



ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

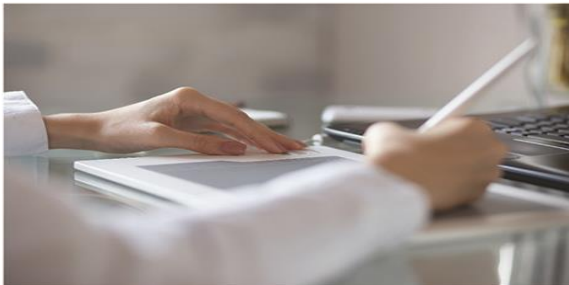
EDUCACIÓN
PROFESIONAL

Diplomado en Big Data y Ciencias de Datos

Minería de Datos Clasificación

Educación Profesional - Escuela de Ingeniería UC

Sebastián Raveau



Clasificación

Interesa definir un sistema capaz de identificar automáticamente la clase a la cual pertenece cada objeto u observación de interés

Las técnicas de clasificación son métodos de aprendizaje supervisado

Requiere de:

- un conjunto de datos de entrenamiento previamente clasificados

- un algoritmo de clasificación (e.g. KNN)

Ejemplo



gato



gato



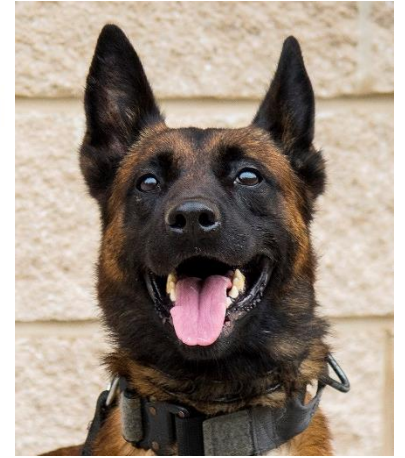
no gato



no gato



gato



no gato



¿Es un gato?

Tiene orejas puntiagudas y bigotes

No tiene alas ni plumas

Luce desconfiado y malvado

¡Es un gato!

Clasificación de imágenes

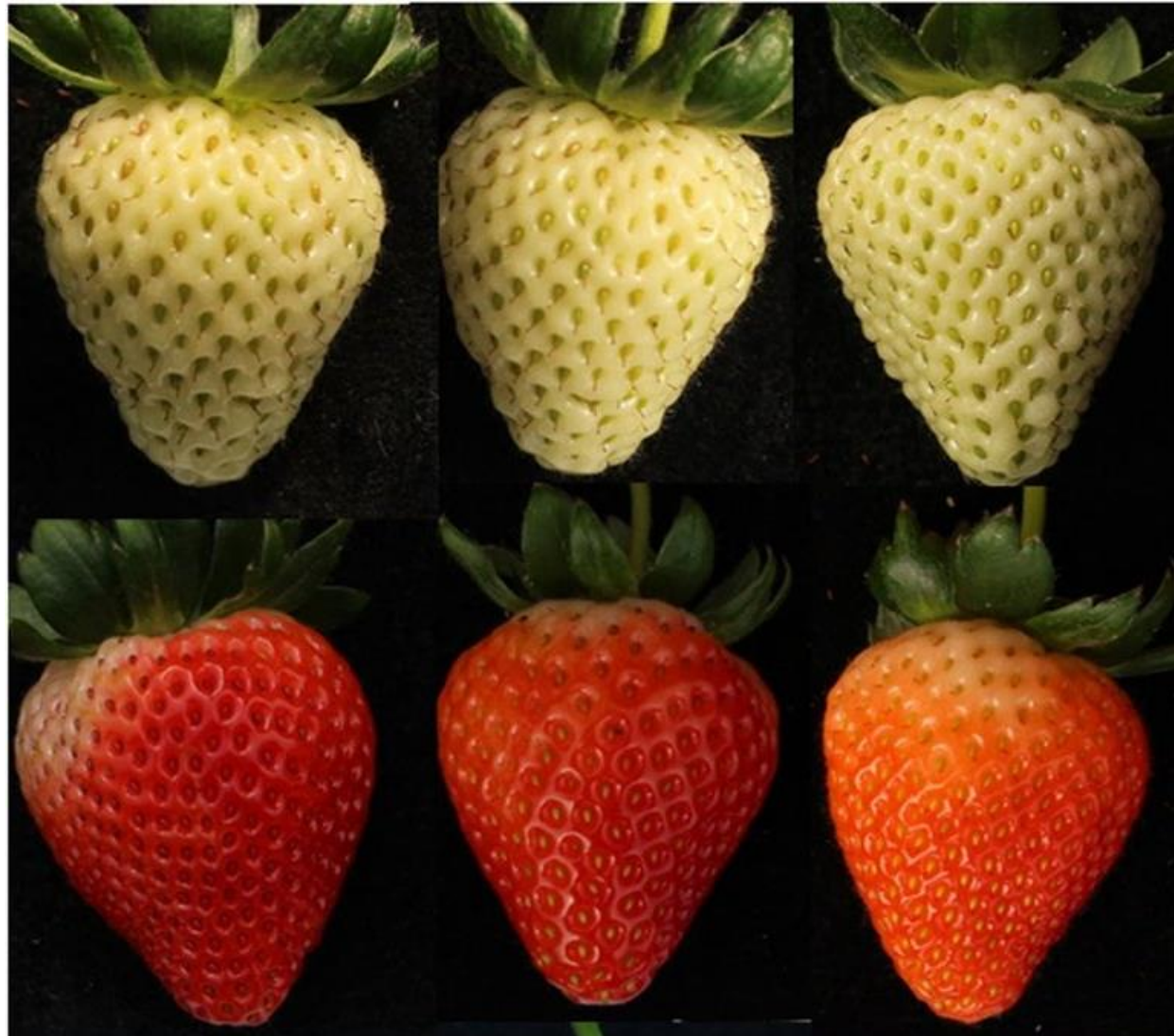


¿chihuahua o muffin?



¿perro ovejero o trapeador?

Clasificación automática



Tests médicos



Reconocimiento de patrones



Reconocimiento facial



K vecinos más cercanos
(K-nearest neighbours, KNN)

KNN

Este algoritmo se basa en buscar los datos más “similares” para clasificar

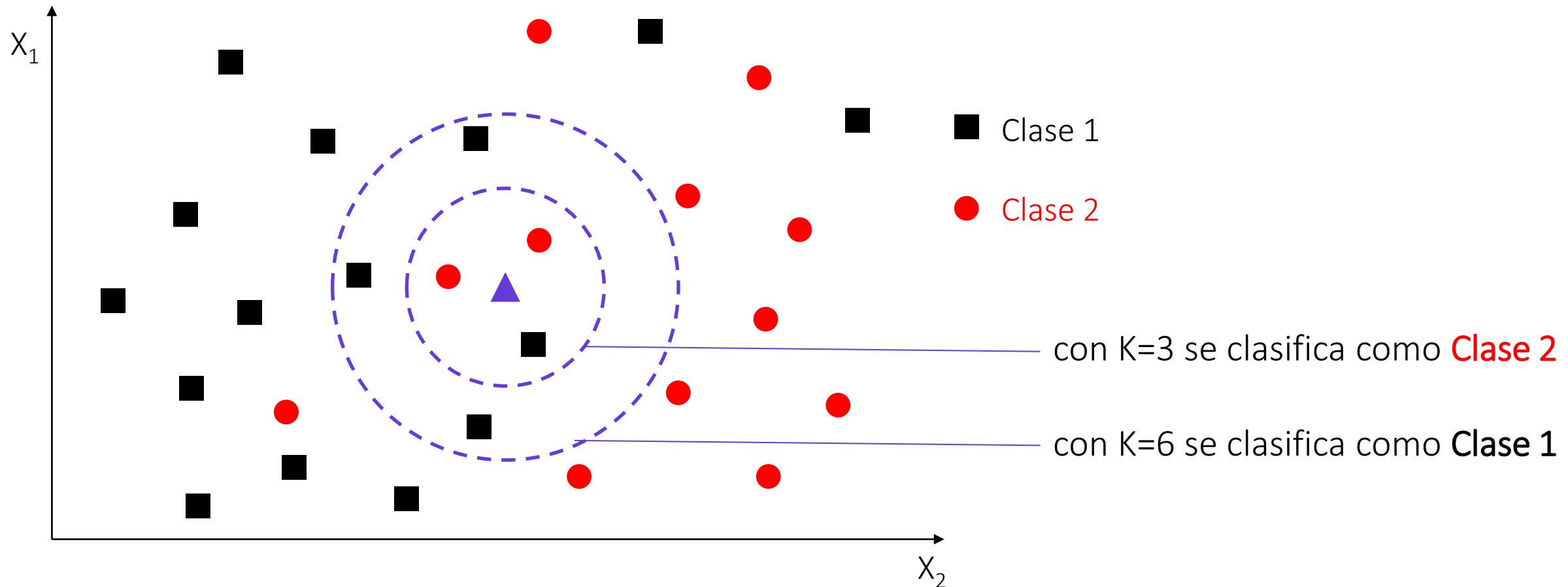
Un dato “similar” es aquel que se encuentra a poca distancia (i.e. un vecino)

Es necesario definir una métrica de distancia y también la cantidad de vecinos a considerar (i.e. K)

Cada vecino vota para clasificar y se elige por mayoría simple

KNN

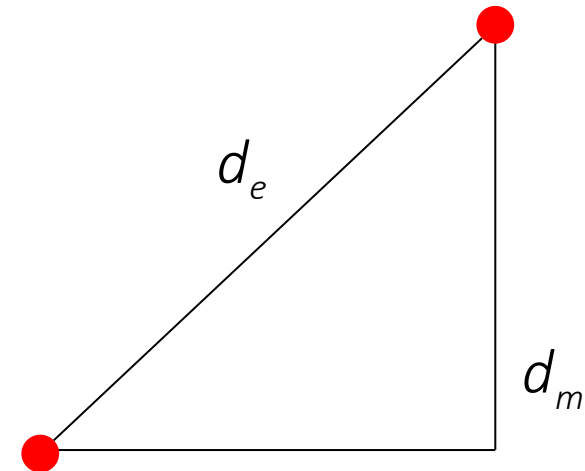
Supongamos que queremos clasificar entre dos clases, a partir de dos atributos



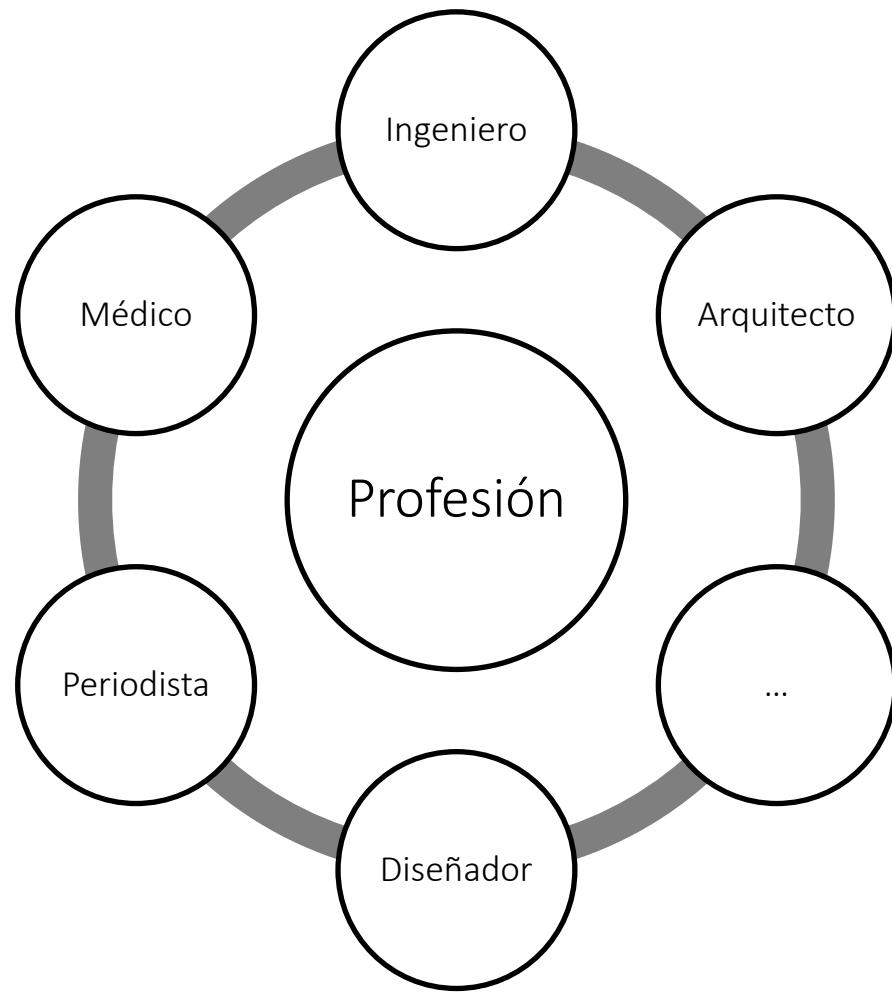
¿Cómo calcular la distancia entre dos observaciones?

Distancia Euclidiana $d_e(X, Y) = \sqrt{\sum_i (x_i - y_i)^2}$

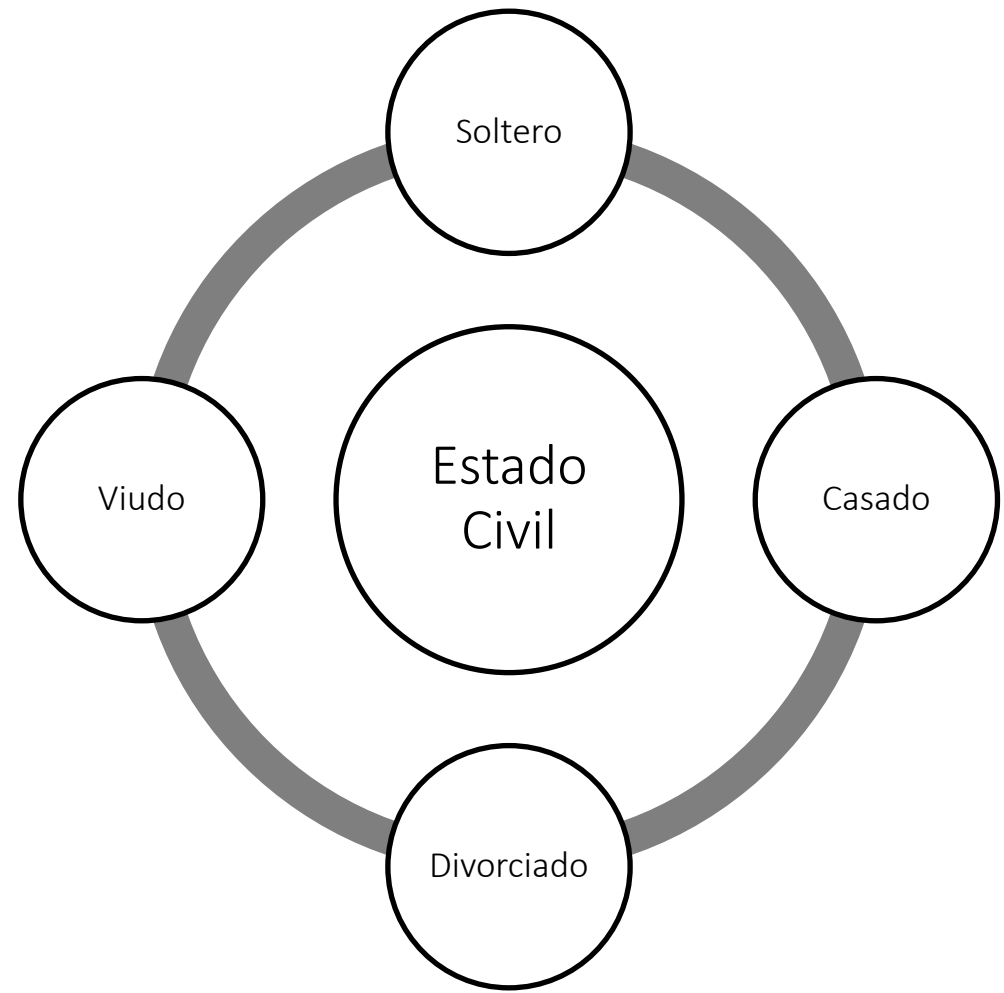
Distancia Manhattan $d_m(X, Y) = \sum_i |x_i - y_i|$



¿Qué hacemos con los atributos cualitativos o categóricos?



100 categorías



4 categorías

¿Qué hacemos con los atributos cualitativos?

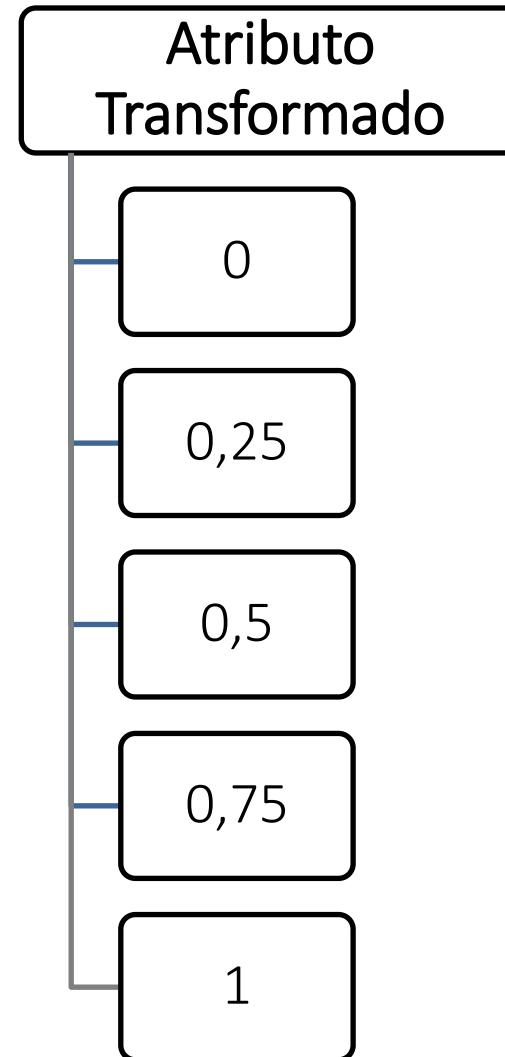
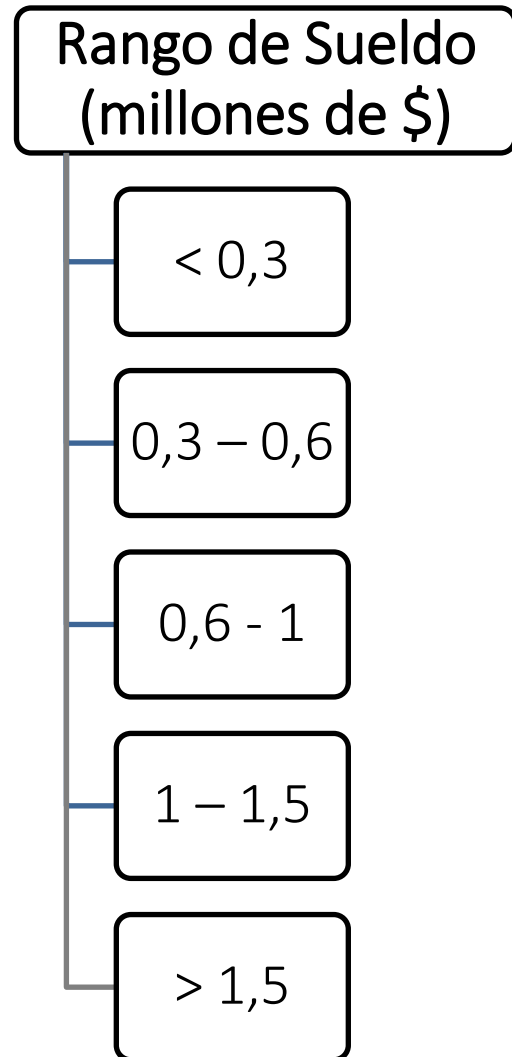
Opción 1 – Transformarlos en N variables binarias

$$d(\text{Arquitecto}, \text{Médico}) = 1 \qquad d(\text{Soltero}, \text{Viudo}) = 1$$

Opción 2 – Considerar la cantidad de posibles categorías

$$d(\text{Arquitecto}, \text{Médico}) = \frac{1}{100} \qquad d(\text{Soltero}, \text{Viudo}) = \frac{1}{4}$$

¿Qué hacemos con los atributos ordinales?



¿Qué hacemos con atributos en distintas escalas?

Normalizamos todos los atributos:

$$z_{ik} = \frac{x_{ik} - \min(X_k)}{\max(X_k) - \min(X_k)}$$

Todas las variables z_{ik} tendrán valores entre 0 y 1

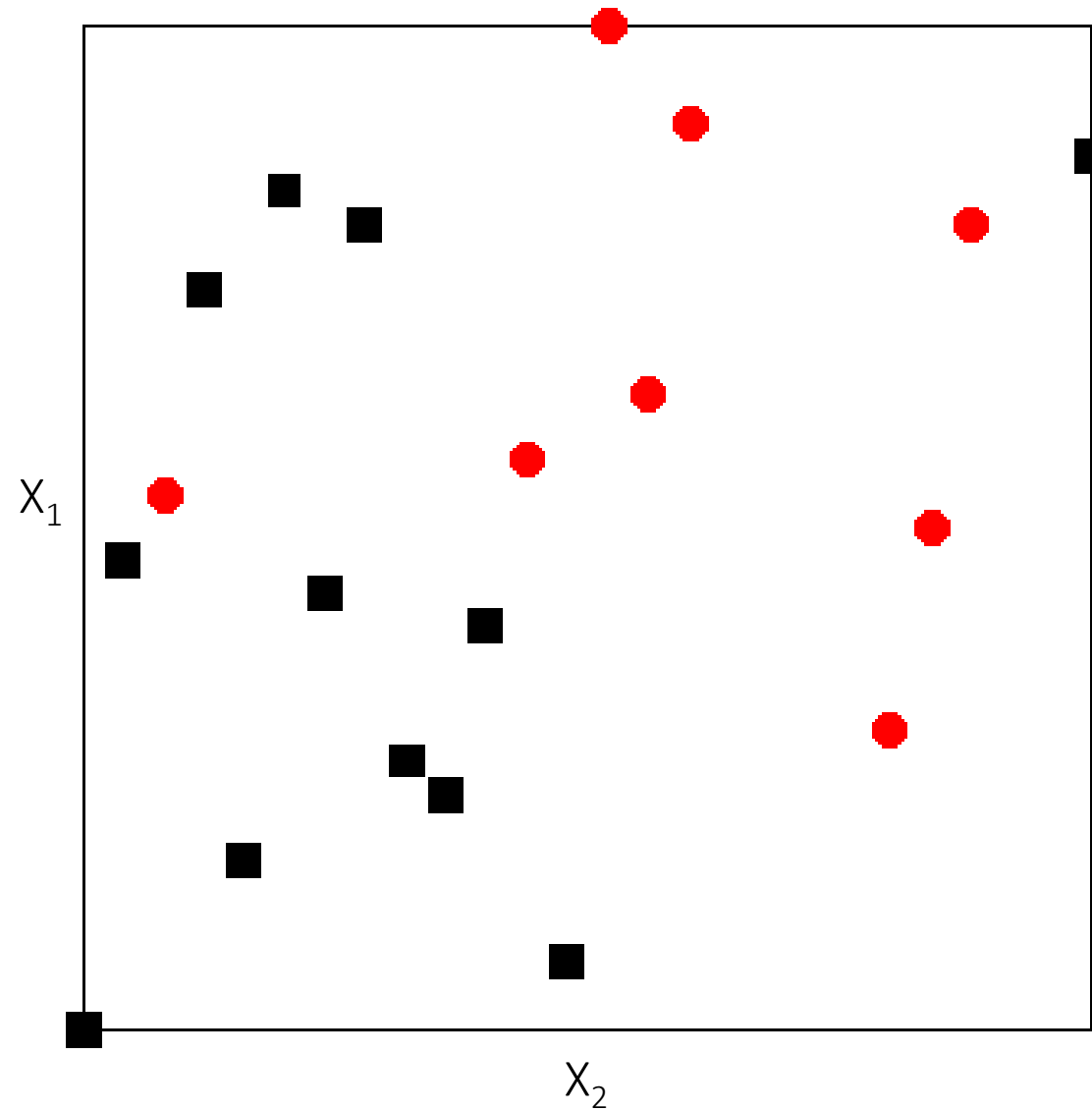
¿Qué hacemos si hay empate en los votos?

Opción 1 – Elegir un K distinto

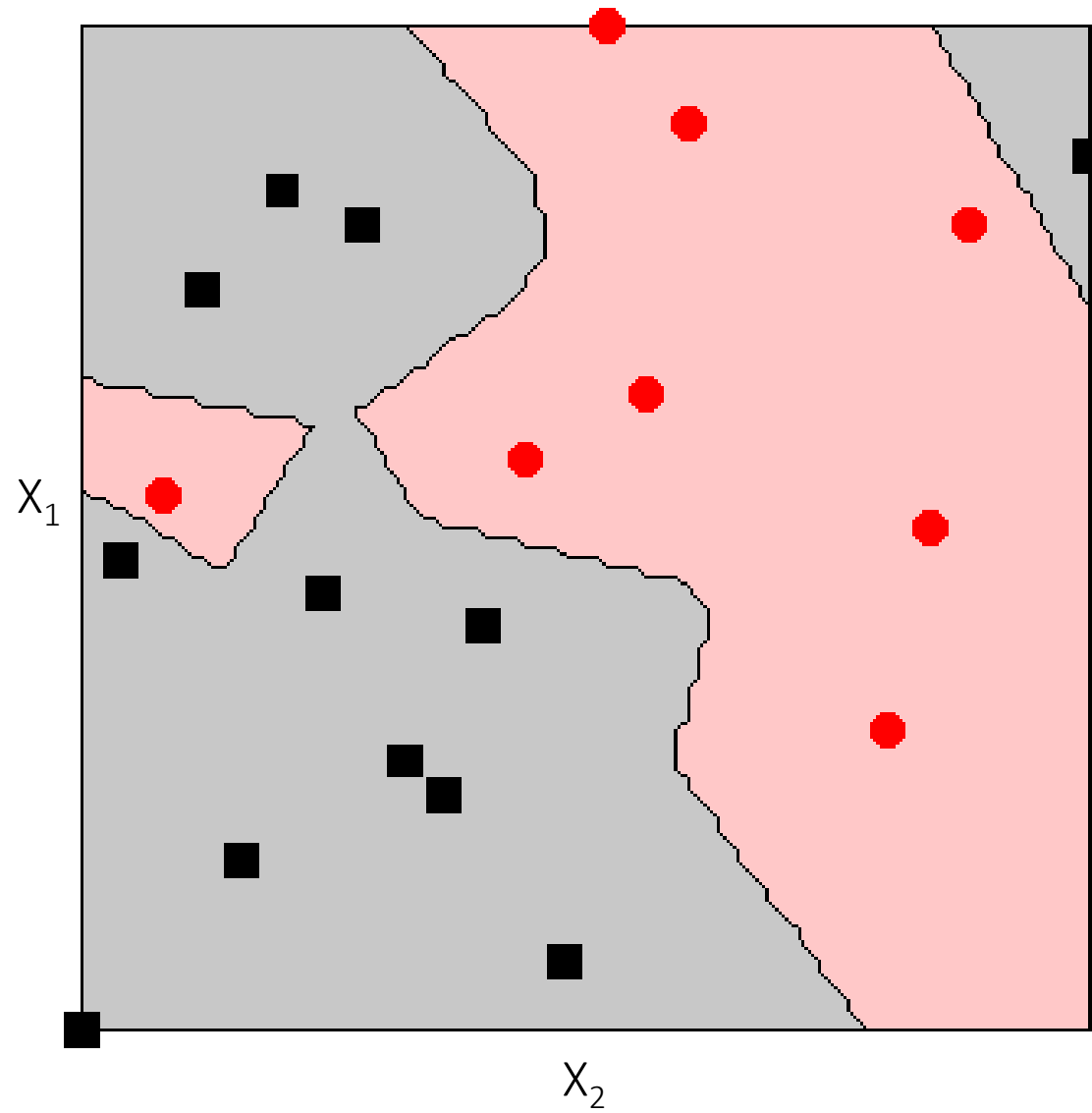
Opción 2 – Clasificar al azar entre las clases empatadas

Opción 3 – Ponderar los votos según la distancia

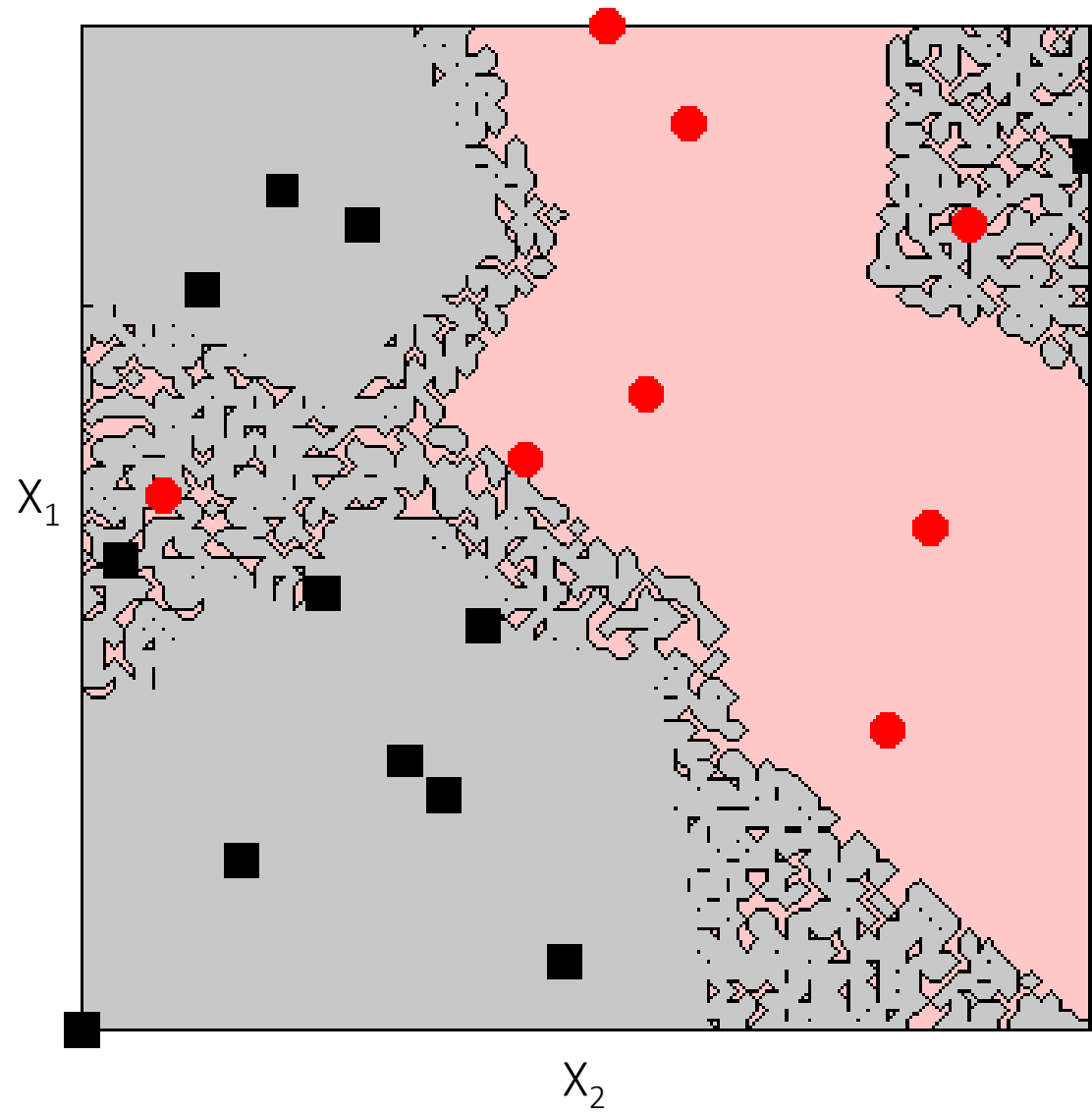
Consideremos los siguientes datos



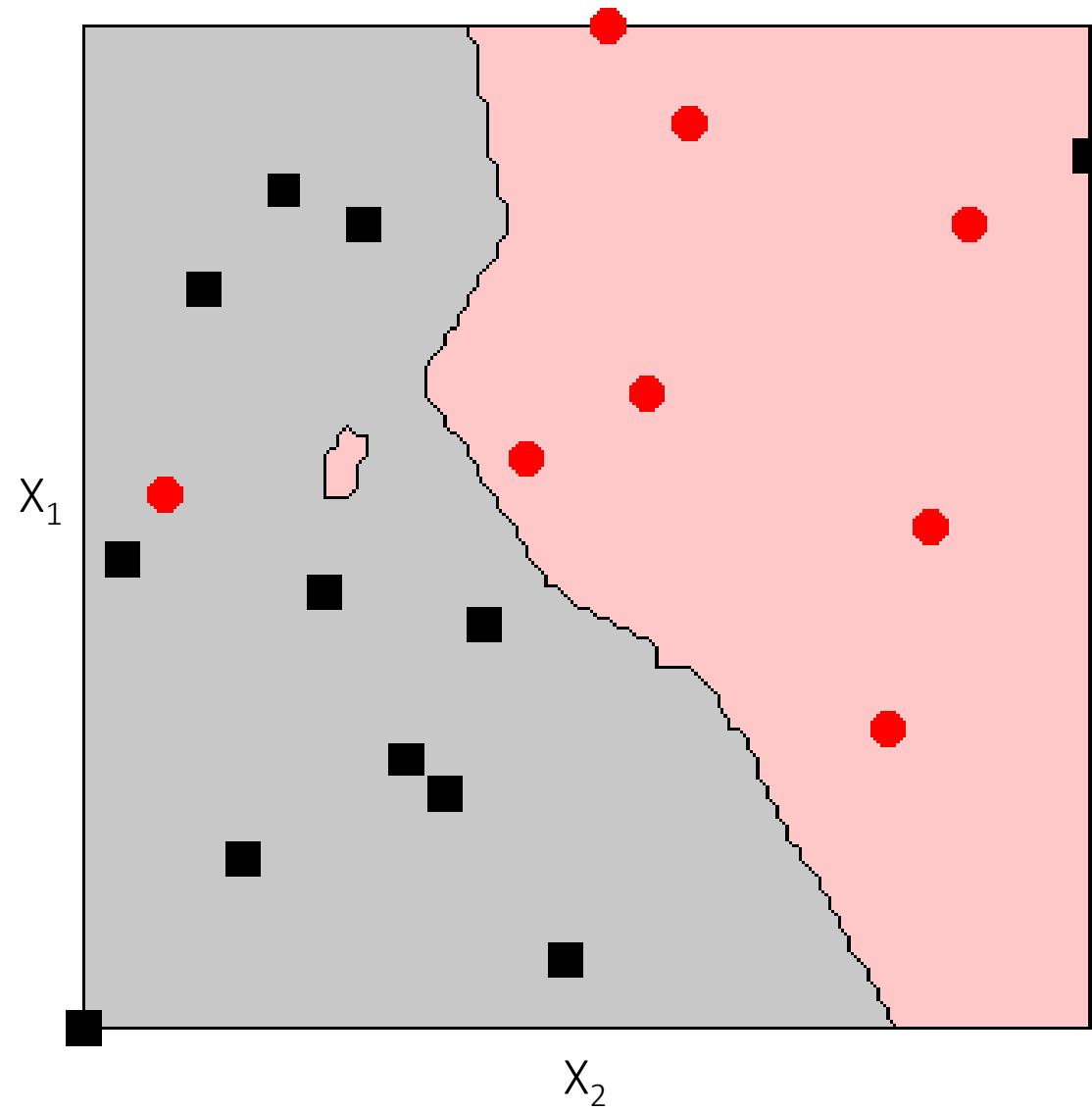
$$K = 1$$



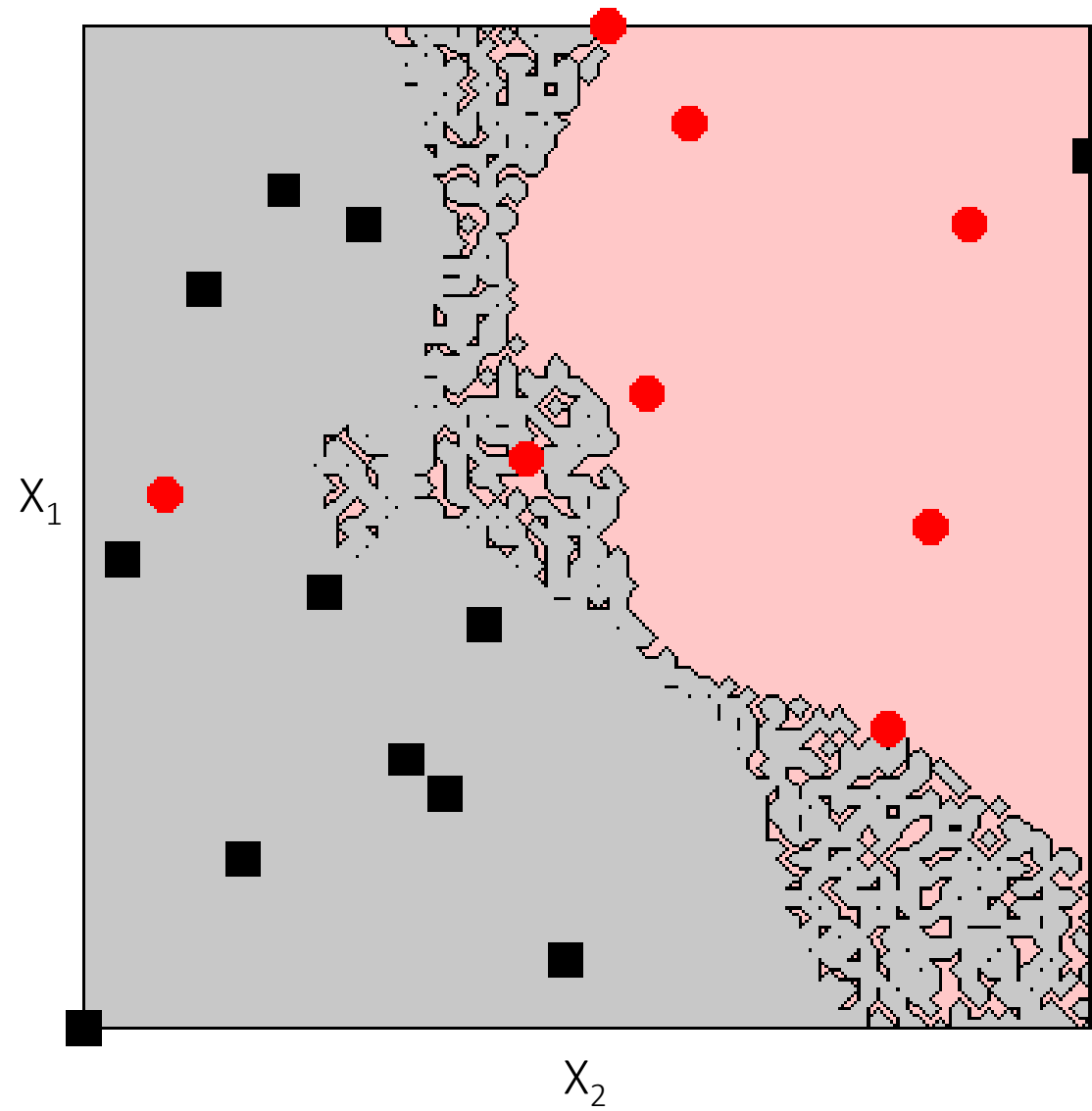
$K = 2$



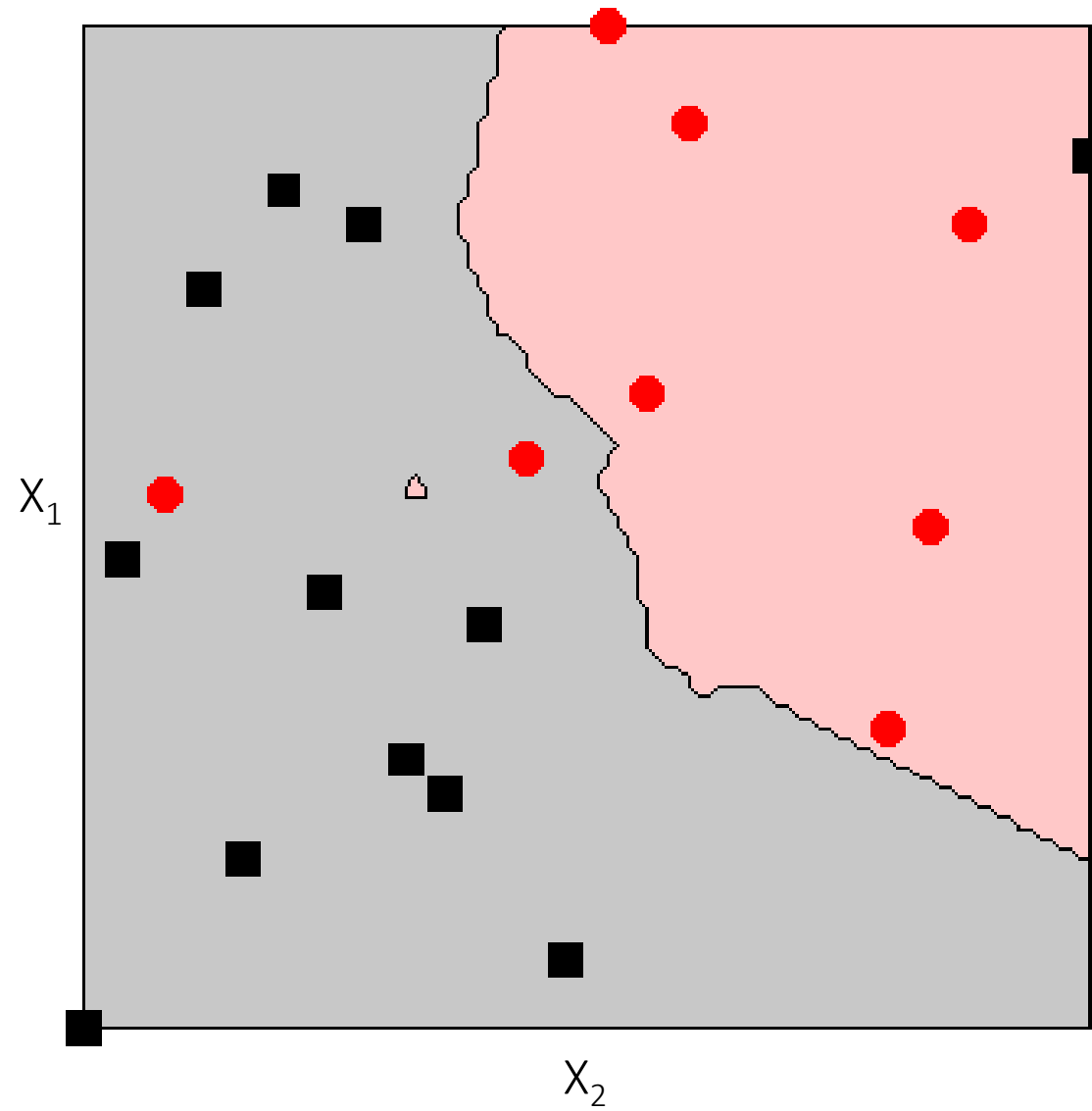
$K = 3$



$K = 4$



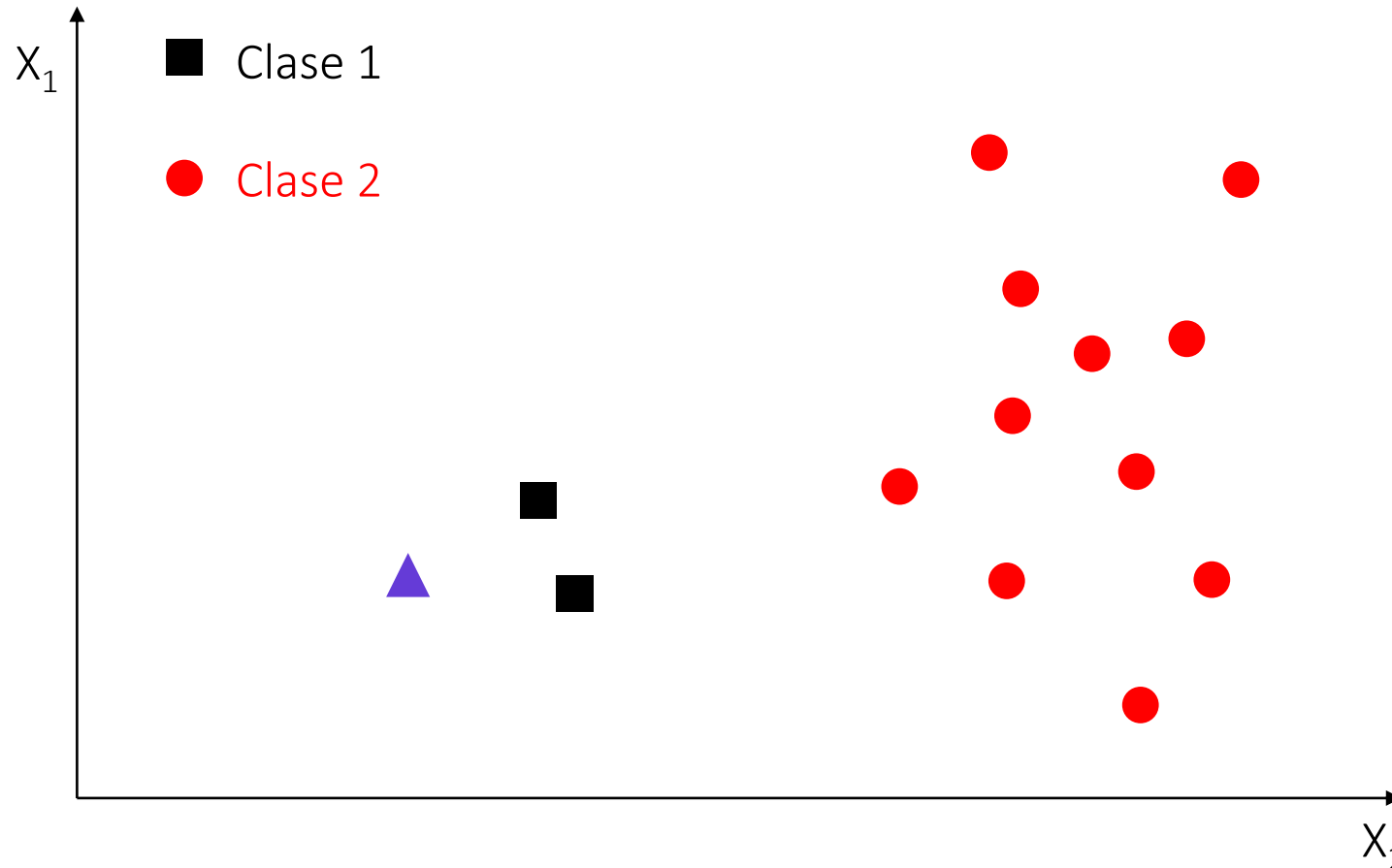
$K = 5$



Variantes de KNN

Resultados poco intuitivos

Supongamos la siguiente situación



¿Cómo podemos mejorar el algoritmo?

Podemos aplicar un peso a cada voto, en función de la distancia

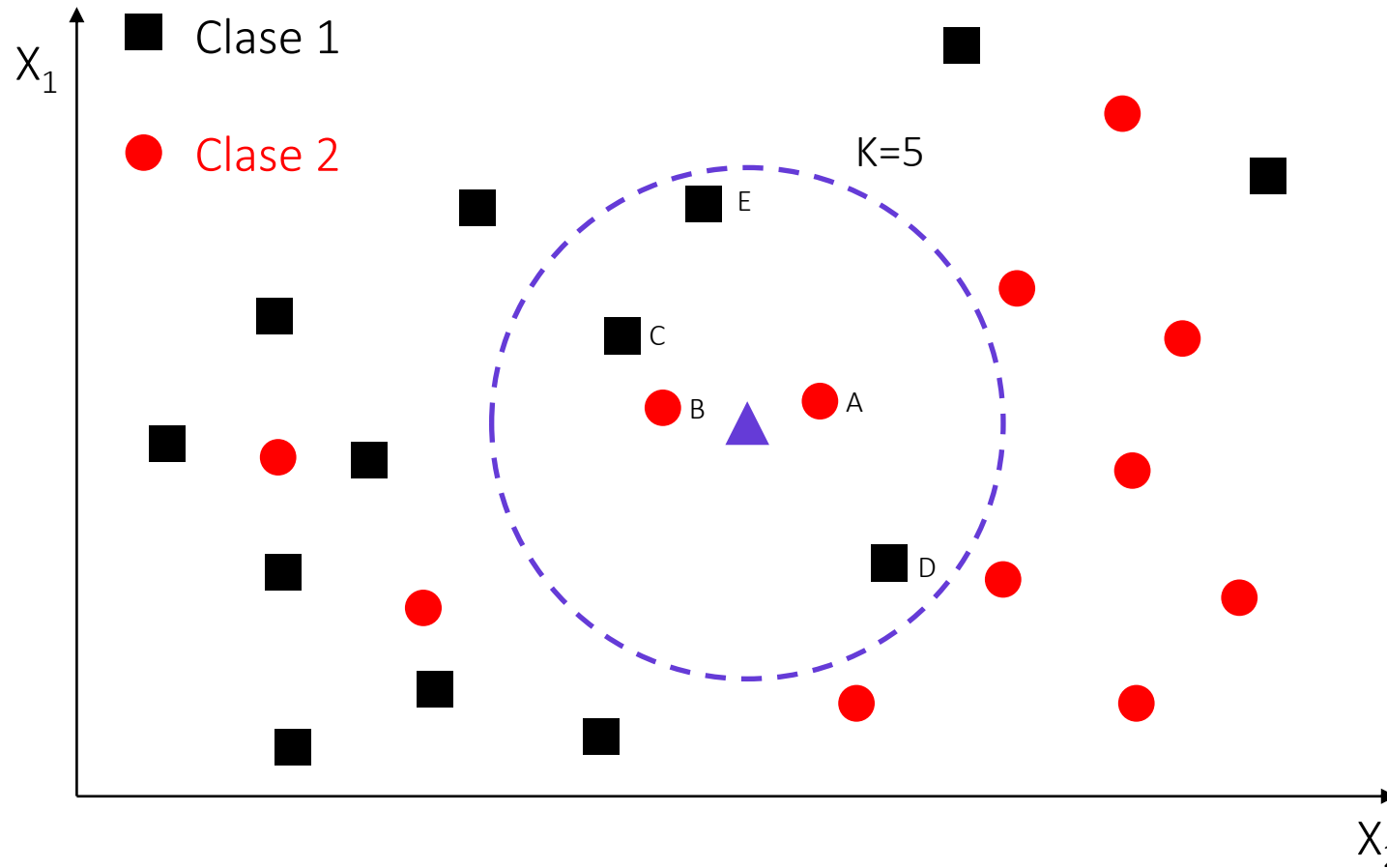
Los pesos deben inversamente proporcionales a la distancia, de modo que el voto de observaciones más lejanas importa menos en la clasificación

Opciones de peso w del voto de Y para clasificar X :

$$w_{xy} = \frac{1}{d(X,Y)}$$

$$w_{xy} = \exp\left(-\frac{d(X,Y)^2}{2}\right)$$

KNN ponderado



Punto	Distancia	Peso ($1/d$)
A	0,3	3,33
B	0,4	2,50
C	0,9	1,11
D	1,2	0,83
E	1,5	0,67

La **Clase 2** tiene 5,83 votos ponderados, mientras que la **Clase 1** tiene 2,61 votos ponderados

Comentarios finales

Ventajas de KNN

Método simple e intuitivo

Se puede utilizar para clasificar todo tipo de datos

Buen método de clasificación si se cuenta con hartos datos

Desventajas de KNN

Seleccionar K no es trivial

El método radica por completo en la etapa de entrenamiento

La etapa de validación es computacionalmente más demandante que la etapa de entrenamiento (contrario a lo habitual)

Clasificación supervisada

Existen distintos enfoques para responder la misma pregunta

En este curso veremos tres:

Árboles de clasificación

KNN

Regresión logística