El método k vecinos más cercanos (KNN)
Aprendizaje Estadístico, 2024/2025

UNIVERSIDAD DE
MURCIA

Semana del 17 de Noviembre, 2024
Aprendizaje Estadístico

Máster en Tecnología de Análisis de Datos Masivos - Big Data.

Juan A. Botía (juanbot@um.es)

Introducción

UNIVERSIDAD DE MURCIA

- Métodos sin modelo (model free)
  - No se requiere de una fase de aprendizaje
- Muy útil para clasificación y como algoritmo *baseline*
- *No trabaja bien en problemas de alta dimensionalidad*
- *Cercano al clasificador bayesiano, con el que estimamos P(Y|X)*
- *Knn también*
  - *trabaja mediante estimaciones on-the-fly de P(Y|X)*
  - *utiliza la regla del clasificador de bayes para la inferencia*
    - *La clase C para la observación o es la que genera un P(C|X) máximo*

El método KNN

UNIVERSIDAD DE MURCIA

- Dados el valor de K y una observación de test x_0 la inferencia funciona como sigue
  - Obtenemos los k puntos de los datos de training más próximos a x_0
  - Estimamos la probabilidad condicionada $P(Y_j|x_0)$ para la clase j

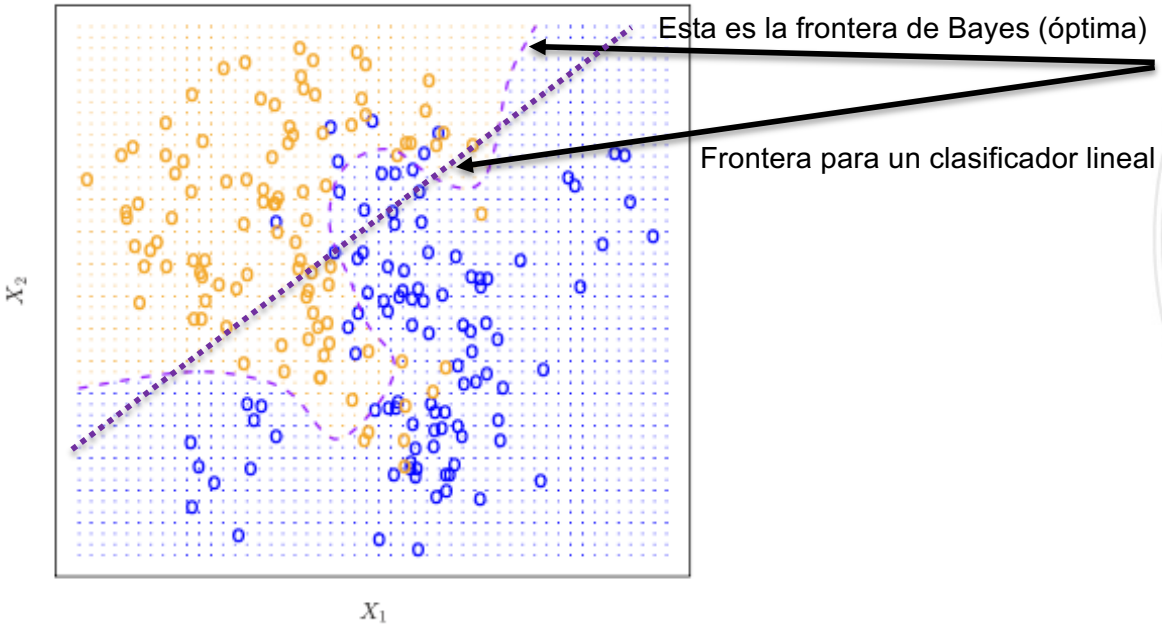
$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

  - Aplicamos la regla de Bayes (x_0 pertenece a la clase cuya probabilidad condicionada es máxima)
- ¿Cómo localizar los puntos más próximos?
  - *Si los predictores están definidos en el espacio de los reales, la distancia Euclidiana es una buena opción*
  - *Necesario normalizar los predictores*

El método KNN (y II)

UNIVERSIDAD DE MURCIA

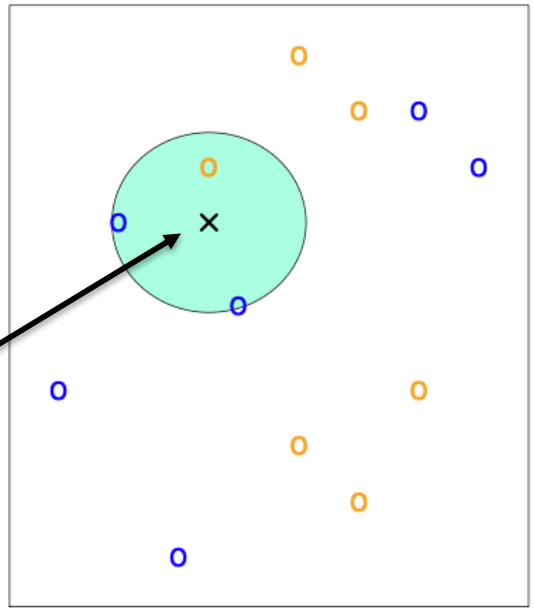
- Cuando las fronteras de decisión (*decision boundaries*) son irregulares, KNN se comporta muy bien, es muy flexible



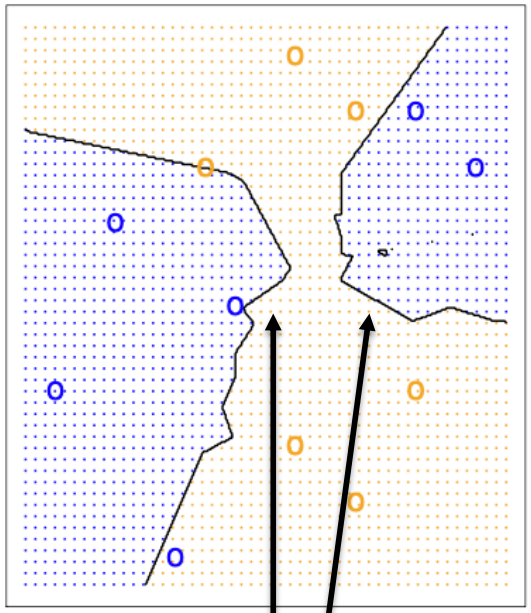
fronteras de decision para un problema de clasificación

Figura 2.13. Ejemplo simulado (100 puntos para cada clase)

Ejemplo: K=3



La clase más probable es la azul para el punto x

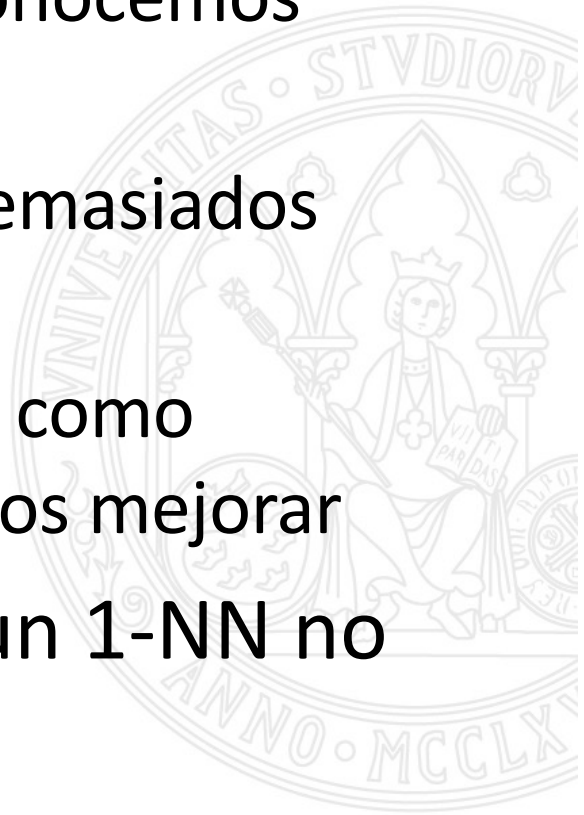


Fronteras de decision para K=3

KNN como algoritmo baseline

UNIVERSIDAD DE MURCIA

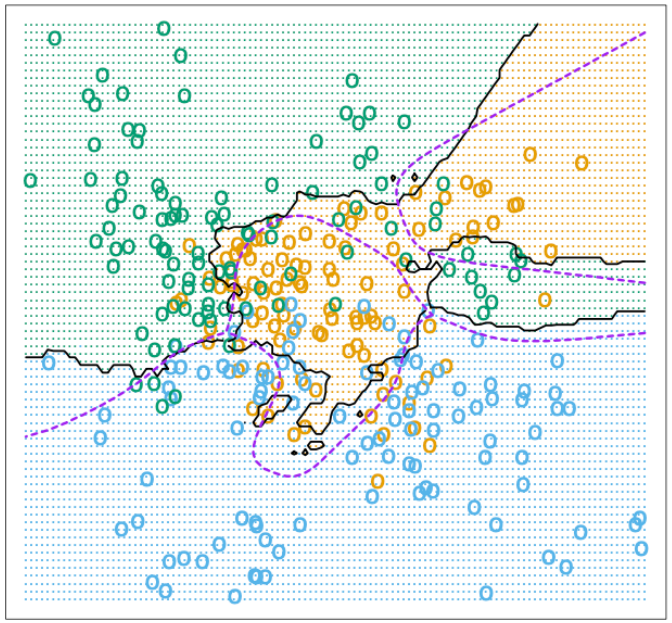
- Un algoritmo baseline:
  - sus sesgos, ventajas y desventajas conocemos muy bien
  - funciona razonablemente bien sin demasiados *ajustes*
  - Es bueno usar un algoritmo baseline como referencia, a partir de él solo podemos mejorar
- Sabemos que el ratio de error de un 1-NN no supera dos veces el error de Bayes
  - El clasificador de Bayes es óptimo



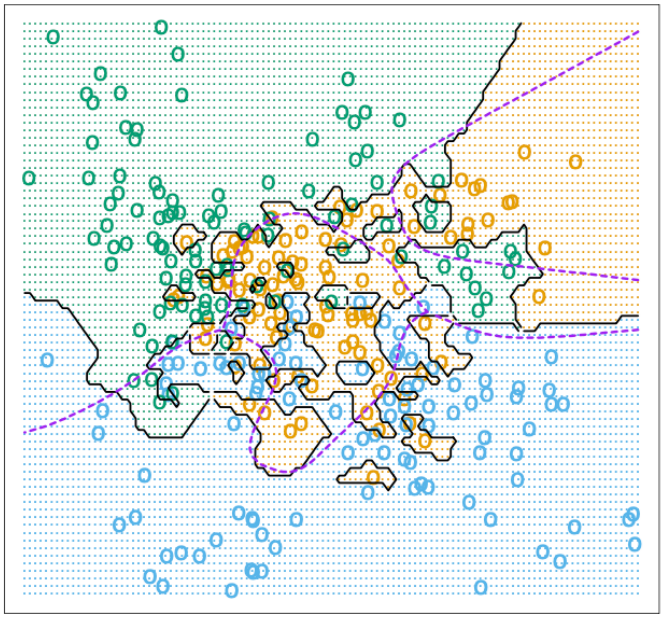
Error de Bayes y 1-NN

UNIVERSIDAD DE MURCIA

15-Nearest Neighbors



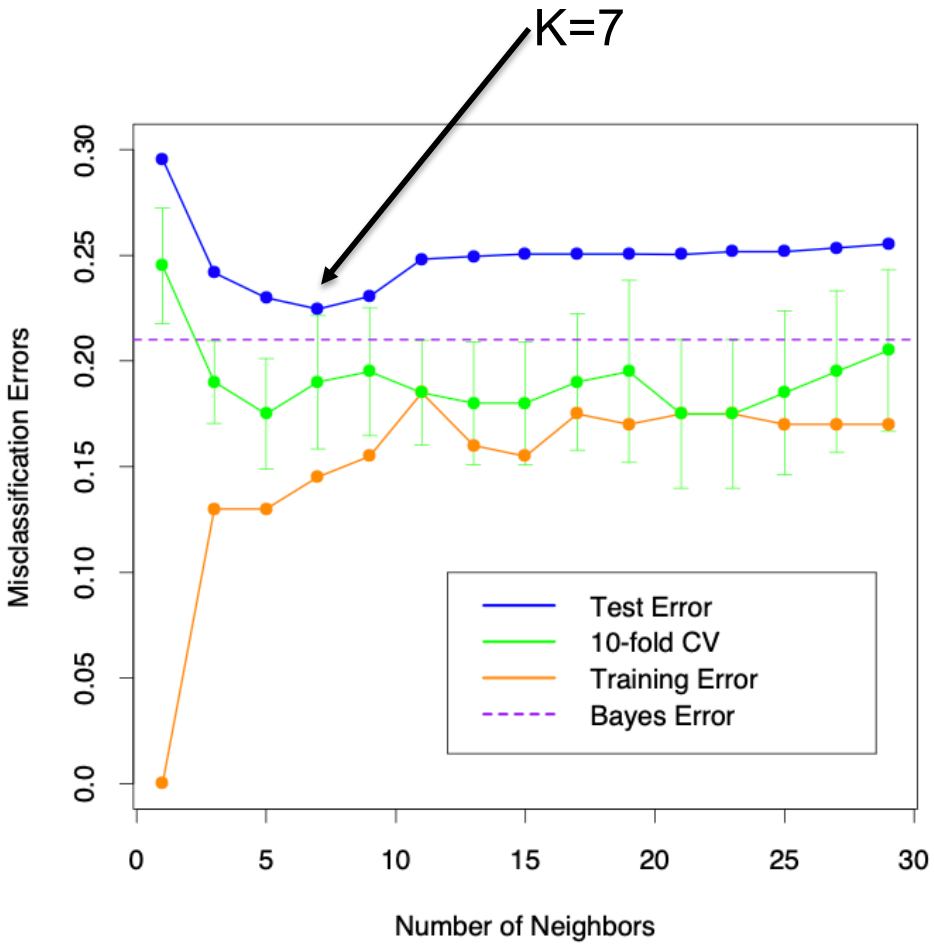
1-Nearest Neighbor

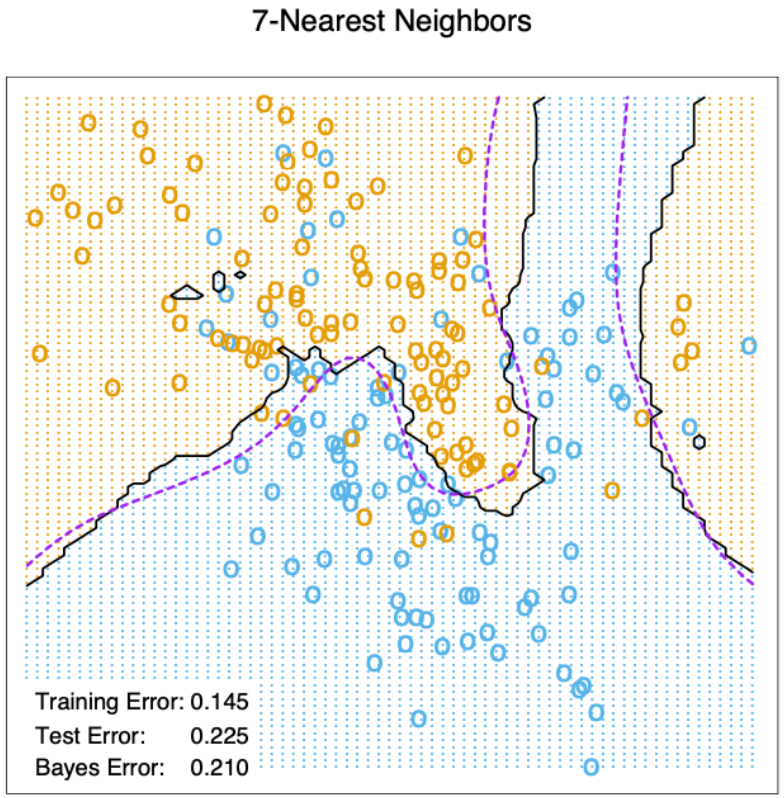


Frontera de decision de Bayes en púrpura

Error como función de K

UNIVERSIDAD DE MURCIA





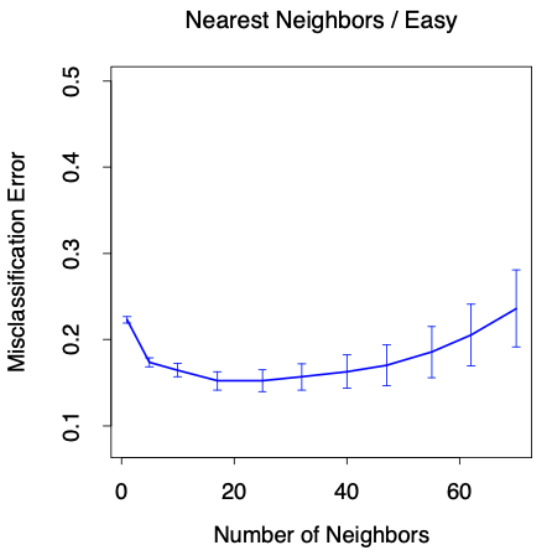
La importancia de la validación cruzada para la estimación de K

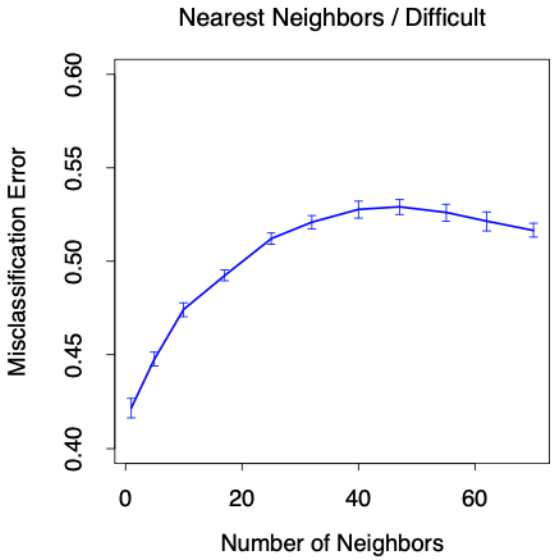
UNIVERSIDAD DE MURCIA

- ¿Cuál es el valor de K óptimo para estos dos problemas ejemplo (chap. 13 Hastie et al.)?

$$Y = I\left(X_1 > \frac{1}{2}\right); \quad \text{problem 1: "easy"},$$

$$Y = I\left(\text{sign}\left\{\prod_{j=1}^{3}\left(X_j - \frac{1}{2}\right)\right\} > 0\right); \quad \text{problem 2: "difficult."}$$





- 10 predictores independientes en ambos problemas, uniformemente distribuidos en [0,1]
- En 1, los dos casos son separables por el hiperplano X_1=1/2
- En 2, el hipercubo forma un patrón de tablero de ajedrez

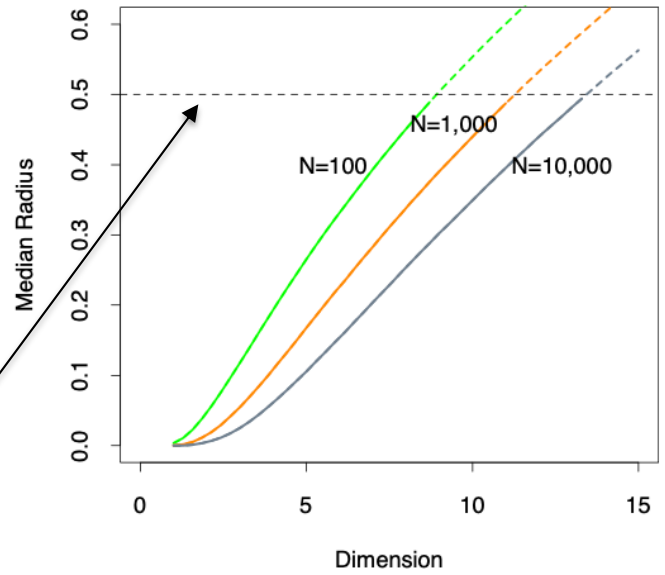
KNN en problemas de alta dimensionalidad

UNIVERSIDAD DE MURCIA

- Sean N puntos distribuidos en el espacio formado por un cubo unidad
  - Puntos confinados en $[-½, ½]^p$
  - Sea R la distancia de un único vecino más próximo al centro del cubo, ¿cómo varia esa distancia según p?

$$\text{median}(R) = v_p^{-1/p}\left(1 - \frac{1}{2}^{1/N}\right)^{1/p}$$

Cuando las dimensiones crecen, aunque sea un poco, los vecinos se distancian alcanzando rápidamente el borde del cubo



Referencias

UNIVERSIDAD DE MURCIA

- G. James, D. Witten, T. Hastie, y R. Tibshirani, An Introduction to Statistical Learning with Applications in R. Springer, 2013.
  - Chapter 2, pp. 39-42
- T. Hastie, R. Tibshirani, y J. Friedman, The elements of statistical learning. Springer, 2009.
  - Chapter 13, Sec. 13.3-13.5
- Transparencias basadas en el material de Manuel Mucientes.