pero los gradientes y jacobianos necesarios para estos métodos no se pueden calcular utilizando el algoritmo de retropropagación estándar. En su lugar, se utilizan algoritmos dinámicos de retropropagación como la retropropagación en el tiempo (BPTT) y el aprendizaje recurrente en tiempo real (RTRL) para calcular los gradientes. BPTT calcula la respuesta de la red para todos los puntos de tiempo y luego calcula el gradiente desde el último punto de tiempo hacia atrás en el tiempo. Aunque es eficaz para los cálculos de gradiente, BPTT resulta complicado de implementar en línea, ya que funciona en orden cronológico inverso. Por otro lado, RTRL estima el gradiente

simultáneamente con la respuesta de la red comenzando en el primer punto de tiempo y avanzando en el tiempo. Si bien RTRL requiere más cálculos que BPTT para el cálculo del gradiente, ofrece un marco conveniente para la implementación en línea. Cuando se trata de cálculos jacobianos, el algoritmo RTRL es generalmente más eficiente que BPTT.

Su estructura

Las redes neuronales son capaces de tener dos tipos de conexiones: conexiones feedforward, que son unidireccionales y solo van en una dirección, y conexiones feedforward combinadas con retroalimentación o conexiones recurrentes, que permiten que la información regrese dentro de la red.

Las redes estáticas consisten exclusivamente en conexiones feedforward, mientras que las redes dinámicas abarcan dos tipos distintos.

- Sólo con conexiones prealimentadas
- Con uniones directas y uniones inversas o recurrentes (RNN).

La innovación financiera

La industria financiera está experimentando una transformación significativa debido a los avances de la tecnología digital. Estos avances han revolucionado los servicios de pago, ahorros, préstamos e inversiones, así como las entidades que ofrecen estos servicios. Las empresas de tecnología financiera y los gigantes tecnológicos se han convertido en competidores de los bancos tradicionales y otras instituciones establecidas en diversos mercados.

Asimismo, la introducción de monedas digitales promete transformar completamente el concepto fundamental de dinero dentro del sistema financiero. Si bien, es esencial examinar hasta qué punto la tecnología ha promovido realmente la inclusión financiera. Sin duda, las finanzas digitales han desempeñado un papel crucial para ayudar a los hogares y las empresas a afrontar los desafíos impuestos por la pandemia de COVID-19 el año pasado. Además, ha brindado a los gobiernos nuevas vías para brindar apoyo a quienes lo necesitan. En general, la tecnología digital ha marcado el comienzo de una nueva era en la industria financiera, revolucionando la forma en que se prestan y acceden a los servicios financieros. No solo ha creado oportunidades para que

nuevos actores ingresen al mercado, sino que también ha desempeñado un papel fundamental para garantizar la resiliencia financiera en tiempos de crisis.

Los avances logrados hasta ahora han sido notablemente notables. Aun cuando, para aprovechar eficazmente todo su potencial para mejorar la inclusión financiera, es crucial que la innovación del sector privado integre suficientes bienes públicos. Esta integración desempeña un papel vital al influir en todas las facetas de la actividad económica. Los bienes públicos sirven como pilares fundamentales sobre los que prospera el concepto de inclusión financiera.

Inclusión

La inclusión financiera se refiere a la disponibilidad generalizada de servicios financieros asequibles y ha logrado avances significativos durante la última década, a pesar de las fluctuaciones económicas y la pandemia de COVID-19. Los datos del Banco Mundial revelan que entre 2011 y 2017, 1.200 millones de adultos obtuvieron acceso a cuentas comerciales, un progreso atribuido en gran medida al surgimiento de nuevas tecnologías digitales. Un ejemplo destacado de esto es el dinero móvil, ejemplificado por M-Pesa en Kenia y aplicaciones similares.

Estas plataformas permiten a los usuarios enviar y recibir pagos a través de cualquier teléfono móvil. Con el tiempo, los proveedores de servicios han ampliado su oferta para incluir microcréditos, cuentas de ahorro y seguros contra diversos riesgos, como malas cosechas. De hecho, en 2019, el 79% de los adultos en Kenia tenían una cuenta móvil. Esta tendencia también está ganando terreno en África, Medio Oriente y América Latina. Mientras tanto, en la India, las iniciativas del gobierno para proporcionar infraestructura básica han tenido un profundo impacto.

El programa Aadhaar, que ofrece identidad digital (ID), ha otorgado a 1.300 millones de personas acceso a identificaciones confiables, facilitando la apertura de cuentas bancarias y el acceso a otros servicios. Además, un nuevo sistema introducido a través de esta iniciativa permite a los usuarios realizar pagos de bajo costo en tiempo real. Según estudios del Banco de Pagos Internacionales, el acceso a la banca en la India ha aumentado del 10% de la población en 2008 a más del 80% en la actualidad. Este rápido progreso, logrado a través de la tecnología, supera lo que los procesos de crecimiento tradicionales habrían logrado en medio siglo.

La pandemia de COVID-19 y las posteriores medidas de distanciamiento social han puesto de relieve la importancia de los pagos digitales. Para muchas personas, las plataformas digitales se volvieron esenciales para realizar pagos, comprar artículos necesarios como cilindros de oxígeno y apoyar a las pequeñas empresas. Además, la tecnología desempeñó un papel crucial a la hora de cerrar brechas y facilitar transferencias de dinero rápidas y asequibles a familiares. Solo en Filipinas, desde mediados de marzo hasta finales de abril de 2020, se abrieron cuatro millones de cuentas digitales.

Los gobiernos de todo el mundo han recurrido a nuevas infraestructuras digitales para llegar a los hogares y a los trabajadores informales durante la pandemia. Perú, por ejemplo, implementó el proyecto Mobile Wallet, integrando compañías de telefonía móvil y bancos para facilitar los pagos. Tailandia también adoptó el sistema de pago rápido PromptPay, lo que demuestra la eficacia de este tipo de iniciativas. Estos ejemplos contrastan marcadamente con las prácticas en economías avanzadas como Estados Unidos, donde los pagos tradicionales con cheques por correo todavía prevalecen.

La innovación digital en el sector económico

A pesar de que la pandemia dejó graves consecuencias para la economía, una mayor desigualdad, hay un lado positivo en forma de un impulso en la adopción de tecnologías digitales. Estas tecnologías desempeñan un papel crucial en la promoción de la inclusión financiera y la creación de oportunidades económicas. Si bien, es importante señalar que la tecnología por sí sola no puede garantizar el éxito. Para comprender plenamente el potencial de la tecnología para facilitar la inclusión digital y dar forma a las políticas, es necesario examinar las condiciones económicas fundamentales que subyacen a estos avances.

Otro factor crucial que contribuye al desarrollo de innovaciones digitales es la capacidad de almacenar y procesar grandes cantidades de datos digitales. Con el aumento exponencial de la generación de datos, se ha hecho necesario contar con capacidades eficientes de almacenamiento y procesamiento para gestionar y analizar esta información de forma eficaz. Esta capacidad permite a las empresas y organizaciones extraer información valiosa, tomar decisiones informadas y desarrollar soluciones innovadoras basadas en el análisis de estos datos.

Asimismo, los continuos avances en tecnologías como la computación en la nube, el aprendizaje automático, la tecnología de contabilidad, los sistemas distribuidos y las tecnologías biométricas han mejorado aún más el panorama de las innovaciones digitales. La computación en la nube permite la implementación flexible y escalable de software y servicios, mientras que los algoritmos de aprendizaje automático permiten la automatización de procesos y la toma de decisiones inteligentes basadas en patrones de datos. La tecnología de contabilidad, comúnmente conocida como blockchain, proporciona registros de transacciones seguros y transparentes, lo que garantiza la confianza y la responsabilidad en las interacciones digitales. Los sistemas distribuidos permiten la gestión descentralizada de recursos y redes, mientras que las tecnologías biométricas ofrecen métodos de autenticación seguros y personalizados.

Estos avances en tecnología aportan sus propias fortalezas y capacidades únicas al ámbito de las innovaciones digitales. En conjunto, mejoran la eficiencia, la seguridad, la accesibilidad y la escalabilidad, fomentando así el crecimiento y el éxito de las innovaciones digitales en diversas industrias y sectores. Es a través de la convergencia de estos factores que se puede aprovechar todo el potencial de las innovaciones digitales, revolucionando la forma en que vivimos, trabajamos e interactuamos en la era digital.

El éxito y el progreso de las innovaciones digitales dependen en gran medida de varios factores clave que desempeñan un papel vital a la hora de hacer posible esta tecnología. Uno de estos factores es el uso generalizado de teléfonos móviles e Internet, ya que sirven como medio principal para conectar a personas, empresas y organizaciones con proveedores de servicios financieros e información. Estos avances tecnológicos han facilitado significativamente la comunicación, el acceso a los recursos y las transacciones financieras, desempeñando así un papel crucial en el crecimiento de las innovaciones digitales.

Empero, el aspecto crucial detrás del éxito de estas innovaciones radica en la capacidad de la ciudad-capa para recopilar información y conectarse con los usuarios a un precio increíblemente asequible. Los expertos en el campo de la economía han examinado a fondo los numerosos gastos específicos que se han reducido significativamente debido a la llegada de las tecnologías digitales. Hay dos atributos económicos importantes asociados con la tecnología digital que demuestran vívidamente la inmensa influencia de estos factores, así como los riesgos potenciales que implican.

Para empezar, las plataformas digitales poseen la ventaja de ser altamente adaptables y servir como convenientes diarios "intermedios" que facilitan la interacción entre varios grupos de usuarios. Tomemos, por ejemplo, el caso de un proveedor de servicios de billetera digital como PayPal, que conecta de manera efectiva a los comerciantes con los clientes que buscan soluciones de pago seguras. A medida que aumenta el número de clientes que utilizan una opción de pago específica, resulta cada vez más beneficioso para los comerciantes ofrecerla y viceversa. Esto ejemplifica el concepto de economías de escala en el ámbito digital, lo que permite a los proveedores experimentar un rápido crecimiento.

Como resultado, potencias tecnológicas como Amazon y Alibaba, particularmente en China, pueden servir como intermediarios que conectan a compradores y vendedores en el mercado de bienes. Además, estas empresas también tienen la capacidad de establecer conexiones entre comerciantes y proveedores, facilitando no sólo las transacciones sino también brindando acceso al crédito y otros servicios varios. La amplia gama de servicios que ofrecen estos gigantes, que va más allá de los financieros, les permite aprovechar sus ofertas financieras de manera efectiva. Esto sirve como un excelente ejemplo de economías de alcance, en las que las empresas involucradas en diversos sectores se ven favorecidas y pueden cosechar los beneficios de sus variadas operaciones.

De igual forma, la utilización de tecnologías digitales puede mejorar en gran medida la evaluación del riesgo al aprovechar los datos secundarios obtenidos de las actividades en línea de las personas. Esto es especialmente ventajoso para diversos servicios como préstamos, inversiones y seguros. Las calificaciones crediticias que se generan utilizando big data y algoritmos de aprendizaje automático han demostrado ser más precisas que las evaluaciones tradicionales, especialmente para individuos o pequeñas empresas con un historial crediticio formal limitado. Una investigación realizada por el BIS revela que casi un tercio de los clientes de Mercado Libre, una destacada empresa de préstamos tecnológicos en Argentina, no habrían podido obtener crédito de un banco tradicional. Además, las empresas a las que Mercado Libre les concedió préstamos demostraron mejores ventas y oferta de productos un año después. Los datos del Ant Group indican además que las grandes corporaciones impulsadas por la tecnología, al utilizar big data, pueden requerir menos garantías colaterales. Esto tiene

el potencial de ampliar las oportunidades de crédito a los prestatarios que carecen de bienes raíces u otros activos que puedan usarse como garantía, al tiempo que reduce la vulnerabilidad de los préstamos a las fluctuaciones en los precios de los activos.

La combinación de economías de escala y alcance, junto con capacidades predictivas mejoradas, tiene el potencial de mejorar significativamente la inclusión financiera. En particular, el crédito de los gigantes tecnológicos ha experimentado un aumento sin precedentes a nivel mundial durante la última década, alcanzando aproximadamente 572 mil millones de dólares en 2019 (Frost, 2021). Estos préstamos desempeñan un papel crucial, especialmente en economías como China, Kenia e Indonesia, donde han superado la importancia de los mercados crediticios tradicionales. Además, su crecimiento se extiende más allá de estas regiones, con indicios de que pueden haber aumentado incluso ligeramente durante la pandemia, cuando los gigantes tecnológicos intervinieron para facilitar la distribución de préstamos a las empresas.

Aun cuando, como cualquier otro avance, los avances que permite el uso de big data también tienen sus inconvenientes, en particular la inclinación hacia los monopolios. En ciertas economías, los gigantes tecnológicos dominantes que brindan servicios de pagos y préstamos se han vuelto tan cruciales para el sistema que son esencialmente "demasiado grandes para quebrar". Esta tendencia a absorber competidores puede obstaculizar la innovación. Además, existe un riesgo significativo de que se produzca un uso indebido de los datos y violaciones importantes que infrinjan la privacidad de las personas. Es fundamental implementar políticas públicas inteligentes que puedan mitigar eficazmente estos riesgos sin limitar innecesariamente el potencial de las tecnologías digitales.

Conclusiones

Para navegar y prosperar eficazmente en este mundo en rápida evolución, las personas y los órganos de gobierno deben adoptar un enfoque proactivo. Es fundamental comprender cómo aprovechar el poder de la innovación digital para fomentar la inclusión y abordar simultáneamente los riesgos potenciales asociados con la estabilidad financiera y los derechos de los consumidores. Para lograrlo, es evidente que se deben implementar cinco categorías distintas de políticas.

- Crear infraestructuras digitales inclusivas es crucial para avanzar hacia cuentas y servicios más avanzados. Iniciativas como Aadhaar en India sirven como punto de partida para este desarrollo. Es esencial establecer sistemas rápidos de pagos minoristas que se basen en una infraestructura pública abierta para garantizar la igualdad de oportunidades para todos. Ejemplos de tales sistemas incluyen el pago rápido en Rusia, CoDi en México y PIX en Brasil, que permiten pagos digitales instantáneos con un costo mínimo o sin costo para individuos, empresas o gobiernos. Además, las monedas digitales de los bancos centrales, que actualmente se están probando en China y otros países y que ya están en circulación en las Bahamas, también pueden servir como plataforma compartida para diversos servicios ofrecidos por proveedores privados.
- Para fomentar la competencia en el mundo digital, es crucial establecer reglas comunes que incentiven el juego limpio. Numerosos países han reconocido la necesidad de desafiar los monopolios digitales y han implementado regulaciones que permiten a los usuarios transferir sus datos entre varias plataformas. Este concepto, conocido como interoperabilidad, permite que diferentes proveedores se conecten entre sí sin problemas, ampliando así las opciones de los consumidores y fomentando una competencia sana. Al igual que los protocolos fundamentales de Internet que facilitan una comunicación fluida, estos estándares comunes sirven como un bien público vital que es indispensable para el crecimiento y avance de los mercados privados.
- En la actual era digital, puede ser necesario revisar y actualizar las políticas de competencia. Es posible que los métodos tradicionales para promover la competencia y prevenir los monopolios ya no sean eficaces. Esto se debe a que las

prácticas monopolísticas ahora pueden adoptar la forma de control de datos, en lugar de únicamente precios elevados. Sin regulaciones adecuadas, podrían surgir nuevas barreras de entrada al mercado y prácticas anticompetitivas que obstaculicen la competencia leal. El creciente escrutinio de las fusiones y adquisiciones, así como el surgimiento de "guardianes" digitales, sugiere que se requieren medidas innovadoras y con visión de futuro para garantizar la competitividad y la disputabilidad de los mercados financieros digitales.

- Para mejorar la privacidad de los datos, es necesario abordar la falta de claridad en las leyes existentes sobre los datos generados por los servicios digitales. Actualmente, las empresas de tecnología tienen un control significativo sobre los datos confidenciales, lo que no es ideal. Para rectificar esto, los usuarios deberían tener más control y autonomía sobre sus datos. La Unión Europea ha implementado leyes de privacidad que pueden servir como modelo útil, así como el proyecto India Stack, que ha implementado prácticas efectivas para controlar los datos de los usuarios. Estudios recientes indican que existen diferencias en la disposición a compartir datos en función de factores como el género y la edad. Por ejemplo, los hombres suelen estar más dispuestos que las mujeres a compartir sus datos a cambio de mejores servicios financieros, y los jóvenes están más inclinados a compartir sus datos en comparación con los adultos. Dada la diversidad de preferencias y perspectivas en la sociedad, será un desafío establecer reglas universalmente aceptables para el uso de datos. Sin embargo, es probable que sean necesarias leyes para regular y proteger la privacidad de los datos.
- Coordinación del trabajo entre diversos órganos, así las tecnologías digitales en las finanzas conciernen no sólo bancos centrales y reguladores financieros, sino también a las autoridades que brindan competencia y seguridad de los datos; ambas las partes deben trabajar juntas. También es muy probable que la política del país influya en usuarios de otros países.

Con definiciones claras y precisas de bienes públicos y una cooperación efectiva de las autoridades, la utilización de la tecnología digital tiene el potencial de mejorar significativamente la inclusión dentro de nuestros sistemas, particularmente en el sector financiero, al llegar a quienes están económicamente desfavorecidos. La adopción generalizada de tecnología tiene la capacidad no solo de mejorar la eficiencia general

dentro de las sociedades, sino también de fomentar una mayor igualdad y mejorar la preparación para la próxima era digital. Es imperativo que la innovación esté diseñada para beneficiar a todos los individuos, asegurando una distribución equitativa de sus ventajas.

Bibliografía

- Aroca, P. R., García, C. L., & López, J. J. G. (2009). Estadística descriptiva e inferencial. *Revista el auge de la estadística en el siglo XX*, 22, 165-176.
- Banco de España. (2020). Plan Estratégico 2024. *Eurosistema*.
- Banco de Pagos Internacionales. (2015). Orientaciones sobre riesgo de crédito y contabilidad de pérdidas crediticias esperadas. *Comité de Supervisión Bancaria de Basilea*.
- Basogain Olabe, X. (s.f.). Redes neuronales artificiales y sus aplicaciones.
- Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3), 602-613.
- Camino, J. R., y de Garcillán López-Rua, M. (2014). *Marketing sectorial. Principios y aplicaciones*. Esic Editorial.
- Chui, M., Manyika, J., & Miremadi, M. (2015). Four fundamentals of workplace automation. *McKinsey Quarterly*, 29(3), 1-9
- Elvery, J. (2019). Changes in the Occupational Structure of the United States: 1860 to 2015. *Economic Commentary*, (2019-09).
- Francés Monedero, T. (2020). *Impacto del machine learning en el sistema financiero* [Trabajo final de grado]. Comillas, Universidad Pontificia.
- Fernández, A. (2019). Inteligencia artificial en los servicios financieros. *Boletín Económico*, (JUN).
- Frost, J., Gambacorta, L., y Song Shin, H. (2021). De la innovación financiera a la inclusión. *Finanzas y desarrollo*, 58(1), 14-18.
- Heros Cárdenas, L. F. (2022). *Aprendizaje automático para el desarrollo de procesos en las instituciones financieras* (Bachelor's thesis, Benemérita Universidad Autónoma de Puebla).

- Hollmén, J. y Simula, O. (1996). Prediction models and sensitivity analysis of industrial process parameters by using the self-organizing map. *Proceedings of IEEE Nordic Signal Processing Symposium (NORSIG'96)*, 79-82.
- IBM Documentation. (2021, diciembre 7). Ibm.com. <https://www.ibm.com/docs/es/spss-statistics/beta?topic=types-time-plots>
- Kohonen, T. (1982a). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 59-69.
- Kohonen, T. (1982b). Analysis of a simple self-organizing process. *Biological Cybernetics*, 44, 135-140.
- Kohonen, T. (1989). *Self-organization and associative* Springer-Verlag.
- Kohonen, T. (1990). *The self-organ* 1480.
- Masters, T. (1993). *Practical neural networks recipes in C++*. London: Academic Press.
- Marinai, S., & Fujisawa, H. (Eds.). (2007). *Machine learning in document analysis and recognition*. Springer, 90, 1-20.
- Martínez, F. (2010). Robots financieros, los nuevos señores del corto plazo. Recuperado el 22 de enero de 2020, de https://cincodias.elpais.com/cincodias/2010/11/20/mercados/1290218503_850215.html
- McCarthy, J. (1960). *Programs with common sense*. RLE and MIT computation center.
- Niederhoffer, V., & Osborne, M. F. M. (1966). Market making and reversal on the stock exchange. *Journal of the American Statistical Association*, 61(316), 897-916.
- Ortega, C. (2021, junio 29). *Gráfica de dispersión. Qué es y cuáles son sus características*. Tudashboard.com. <https://tudashboard.com/grafica-de-dispersion/>
- Palmer, A., Montaña, J.J. y Jiménez, R. (2002). Tutorial sobre Redes Neuronales Artificiales: Los Mapas Autoorganizados de Kohonen. *REVISTA ELECTRÓNICA DE PSICOLOGÍA*, 6(1).

- Ruiz, F. (2020, agosto 7). *5 usos del aprendizaje automático en los servicios financieros*. Finerio Connect. <https://blog.finerioconnect.com/usos-del-aprendizaje-automatico-en-servicios-financieros/>
- Sosa Sierra, M. D., (2007). Inteligencia artificial en la gestión financiera empresarial. *Pensamiento & Gestión*, (23), 153-186.
- United Consulting Group. (2018). Artificial Intelligence Effects on the Financial Services Sector.
- Vercellis, C. (2011). *Business intelligence: data mining and optimization for decision making*. John Wiley & Sons.
- Vorobioff, j., Cerrotta, S., Eneas Morel, N., y Amadio, A. (2022). *Inteligencia Artificial y Redes Neuronales fundamentos, ejercicios y aplicaciones con Python y Matlab*. edUTecNe – Editorial de la Universidad Tecnológica Nacional.
- Wang, J., Wang, J. (2017). Forecasting stochastic neural network based on financial empirical mode decomposition. *Neural Networks*, 90, 8-20

EST. 2021 | **EMC**
EDITORIAL MAR CARIBE

**MÉTODOS ESTADÍSTICOS DESCRIPTIVOS Y
DE APRENDIZAJE AUTOMÁTICO PARA LAS
FINANZAS**

EST. 2021 **EMC**
EDITORIAL MAR CARIBE

**ROLANDO WILBER ORDOÑEZ NEYRA
BETO PUMA HUAMÁN
EDGARDO MARTIN FIGUEROA DONAYRE
ROGGER HUMPIRI FLORES
HECTOR ITO MAMANI
CESAR EUSEBIO PACORI MAMANI**

ISBN: 978-9915-9682-2-3



ISBN: 978-9915-9682-2-3