Informática y comunicaciones

ies azarquiel

Curso de especialización inteligencia artificial y big data

Tema 6: K-nn

Profesor: Sebastián Rubio Valero

Noviembre 2025

Sistema aprendizaje automático

El algoritmo k vecinos más cercanos (kNN) 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. Es uno de los clasificadores de clasificación y regresión más populares y sencillos que se utilizan en machine learning hoy en día.

Se toma el promedio de los k vecinos más cercanos para hacer una predicción sobre una clasificación.Aquí la clasificación se usa para valores discretos, mientras que la regresión se usa para valores continuos. Pero antes de hacer una clasificación, se debe definir la distancia.

Para determinar qué puntos de datos están más cerca de otro punto es necesario calcular la distancia entre el punto de consulta y los demás puntos de datos. Estas métricas de distancia ayudan a formar límites de decisión,que dividen los puntos consulta en diferentes regiones. También puede ser utilizado para regresión, aunque esencialmente es un clasificador

Tenemos varias formas de calcular la distancia:

- Distancia euclídea (p=2)

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

- Distancia Manhatan (p=1)

$$d(x,y) = \left(\sum_{i=1}^{m}|x_i - y_i|\right) \tag{2}$$

  El parámetro p permite la creación de otras métricas

- Distancia de Minkowski

$$d(x,y) = \left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{1/p} \tag{3}$$

Vamos a ver un ejemplo de aplicación Partiendo de los datos anteriores queremos saber a que clase (manzana o naranja) pertenece la siguiente fruta:

- Peso=160 g

Peso (g)	Textura	Clase
150	0	Manzana
170	0	Manzana
140	1	Naranja
130	1	Naranja

Table 1: Peso, textura y clase de frutas

- Textura=0

Las diferentes fases del calculos son:

1. (150,0) → Manzana

$$\sqrt{(160-150)^2+(0-0)^2} = \sqrt{100} = 10 \tag{4}$$

2. (170,0) → Manzana

$$\sqrt{(160-170)^2+(0-0)^2} = \sqrt{100} = 10 \tag{5}$$

3. (140,1) → Naranja

$$\sqrt{(160-140)^2+(0-1)^2} = \sqrt{400+1} \approx 20.2 \tag{6}$$

4. (130,1) → Naranja

$$\sqrt{(160-130)^2+(0-1)^2} = \sqrt{900+1} \approx 30.01 \tag{7}$$

Si elegimos un k=3, entonces los tres vecinos más cercanos son: Entonces clasifica-

Punto	Distancia	Clase
(150,0)	10	Manzana
(170,0)	10	Manzana
(140,1)	20.02	Naranja

Table 2: Peso, textura y clase de frutas

ciones por mayoría, *Manzana* Si tomamos un k=4 se produce un empate, la solución pasa por reducir k o elegir una clase aleatoriamente.

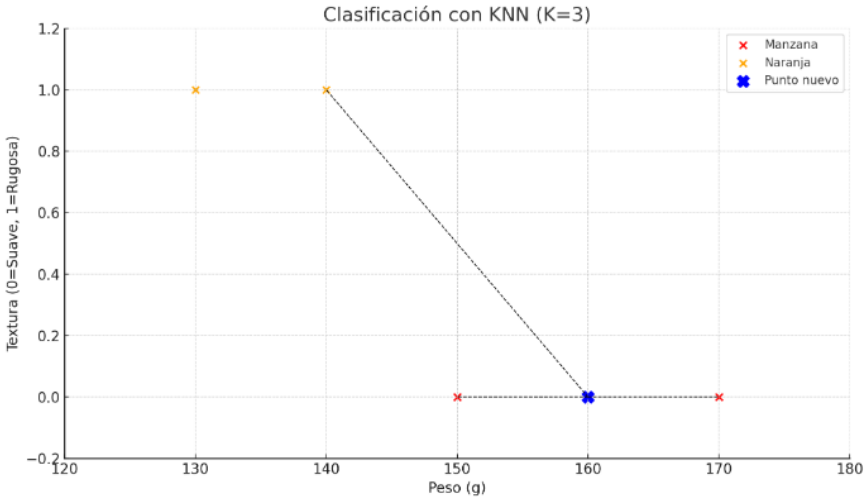

Figure 1: Enter Caption

Lo ideal es escoger un valor de k impar para evitar los posibles empates. Indudablemente sabemos de antemano las clases que hay con una característica (puden ser más de una) Si k es muy pequeña,elresultado es muy sensible al ruido. Si k es demasiado grande,se tendrán encuenta elementos de clases lejanas. Por lo que se debe encontrar un valor adecuado.

Ventajas: Simple de implementar
Desventajas: Sólo es aplicable a características numéricas, y si son contínuas mejor Si hay muchos datos será necesario mucha capacidad de computo No es robusto para clases poco balanceadas.

Estos modelos no necesitan un entrenamiento previo, por lo general usaremos validación cruzada para que se encuentre el mejor valor de K, pero si lo vamos hacer a mano entonces empezaremos con un valor de $k = \sqrt{n}$ donde n será el número de ejemplos de entrenamiento.

Otro aspecto ha tener en cuenta son los valores faltantes y los atípicos que afectan mucho, por lo tanto hay que tenerlos debemos actuar sobre ellos.
$EJERCICIO.1$: Dada el siguiente conjunto de datos, realizar los cálculos necesarios para decidir a que clase pertenece una flor con largo=5 y ancho=1.7
$EJERCICIO.2$ Repetir el ejercicio anterior utilizando otra métrica de las estudiadas.

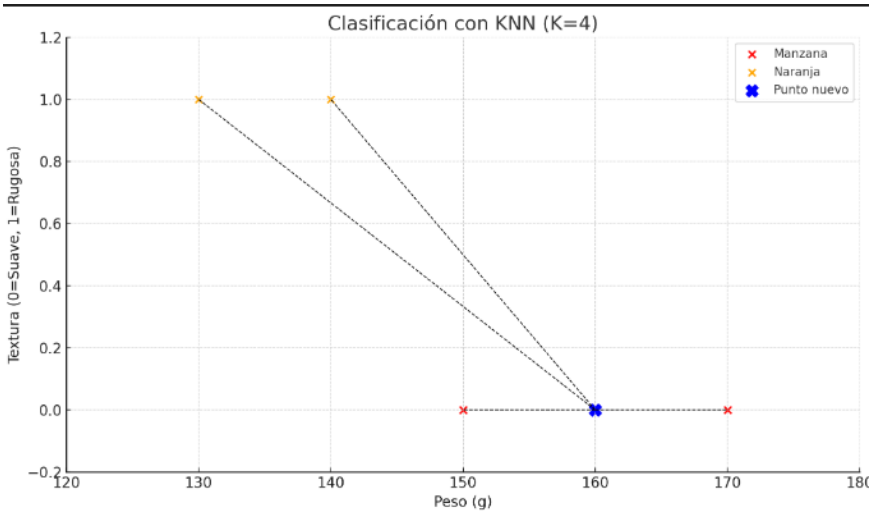

Figure 2: Enter Caption

	sepal_length	sepal_width	petal_length	petal_width	species
114	5.8	2.8	5.1	2.4	virginica
62	6.0	2.2	4.0	1.0	versicolor
33	5.5	4.2	1.4	0.2	setosa
107	7.3	2.9	6.3	1.8	virginica
7	5.0	3.4	1.5	0.2	setosa

Figure 3: Enter Caption