ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN
PROFESIONAL

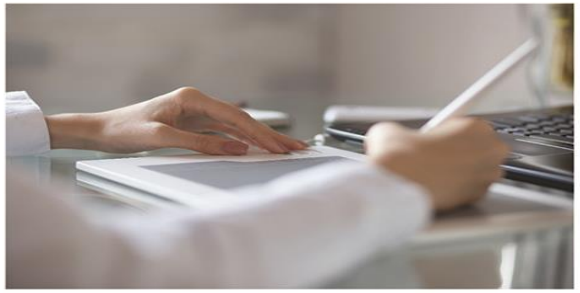
Diplomado en Big Data y Ciencias de Datos

Minería de Datos

Clasificación

Educación Profesional - Escuela de Ingeniería UC

Sebastián Raveau




Recordemos las técnicas de Minería de Datos

Técnicas Predictivas – Aprendizaje Supervisado

Regresión	ajustar variables/relaciones continuas
Clasificación	ajustar variables/relaciones discretas

Técnicas Descriptivas – Aprendizaje No Supervisado

Clustering	agrupar datos similares
Asociación	identificar patrones y coocurrencias

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)

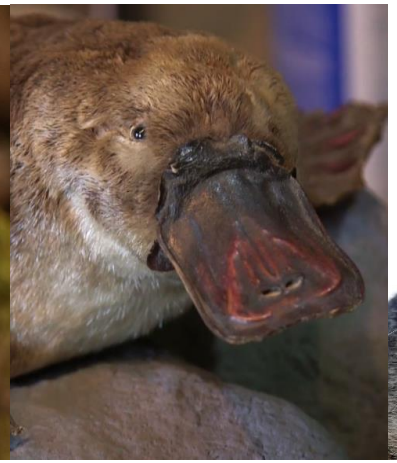
K vecinos más cercanos
(K-nearest neighbours, KNN)

Cercanía = similitud

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!

KNN

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

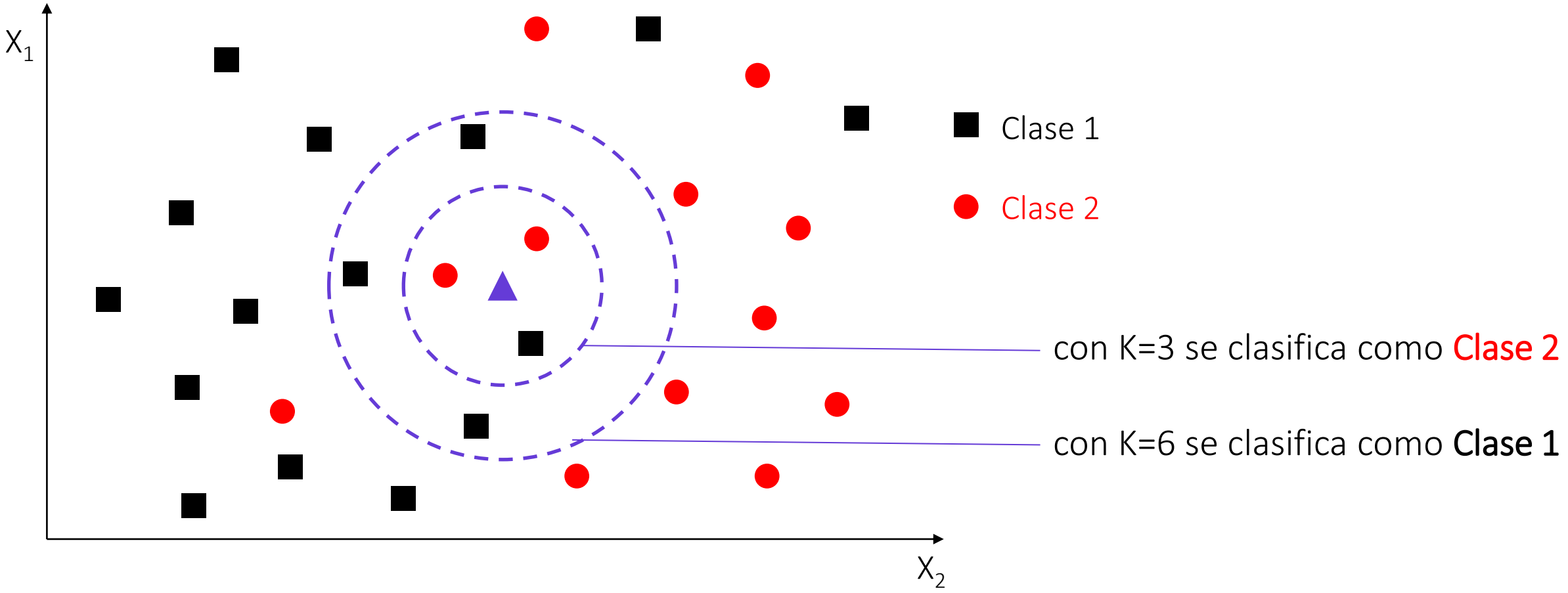
Un dato “similar” es aquel que se encuentra a poca distancia (i.e. un vecino)

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

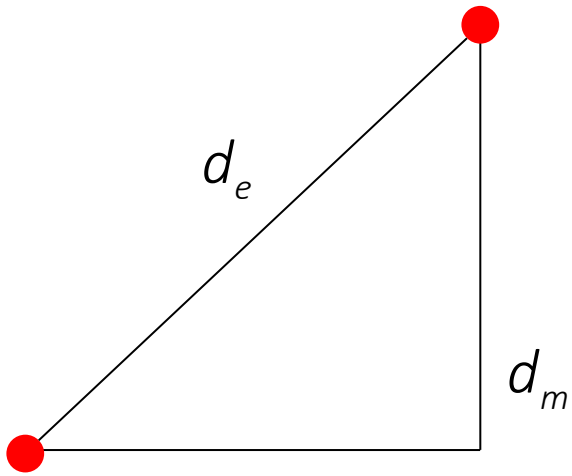
Supongamos que queremos clasificar entre dos clases, a partir de dos atributos

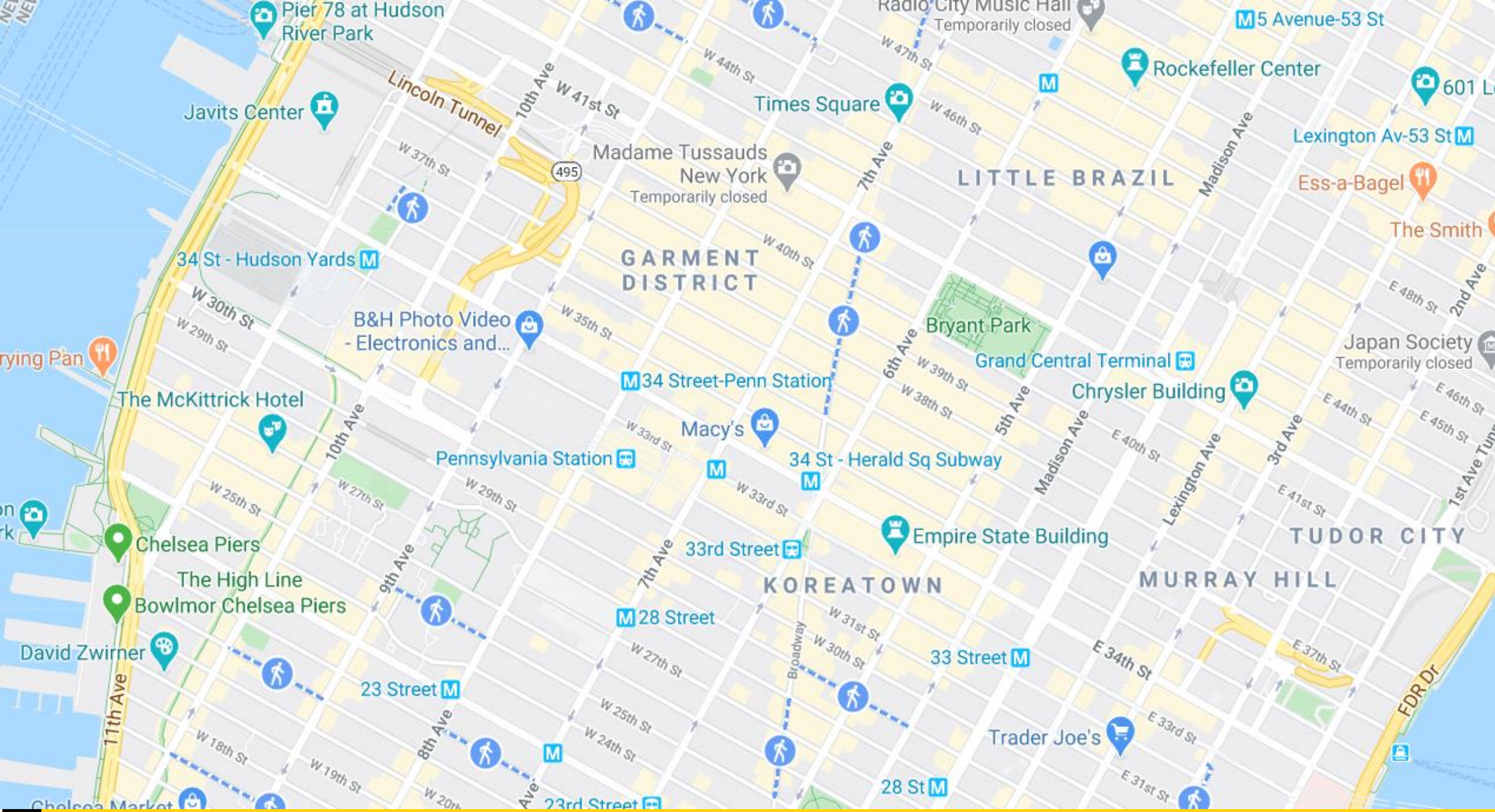


¿Cómo calcular la distancia entre dos observaciones?

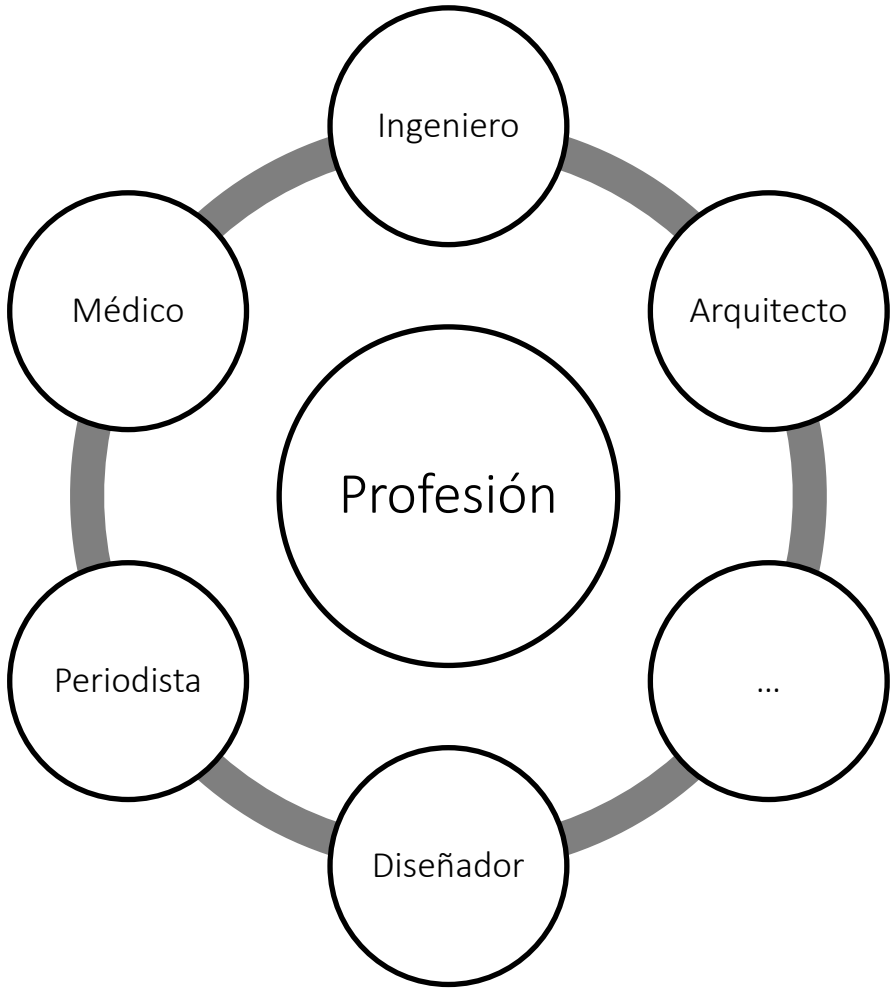
Distancia Euclidiana $$d_e(X,Y) = \sqrt{\sum_i (x_i - y_i)^2}$$

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

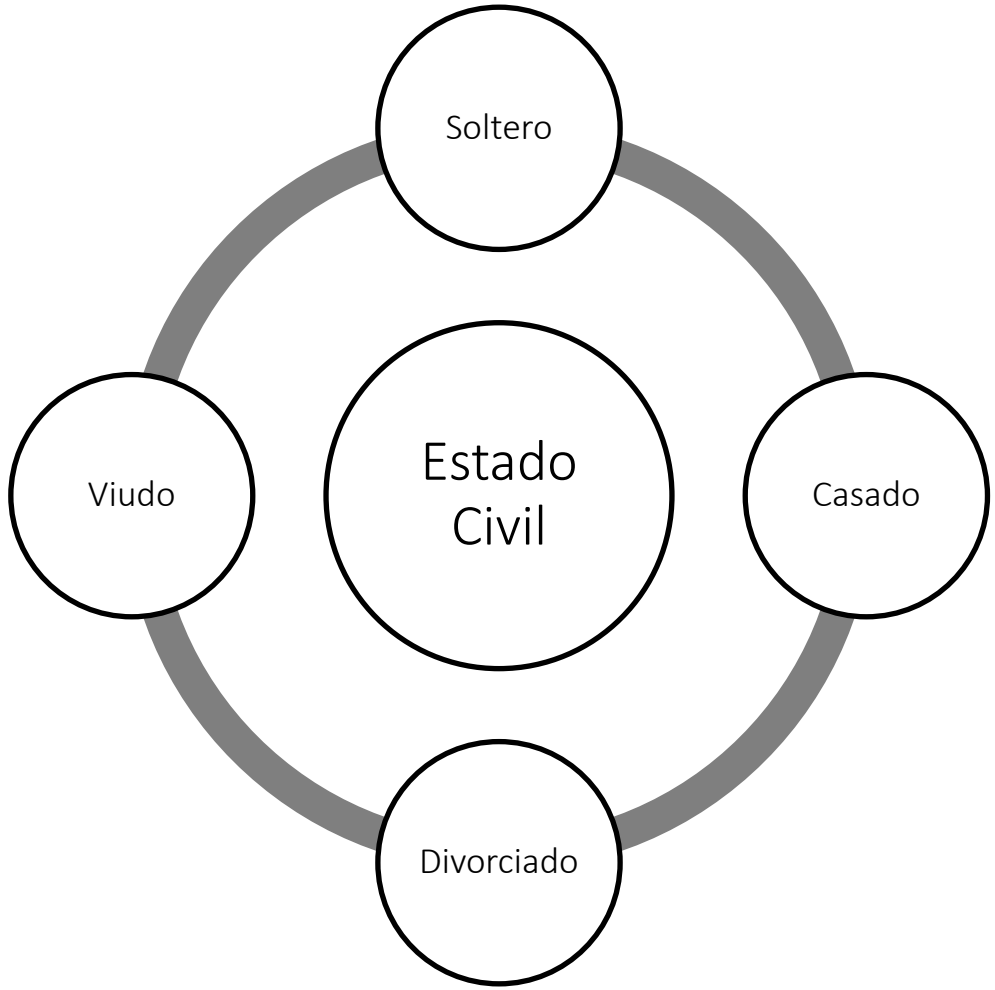



Pier 78 at Hudson River Park
Lincoln Tunnel
Javits Center
10th Ave
W 41st St
W 44th St
W 47th St
Radio City Music Hall
Temporarily closed
5 Avenue-53 St
Rockefeller Center
Times Square
W 46th St
601 L
Lexington Av-53 St
Madame Tussauds New York
Temporarily closed
495
W 37th St
7th Ave
LITTLE BRAZIL
Madison Ave
Ess-a-Bagel
The Smith
34 St - Hudson Yards
GARMENT DISTRICT
W 40th St
W 30th St
W 29th St
W 35th St
B&H Photo Video - Electronics and...
Bryant Park
Grand Central Terminal
Japan Society
Temporarily closed
E 48th St
2nd Ave
6th Ave
W 39th St
W 38th St
5th Ave
Chrysler Building
E 46th St
E 45th St
E 44th St
34 Street-Penn Station
The McKittrick Hotel
Macy's
Pennsylvania Station
W 33rd St
10th Ave
W 27th St
W 29th St
34 St - Herald Sq Subway
Madison Ave
E 40th St
Lexington Ave
3rd Ave
1st Ave Tunn
E 41st St
W 25th St
Chelsea Piers
Empire State Building
TUDOR CITY
33rd Street
7th Ave
9th Ave
The High Line
Bowlmor Chelsea Piers
KOREATOWN
MURRAY HILL
28 Street
W 31st St
W 30th St
Broadway
E 34th St
E 37th St
David Zwirner
33 Street
23 Street
W 27th St
W 25th St
W 24th St
8th Ave
11th Ave
W 18th St
W 19th St
Trader Joe's
E 33rd St
FDR Dr
E 31st St
28 St
23rd Street

¿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

Variables de profesión toman valores 1 o 0

Variables de estado civil toman valores 1 o 0

$$d(\text{Arquitecto},\text{Médico}) = \sqrt{(0-0)^2 + \ldots + (1-0)^2 + (0-1)^2 + \ldots + (0-0)^2}$$
$$= 1.414$$

$$d(\text{Soltero},\text{Viudo}) = \sqrt{(0-0)^2 + (1-0)^2 + (0-1)^2 + (0-0)^2}$$
$$= 1.414$$

¿Qué hacemos con los atributos cualitativos?

Opción 2 – Considerar la cantidad de posibles categorías

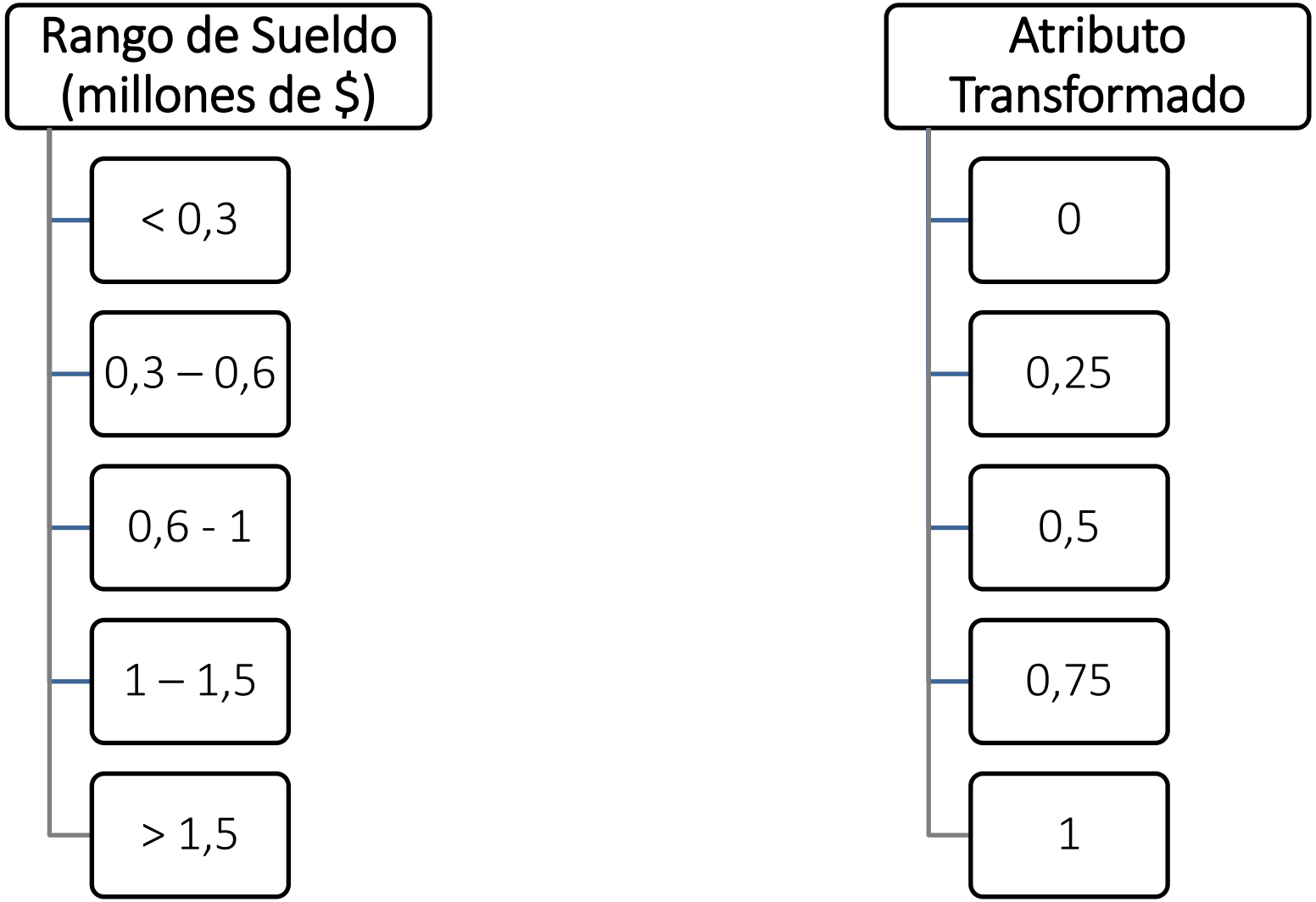
Variables de profesión toman valores 1/100 o 0

Variables de estado civil toman valores 1/4 o 0

$$d(\text{Arquitecto},\text{Médico}) = \sqrt{(0-0)^2 + \ldots + (1/100-0)^2 + (0-1/100)^2 + \ldots + (0-0)^2}$$
$$= 0.0141$$

$$d(\text{Soltero},\text{Viudo}) = \sqrt{(0-0)^2 + (1/4-0)^2 + (0-1/4)^2 + (0-0)^2}$$
$$= 0.353$$

¿Qué hacemos con los atributos ordinales?



¿Qué hacemos con atributos en distintas escalas?

Normalizamos o estandarizamos 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

$$z_{ik} = \frac{x_{ik} - \bar{X}_k}{s_k}$$

Todas las variables z_{ik} tendrán media 0 y desviación estándar 1

Selección de variables

Todas las variables que incluyamos en el cálculo de la distancia serán consideradas por el algoritmo para realizar la clasificación

A diferencia de otros algoritmos (como árboles de decisión o modelos de regresión), KNN no identifica cuáles variables son relevantes para ser utilizadas

Podemos basarnos en criterios teóricos y métricas de rendimiento para seleccionar las variables a incluir en el algoritmo

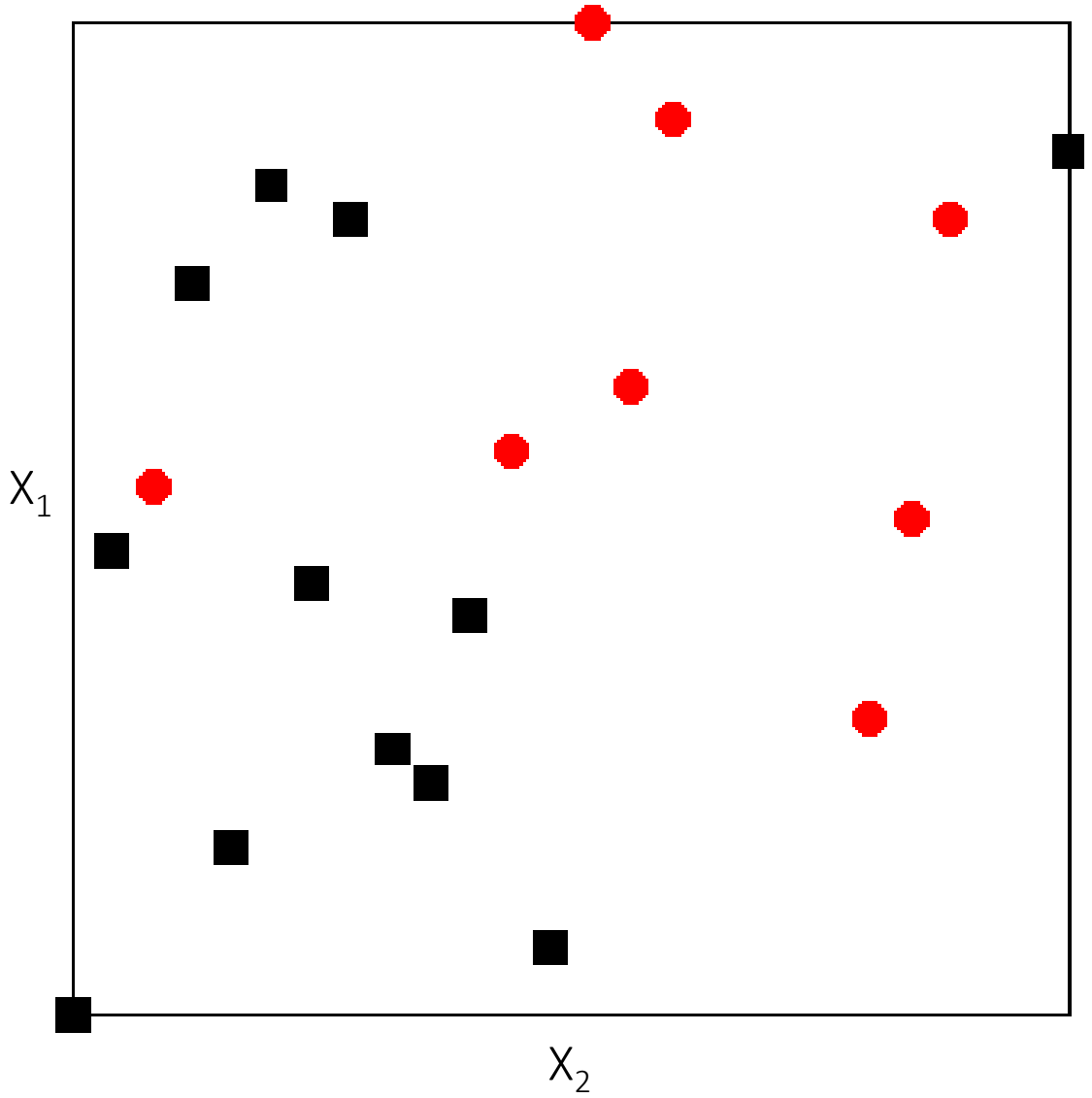
¿Qué hacemos si hay empate en los votos?

Opción 1 – Elegir un K distinto

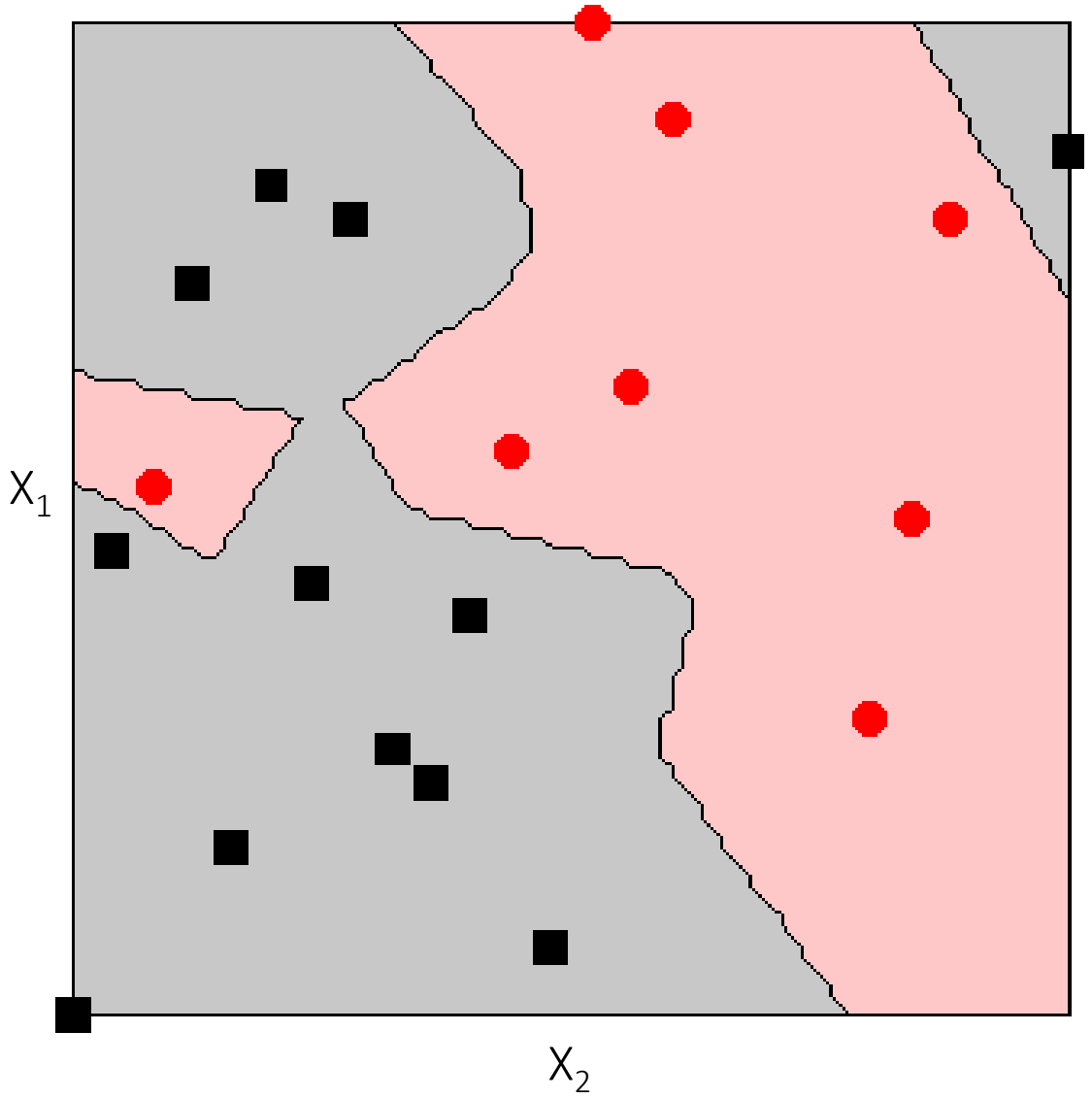
Opción 2 – Clasificar al azar entre las clases empatadas

Opción 3 – Ponderar los votos según la distancia

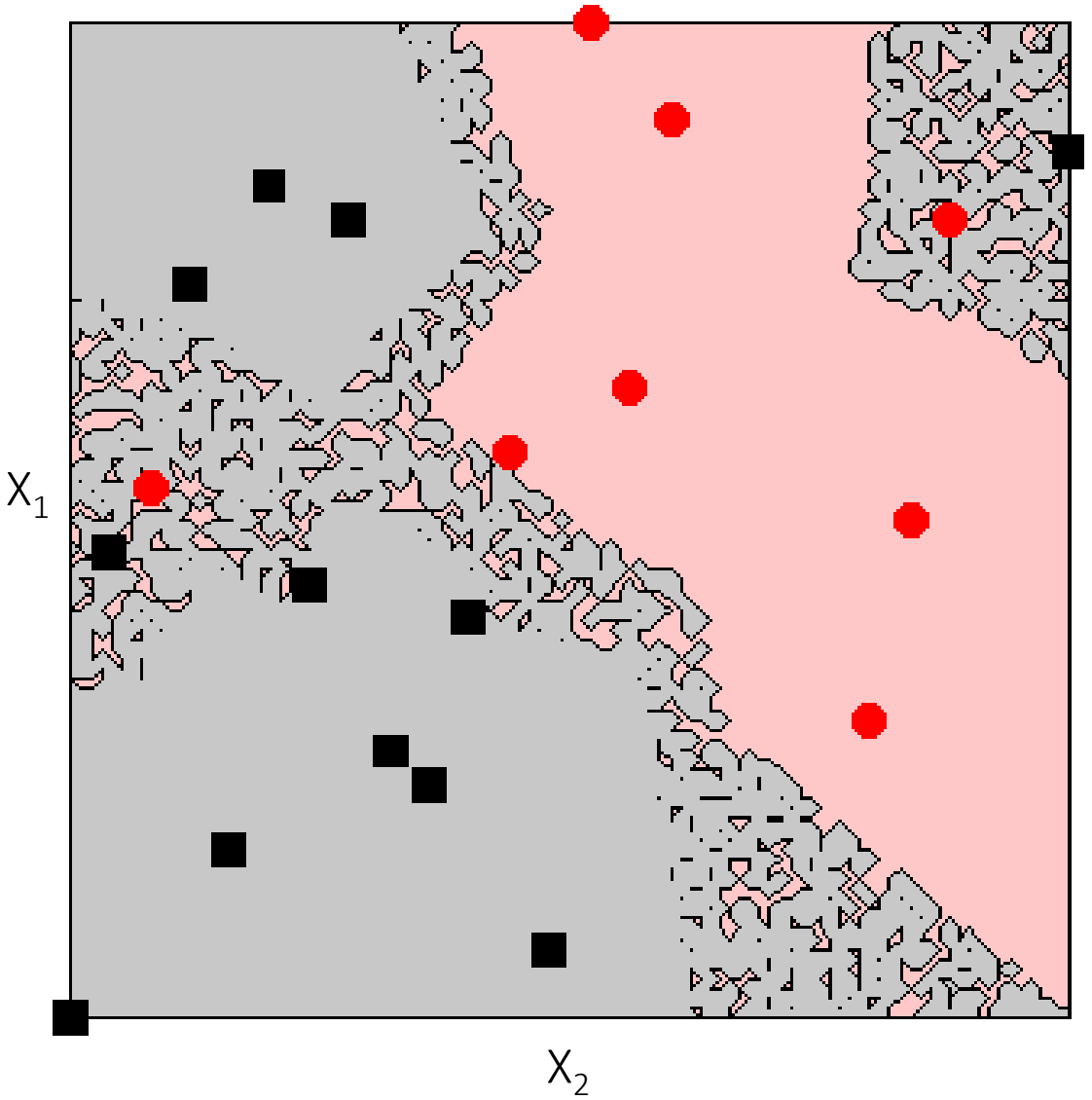
Consideremos los siguientes datos



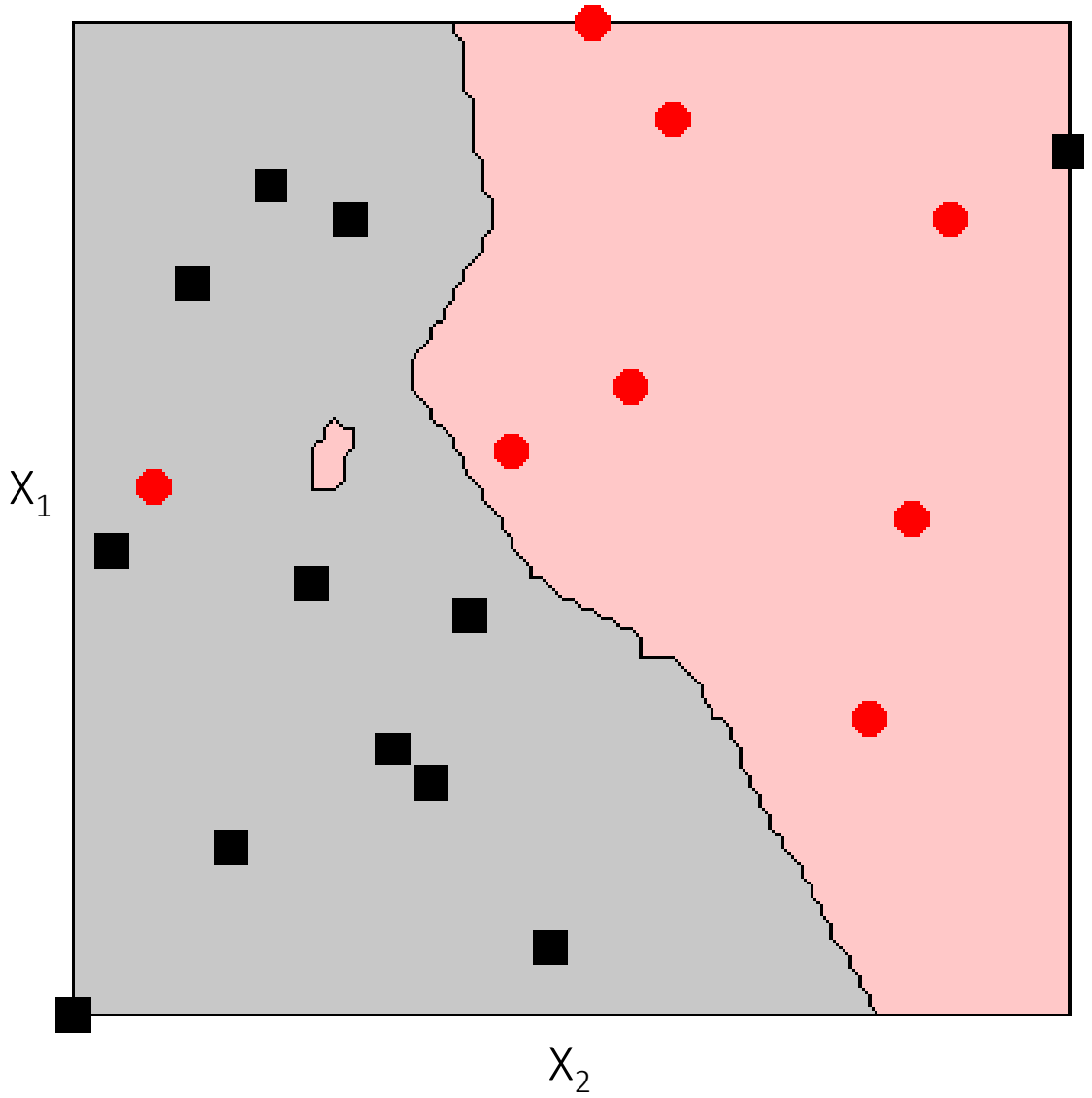
K = 1



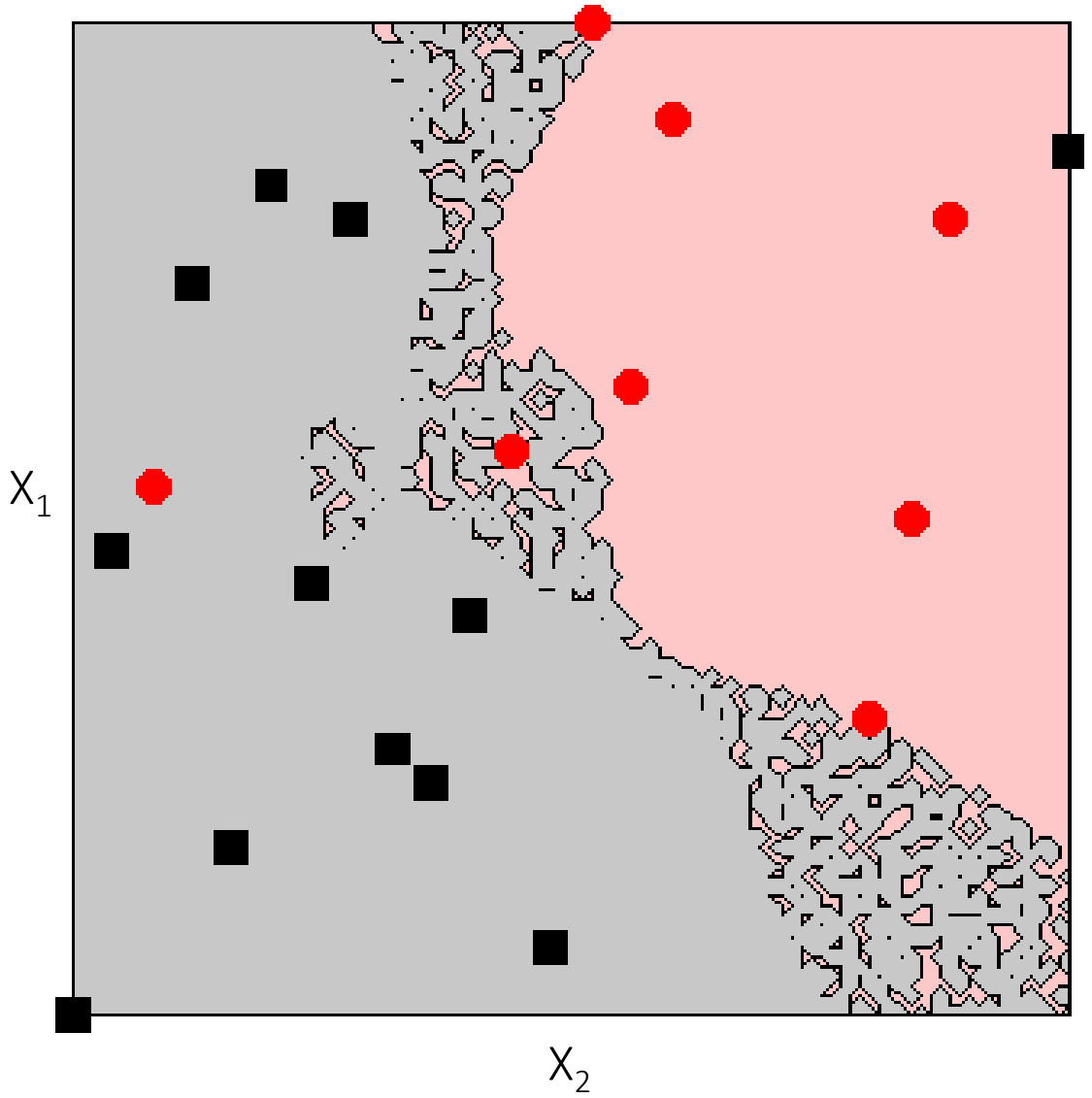
K = 2



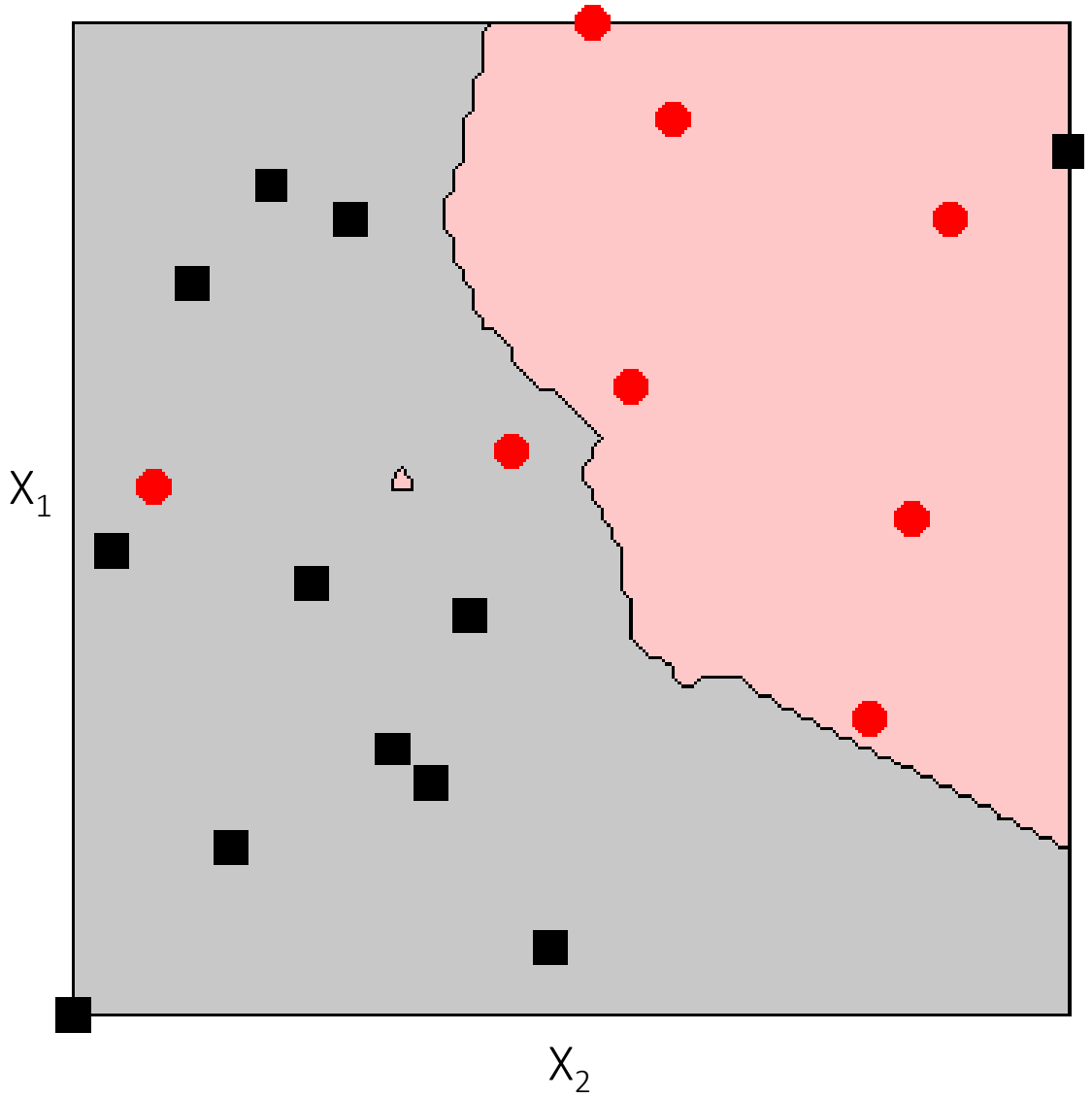
K = 3



K = 4



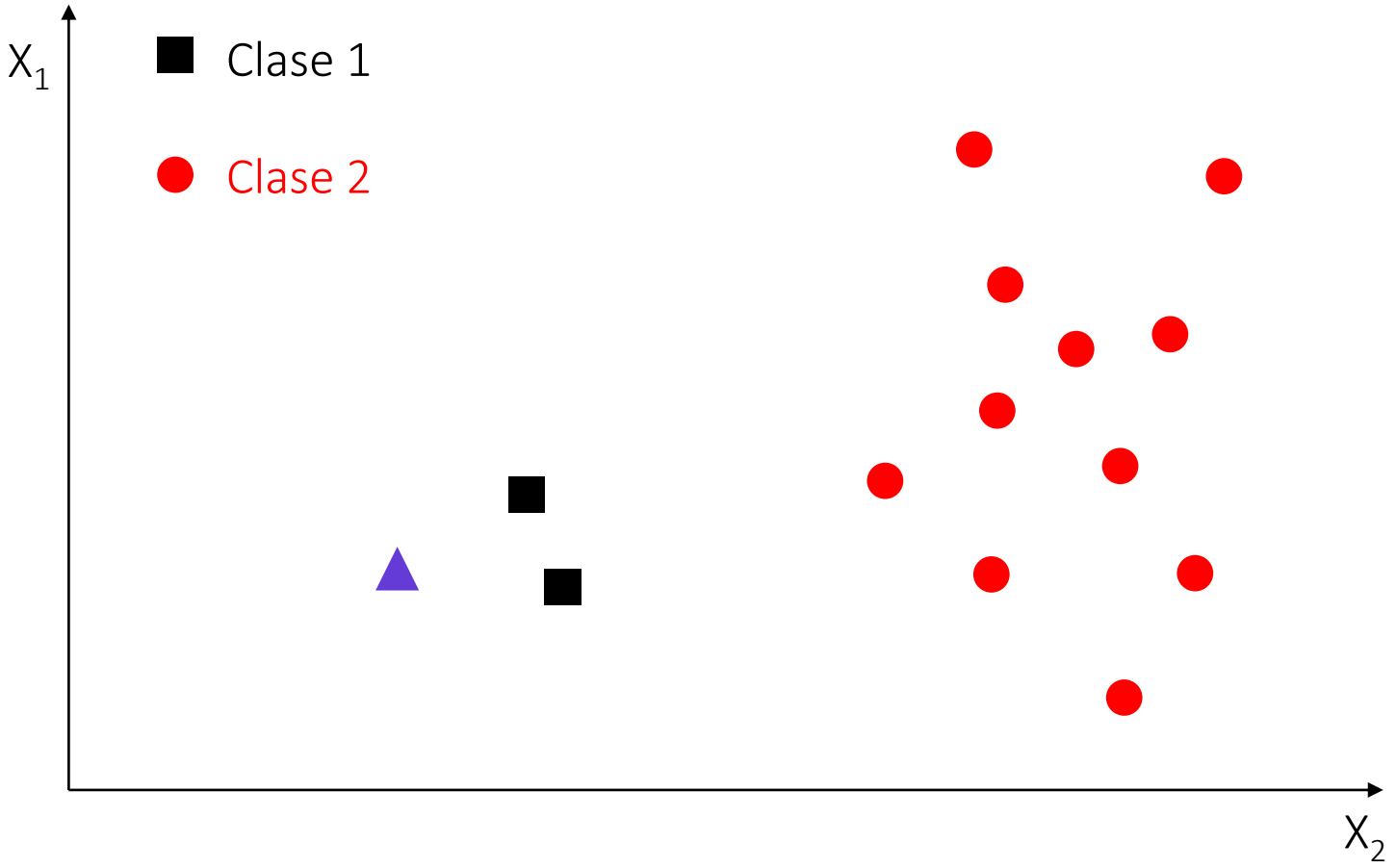
K = 5



Variantes de KNN

Resultados poco intuitivos

Supongamos la siguiente situación



con K=5 se clasifica como **Clase 2**, aún cuando una clasificación más razonable sería **Clase 1**

¿Cómo podemos mejorar el algoritmo?

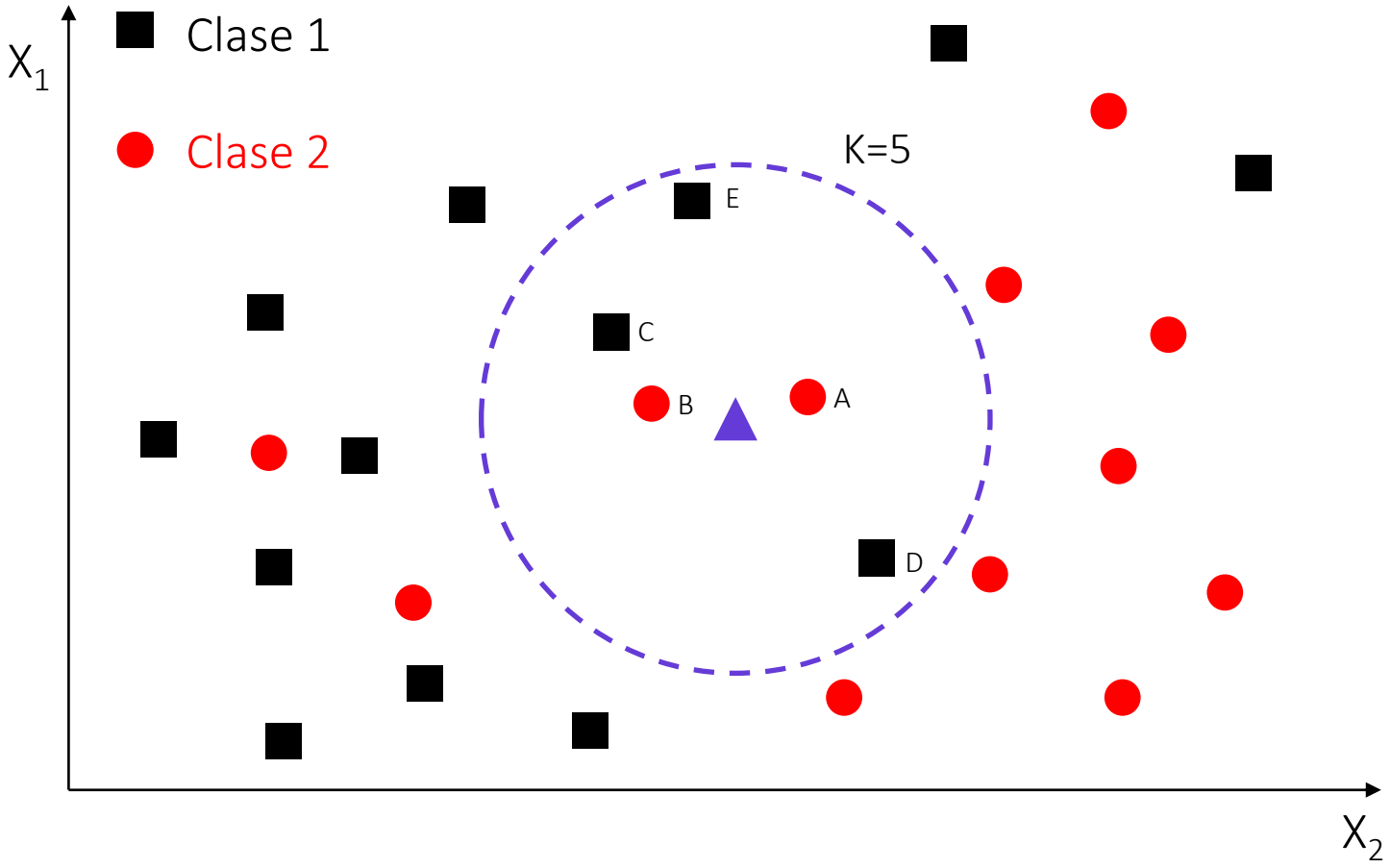
Podemos aplicar un peso a cada voto, en función de la distancia

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*:

$$w_{XY} = \frac{1}{d(X,Y)} \qquad w_{XY} = \exp\left(-\frac{d(X,Y)^2}{2}\right)$$

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 y seleccionar las variables 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 vemos tres:

- Árboles de clasificación
- KNN
- Regresión logística (asumiendo una regla de clasificación)