Inteligencia Artificial

Juan Pablo Restrepo Uribe

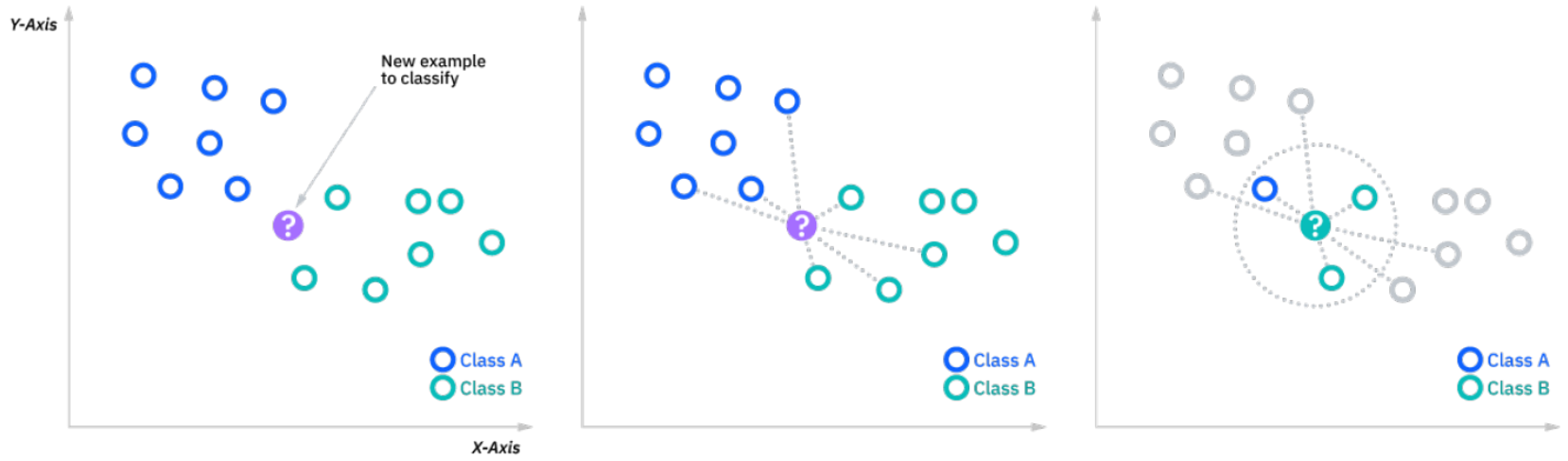
Ing. Biomedico - MSc. Automatización y Control Industrial

jprestrepo@correo.iue.edu.co

Institución Universitaria de Envigado

KNN

También conocido como k vecinos más cercanos es un clasificador de aprendizaje supervisado no paramétrico, que utiliza la proximidad para hacer clasificaciones o predicciones sobre la agrupación de un punto de datos individual.

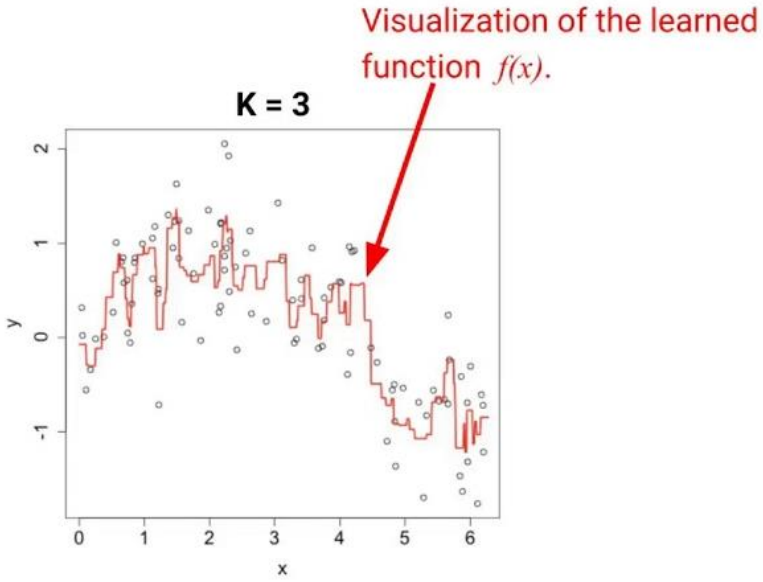


KNN VS Voto de mayoría

- El KNN asigna una etiqueta de clase sobre la base de un voto mayoritario (frecuencia alrededor de un punto de datos determinado)
- Esto técnicamente se considera "voto por mayoría
- La distinción entre estas terminologías es que "voto mayoritario" técnicamente requiere una mayoría superior al 50 %, lo que funciona principalmente cuando solo hay dos categorías. Cuando tiene varias clases no necesita necesariamente el 50 % de los votos para llegar a una conclusión sobre una clase.

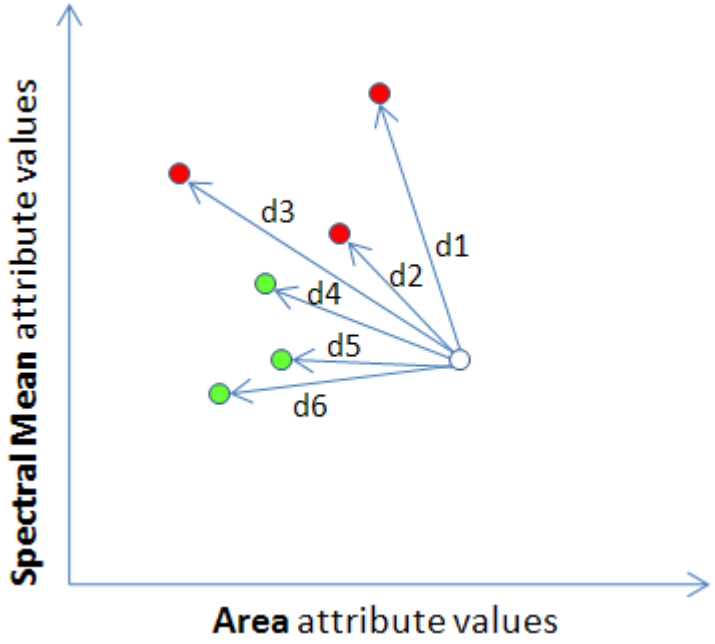
KNN Regresión

Los problemas de regresión usan un concepto similar al de los problemas de clasificación, pero en este caso, se toma el promedio de los k vecinos más cercanos para hacer una predicción sobre una clasificación.



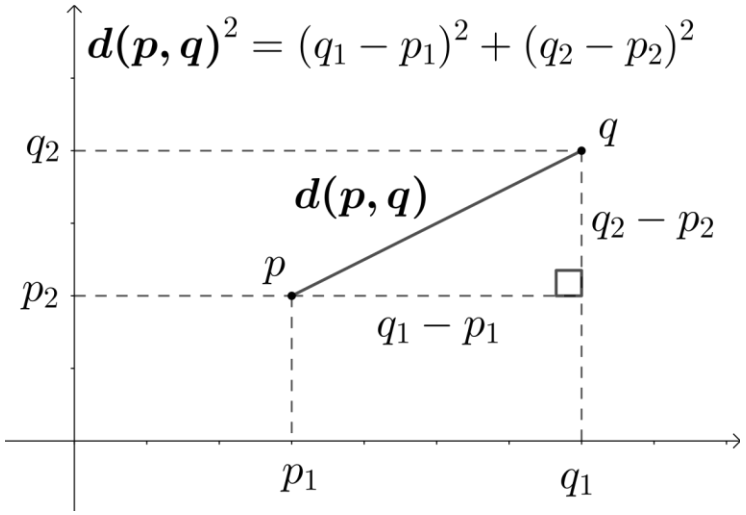
Métricas de distancia

El objetivo del algoritmo del vecino más cercano es identificar los vecinos más cercanos de un punto de consulta dado



Distancia euclidiana (p=2)

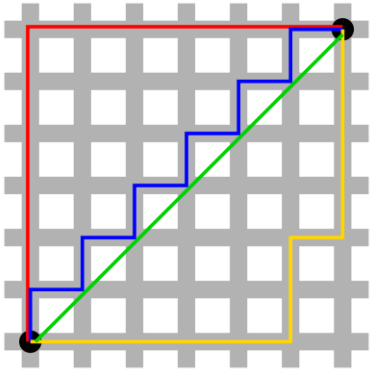
Esta es la medida de distancia más utilizada y está limitada a vectores de valor real. Mide una línea recta entre el punto de consulta y el otro punto que se mide.



$$d(x,y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

Distancia Manhattan (p=1)

Mide el valor absoluto entre dos puntos. También se conoce como distancia de taxi o distancia de cuadra de la ciudad, ya que comúnmente se visualiza con una cuadrícula, que ilustra cómo se puede navegar de una dirección a otra a través de las calles de la ciudad.

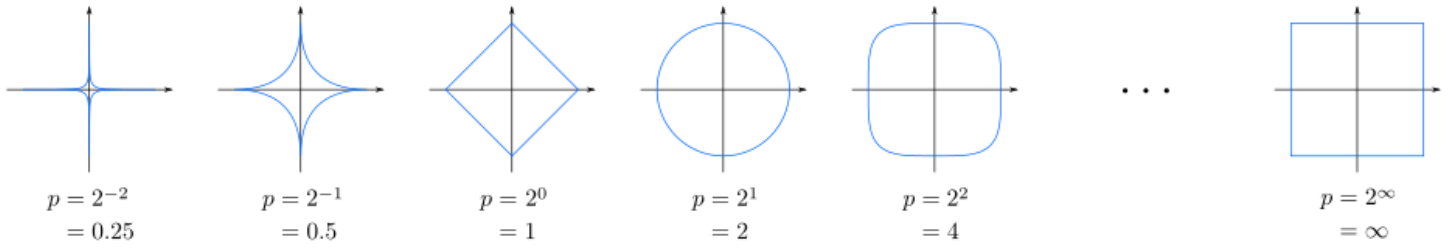


$$d(x,y) = \left(\sum_{i=1}^{m} \left| x_i - y_i \right| \right)$$

$$\left(\sum_{i=1}^{n} \left|x_i - y_i\right|\right)^{1/p}$$

Distancia Minkowski

Esta medida de distancia es la forma generalizada de las métricas de distancia Euclidiana y Manhattan. El parámetro, p, en la fórmula a continuación, 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.

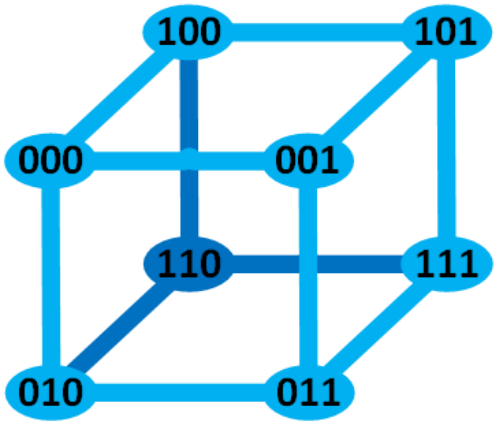


Distancia de hamming

Esta técnica se usa típicamente con vectores booleanos o de cadena, identificando los puntos donde los vectores no coinciden. Como resultado, también se la conoce como la métrica de superposición. Esto se puede representar con la siguiente fórmula:

3 Dimensiones

0 0 0
0 0 1
0 1 0
0 1 1
1 0 0
1 0 1
1 1 0
1 1 1



$$\text{Hamming Distance} = D_H = \left(\sum_{i=1}^{k} |x_i - y_i| \right)$$

$x=y \quad D=0$

$x \neq y \quad D \neq 1$

Definiendo k

- El valor k en el algoritmo k-NN define cuántos vecinos se verificarán para determinar la clasificación de un punto de consulta específico. Por ejemplo, si k=1, la instancia se asignará a la misma clase que su vecino más cercano.
- Los valores más bajos de k pueden tener una varianza alta, pero un sesgo bajo, y los valores más grandes de k pueden generar un sesgo alto y una varianza más baja.
- En general, se recomienda tener un número impar para k para evitar empates en la clasificación, y las tácticas de validación cruzada pueden ayudarlo a elegir la k óptima para su conjunto de datos.

Aplicaciones de k-NN en machine learning

- Preprocesamiento de datos: Los conjuntos de datos suelen tener valores faltantes, pero el algoritmo KNN puede estimar esos valores en un proceso conocido como imputación de datos faltantes.
- Motores de recomendación : utilizando datos de flujo de clics de sitios web, el algoritmo KNN se ha utilizado para proporcionar recomendaciones automáticas a los usuarios sobre contenido adicional.
- Finanzas: Se ha KNN en datos crediticios para ayudar a los bancos a evaluar el riesgo de un préstamo para una organización o individuo. Se utiliza para determinar la solvencia crediticia de un solicitante de préstamo.
- Cuidado de la salud: KNN se ha aplicado dentro de la industria de la salud, haciendo predicciones sobre el riesgo de ataques cardíacos y cáncer de próstata.
- Reconocimiento de patrones: KNN también ha ayudado a identificar patrones, como en texto y clasificación de dígitos

Referencias interesantes

- https://pages.stat.wisc.edu/~sraschka/teaching/stat479-fs2018/
- https://apps.dtic.mil/sti/pdfs/ADA800276.pdf
- https://isl.stanford.edu/~cover/papers/transIT/0021cove.pdf
- https://iopscience.iop.org/article/10.1088/1742-6596/1025/1/012114/pdf
- https://www.ijera.com/papers/Vol3_issue5/DI35605610.pdf
- https://developer.ibm.com/tutorials/learn-classification-algorithms-using-python-and-scikit-learn/

Variante del algoritmo básico (Vecinos más cercanos con distancia ponderada)

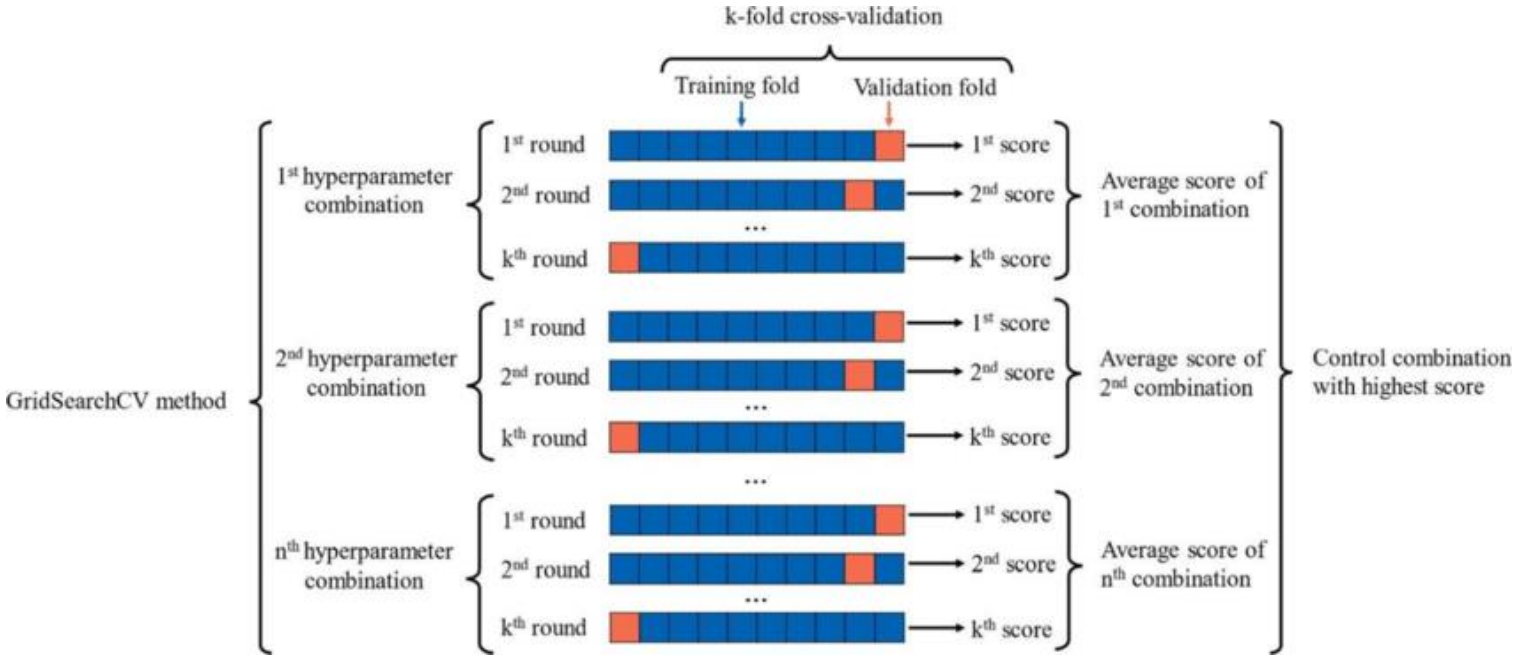
Se puede ponderar la contribución de cada vecino de acuerdo con la distancia entre él y el ejemplar a ser clasificado, dando mayor peso a los vecinos más cercanos. Por ejemplo, podemos ponderar el voto de cada vecino de acuerdo con el cuadrado inverso de sus distancias

$$w_i \equiv \frac{1}{d(x_q, x_i)^2}$$

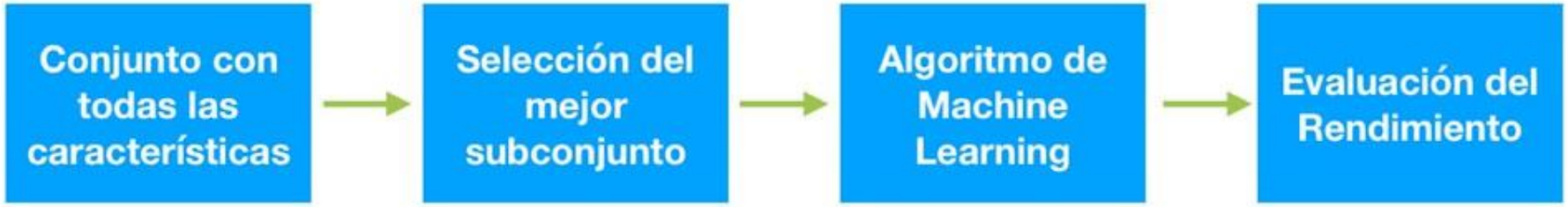
$$\hat{f}(x_q) \leftarrow \arg\max_{v \in V} \sum_{i=1}^{k} w_i [v = f(x_i)]$$

Búsqueda exhaustiva (GridSearchCV)

Búsqueda exhaustiva sobre valores de parámetros específicos para un estimador.



Selección de características (Métodos de Filtro)



Selección de características (Métodos de envoltura)

Selección de características (SelectKBest)

Seleccione características de acuerdo con las k puntuaciones más altas.