¿Qué es el algoritmo KNN?

El algoritmo de k-vecinos más cercanos (KNN) es un clasificador de aprendizaje supervisado no paramétrico, que utiliza la proximidad para realizar clasificaciones o predicciones sobre la agrupación de un punto de datos individual. Es uno de los clasificadores de clasificación y regresión más populares y simples que se utilizan en el aprendizaje automático en la actualidad.

Si bien el algoritmo KNN se puede utilizar para problemas de regresión o clasificación, normalmente se utiliza como algoritmo de clasificación, partiendo del supuesto de que se pueden encontrar puntos similares cerca unos de otros.

Para los problemas de clasificación, se asigna una etiqueta de clase sobre la base de un voto mayoritario; es decir, se utiliza la etiqueta que se representa con mayor frecuencia alrededor de un punto de datos determinado. Si bien esto se considera técnicamente "voto plural", el término "voto mayoritario" se usa más comúnmente en la literatura. La distinción entre estas terminologías es que la "votación mayoritaria" técnicamente requiere una mayoría superior al 50%, lo que funciona principalmente cuando solo hay dos categorías. Cuando tienes varias clases (por ejemplo, cuatro categorías), no necesariamente necesitas el 50% de los votos para llegar a una conclusión sobre una clase; se podría asignar una etiqueta de clase con un voto superior al 25%.

Los problemas de regresión utilizan un concepto similar al problema de clasificación, pero en este caso, se toma el promedio de los k vecinos más cercanos para hacer una predicción sobre una clasificación. La principal distinción aquí es que la clasificación se usa para valores discretos, mientras que la regresión se usa para valores continuos. Sin embargo, antes de poder realizar una clasificación, se debe definir la distancia. La distancia euclidiana es la más utilizada, en la que profundizaremos más a continuación. También vale la pena señalar que el algoritmo KNN también es parte de una familia de modelos de "aprendizaje diferido", lo que significa que solo almacena un conjunto de

datos de entrenamiento en lugar de pasar por una etapa de entrenamiento. Esto también significa que todo el cálculo ocurre cuando se realiza una clasificación o predicción. Dado que depende en gran medida de la memoria para almacenar todos sus datos de entrenamiento, también se lo conoce como método de aprendizaje basado en instancias o basado en memoria.
A Evelyn Fix y Joseph Hodges se les atribuyen las ideas iniciales sobre el modelo KNN en este artículo de 1951 (el enlace se encuentra fuera de ibm.com), mientras que Thomas Cover amplía su concepto en su investigación (el enlace se encuentra fuera de ibm.com), “Nearest Neighbor Pattern”. Clasificación." Si bien no es tan popular como antes, sigue siendo uno de los primeros algoritmos que uno aprende en ciencia de datos debido a su simplicidad y precisión.

Sin embargo, a medida que crece un conjunto de datos, KNN se vuelve cada vez más ineficiente, lo que compromete el rendimiento general del modelo. Se utiliza comúnmente para sistemas de recomendación simples, reconocimiento de patrones, extracción de datos, predicciones de mercados financieros, detección de intrusiones y más.

En resumen, el objetivo del algoritmo de k vecinos más cercanos es identificar los vecinos más cercanos de un punto de consulta determinado, de modo que podamos asignar una etiqueta de clase a ese punto. Para hacer esto, KNN tiene algunos requisitos:

Calcular KNN: métricas de distancia

Determina tus métricas de distancia

Para determinar qué puntos de datos están más cerca de un punto de consulta determinado, será necesario calcular la distancia entre el punto de consulta y los otros puntos de datos. Estas métricas de distancia ayudan a formar límites de decisión, que dividen los puntos de consulta en diferentes regiones. Normalmente verá límites de decisión visualizados con diagramas de Voronoi.

Si bien hay varias medidas de distancia entre las que puede elegir, este artículo solo cubrirá lo siguiente:

Distancia euclidiana (p=2): esta es la medida de distancia más utilizada y está limitada a vectores de valor real. Usando la siguiente fórmula, mide una línea recta entre el punto de consulta y el otro punto que se está midiendo.

$$d_{(p1,p2)} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Distancia de Manhattan (p=1) : Esta es también otra métrica de distancia popular, que mide el valor absoluto entre dos puntos. También se la conoce como distancia en taxi o distancia a una cuadra de la ciudad, ya que comúnmente se visualiza con una cuadrícula, que ilustra cómo uno puede navegar de una dirección a otra a través de las calles de la ciudad.

Distancia de Minkowski : esta medida de distancia es la forma generalizada de las métricas de distancia euclidiana y de Manhattan. El parámetro p, en la siguiente fórmula, permite la creación de otras métricas de distancia. La distancia euclidiana se representa mediante esta fórmula cuando p es igual a dos, y la distancia de Manhattan se denota con p igual a uno.