Conceptos Básicos de Machine Learning

Felipe Sánchez Soberanis

27 de octubre de 2022

Índice

Machine learning

Machine Learning (aprendizaje automático) es una técnica de análisis de datos que enseña a los ordenadores a hacer lo que resulta natural para las personas y los animales: aprender de la experiencia. Los algoritmos de aprendizaje automático emplean métodos de cálculo para "aprender" información directamente de los datos sin depender de una ecuación predeterminada como modelo. Los algoritmos mejoran su rendimiento de forma adaptativa a medida que aumenta el número de muestras disponibles para el aprendizaje.

Por qué es importante el machine learning

Con el aumento de la cantidad de big data, el Machine Learning se ha convertido en una técnica clave para resolver problemas en áreas tales como:

- Finanzas computacionales: para la calificación crediticia y el trading algorítmico
- Procesamiento de imágenes y visión artificial: para el reconocimiento facial, la detección de movimiento y la detección de objetos
- Biología computacional: para la detección de tumores, el descubrimiento de fármacos y la secuenciación del ADN
- Producción de energía: para la previsión de la carga y el precio
- Automoción, sector aeroespacial y fabricación: para el mantenimiento predictivo
- Procesamiento del lenguaje natural: para aplicaciones de reconocimiento de voz

Los algoritmos de Machine Learning encuentran patrones naturales en los datos que generan conocimiento y contribuyen a tomar mejores decisiones y a realizar mejores predicciones. Se utilizan a diario para tomar decisiones cruciales en diagnósticos médicos, trading de acciones, previsión de la carga energética, etc. Por ejemplo, los sitios multimedia confían en el aprendizaje automático para cribar millones de opciones con objeto de ofrecerle recomendaciones sobre canciones o películas. Los minoristas lo utilizan para obtener información sobre el comportamiento de compra de sus clientes.

Considere el uso de Machine Learning cuando tenga una tarea o un problema complejos que impliquen una gran cantidad de datos y muchas variables, pero no disponga de ninguna fórmula o ecuación.

Cómo funciona el machine learning

El aprendizaje automático emplea dos tipos de técnicas: el aprendizaje supervisado, que entrena un modelo con datos de entrada y salida conocidos para que pueda predecir salidas futuras, y el aprendizaje no supervisado, que encuentra patrones ocultos o estructuras intrínsecas en los datos de entrada.

Aprendizaje supervisado

El aprendizaje automático supervisado crea un modelo que realiza predicciones en función de las pruebas en presencia de una incertidumbre. Un algoritmo de aprendizaje supervisado toma un conjunto conocido de datos de entrada y respuestas conocidas para estos datos (salidas) y entrena un modelo con objeto de generar predicciones razonables como respuesta a datos nuevos.

El aprendizaje supervisado emplea técnicas de clasificación y regresión para desarrollar modelos predictivos.

Las técnicas de clasificación predicen respuestas discretas; por ejemplo, si un correo electrónico es legítimo o es spam, o bien si un tumor es cancerígeno o benigno. Los modelos de clasificación organizan los datos de entrada en categorías. Las aplicaciones más habituales son las imágenes médicas, el reconocimiento de voz y la calificación crediticia.

La clasificación se debe utilizar si los datos se pueden etiquetar, categorizar o dividir en grupos o clases concretos. Por ejemplo, las aplicaciones para el reconocimiento de la escritura emplean la clasificación para reconocer letras y números. En el procesamiento de imágenes y la visión artificial, se emplean técnicas de reconocimiento de patrones sin supervisión para la detección de objetos y la segmentación de imágenes.

Algunos algoritmos habituales para realizar la clasificación son: máquina de vectores de soporte (SVM), árboles de decisión boosted y bagged, k-vecino más cercano, clasificadores bayesianos (Naïve Bayes), análisis discriminante, regresión logística y redes neuronales.

Las técnicas de regresión predicen respuestas continuas; por ejemplo, cambios de temperatura o fluctuaciones en la demanda energética. Las aplicaciones más habituales son la predicción de la carga eléctrica y el trading algorítmico.

Se deben utilzar técnicas de regresión si se trabaja con un intervalo de datos o si la naturaleza de la respuesta es un número real, como la temperatura o el tiempo que tardará una pieza de equipamiento en fallar.

Algunos algoritmos habituales de regresión son: modelo lineal, modelo no lineal, regularización, regresión por pasos, árboles de decisión boosted y bagged, redes neuronales y aprendizaje neurodifuso adaptativo.

Aprendizaje no supervisado

El aprendizaje no supervisado halla patrones ocultos o estructuras intrínsecas en los datos. Se emplea para inferir información a partir de conjuntos de datos que constan de datos de entrada sin respuestas etiquetadas.

El clustering es la técnica de aprendizaje no supervisado más común. Se emplea para el análisis de datos exploratorio, con objeto de encontrar patrones o agrupaciones ocultos en los datos. Entre las aplicaciones del análisis de clusters están el análisis de secuencias genéticas, la investigación de mercados y el reconocimiento de objetos.

Por ejemplo, si una empresa de telefonía móvil quiere optimizar las ubicaciones donde construir antenas, puede recurrir al aprendizaje automático para calcular el número de clusters de personas que utilizan sus antenas. Un teléfono solo puede comunicarse con una antena en cada ocasión, de modo que el equipo emplea algoritmos de clustering para diseñar la mejor ubicación de antenas a fin de optimizar la recepción de la señal para grupos (o clusters) de clientes.

Algunos algoritmos habituales para realizar el clustering son: k-means y k-medoids, clustering jerárquico, modelos de mezclas gaussianas, modelos de Markov ocultos, mapas autoorganizados, clustering difuso de c-means y clustering sustractivo.