Minería de Datos

Técnicas de Minería de datos – Métodos basados en vecindad

Agenda:

- **Clasificación con métodos basados en vecindad**
- **Algoritmo K-vecinos**

Clasificación con métodos basados en vecindad

- **La predicción se basa en la utilización de los datos o ejemplos "vecinos" al dato que hay que procesar.**
- **Idea: ante una nueva situación, se podría actuar como se hizo en situaciones anteriores parecidas y similares, si éstas fueron exitosas.**
- **La similitud o distancia entre cada ejemplo y el dato a procesar es esencial en el proceso**

Ejemplo:

☞ **En clasificación:**
Asignar una clase a un nuevo dato, observando la clase de datos similares

☞ **En regresión:**
El valor numérico predicho para un nuevo dato, se obtiene de los valores obtenidos para ejemplos similares

Clasificación con métodos basados en vecindad

Dos aspectos importantes:

a) **¿Qué se entiende por similitud?**

b) **¿Cuándo se explota dicha similitud?**

a) ¿Qué se entiende por similitud?

Similitud

- **Es una medida numérica que indica el grado al cual dos objetos se parecen**
- **A más alto este valor más parecidos los objetos**
- **Es no negativa y generalmente entre 1 (similitud máxima) y 0 (no hay similitud)**

Sin embargo, es común utilizar la distancia (inverso de la similitud), también conocida como disimilitud:

Clasificación con métodos basados en vecindad

Distancia

- **Es una medida numérica que indica el grado al cual dos objetos son diferentes**
- **Mientras más bajo este valor, más parecidos**
- **Puede asumir valores entre [0, 1] o entre [0, ∞]**

Las medidas que satisfacen las siguientes tres propiedades

1. **Positividad:** $d(X, Y) \geq 0 \quad \forall \; X, Y$

   $d(X, Y) = 0 \longrightarrow X = Y$

   ➡ **Métricas o distancias**

2. **Simetría:** $d(X, Y) = d(Y, X) \quad \forall \; X, Y$

3. **Desigualdad triangular:** $d(X, Z) \leq d(X, Y) + d(Y, Z) \quad \forall \; X, Y, Z$

Sin embargo, muchas medidas no satisfacen estas 3 propiedades pero son muy útiles

Para las medidas de similitud la desigualdad triangular generalmente no se cumple

Clasificación con métodos basados en vecindad

Algunas medidas de similitud:

- Para datos numéricos

$$Cos(X, Y) = \frac{X \,.\, Y}{||X|| + ||Y||}$$

Ejemplo:

Si X = (2,1), Y = (3,2), Z = (5,1)

$$Cos(X, Y) = \frac{(2 \times 3) + (1 \times 2)}{(2.23) + (3.60)} = 1.37$$

$$Cos(X, Z) = \frac{(2 \times 5) + (1 \times 1)}{(2.23) + (5.09)} = 1.50 \leftarrow$$

X, Z son más parecidos

Clasificación con métodos basados en vecindad

- Para datos categóricos

$$\text{Similitud}(X, Y) = \frac{\text{comunes}}{\text{comunes} + \text{no\_comunes}}$$

Ejemplo:

Si X = (Rojo, Alto, Maracay, Pequeño, Redondo)

Y = (Rojo, Bajo, Maracay, Grande, Redondo)

Z = (Verde, Bajo, Caracas, Grande, Cuadrado)

$$\text{Similitud}(X, Y) = \frac{3}{3 + 2} = 0.6$$

X, Y son más parecidos

$$\text{Similitud}(Z,Y) = \frac{2}{3 + 2} = 0.4$$

Clasificación con métodos basados en vecindad

Algunas medidas de distancia:

- Para datos numéricos

Sea $X = (x_1, x_2, x_3, \ldots\ldots, x_d)$

$Y = (y_1, y_2, y_3, \ldots\ldots, y_d)$

Distancia Euclídea = $\sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$

Distancia Manhattan = $\sum_{i=1}^{d} \left| (x_i - y_i) \right|$

Distancia Canberra = $\sum_{i=1}^{d} \dfrac{\left| (x_i - y_i) \right|}{\left| x_i \right| + \left| y_i \right|}$

También se han definido medidas de similitud o distancias para datos binarios, datos complejos como cadenas de caracteres, grafos, árboles, medidas difusas, entre otras

Clasificación con métodos basados en vecindad

b) ¿Cuándo se explota esta similitud?

En el marco de clasificación mostrado hasta el momento

➡ **▪ Se realiza un paso inductivo para construir un modelo a partir de los datos.**

▪ Luego se aplica un paso deductivo para aplicar el modelo a los datos de test.

⟶ ***Métodos anticipados***

Otro esquema:

▪ Esperar a que se plantee una predicción sobre un nuevo ejemplo.

▪ En este momento, determinar los casos o ejemplos más parecidos (similares) y utilizar estos datos para responder.

⟶ ***Métodos retardados o perezosos***

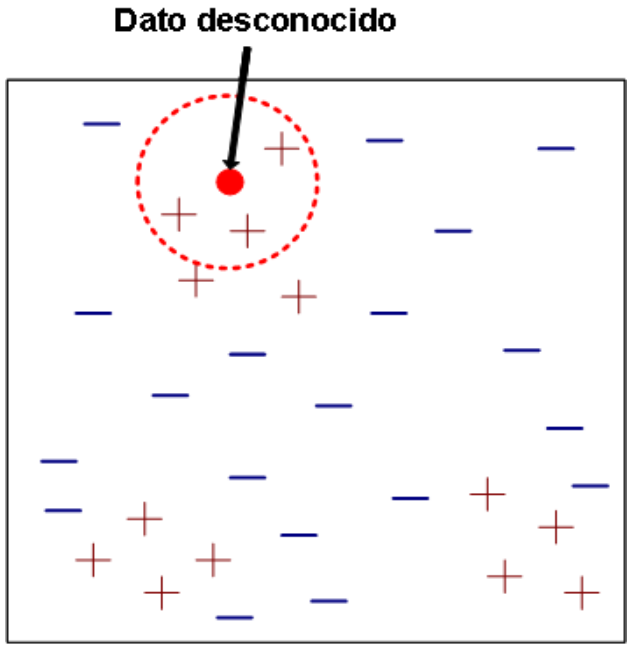
Clasificación con métodos basados en vecindad

- ***Métodos anticipados:***
  - **Construyen un modelo antes de realizar una tarea de predicción o generalización**
  - **Se construye una aproximación global utilizando la totalidad del conjunto de datos**
  - **Ejemplo: algoritmos de árboles de decisión**

- ***Métodos retardados***
  - **No construyen un modelo y retrasan la decisión de predicción hasta el instante en que se recibe un nuevo dato a procesar**
  - **Realizan una aproximación local al dato a generalizar (hace predicciones basado en información local).**
  - **Ejemplo: algoritmo k-vecinos**

Algoritmo k-vecinos

- **Idea básica: Encontrar los K ejemplos de entrenamiento que son más similares al ejemplo test**
- **Estos ejemplos = vecinos más próximos**
- **Se utilizan para determinar la clase del ejemplo test**



Requiere:

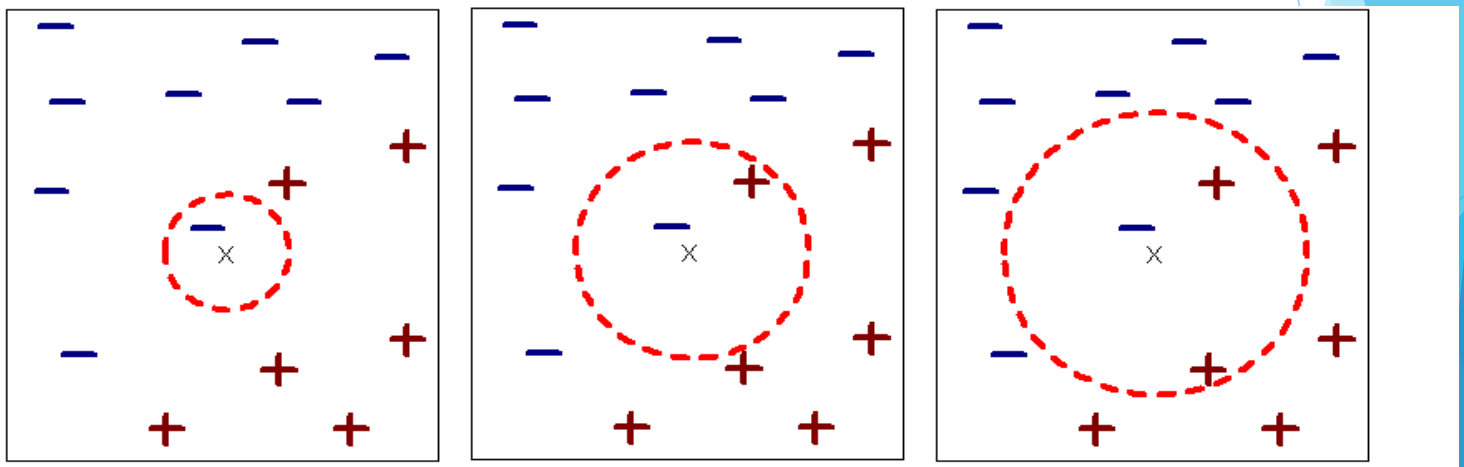
- **Un conjunto de registros almacenados**
- **Una medida de distancia o similitud**
- **El valor de K, el número de vecinos a recuperar**

Algoritmo k-vecinos

Para clasificar un nuevo dato:

- **Calcular la distancia a los registros de entrenamiento**
- **Determinar los k vecinos (más parecidos)**
- **Utilizar la etiqueta de clase de los vecinos para determinar la clase del nuevo dato (por ejemplo, por mayoría)**

Ejemplo:



a) 1- vecino más próximo — Clase = —

b) 2- vecinos más próximos — Clase = ?

c) 3- vecinos más próximos — Clase = +

Algoritmo k-vecinos

Algoritmo:

{Entrada: D (conjunto de entrenamiento), K (número de vecinos)}

Para cada ejemplo z_i

Para j = 1 hasta N

$d_i(z_i, X_j)$ = distancia entre z_i y el ejemplo de entrenamiento X_j

Fin_Para

D_z = conjunto de los K ejemplos más cercanos a z_i (lista de vecinos)

$$y_i = \underset{v}{\operatorname{argmax}} \sum_{(x_k, y_k) \in D_z} I(v = y_k)$$

% Clasifica el ejemplo de acuerdo a la clase mayoritaria de sus vecinos

Fin_Para

{Salida: conjunto de ejemplos test clasificados}

Donde,

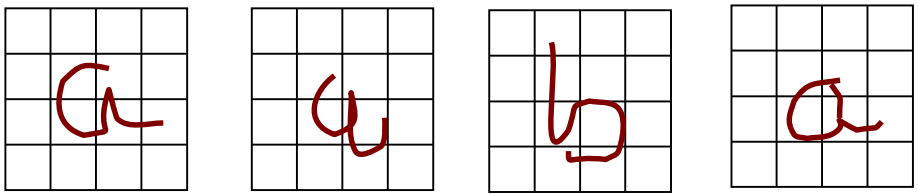
Función indicadora = $\begin{cases} 1 \text{ si } v = y \\ 0 \text{ si } v \neq y \end{cases}$

$$y_i = \underset{v}{\operatorname{argmax}} \sum_{(x_k, y_k) \in D_z} I(v = y_k)$$

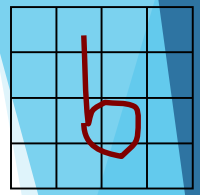
Etiqueta de clase

Algoritmo k-vecinos

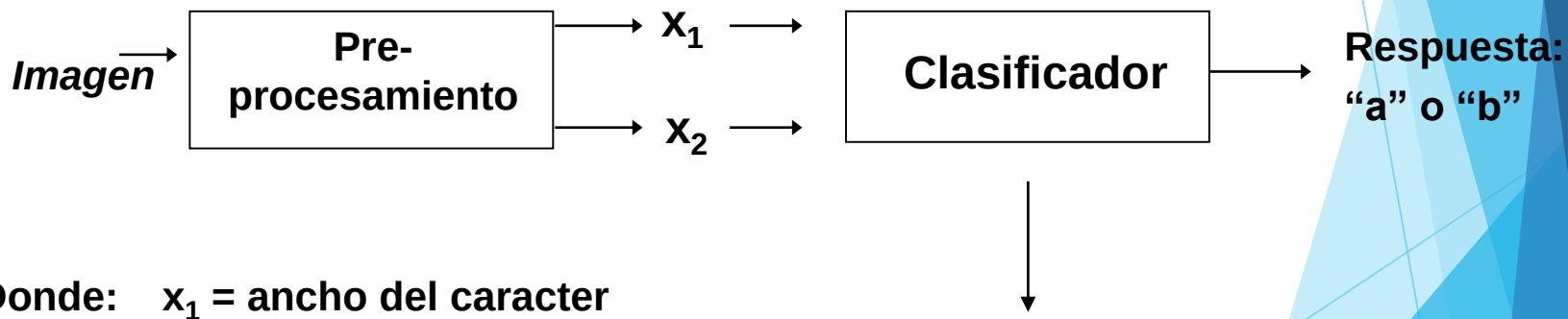
Ejemplo: **Reconocimiento de caracteres manuscritos a partir de imágenes**



“a” “a” “b” “a”



“b”

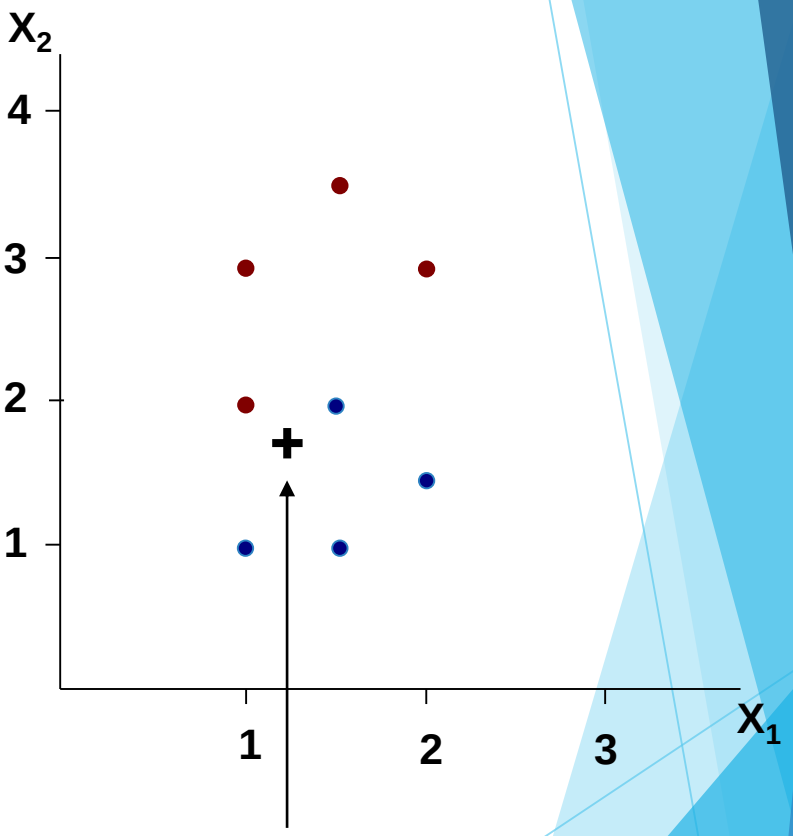


Donde: x_1 = ancho del caracter

x_2 = alto del caracter

Algoritmo k-vecinos

X_1	X_2	CLASE
1.0	1.0	a
1.5	1.0	a
1.5	2.0	a
2.0	1.5	a
1.0	2.0	b
1.0	3.0	b
1.5	3.5	b
2.0	3.0	b



Ejemplo test: z = (1.2, 1.8)

¿Cómo se clasifica?

Algoritmo k-vecinos

Si K = 3

1. Calcular la distancia de z a cada uno de los ejemplos de D utilizando, por ejemplo, distancia euclídea

$d(x^1, z) = 0.82$ $d(x^5, z) = 0.28$

$d(x^2, z) = 0.85$ $d(x^6, z) = 1.21$

$d(x^3, z) = 0.36$ $d(x^7, z) = 1.72$

$d(x^4, z) = 0.85$ $d(x^8, z) = 1.44$

2. Determinar la lista de los 3 vecinos más próximos

$$Dz = \{(x^1, y_1), (x^3, y_3), (x^5, y_5)\}$$

Algoritmo k-vecinos

3. Clasificar z de acuerdo a la clase mayoritaria

$y_1 = a$

$y_3 = a$

$y_5 = b$

➡ **Votos para “a” = 2**

Votos para “b” = 1

➡ **Se asigna a z la clase “a”**