Métodos No Lineales

Universidad Católica del Uruguay

K-vecinos más cercanos (K-nearest neighbors KNN)

Universidad Católica del Uruguay

- Clasificación o regresión
- La predicción de un nuevo ejemplo se define por los valores de los k ejemplos más cercanos en el dataset
- ¿Qué definición de “cercano” usar?
  - Ej: distancia euclídea

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \bullet\bullet\bullet + (x_n - y_n)^2}$$

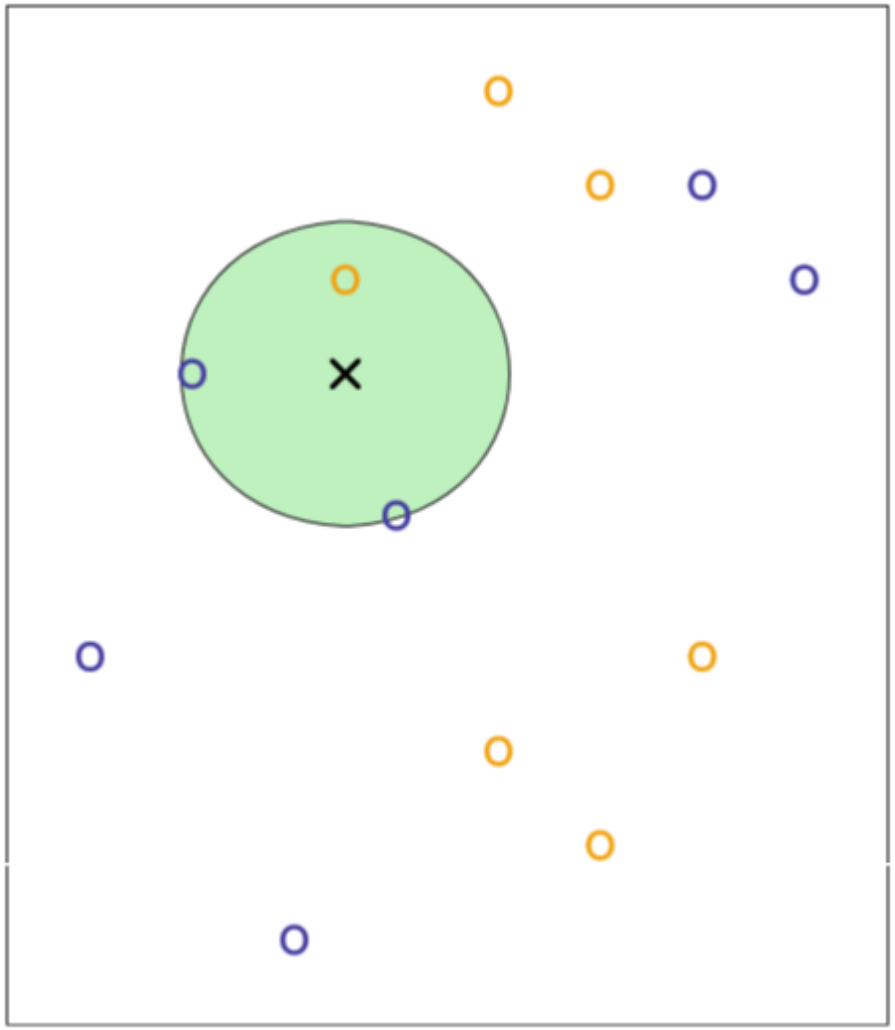
KNN

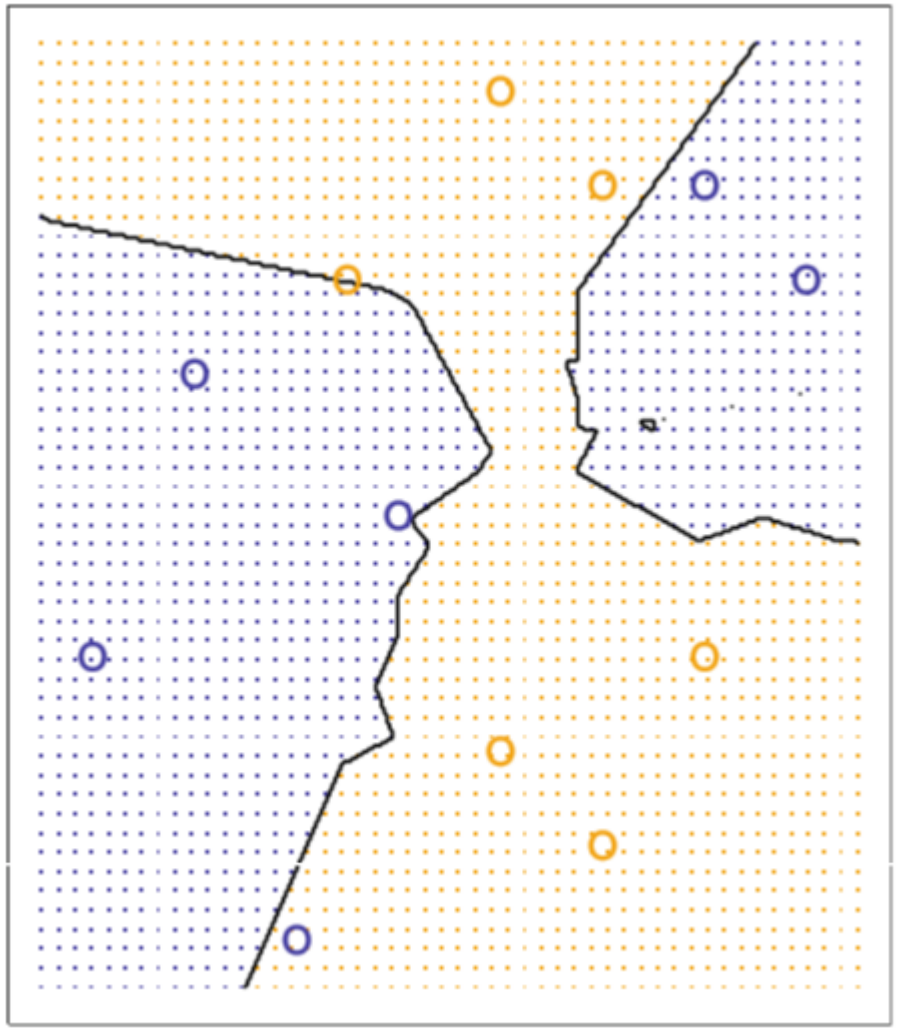
- Clasificación
  - se elige la clase más frecuente de los k vecinos
- Regresión
  - Media (o por ej. mediana) de los k vecinos
- Puede definirse un peso para cada vecino en función de su distancia:

$$w_i = \frac{e^{-d(x,n_i)}}{\sum_{i=1}^{k} e^{-d(x,n_i)}}$$

KNN

Ejemplo: KNN para clasificación con k=3



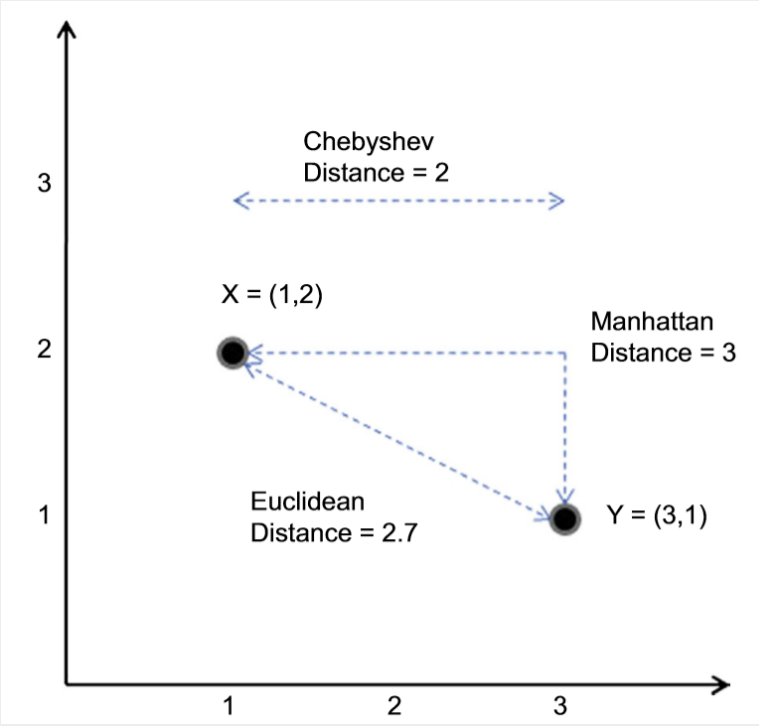


KNN

- Consideraciones
  - valores pequeños de k (ej k=1) tienden a "sobreajustar"
  - Variables independientes deben ser normalizadas (transformación Z)
  - ¿Cómo medir distancia entre variables categóricas? Si es la misma categoría la distancia es 1 / 0 si son diferentes
    - Categóricas ordinales pueden convertirse a números manteniendo las distancias

KNN

- Consideraciones
  - Otras medidas de distancia: Manhattan, Chebyshev
  - también existen para medir similitudes en textos: Jackard, Cosine simmilarity

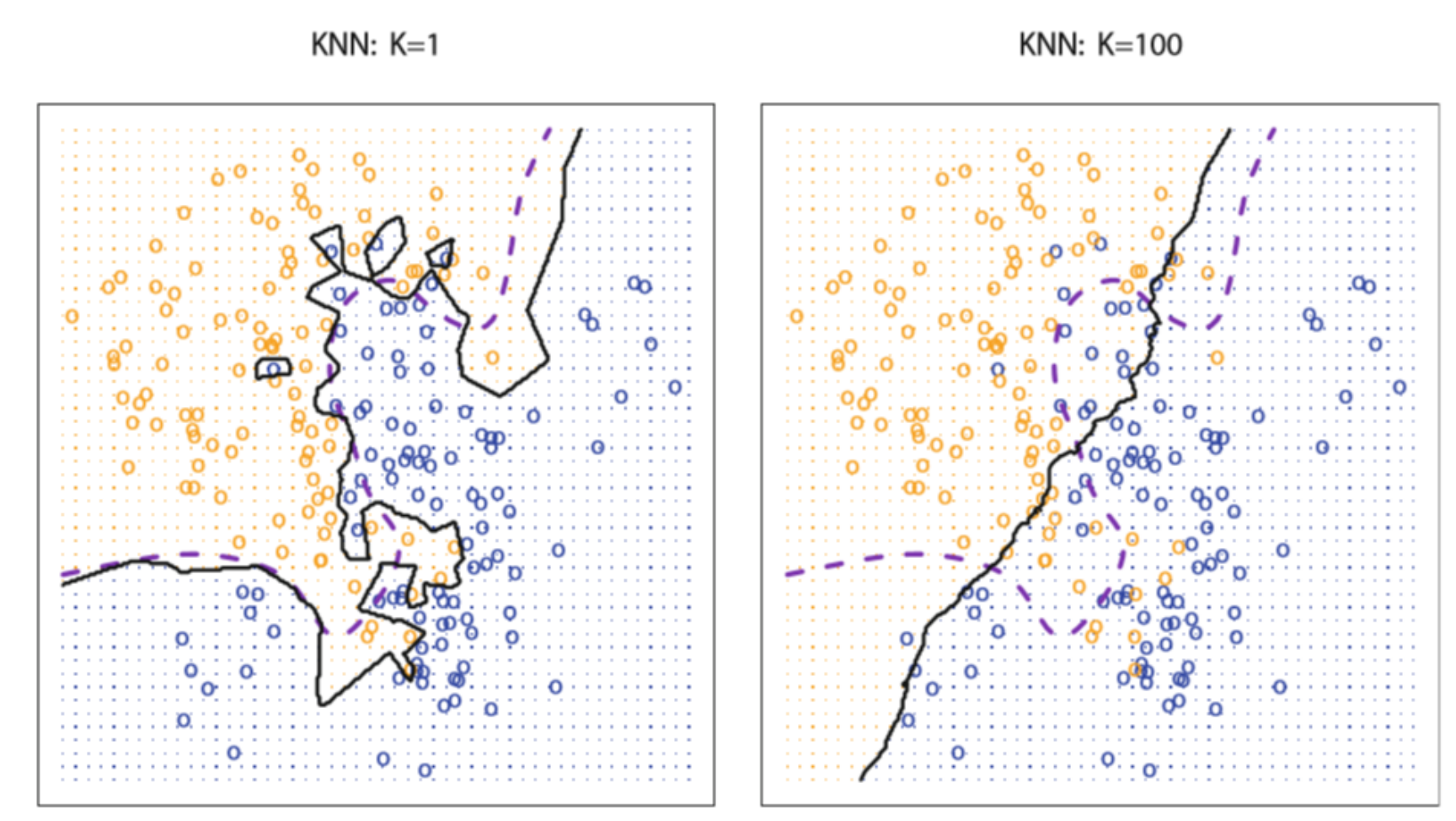


- Tanto el valor de k como la medida de distancia deben definirse con CV al momento de utilizar KNN para predicción

KNN

Universidad Católica del Uruguay

Flexibilidad de los modelos según k



KNN

Tanto k como la medida de distancia deben definirse mediante validación (Ej: CV) al momento de utilizar KNN para predicción