¿Qué es el algoritmo de k vecinos más cercanos?

URL	https://www.ibm.com/es-es/topics/knn

¿Qué es el algoritmo KNN?

- Clasificador de aprendizaje supervisado no paramétrico.
- Se suele usar como un algoritmo para problemas de regresión o clasificación.
- Objetivo: identificar los vecinos más cercanos de un punto de consulta determinado para asignarle una etiqueta de clase.
- Parte de la suposición de que se pueden encontrar puntos similares cerca unos de otros.
- La etiqueta de clase se asigna en función de la mayoría de votos.
    - Ej: si hay dos categorías, la mayoría debe ser superior al 50%.
    - Ej2: si hay cuatro categorías, el 25% es suficiente.

- Se toma el promedio de los k vecinos más cercanos para hacer una predicción sobre una clasificación.

- Distinción con la regresión: la clasificación usa valores discretos, la regresión usa valores continuos.
- Antes de hacer una clasificación, se debe definir la distancia.
  - La más utilizada es la distancia euclidiana.
- Es parte de la familia de los modelos de “aprendizaje vago”: solo almacena un conjunto de datos de entrenamiento, no se somete a una etapa de entrenamiento.
  - Esto significa que todo el cálculo se produce cuando se está haciendo una clasificación o predicción.
- También conocido como un método de aprendizaje basado en instancias o en memoria porque depende en gran medida de la memoria para almacenar todos sus datos de entrenamiento.
- Créditos:
  - Evelyn Fix y Joseph Hodges, 1951: ideas iniciales.
  - Thomas Cover: amplía su concepto de investigación, “Nearest Neighbor Pattern Clasification”.
- Usos comunes: recomendaciones simples, reconocimiento de patrones, minería de datos, predicciones de mercados financieros, detección de intrusiones…

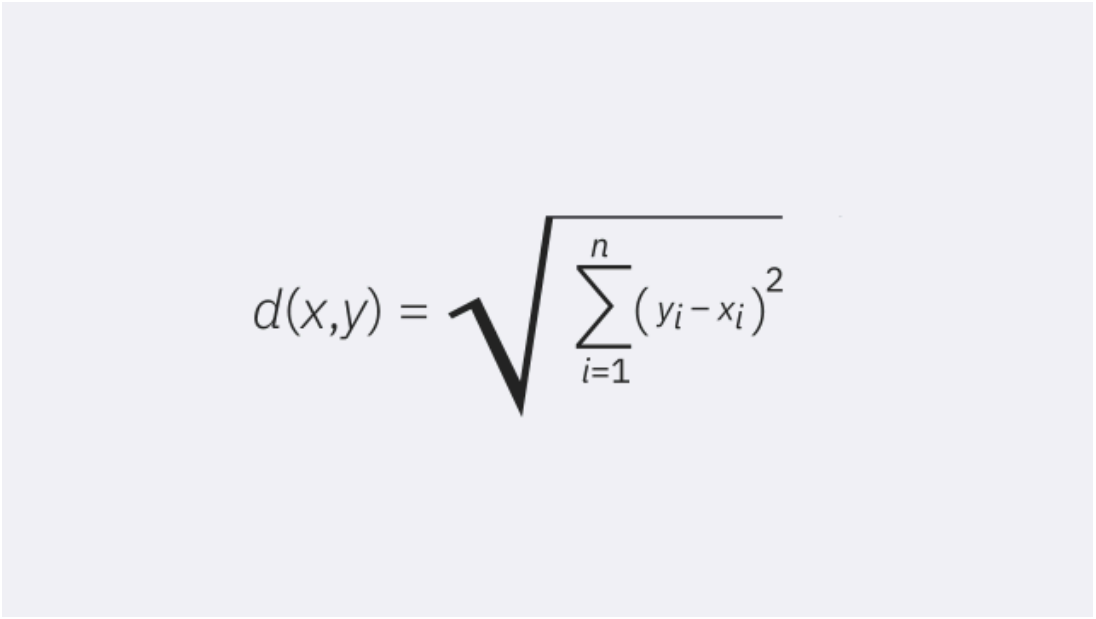
Calcular KNN: métricas de distancia

- Calcular la distancia entre el punto de consulta y los demás puntos de datos.
- Ayudan a formar límites de decisión.
- Los limites de decisión de visualizan frecuentemente con diagramas de Voronoi.

Distancia más comunes:

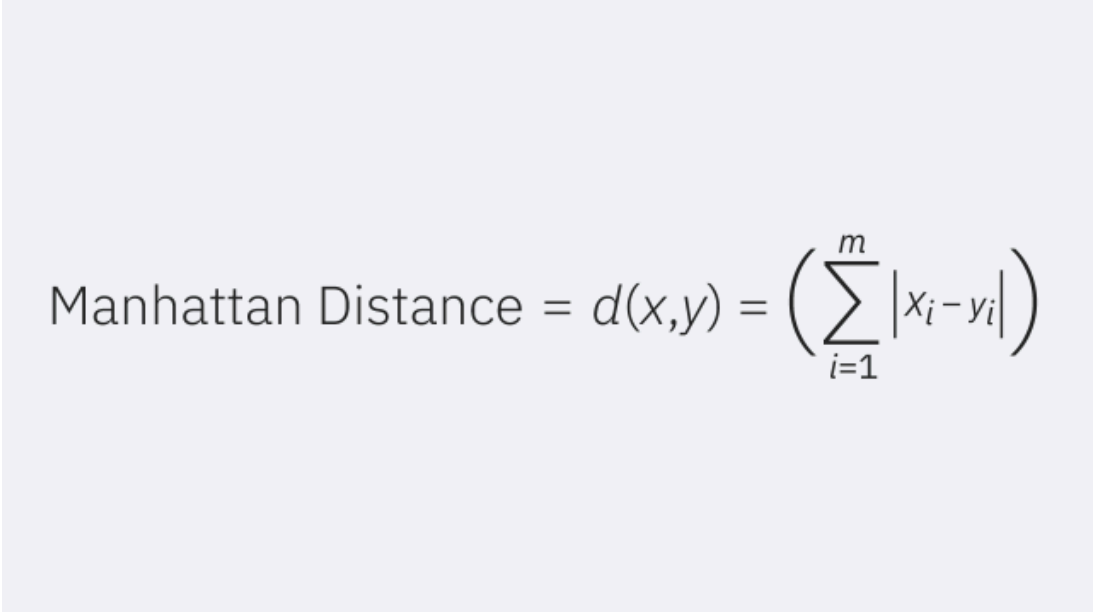
▼ Distancia euclídea (p=2)

- La más utilizada.
- Se limita a vectores de valor real.
- Mide una línea recta entre el punto de consulta y el otro punto medido.



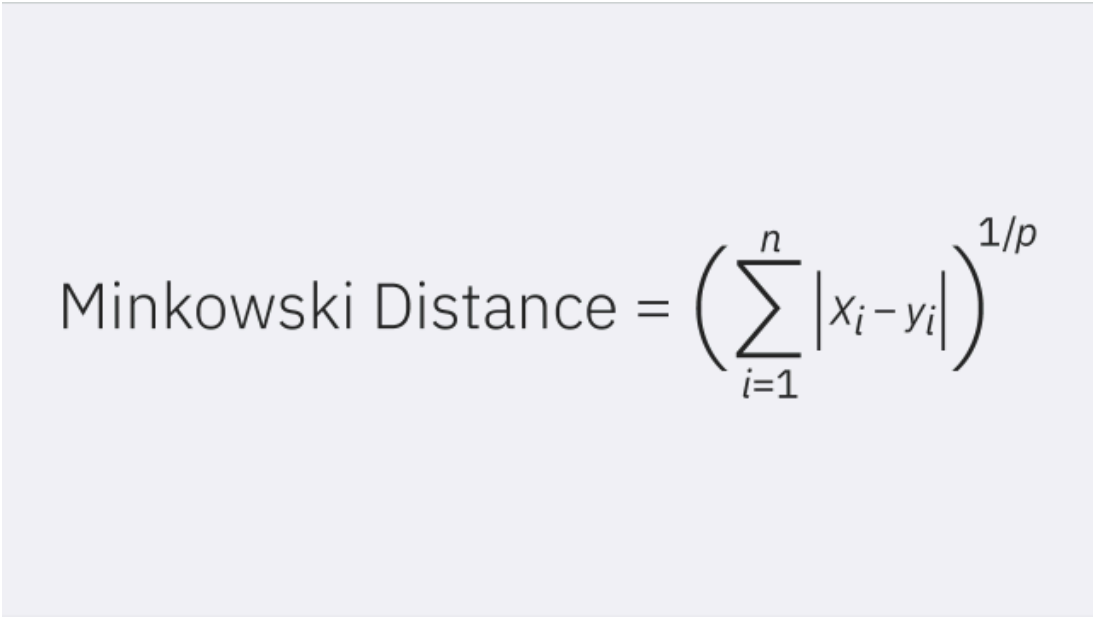
▼ Distancia de Manhattan (p=1)

- Mide el valor absoluto entre dos puntos.
- También conocida como distancia de taxi o distancia de cuadra de la ciudad porque comúnmente se visualiza con una cuadrícula.



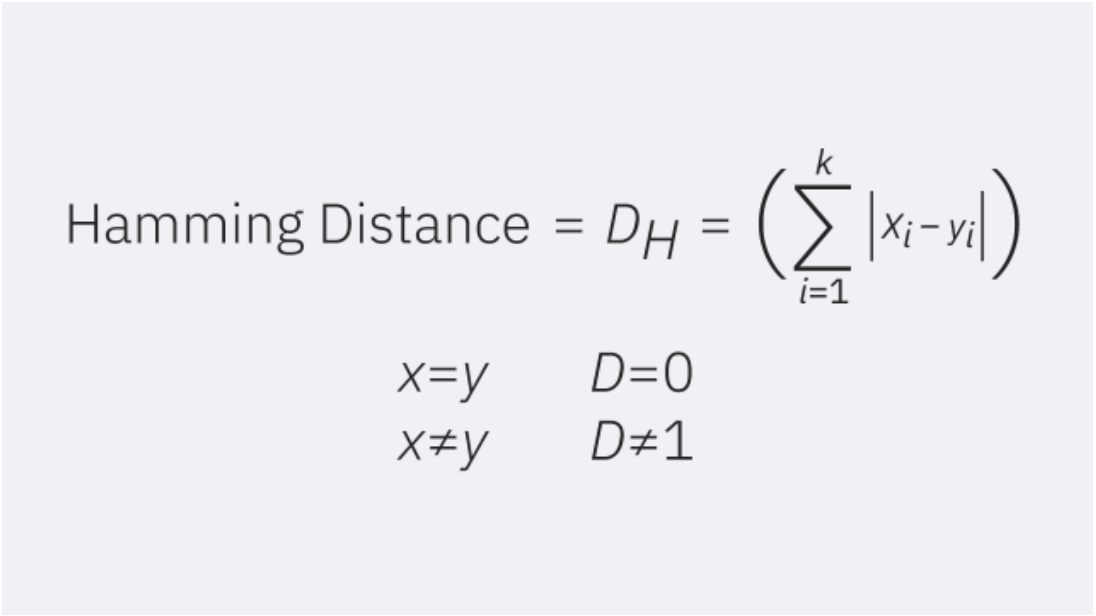
▼ Distancia de Minkowski

- Forma generalizada de las métricas de distancia euclidianas y de Manhattan.
- p: permite la creación de otras métricas de distancia.

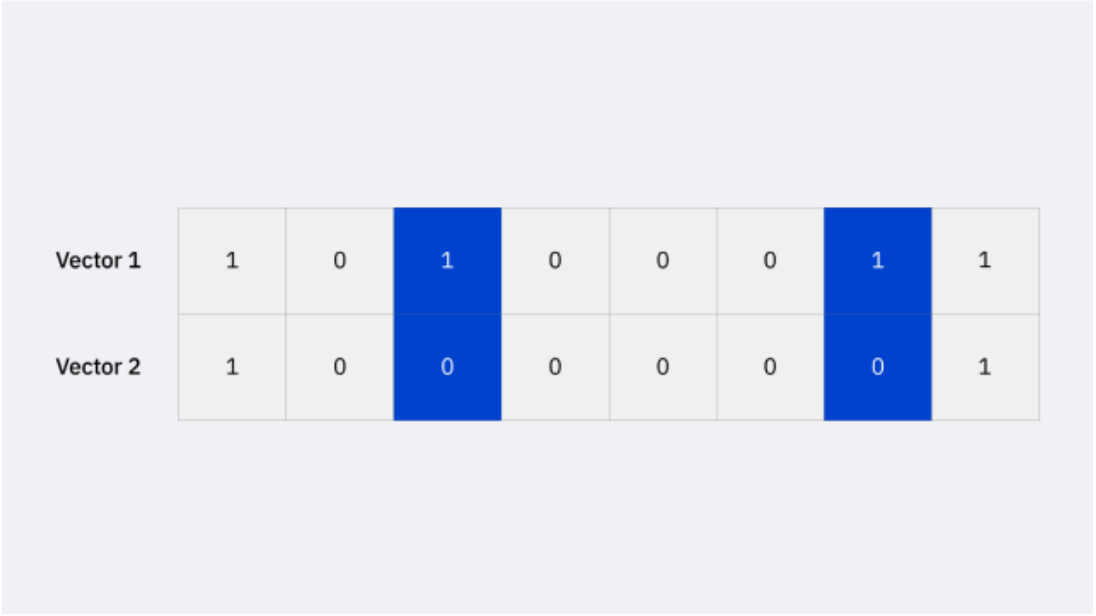


▼ Distancia de Hamming

- Se utiliza con vectores booleanos o de cadena para identificar los puntos en los que los vectores no coinciden.
- También denominado métrica de superposición.



- Por ejemplo, con las siguientes cadenas la distancia de hamming sería 2 ya que solo dos de los valores difieren:



Calcular KNN: definir k

- Define cuántos vecinos se comprobarán para determinar la clasificación del punto de consulta.
  - Ej: k=1, la instancia se asignará a la misma clase que su único vecino más cercano.
- La elección de k dependerá de los datos de entrada.
  - Los valores atípicos o ruido funcionarán mejor con valores más altos de k.
- Se recomienda tener un número impar para k para evitar empates en la clasificación.
- Las tácticas de validación cruzada pueden ayudar a elegir el k óptimo para el conjunto de datos.

Aplicaciones de KNN en el ML

- Preprocesamiento de datos: a los conjuntos de datos les faltan valores. Con el algoritmo KNN se pueden estimar (proceso conocido como imputación de datos faltantes).
- Motores de recomendación: ofrecer recomendaciones automáticas a los usuarios sobre contenido pero no es óptimo para conjuntos de datos grandes.
- Finanzas: determinar la solvencia crediticia de un solicitante de préstamo, previsiones del mercado bursátil, tipos de cambio de divisas, futuros de trading, análisis del blanqueo de dinero…

- Sanidad: predicciones sobre el riesgo de ataques cardiacos, de cáncer de próstata... Funciona calculando las expresiones genéticas más probables.
- Reconocimiento de patrones: identificar números manuscritos.

Ventajas y desventajas del algoritmo KNN

Ventajas

- Fácil de implementar.
- Se adapta fácilmente: cuando se añaden nuevas muestras de entrenamiento, el algoritmo se ajusta para tener en cuenta los nuevos datos, ya que todos los datos de entrenamiento se almacenan en memoria.
- Pocos hiperparámetros: solo requiere el valor k y una métrica de distancia.

Desventajas

- Mala escalabilidad: al ser un algoritmo vago ocupa más memoria y almacenamiento de datos que otros clasificadores.
- La maldición de la dimensionalidad: no funciona bien con entradas de datos de alta dimensión. Cuando el algoritmo alcanza el número óptimo de características, las características adicionales aumentan la cantidad de errores de clasificación.
- Propenso al sobreajuste: los valores más bajos de k pueden sobreajustar los datos, mientras que los valores más altos “suavizan” los valores de predicción. Si el valor de k es demasiado alto, puede infraajustarse a los datos.