Aprendizaje automático, *machine learning*

Big data e inteligencia artificial

CONTENIDO

Objetivo

Aprender que es el *machine learning* y sus aplicaciones.

El *machine learning*

El *machine learning* es una disciplina del campo de la Inteligencia Artificial que, a través de algoritmos, dota a los ordenadores de la capacidad de identificar patrones en datos masivos para hacer predicciones. Este aprendizaje permite a los computadores realizar tareas específicas de forma autónoma, es decir, sin necesidad de ser programados.

El término se utilizó por primera vez en 1959. Sin embargo, ha ganado relevancia en los últimos años debido al aumento de la capacidad de computación y al *boom* de los datos. Las técnicas de aprendizaje automático son, de hecho, una parte fundamental del *big data*.

Debido a nuevas tecnologías de cómputo, hoy día el *machine learning* no es como el del pasado. Nació del reconocimiento de patrones y de la teoría que dice que las computadoras pueden aprender sin ser programadas para realizar tareas específicas; investigadores interesados en la inteligencia artificial deseaban saber si las computadoras podían aprender de datos. El aspecto iterativo del *machine learning* es importante porque a medida que los modelos son expuestos a nuevos datos, estos pueden adaptarse de forma independiente. Aprenden de cálculos previos para producir decisiones y resultados confiables y repetibles. Es una ciencia que no es nueva, pero que ha cobrado un nuevo impulso.

Aunque muchos algoritmos de aprendizaje basado en máquina han existido por largo tiempo, la posibilidad de aplicar automáticamente cálculos matemáticos complejos al *big data* (una y otra vez, cada vez más rápido) es un logro reciente. Estos son algunos ejemplos ampliamente publicados de aplicaciones de *machine learning* con los que quizá muchas personas estén familiarizadas:

- **El automóvil de conducción autónoma de Google tan publicitado:** la esencia del *machine learning*.
- **Ofertas de recomendación en línea como las de Amazon y Netflix:** aplicaciones de *machine learning* para la vida diaria.
- **Saber lo que los clientes dicen acerca de usted en Twitter:** *machine learning* combinado con creación de reglas lingüísticas.
- **Detección de fraudes:** uno de los usos más obvios e importantes actualmente en el mundo.

¿Por qué es importante el *machine learning*?

El resurgimiento del interés en el aprendizaje basado en máquina se debe a los mismos factores que han hecho la minería de datos y el análisis bayesiano más populares que nunca. Cosas como los volúmenes y variedades crecientes de datos disponibles, procesamiento computacional más económico y poderoso y almacenaje de datos asequible.

Todas estas cosas significan que es posible producir modelos de manera rápida y automática que puedan analizar datos más grandes y complejos; producir resultados más rápidos y precisos, incluso, en una escala muy grande.

Con la construcción de modelos precisos, una organización tiene una mejor oportunidad de identificar oportunidades rentables o de evitar riesgos desconocidos.

¿Quién lo utiliza?

- **Servicios financieros**

  Los bancos y otras empresas de la industria financiera utilizan la tecnología del aprendizaje basado en máquina para dos fines principales: identificar *insights* importantes en los datos y prevenir el fraude. Los *insights* pueden identificar oportunidades de inversión o bien ayudar a los inversionistas a saber cuándo vender o comprar. La minería de datos también puede identificar clientes con perfiles de alto riesgo o bien utilizar la cíber vigilancia para detectar signos de advertencia de fraude.

- **Gobierno**

  Dependencias de gobierno como seguridad pública y los servicios públicos tienen una necesidad particular del *machine learning* porque tienen múltiples fuentes de datos de las que se pueden extraer *insights*. Por ejemplo, el análisis de datos de sensores identifica formas de incrementar la eficiencia y ahorrar dinero. Asimismo, el aprendizaje basado en máquina puede ayudar a detectar fraude y minimizar el robo de identidad.

- **Atención a la saludo**

  El *machine learning* es una tendencia en rápido crecimiento en la industria de atención a la salud, gracias a la aparición de dispositivos y sensores de vestir que pueden usar datos para evaluar la salud de un paciente en tiempo real. Asimismo, la tecnología puede ayudar a expertos médicos a analizar datos para identificar tendencias o banderas rojas que puedan llevar a diagnósticos y tratamientos mejorado.

- ***Marketing* y ventas**

  Los sitios web que le recomiendan artículos que podrían gustarle con base en compras anteriores, utilizan el *machine learning* para analizar su historial de compras y promocionar otros artículos que podrían interesarle. Esta capacidad de capturar datos, analizarlos y usarlos para personalizar una experiencia de compra (o implementar una campaña de *marketing*) es el futuro del comercio detallista.

APLICACIONES PRÁCTICAS DEL *MACHINE LEARNING*

El *machine learning* es uno de los pilares sobre los que descansa la transformación digital. En la actualidad, ya se está utilizando para encontrar nuevas soluciones en diferentes campos, entre los que cabe destacar:

- **Recomendaciones:** permite hacer sugerencias personalizadas de compra en plataformas online o recomendar canciones. En su forma más básica analiza el historial de compras y reproducciones del usuario y lo compara con lo que han hecho otros usuarios con tendencias o gastos parecidos.
- **Vehículos inteligentes:** según el informe *Automotive 2025: industry without borders* de IBM, en 2025 se verán coches inteligentes en las carreteras. Gracias al aprendizaje automático, estos vehículos podrán ajustar la configuración interna (temperatura, música, inclinación del respaldo, etc.) de acuerdo con las preferencias del conductor e, incluso, mover el volante solos para reaccionar al entorno.
- **Redes sociales:** Twitter, por ejemplo, se sirve de algoritmos de *machine learning* para reducir en gran medida el spam publicado en esta red social mientras que Facebook, a su vez, lo utiliza para detectar tanto noticias falsas como contenidos no permitidos en retransmisiones en directo que bloquea automáticamente.

- **Procesamiento de lenguaje natural (PLN):** a través de la comprensión del lenguaje humano, asistentes virtuales como Alexa o Siri pueden traducir instantáneamente de un idioma a otro, reconocer la voz del usuario e incluso analizar sus sentimientos. Por otro lado, el PLN también se utiliza para otras tareas complejas como traducir la jerga legal de los contratos a un lenguaje sencillo o ayudar a los abogados a ordenar grandes volúmenes de información relativos a un caso.
- **Búsquedas:** los motores de búsqueda se sirven del aprendizaje automático para optimizar sus resultados en función de su eficacia, midiendo la misma a través de los clics del usuario.
- **Medicina:** investigadores del Instituto de Tecnología de Massachusetts (MIT) ya utilizan el *machine learning* para detectar con mayor antelación el cáncer de mama, algo de vital importancia ya que su detección temprana aumenta las probabilidades de curación. Asimismo, también se utiliza con una alta eficacia para detectar neumonía y enfermedades de la retina que pueden provocar ceguera.
- **Ciberseguridad:** los nuevos antivirus y motores de detección de malware ya se sirven del aprendizaje automático para potenciar el escaneado, acelerar la detección y mejorar la habilidad de reconocer anomalías.

Aprendizaje supervisado y no supervisado

Los algoritmos de *machine learning* se dividen en tres categorías, siendo las dos primeras las más comunes:

- **Aprendizaje supervisado:** estos algoritmos cuentan con un aprendizaje previo basado en un sistema de etiquetas asociadas a unos datos que les permiten tomar decisiones o hacer predicciones. Un ejemplo es un detector de spam que etiqueta un e-mail como spam o no dependiendo de los patrones que ha aprendido del histórico de correos (remitente, relación texto/imágenes, palabras clave en el asunto, etc.).
- **Aprendizaje no supervisado**: estos algoritmos no cuentan con un conocimiento previo. Se enfrentan al caos de datos con el objetivo de encontrar patrones que permitan organizarlos de alguna manera. Por ejemplo, en el campo del *marketing* se utilizan para extraer patrones de datos masivos provenientes de las redes sociales y crear campañas de publicidad altamente segmentadas.
- **Aprendizaje por refuerzo:** su objetivo es que un algoritmo aprenda a partir de la propia experiencia. Esto es, que sea capaz de tomar la mejor decisión ante diferentes situaciones de acuerdo con un proceso de prueba y error en el que se recompensan las decisiones correctas. En la actualidad se está utilizando para posibilitar el reconocimiento facial, hacer diagnósticos médicos o clasificar secuencias de ADN.

Otros ejemplos, más específicos, se mencionan a continuación:

- **Supervisados de predicción de variable continua**
  - Predicción del precio de las casas.
  - Ventas por una empresa para comprobar nivel de stocks
  - Entender mejor a los usuarios a través de los datos recolectados de tráfico
  - Entender mejor las enfermedades a través de los expedientes médicos.
- **Supervisados de predicción de variable discreta**
  - Por cada cliente de un banco y se trata de saber si tiene riesgo de ser moroso o no.
  - Riesgo de fuga o no.
  - Tumor benigno o no, en función de la edad del paciente y el tamaño del tumor.
- **No supervisados - *clustering***

  Un analista de negocios desea clasificar 22 empresas de manufactura pequeñas y medianas en grupos significativos para futuros análisis. El analista recoge datos sobre el número de clientes, la tasa de retorno, las ventas y los años que las empresas han estado en el negocio. Para iniciar el proceso de clasificación, el analista divide a las empresas en tres grupos iniciales: establecidas, crecimiento medio y recientes.

  Segmentación de clientes en *marketing*, de manera que se pueda vender más eficientemente.

- **Diferencia entre un problema de clasificación y un problema de *clustering***

  Un problema de clasificación se asemeja a un problema de análisis de clúster en el hecho de que existen grupos, pero se diferencia en que en este caso concreto se sabe cuántos grupos hay y se conoce a qué grupo pertenece cada dato para un conjunto de datos etiquetado.

Algunos tipos de algoritmos

- **Aprendizaje supervisado**
  - **Árboles de decisión:** un árbol de decisiones es una herramienta de apoyo a la decisión que utiliza un gráfico o un modelo similar a un árbol de decisiones y sus posibles consecuencias, incluidos los resultados de eventos fortuitos, los costos de recursos y la utilidad.

    Desde el punto de vista de la toma de decisiones empresariales, un árbol de decisiones es el número mínimo de preguntas sí / no que uno tiene que hacer, para evaluar la probabilidad de tomar una decisión correcta, la mayoría del tiempo. Este método le permite abordar el problema de una manera estructurada y sistemática para llegar a una conclusión lógica.

  - ***Naïve bayes clasification:*** los clasificadores *Naïve Bayes* son una familia de simples clasificadores probabilísticos basado en la aplicación de bayes (teorema con fuertes (*Naïve*) supuestos de independencia entre las características). La imagen destacada es la ecuación con P (A | B) es probabilidad posterior, P (B | A) es probabilidad, P (A) es probabilidad previa de clase y P (B) predictor probabilidad previa.
  - ***Ordinary least squares regression:*** si se ha estado en contacto con la estadística, probablemente se haya oído hablar de regresión lineal antes. *Ordinary least squares regression* es un método para realizar la regresión lineal. Se puede pensar en la regresión lineal como la tarea de ajustar una línea recta a través de un conjunto de puntos. Hay varias estrategias posibles para hacer esto y la estrategia de "mínimos cuadrados ordinarios" va así: se puede dibujar una línea y, luego, para cada uno de los puntos de datos, medir la distancia vertical entre el punto y la línea y sumarlos; la línea ajustada sería aquella en la que esta suma de distancias sea lo más pequeña posible.

    Linear se refiere al tipo de modelo que está utilizando para ajustar los datos, mientras que los mínimos cuadrados se refieren al tipo de métrica de error que está minimizando.
  - ***Logistic regression:*** la regresión logística es una poderosa manera estadística de modelar un resultado binomial con una o más variables explicativas. Mide la relación entre la variable dependiente categórica y una o más variables independientes estimando las probabilidades utilizando una función logística, que es la distribución logística acumulativa.
  - ***Support vector machines:*** SVM es un algoritmo de clasificación binario. Dado un conjunto de puntos de 2 tipos en el lugar N dimensional, SVM genera un hiperlano (N-1) dimensional para separar esos puntos en dos grupos. Suponga que tiene algunos puntos de dos tipos en un papel que son linealmente separables. SVM encontrará una línea recta que separa esos puntos en dos tipos y situados lo más lejos posible de todos esos puntos.

    En términos de escala, algunos de los mayores problemas que se han resuelto utilizando SVMs (con implementaciones adecuadamente modificadas) son publicidad en pantalla, reconocimiento de sitios de empalme humanos, detección de género basada en imágenes, clasificación de imágenes a gran escala.
  - **Métodos *ensemble*:** los métodos *ensemble* son algoritmos de aprendizaje que construyen un conjunto de clasificadores y, luego, clasifican nuevos puntos de datos tomando un voto ponderado de sus predicciones. El método de conjunto original es bayesiano promediando, pero los algoritmos más recientes incluyen error de corrección de salida de codificación.
- **Aprendizaje no supervisado**
  - **Algoritmos *clustering*:** *clustering* es la tarea de agrupar un conjunto de objetos tales que los objetos en el mismo grupo (clúster) son más similares entre sí que a los de otros grupos.
  - **Análisis de componentes principales:** PCA es un procedimiento estadístico que usa una transformación ortogonal para convertir un conjunto de observaciones de variables posiblemente correlacionadas en un conjunto de valores de variables linealmente no correlacionadas llamadas componentes principales.

Algunas de las aplicaciones de PCA incluyen compresión, simplificación de datos para un aprendizaje más fácil, visualización. Tenga en cuenta que el conocimiento del dominio es muy importante al elegir si seguir adelante con PCA o no. No es adecuado en los casos en que los datos son ruidosos (todos los componentes de PCA tienen una variación bastante alta).

- ***Singular value decomposition:*** en el álgebra lineal, SVD es una factorización de una matriz compleja real. Para una matriz M * n dada, existe una descomposición tal que M = UΣV, donde U y V son matrices unitarias y Σ es una matriz diagonal.
- **Análisis de componentes independientes:** ICA es una técnica estadística para revelar los factores ocultos que subyacen a conjuntos de variables, mediciones o señales aleatorias. ICA define un modelo generativo para los datos multivariados observados, que se suele dar como una gran base de datos de muestras. En el modelo, se supone que las variables de datos son mezclas lineales de algunas variables latentes desconocidas y el sistema de mezcla también es desconocido. Las variables latentes se asumen no gaussianas y mutuamente independientes y se les llama componentes independientes de los datos observados.

ICA está relacionado con PCA, pero es una técnica mucho más poderosa que es capaz de encontrar los factores subyacentes de fuentes cuando estos métodos clásicos fallan por completo. Sus aplicaciones incluyen imágenes digitales, bases de datos de documentos, indicadores económicos y mediciones psicométricas.

Bibliografía

[1] A. Nag, *Pragmatic machine learning with python: learn how to deploy machine learning models in production.*

[2] A. Gron, *Hands-on machine learning with scikit-learn, keras, and tensorflow: concepts, tools, and techniques to build intelligent systems.*