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.

Validación cruzada

La validación cruzada (CV) es una técnica utilizada para evaluar un modelo de aprendizaje automático y comprobar su rendimiento (o precisión). Consiste en reservar una muestra específica de un conjunto de datos en el que el modelo no está entrenado. Posteriormente, el modelo se pone a prueba en esta muestra para evaluarlo.

La validación cruzada se utiliza para proteger un modelo de un sobreajuste, especialmente si la cantidad de datos disponibles es limitada. También se conoce como estimación por rotación o prueba fuera de la muestra y se utiliza principalmente en entornos en los que el objetivo del modelo es la predicción.

Este procedimiento de remuestreo también se utiliza para comparar diferentes modelos de aprendizaje automático y determinar su eficacia para resolver un problema concreto. En otras palabras, la validación cruzada es un método utilizado para evaluar la habilidad de los modelos de aprendizaje automático.

En pocas palabras, en el proceso de validación cruzada, la muestra de datos original se divide aleatoriamente en varios subconjuntos. El modelo de aprendizaje automático se entrena en todos los subconjuntos, excepto en uno. Tras el entrenamiento, el modelo se pone a prueba haciendo predicciones en el subconjunto restante.

En muchos casos, se realizan múltiples rondas de validación cruzada utilizando diferentes subconjuntos, y sus resultados se promedian para determinar qué modelo es un buen predictor.

La validación cruzada es crucial cuando la cantidad de datos disponibles es limitada.

Supongamos que tiene que predecir la probabilidad de que una rueda de bicicleta se pinche. Para ello, ha recopilado datos sobre los neumáticos existentes: la edad del neumático, el número de kilómetros recorridos, el peso del ciclista y si se ha pinchado antes.

Para crear un modelo predictivo, utilizarás estos datos (históricos). Hay dos cosas que tienes que hacer con estos datos: entrenar el algoritmo y probar el modelo.

Como sólo tienes una cantidad limitada de datos disponibles, sería ingenuo utilizar todos los datos en el entrenamiento del algoritmo. Si lo haces, no te quedarían datos para probar o evaluar el modelo.

Reutilizar el conjunto de entrenamiento como conjunto de prueba no es una buena idea, ya que necesitamos evaluar la precisión del modelo en datos en los que no se ha entrenado. Esto se debe a que el objetivo principal del entrenamiento es preparar el modelo para que funcione con datos del mundo real. Y es improbable que tu conjunto de datos de entrenamiento contenga todos los posibles puntos de datos que el modelo pueda encontrar.

Una mejor idea sería utilizar el primer 75 % (tres bloques) de los datos como conjunto de datos de entrenamiento y el último 25 % (un bloque) como conjunto de datos de prueba. Esto le permitirá comparar lo bien que los diferentes algoritmos clasificaron los datos de prueba.

Pero, por supuesto, ¿cómo puede saber que utilizar el primer 75 por ciento de los datos como conjunto de entrenamiento y el 25 por ciento restante como conjunto de prueba es la mejor manera?

En su lugar, puede utilizar el primer 25 por ciento de los datos para la prueba; o bien, puede utilizar el tercer bloque de los datos como conjunto de datos de prueba y el restante como conjunto de datos de entrenamiento.

Como resultado, un tipo de validación cruzada llamado k-fold cross-validation utiliza todas las (cuatro) partes del conjunto de datos como datos de prueba, una por una, y luego resume los resultados.

Traducción realizada con la versión gratuita del traductor www.DeepL.com/Translator

La matriz de confusión

Una matriz de confusión es una técnica de medición del rendimiento de un algoritmo de clasificación de aprendizaje automático. Los científicos de datos la utilizan para evaluar el rendimiento de un modelo de clasificación en un conjunto de datos de prueba cuando se conocen los valores reales. Por ejemplo, la precisión de la clasificación puede ser engañosa, especialmente cuando hay dos o más clases en el conjunto de datos.

En consecuencia, el cálculo de la matriz de confusión ayuda a los científicos de datos a comprender la eficacia del modelo de clasificación.

La matriz de confusión visualiza la precisión de un clasificador comparando los valores reales y los valores predichos. Además, presenta una tabla con los diferentes resultados de la predicción.

Para una tabla de 2x2, ofrece los siguientes resultados:

Verdadero positivo

- El valor predicho por el modelo coincide con el valor real
- El valor real era positivo y el modelo de aprendizaje automático predijo un valor positivo

Verdadero negativo

- El valor predicho por el modelo coincide con el valor real
- El valor real era negativo y el modelo de aprendizaje automático predijo un valor negativo

Falso positivo

- El modelo de aprendizaje automático hizo una predicción falsa
- El valor real era negativo, pero el modelo de aprendizaje automático predijo un valor positivo
- Un falso positivo también se conoce como error de tipo 1

Falso negativo

- El modelo de aprendizaje automático hizo una predicción falsa
- El valor real era positivo, pero el modelo de aprendizaje automático predijo un valor negativo
- Un falso negativo también se conoce como error de tipo 2

Sesgo y varianza

El error puede describirse como una acción inexacta o equivocada. En el aprendizaje automático, el error se utiliza para ver la precisión con la que el modelo puede predecir en los datos que utiliza para aprender, así como en los nuevos datos no vistos. En función del error, se puede seleccionar el modelo de aprendizaje automático que mejor funcione para un conjunto de datos concreto.

Hay dos tipos principales de errores presentes en cualquier modelo de aprendizaje automático. Son los errores reducibles y los errores irreductibles.

Los errores irreductibles son errores que siempre estarán presentes en un modelo de aprendizaje automático, debido a variables desconocidas, y cuyos valores no pueden reducirse.

Los errores reducibles son aquellos cuyos valores pueden reducirse para mejorar un modelo. Se producen porque la función de salida de nuestro modelo no coincide con la función de salida deseada y puede optimizarse.

Podemos dividir a su vez los errores reducibles en dos: sesgo y varianza.

Para hacer predicciones, nuestro modelo analizará nuestros datos y encontrará patrones en ellos. Usando estos patrones, podemos hacer generalizaciones sobre ciertos casos en nuestros datos. Nuestro modelo, tras el entrenamiento, aprende estos patrones y los aplica al conjunto de pruebas para predecirlas.

El sesgo es la diferencia entre nuestros valores reales y los predichos. El sesgo son las simples suposiciones que nuestro modelo hace sobre nuestros datos para poder predecir nuevos datos.

Cuando el Sesgo es alto, las suposiciones que hace nuestro modelo son demasiado básicas, el modelo no puede capturar las características importantes de nuestros datos. Esto significa que nuestro modelo no ha captado los patrones en los datos de entrenamiento y, por tanto, no puede funcionar bien también en los datos de prueba. Si este es el caso, nuestro modelo no puede rendir en los nuevos datos y no puede ser enviado a producción.

Este caso, en el que el modelo no puede encontrar patrones en nuestro conjunto de entrenamiento y, por tanto, falla tanto en los datos vistos como en los no vistos, se denomina infraajuste.

La siguiente figura muestra un ejemplo de infraajuste. Como podemos ver, el modelo no ha encontrado patrones en nuestros datos y la línea de mejor ajuste es una línea recta que no pasa por ninguno de los puntos de datos. El modelo no se ha entrenado correctamente con los datos dados y tampoco puede predecir nuevos datos.

La varianza es lo contrario del sesgo. Durante el entrenamiento, permite a nuestro modelo "ver" los datos un cierto número de veces para encontrar patrones en ellos. Si no trabaja con los datos durante el tiempo suficiente, no encontrará patrones y se producirá el sesgo. Por otro lado, si se permite a nuestro modelo ver los datos demasiadas veces, aprenderá muy bien sólo para esos datos. Capturará la mayoría de los patrones de los datos, pero también aprenderá de los datos innecesarios presentes, o del ruido.

Podemos definir la varianza como la sensibilidad del modelo a las fluctuaciones de los datos. Nuestro modelo puede aprender del ruido. Esto hará que nuestro modelo considere importantes las características triviales.

Para cualquier modelo, tenemos que encontrar el equilibrio perfecto entre sesgo y varianza. Esto sólo asegura que capturemos los patrones esenciales en nuestro modelo mientras ignoramos el ruido presente en él. Esto se denomina equilibrio entre sesgo y varianza. Ayuda a optimizar el error de nuestro modelo y a mantenerlo lo más bajo posible.

Un modelo optimizado será sensible a los patrones de nuestros datos, pero al mismo tiempo será capaz de generalizar a nuevos datos. En este sentido, tanto el sesgo como la varianza deben ser bajos para evitar el sobreajuste y el infraajuste.

Sensibilidad y especificidad

La especificidad propiamente dicha puede describirse como la capacidad del algoritmo/modelo para predecir un verdadero negativo de cada categoría disponible. En la literatura, también se conoce simplemente como la tasa de verdaderos negativos. Formalmente puede calcularse mediante la siguiente ecuación

Especificidad = TN / TN + FP (Verdadero Negativo / Verdadero Negativo + Falso Positivo)

La sensibilidad en el aprendizaje automático puede describirse como la métrica utilizada para evaluar la capacidad de un modelo para predecir los verdaderos positivos de cada categoría disponible. En la literatura, este término también se puede reconocer como tasa de verdaderos positivos y se puede calcular con la siguiente ecuación

Sensibilidad = TP / TP+FN (Verdaderos positivos/verdaderos positivos + falsos negativos)

La especificidad y la sensibilidad son métricas importantes, pero un término no menos importante es la Exactitud del Aprendizaje Automático, que es en realidad la proporción de resultados verdaderos (verdaderos positivos o verdaderos negativos) y que se utiliza comúnmente con los términos de especificidad y sensibilidad del modelo en el área del aprendizaje automático. La precisión del aprendizaje automático se calcula formalmente mediante la siguiente ecuación

Exactitud = TP + TN/TP+FP+TN+FN (Verdadero Positivo + Verdadero Negativo/Verdadero Positivo + Falso Positivo + Verdadero Negativo)

La exactitud, la sensibilidad y la especificidad del aprendizaje automático son partes constitutivas del entrenamiento de la predicción. El modelado predictivo en el aprendizaje automático depende del tamaño del conjunto de datos especificado y de las prestaciones del modelo incluidas en el entrenamiento. El problema potencial puede producirse cuando la relación es desconocida o tal vez ni siquiera existe para los conjuntos de datos elegidos y los rendimientos del modelo. La resolución del problema puede realizarse a través de la prueba de sensibilidad del modelo, que a menudo da como resultado los límites aproximados de los conjuntos de datos incluidos y necesarios para obtener rendimientos efectivos en conjuntos de datos más grandes. Para simplificar, el análisis de sensibilidad del modelo es un procedimiento que puede ayudar a descubrir cómo varía la eficacia de sus conjuntos de datos con menos o más datos incluidos. Es útil para la optimización del modelo y la rentabilidad y el tiempo.

Para concluir, todas las medidas enumeradas anteriormente (sensibilidad, precisión y especificidad) nos proporcionan información igualmente importante sobre el valor real de nuestro modelo de clasificación. Son indicadores importantes que hay que mirar con igual cuidado y dedicación porque, si, por ejemplo, se omite la especificidad pero se incluyen todas las demás medidas, se podría crear un modelo con alta precisión y recuerdo que simplemente reconozca todo como verdadero, y no podría aceptarse como digno de confianza. La sensibilidad métrica es uno de los principales pilares del procesamiento del aprendizaje automático. Ahorre su tiempo y recursos siguiendo las ecuaciones exactas para cada métrica de evaluación para que su trabajo pueda ser representado y reconocido en la comunidad de Machine Learning en todo el mundo.