Modelos Supervisados

Certificación en Ciencia de Datos

Andrés Martínez

28 Junio 2024

Universidad Externado de Colombia
135 Años

Contenido

- Introducción.
- Definición.
- Importancia.
- Métodos.
- Validación cruzada k-fold.
- Validación cruzada Leave-One-Out.
- Comparación de métodos.
- Implementación.
- Aplicaciones.
- Conclusión.

Algoritmo k-Nearest Neighbors (k-NN)

Introducción

El algoritmo k-Nearest Neighbors (k-NN) es un método de aprendizaje supervisado que se utiliza para clasificación y regresión. Su enfoque se basa en la proximidad de los puntos de datos en un espacio de características.

Clasificación en k-NN

Dado un punto de consulta Y, la clasificación en k-NN se realiza de la siguiente manera:

- Calculamos la distancia entre Y y todos los puntos en el conjunto de entrenamiento.
- Seleccionamos los k puntos más cercanos a Y según la medida de distancia.
- Contamos las clases de estos k puntos.
- Asignamos la clase más común a Y como la predicción.

KNN Regresión

Introducción

El algoritmo k-Nearest Neighbors (k-NN) utiliza medidas de distancia para encontrar los k puntos de datos más cercanos a un punto de consulta. Estas medidas son esenciales en el proceso de clasificación y regresión en k-NN.

Regresión en k-NN

Para la regresión en k-NN, el proceso es similar, pero en lugar de contar clases, calculamos un promedio de los valores de los k vecinos más cercanos y asignamos ese promedio como la predicción de Y.

Diapositiva 1: Distancia Euclidiana

Distancia Euclidiana

La distancia euclidiana es una medida de distancia común en k-NN. Se calcula utilizando la fórmula:

$$d(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (X_1^{(i)} - X_2^{(i)})^2}$$

Donde:

- $d(Y, X)$: Distancia entre los puntos Y y X.
- n: Número de características (dimensiones) en los puntos.
- $Y^{(i)}$ y $X^{(i)}$: Valores de la característica i en los puntos Y y X.

Diapositiva 2: Otras Medidas de Distancia

Otras Medidas de Distancia

Existen varias medidas de distancia que se pueden utilizar en k-NN, dependiendo de la naturaleza de los datos y el problema. Algunas de ellas incluyen:

- Distancia de Manhattan:

$$d(Y, X) = \sum_{i=1}^{n} |Y^{(i)} - X^{(i)}|$$

- Distancia de Hamming (para datos categóricos):

$$d(Y, X) = \frac{\sum_{i=1}^{n} \delta(Y^{(i)}, X^{(i)})}{n}$$

- Coeficiente de correlación (para datos relacionados):

$$d(Y, X) = 1 - \frac{\sum_{i=1}^{n} (Y^{(i)} \cdot X^{(i)})}{\sqrt{\sum_{i=1}^{n} (Y^{(i)})^2} \cdot \sqrt{\sum_{i=1}^{n} (X^{(i)})^2}}$$

Diapositiva 3: Cálculo de Distancias en k-NN

Cálculo de Distancias

Para encontrar los k vecinos más cercanos a un punto de consulta Y, calculamos la distancia entre Y y cada punto X_i en el conjunto de entrenamiento. Luego, seleccionamos los k puntos con las distancias más pequeñas.

$$d(Y, X_i) = \sqrt{\sum_{j=1}^{n} (Y^{(j)} - X_i^{(j)})^2}$$

Donde:

- $d(Y, X_i)$: Distancia entre Y y X_i.
- n: Número de características (dimensiones) en los puntos.
- $Y^{(j)}$ y $X_i^{(j)}$: Valores de la característica j en Y y X_i.

K Vecinos más Cercanos (K-NN)

Algoritmo K-NN:

- Calcula la distancia entre el punto y los vecinos más cercanos.
- Clasifica basado en la mayoría de las etiquetas de los vecinos.

Ecuación de Predicción:

$$y = \text{mode}(\text{neighbors})$$

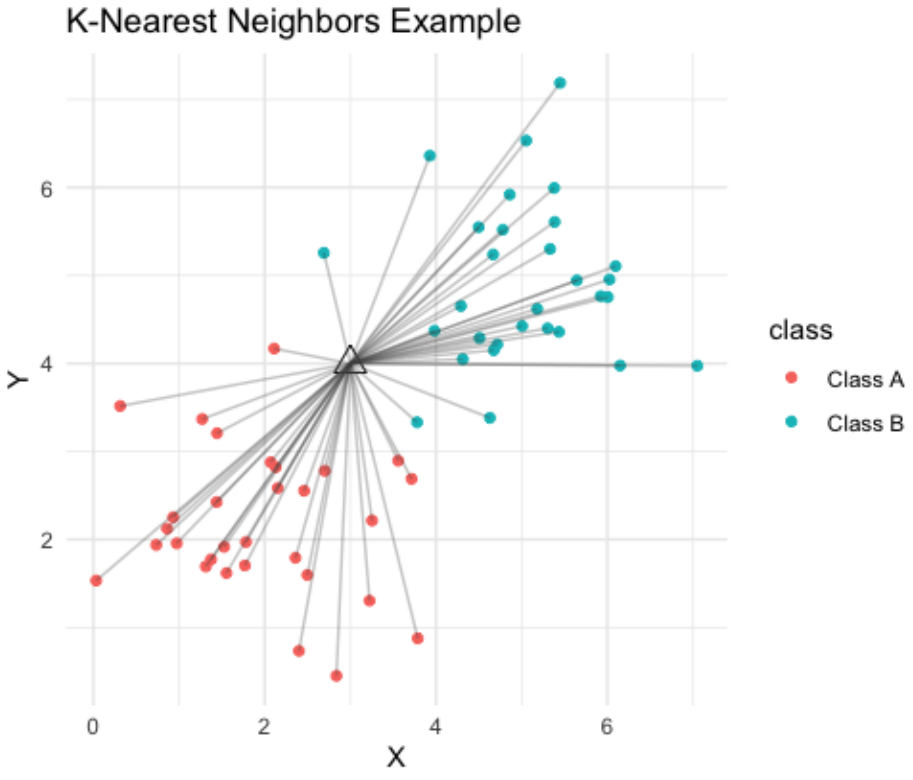
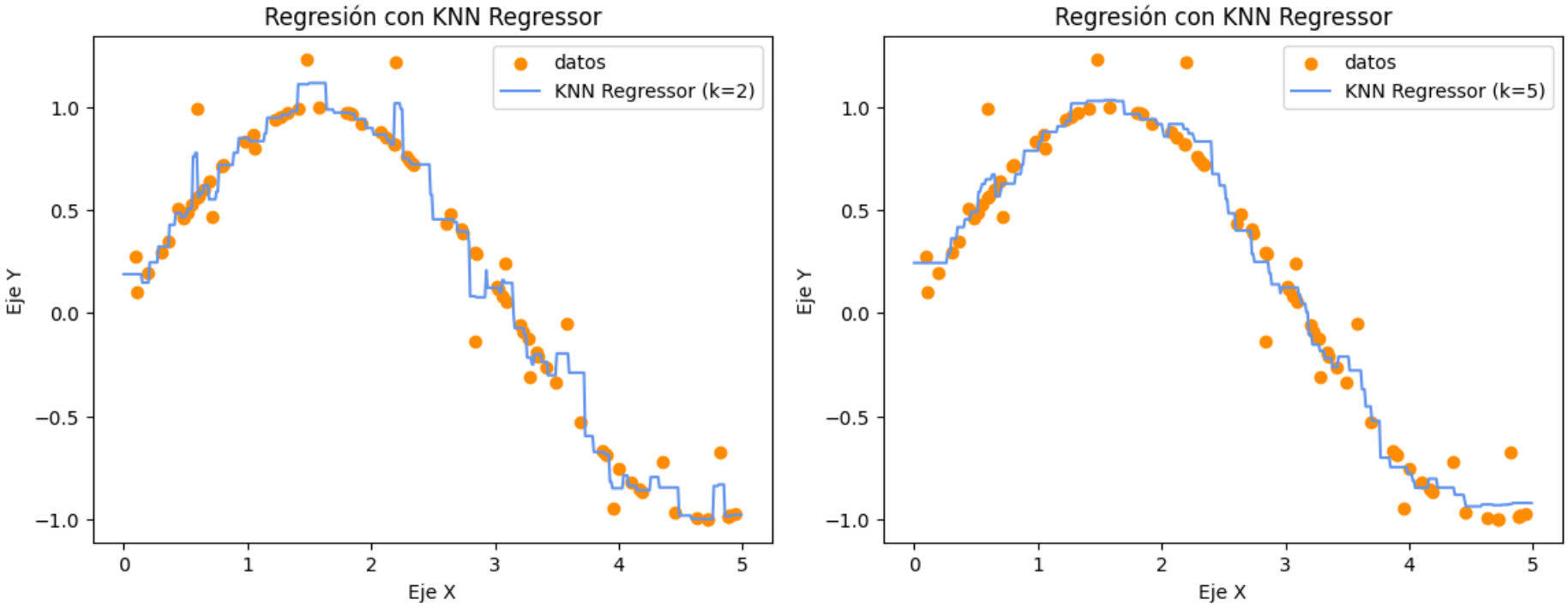


Figure 1: KNN

Calibración de un KNN de regresión

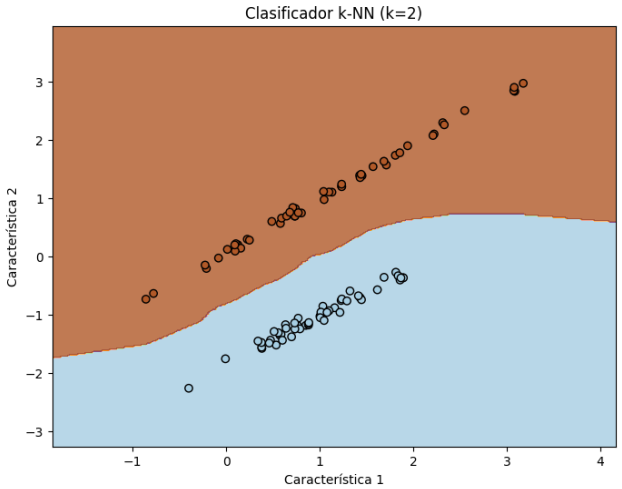


(a) Dos(K) vecinos

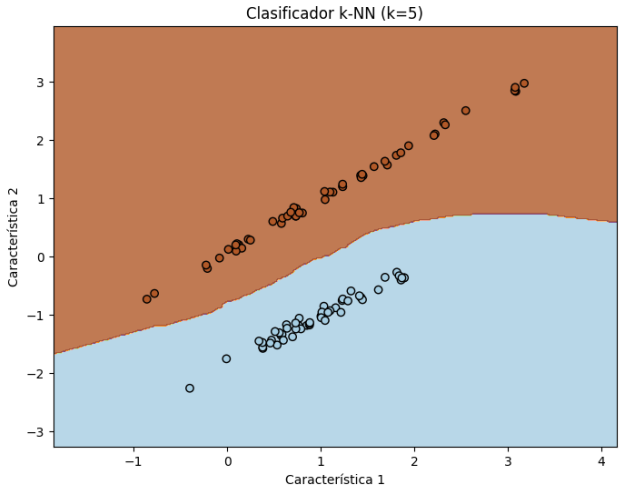
(b) Cinco (k) Vecinos

Figure 2: Comparación de KNN de regresión

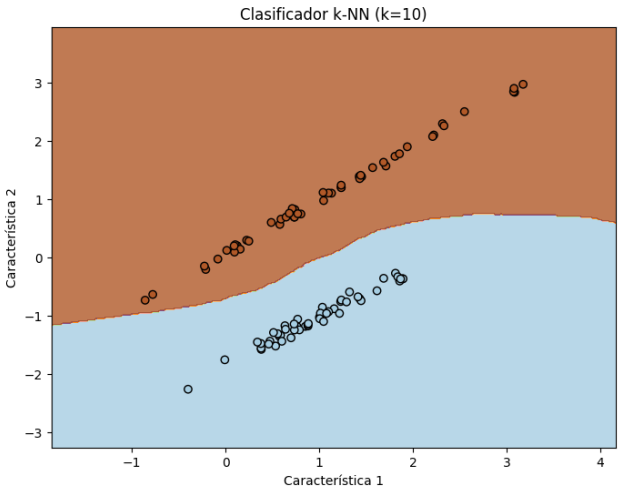
Calibración de un KNN de clasificación



(a) Dos(K) vecinos



(b) Cinco (k) Vecinos



(c) Diez (k) Vecinos