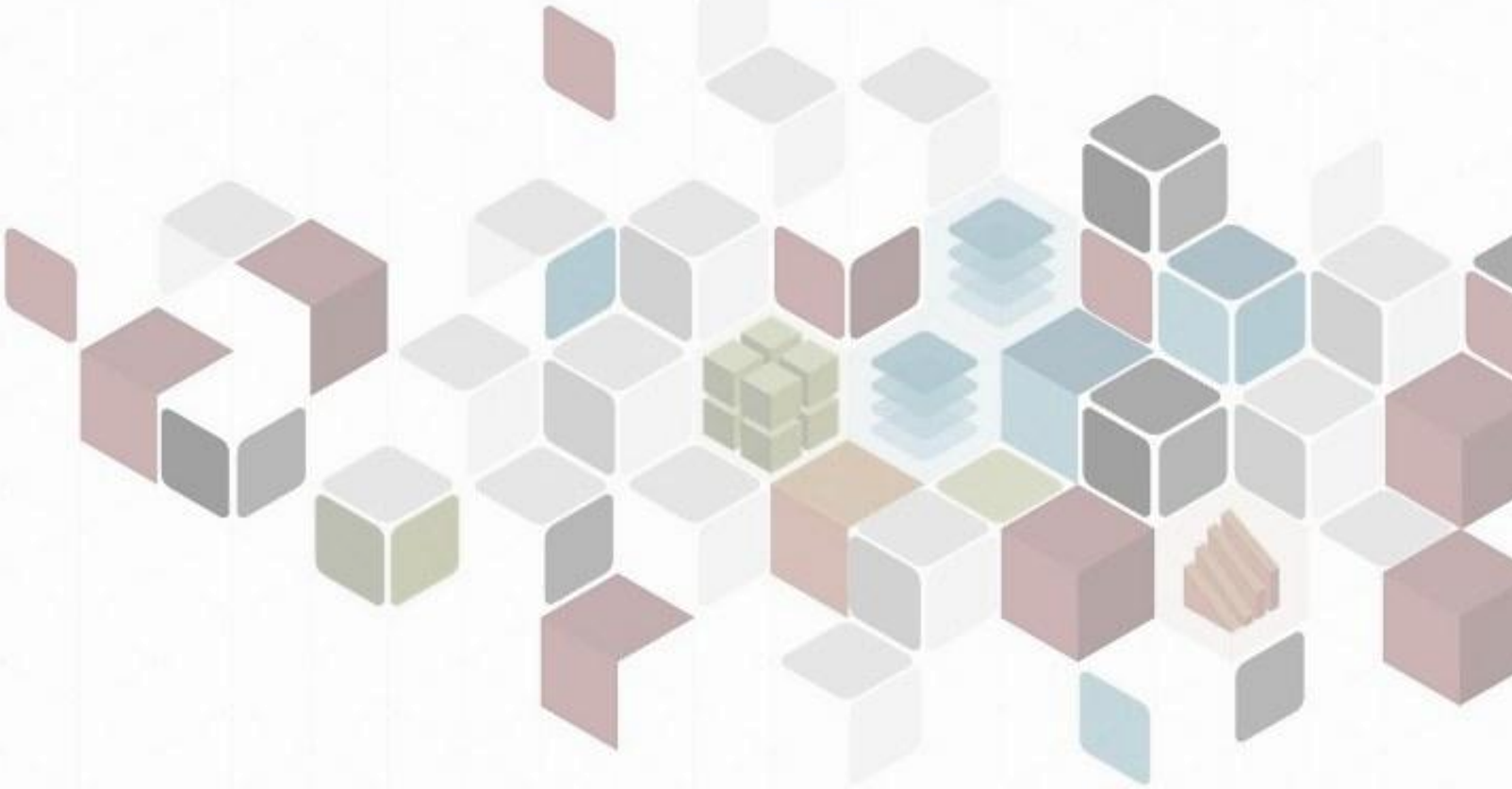
Grandes de Bases de Datos

Modelos de clasificación y predicción

K vecinos más próximos

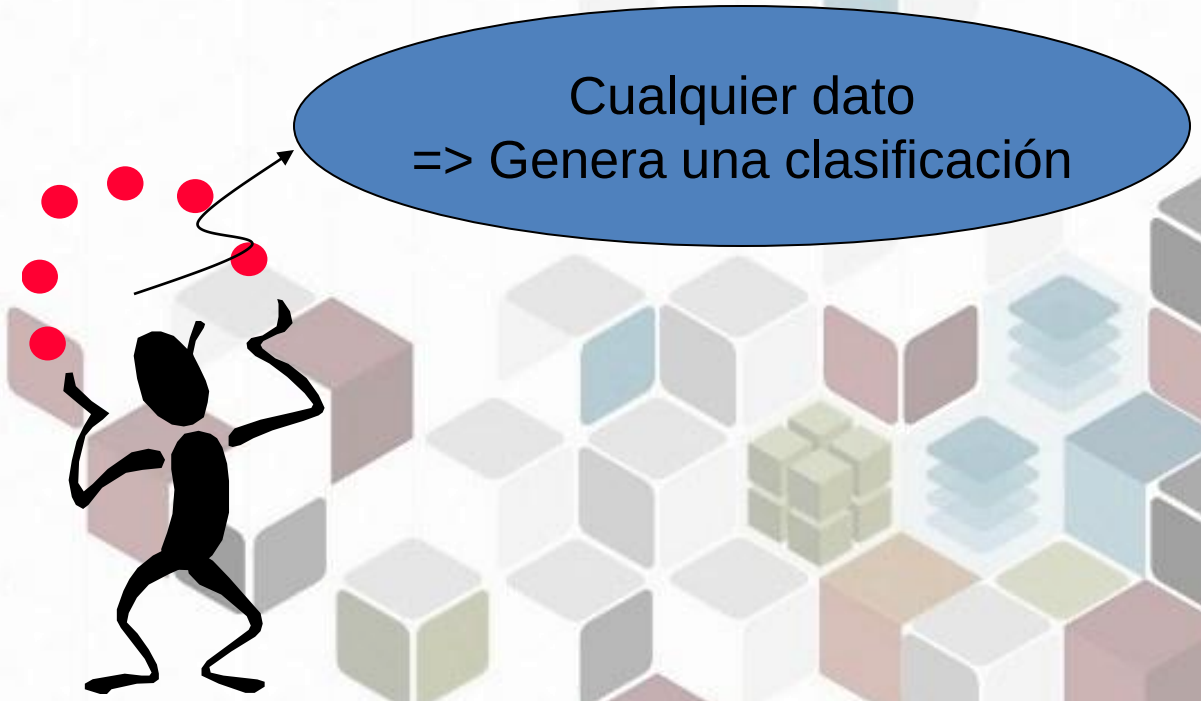


Distintos métodos de aprendizaje

- Aprendizaje “Ansioso”
  - Descripción explicita de la función objetivo basados en el total del conjunto de entrenamiento
- Aprendizaje basado en instancias (Perezoso)
  - Aprendizaje –> Almacenar todas las instancias
  - Clasificación –> Asignar la función objetivo a una nueva instancia

Distintos métodos de aprendizaje

- Aprendizaje “Ansioso”



Distintos métodos de aprendizaje

- Aprendizaje “Perezoso”

Distintos métodos de aprendizaje

- Aprendizaje “Perezoso”

Distintos métodos de aprendizaje

- Idea general

Aprendizaje basado en instancias

K vecinos más próximos

Regresión ponderada

K vecinos más próximos

- Características
  - Todas las instancias corresponden a puntos en un espacio n dimensional
  - La clasificación se retarda hasta que la nueva instancia llega
  - La clasificación se realiza comparando los puntos de los vectores
  - Existe una función objetivo

K vecinos más próximos

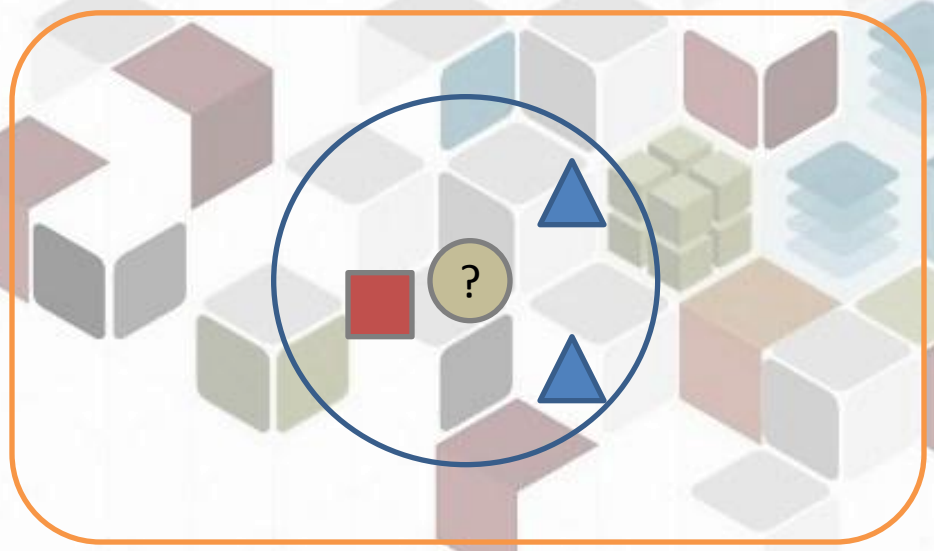
- Usado para clasificar objetos basados en la cercanía presente en elementos de un espacio n - dimensional
- Dentro de los 10 algoritmos de minado más utilizados
  - ICDM – Diciembre 2007
- Aproximación simple pero sofisticada al problema de clasificación

K vecinos más próximos

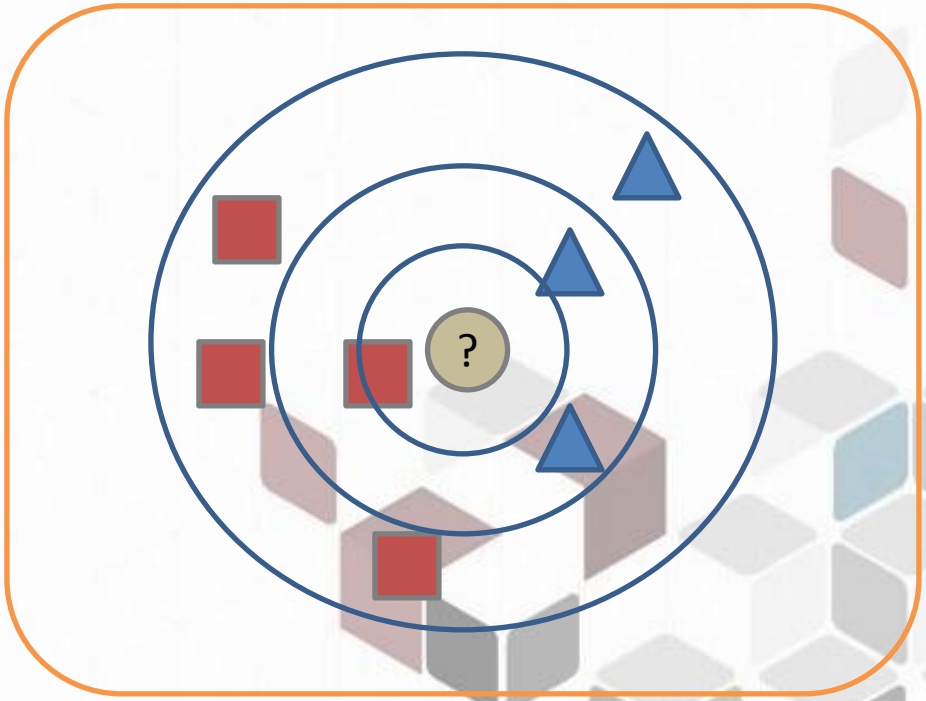
- Cuando un elemento llega, k-NN encuentra los k vecinos más próximos al nuevo ejemplar tomando en cuenta TODO el conjunto existente previamente, basándose en la similitud o distancia.
- Se crea un proceso de votación para elegir la clase perteneciente según los k vecinos más cercanos

K vecinos más próximos

- 3 elementos clave
  - Conjunto inicial de objetos
  - Métrica de similitud o distancia
  - Valor de k, total de vecinos próximos



K vecinos más próximos



$K = 1$

$\Rightarrow$Pertenece a clase Cuadro

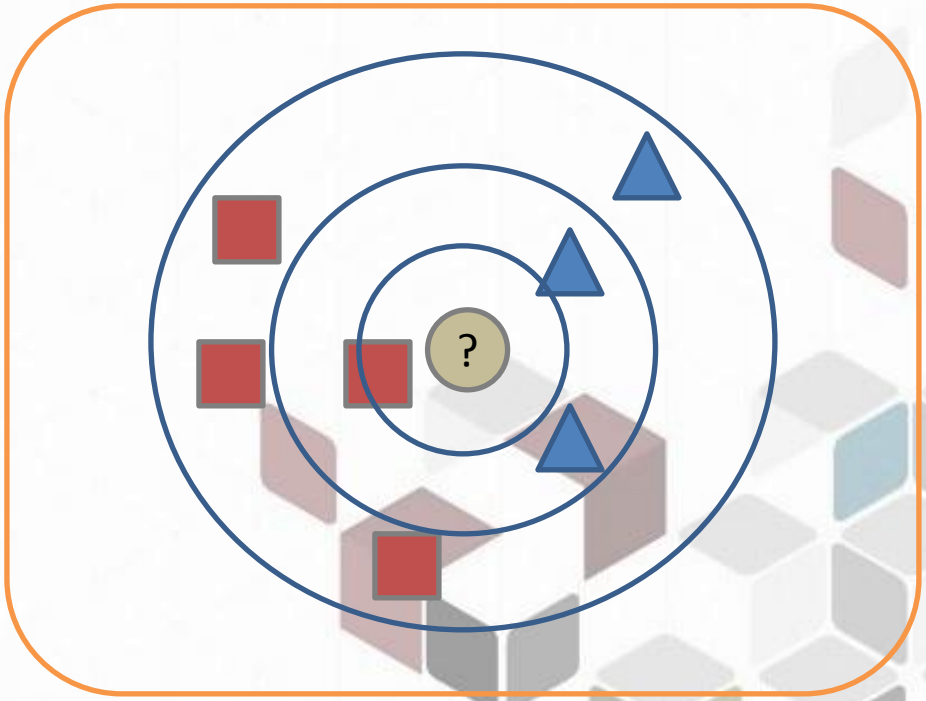
$K = 2$

$\Rightarrow$Pertenece a clase Triangulo

$K = 3$

$\Rightarrow$Pertenece a clase Cuadro

K vecinos más próximos



Valor de *k*:

- K muy pequeño, sensible a puntos extremos
- K muy grande, la vecindad puede incluir puntos de otras clases
- Elección de un valor non para eliminar empates

Problemas

- ¿Cuántos vecinos considerar? ¿Valor de k?
- ¿Cómo medir la distancia?
- ¿Cómo combinar la información de más de una observación?
- ¿El peso de los vecinos debe ser igual? ¿algunos vecinos deben tener mayor influencia que otros?

Métricas de Distancia

Sean x, y y z elementos; se define una métrica o función de distancia *d(.,.)*, si se cumplen

1. No negatividad $d(x, y) \geq 0$

2. Reflexiva $d(x, y) = 0, \quad \Leftrightarrow \quad x = y$

3. Conmutativa $d(x, y) = d(y, x)$

4. Desigualdad del triangulo

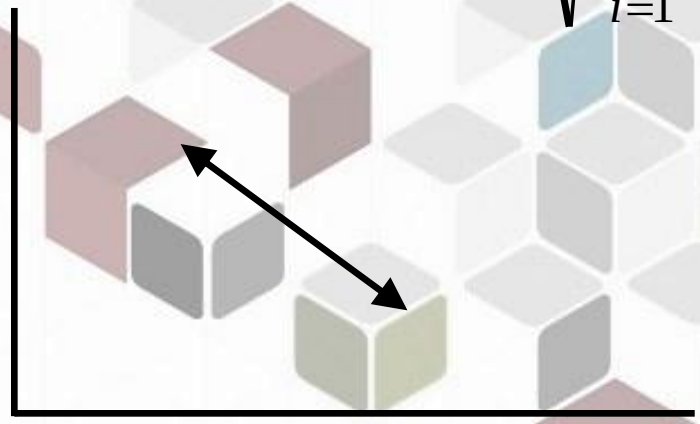
$$d(x, y) + d(y, z) \geq d(x, z)$$

Métricas de Distancia

Ejemplos de métricas

1. Distancia Euclidiana

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

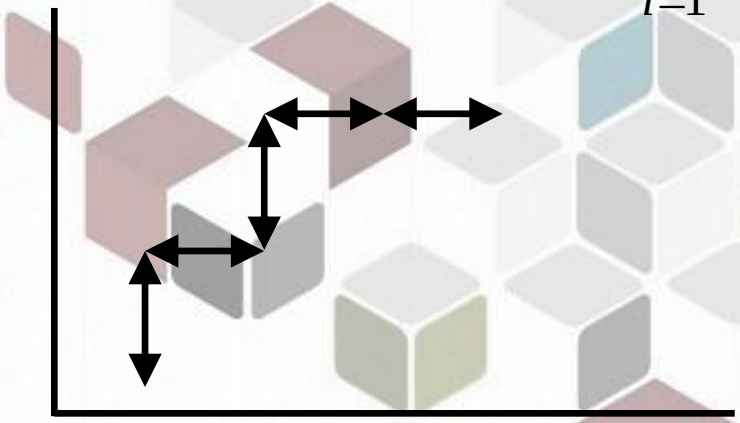


Métricas de Distancia

Ejemplos de métricas

1. Distancia Manhattan

$$d(x, y) = \sum_{i=1}^{n} | x_i - y_i |$$

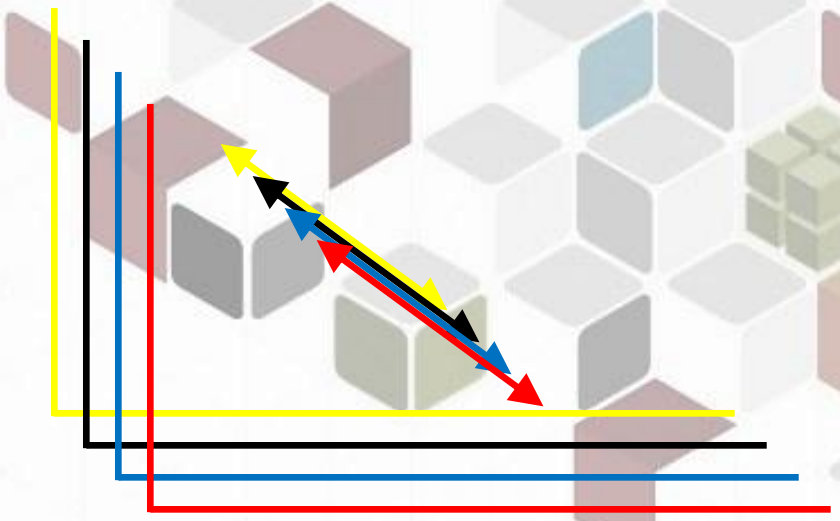


Métricas de Distancia

Ejemplos de métricas

1. Distancia Chebychev

$$d(x, y) = \max_{i = 1 \ldots n} | x_i - y_i |$$



Métricas de Distancia

Ejemplos de métricas

1. Distancia Coseno

$$d(x, y) = \arccos\left(\frac{x^T y}{\| x \| \cdot \| y \|}\right)$$

Métricas de Distancia

Ejemplos de métricas

1. Distancia Mahalanobis

$$d(x, y) = \sqrt{(x - y)^T S^{-1} (x - y)}$$

Métricas de Distancia

Ejemplos de métricas

1. Distancia Hamming

$$d(x, y) = \sum_i x_i \oplus y_i$$

Métricas de Distancia

Ejemplos de métricas

1. Distancia Levenshtein

Mínimo número de modificaciones para transformar una cadena en otra

casa → cama, distancia 1

casa → carro, distancia 3

Métricas de Distancia

Ejemplos de métricas

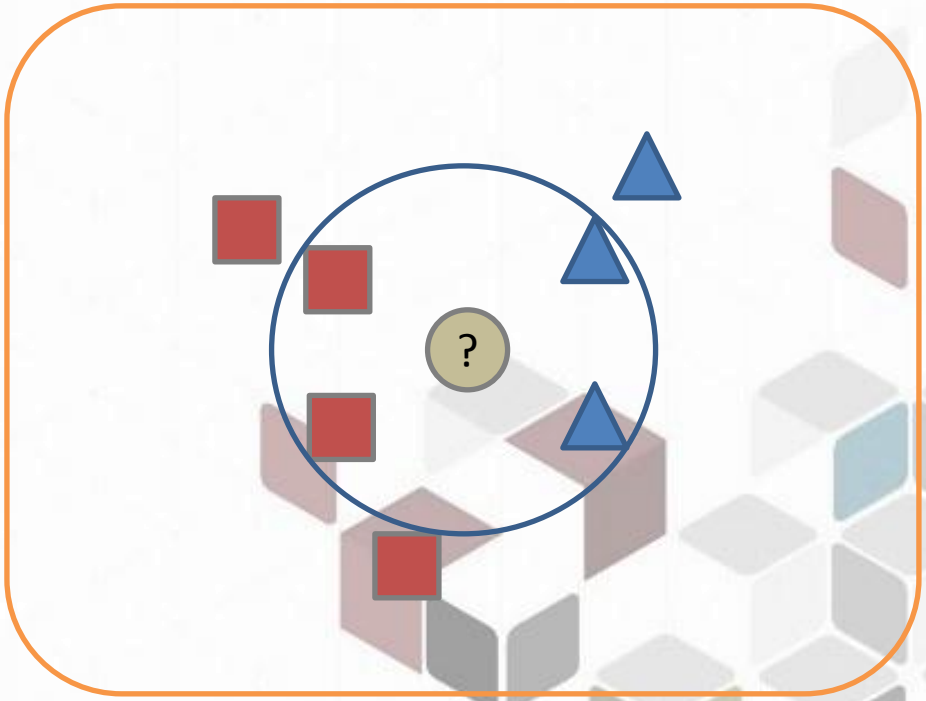
1. Distancia Damerau–Levenshtein

Mínimo número de modificaciones para transformar una cadena en otra… se considera la transposición de caracteres

ca → abc, distancia 2

No es una métrica… ¿por qué?

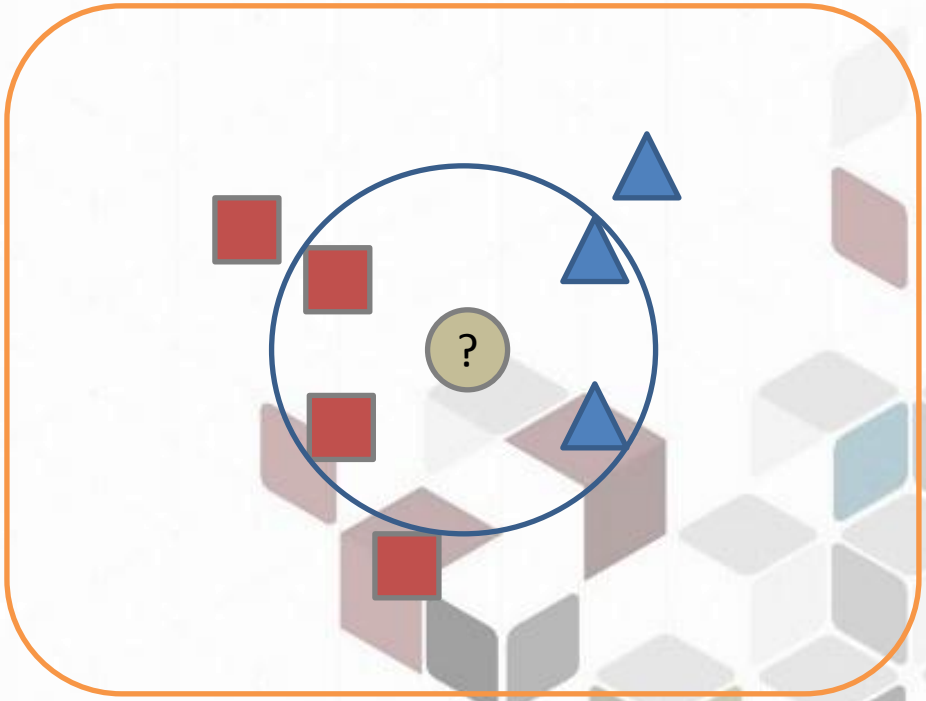
K Vecinos más próximos



¿Qué sucede si valoramos datos como los siguientes?

- Edad
- Peso
- Salario

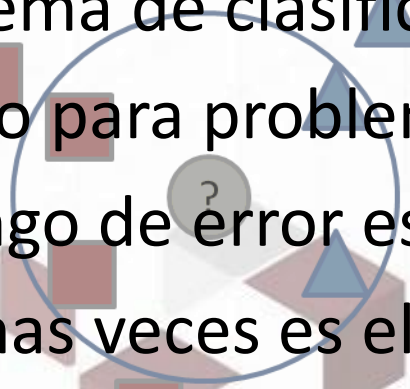
K Vecinos más próximos



Es necesario realizar siempre una normalización de los datos numéricos

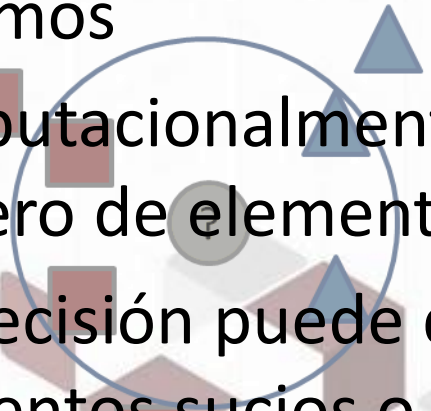
K Vecinos más próximos

- Técnica simple de implementar
- Construir el modelo es relativamente sin costo
- Esquema de clasificación simple
- Bueno para problemas con múltiples clases
- El rango de error es a lo más, el doble que el de Bayes
- Algunas veces es el mejor método
  - ¿Cuándo?



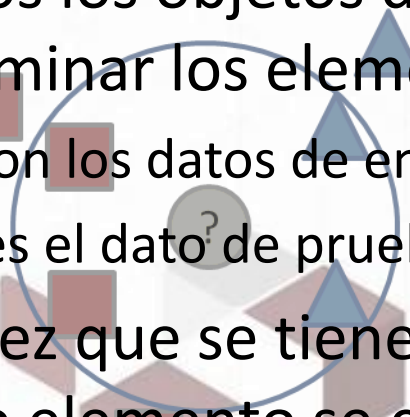
K Vecinos más próximos

- Clasificar nuevos registros es costoso
- Requiere el calculo de distancias para k vecinos más próximos
- Computacionalmente caro especialmente cuando el número de elementos de entrenamiento crece
- La precisión puede degradarse cuando se encuentran elementos sucios o atributos irrelevantes
- Maldición de la multi-dimensionalidad



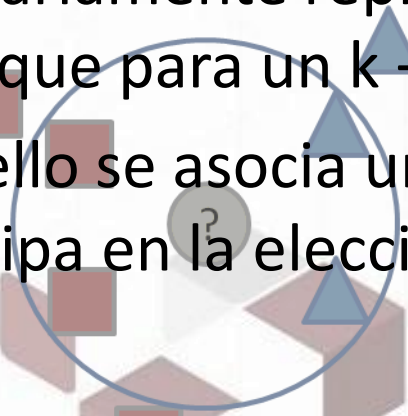
K Vecinos más próximos

- Dado un conjunto inicial de datos D, y un nuevo objeto *x* = *(**x**', y')*, el algoritmo calcula la distancia (o similitud) entre x y todos los objetos del conjunto D *(**x**, y)* ∈ *D* para determinar los elementos más cercanos, *Dz*.
  - **X** son los datos de entrenamiento, mientras *y* es su clase.
  - **x'** es el dato de prueba o clasificación y *y'* su clase.
- Una vez que se tiene la lista de los vecinos más próximos, el nuevo elemento se clasifica basándose en la clase mayoritaria de los vecinos



K Vecinos más próximos

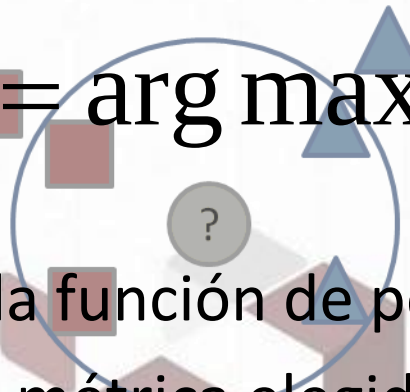
- Puede presentarse el hecho de que los elementos más cercanos para un determinado elemento no necesariamente representen a la clase del nuevo elemento, dado que para un k + 1 la clase del elemento se modifica.
- Para ello se asocia un peso al voto de cada elemento que participa en la elección de la clase.



K Vecinos más próximos

Por ello la clasificación para un elemento x con atributos discretos se da mediante la siguiente ecuación:

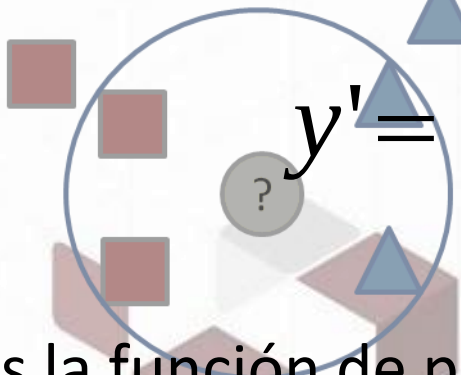
$$y' = \arg\max_{c \in C} \sum_{i=1}^{k} w_i \delta(f(x_i), c)$$



- w_i es la función de peso o ponderación
- f es la métrica elegida
- δ es una función de indicación que regresa el valor
  - 1 si su argumento es similar
  - 0 en caso contrario

K Vecinos más próximos

Por ello la clasificación para un elemento x con atributos continuos se da mediante la siguiente ecuación:



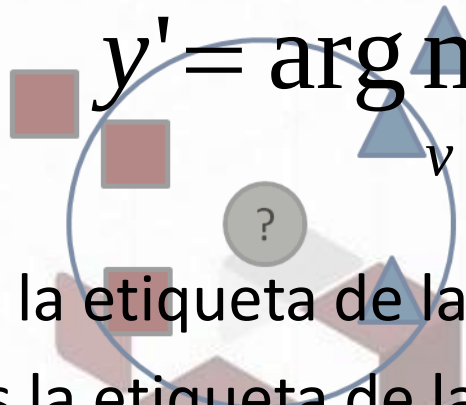
$$y' = \frac{\sum_{i=1}^{k} w_i f(x_i)}{\sum_{i=1}^{k} w_i}$$

- w_i es la función de peso o ponderación
- f es la métrica elegida

K Vecinos más próximos

Por ello la clasificación para un elemento x con atributos discretos se da mediante la siguiente ecuación:

$$y' = \arg\max_{v} \sum_{(x_i, y_i) \in D} I(v = y_i)$$



- *v* es la etiqueta de la clase
- y_i es la etiqueta de la clase para el i-ésimo vecino más cercano,
- *I (·)* es una función de indicación que regresa el valor
  - 1 si su argumento es similar
  - 0 en caso contrario