

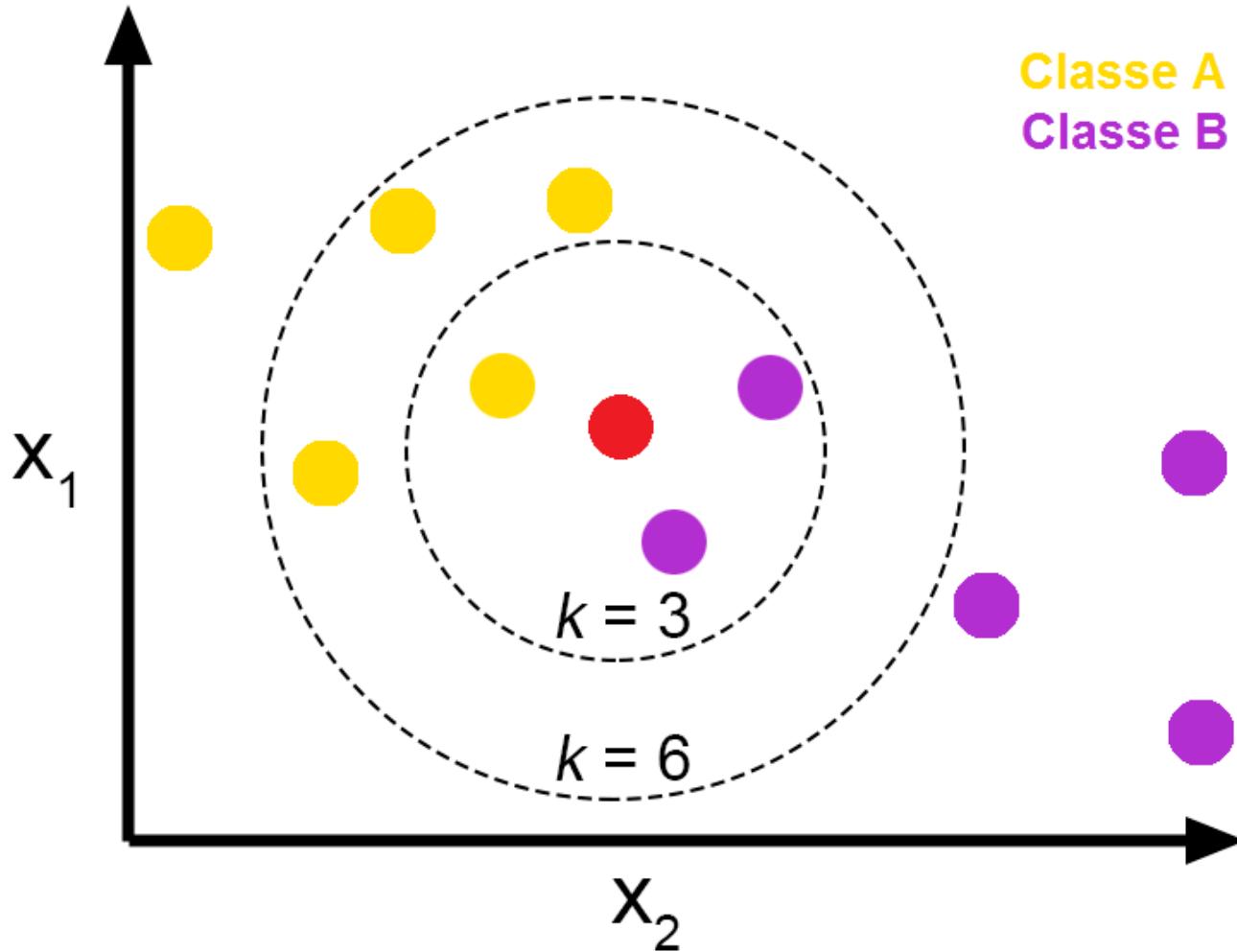


ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

IIC2433 Algoritmo de Clasificación K-NN

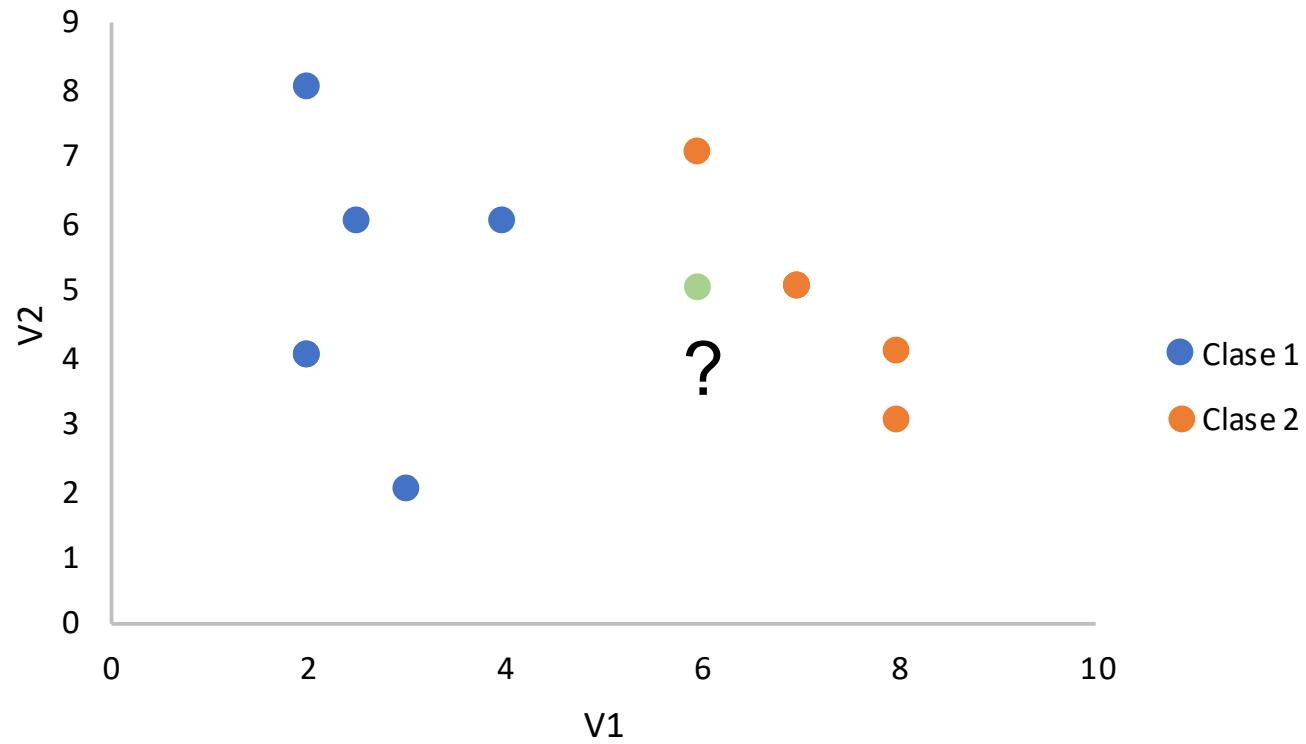
Profesor: Mauricio Arriagada
Minería de Datos

Clasificación: Vecinos cercanos (K-NN)



K vecinos cercanos

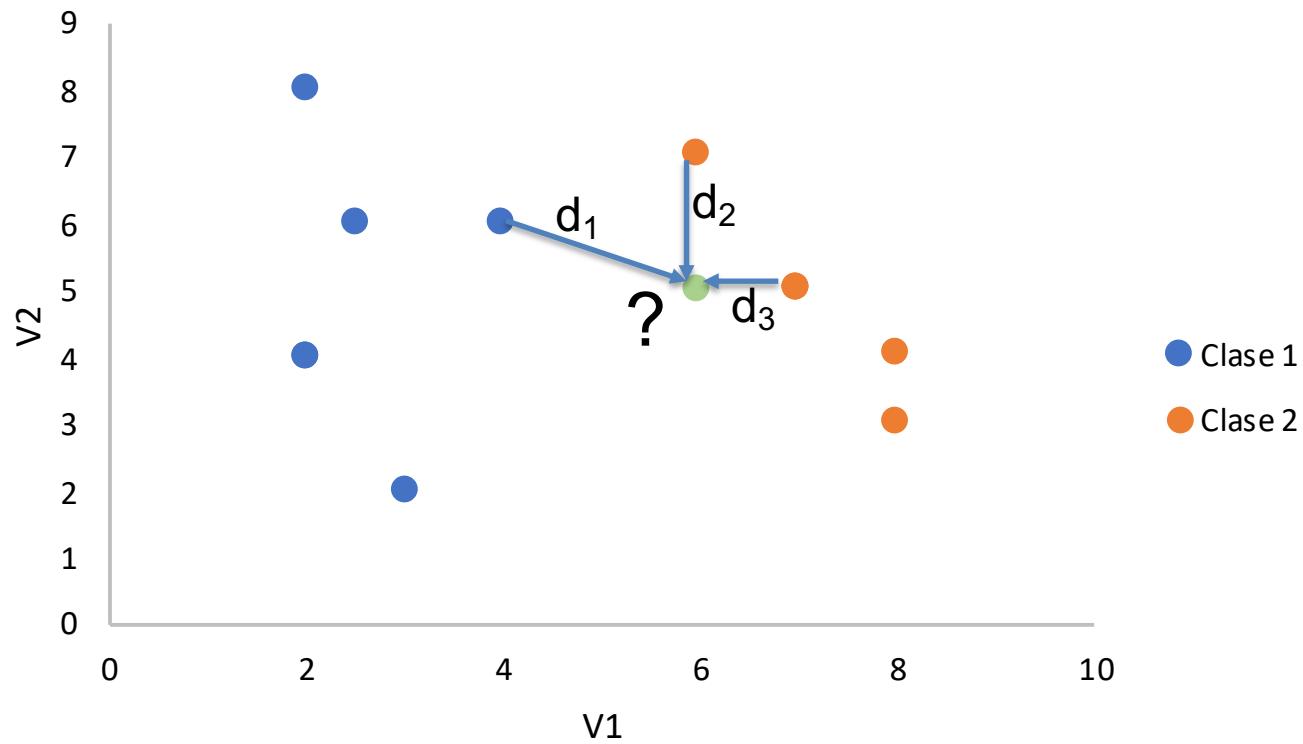
- Algoritmo que se basa en usar los datos más parecidos para poder clasificar



K vecinos cercanos

- ¿Qué dato es el más similar?

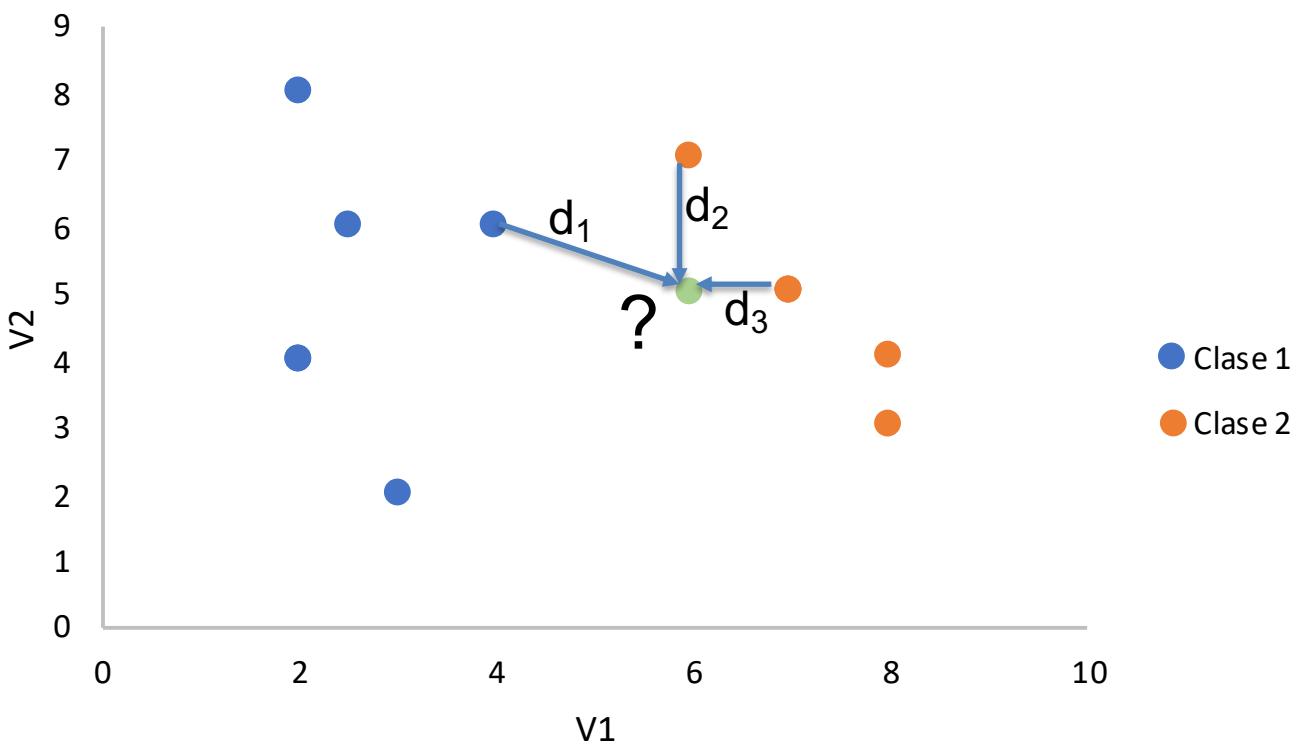
– Se necesita definir previamente una noción de distancia entre datos a sumiendo que en los datos más parecidos son los que están más cerca en el espacio de las variables V1 y V2



K vecinos cercanos

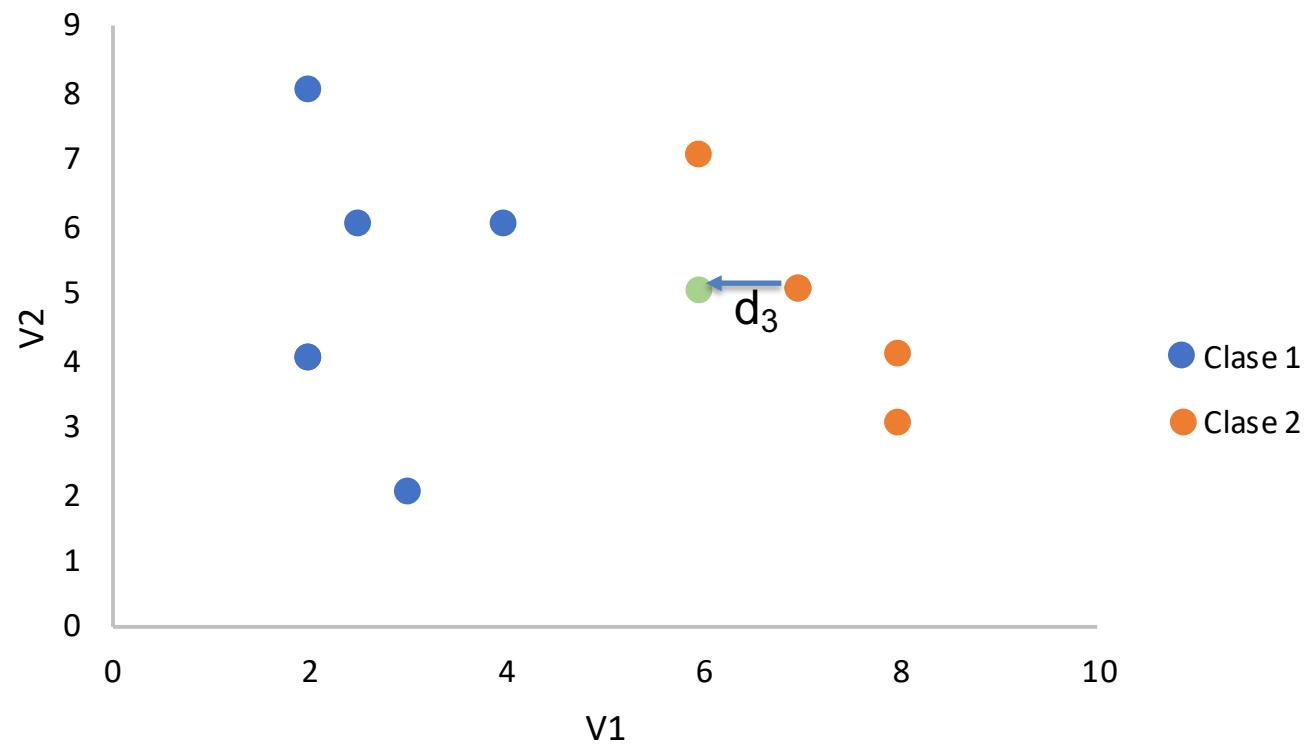
$$d = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

Distancia
Euclíadiana



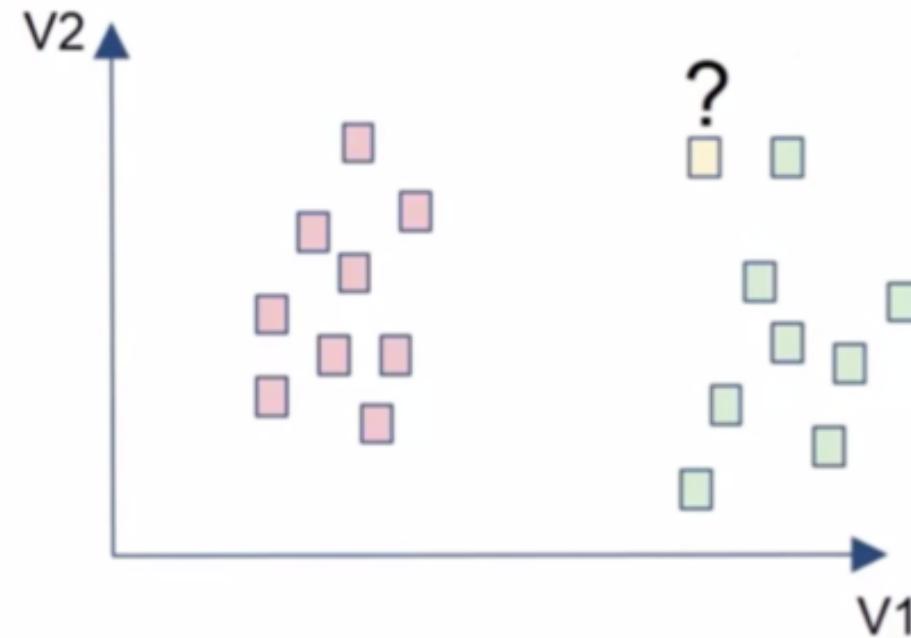
1 vecino cercano

- El dato más parecido es aquel que tenga las menor distancia
- Por lo tanto la clasificación sería naranja



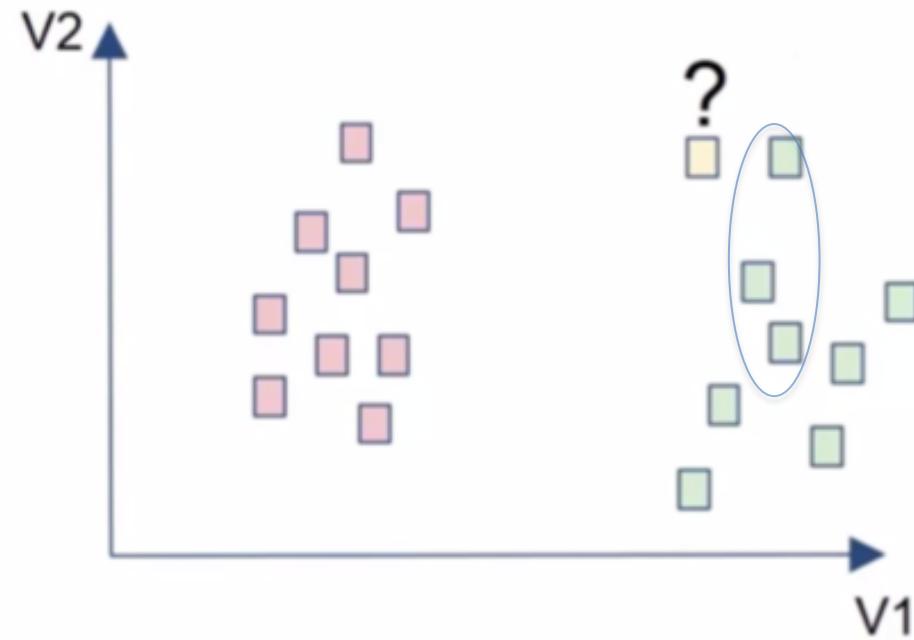
K vecinos cercanos

- Se considera el voto de la mayoría de las k instancias que más se parecen el dato en cuestión



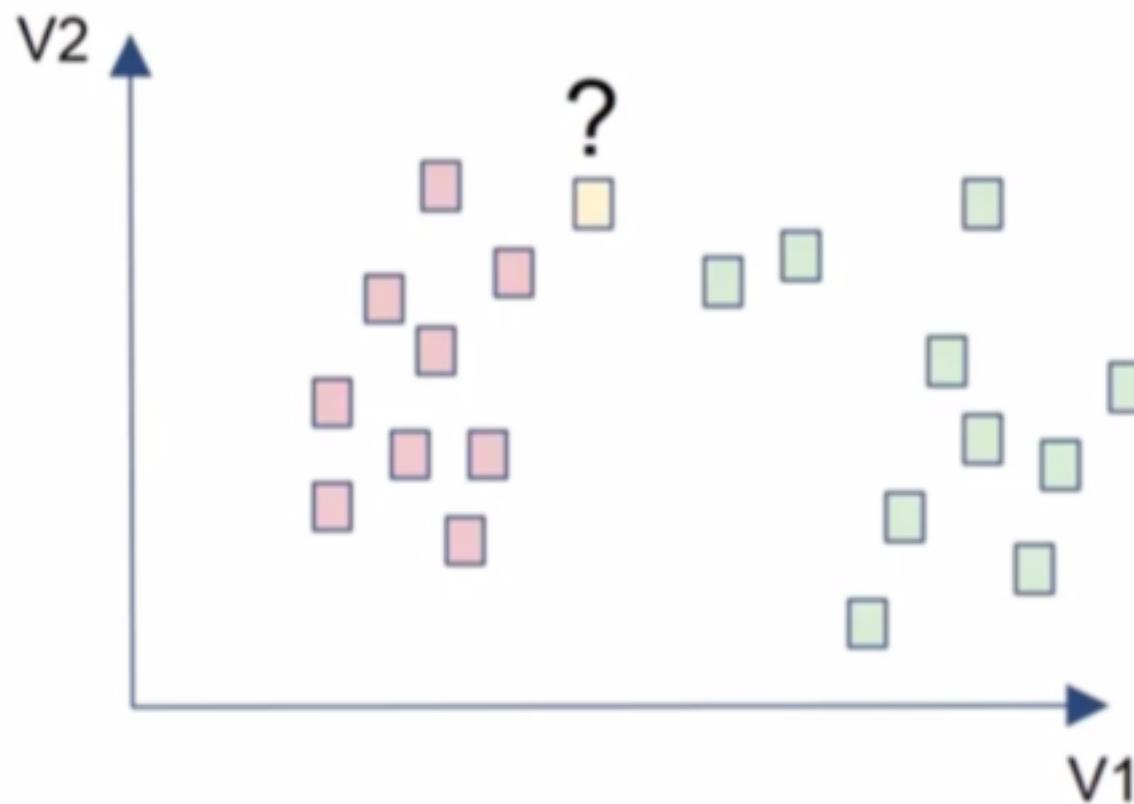
K vecinos cercanos

- Ejemplo: si K=3



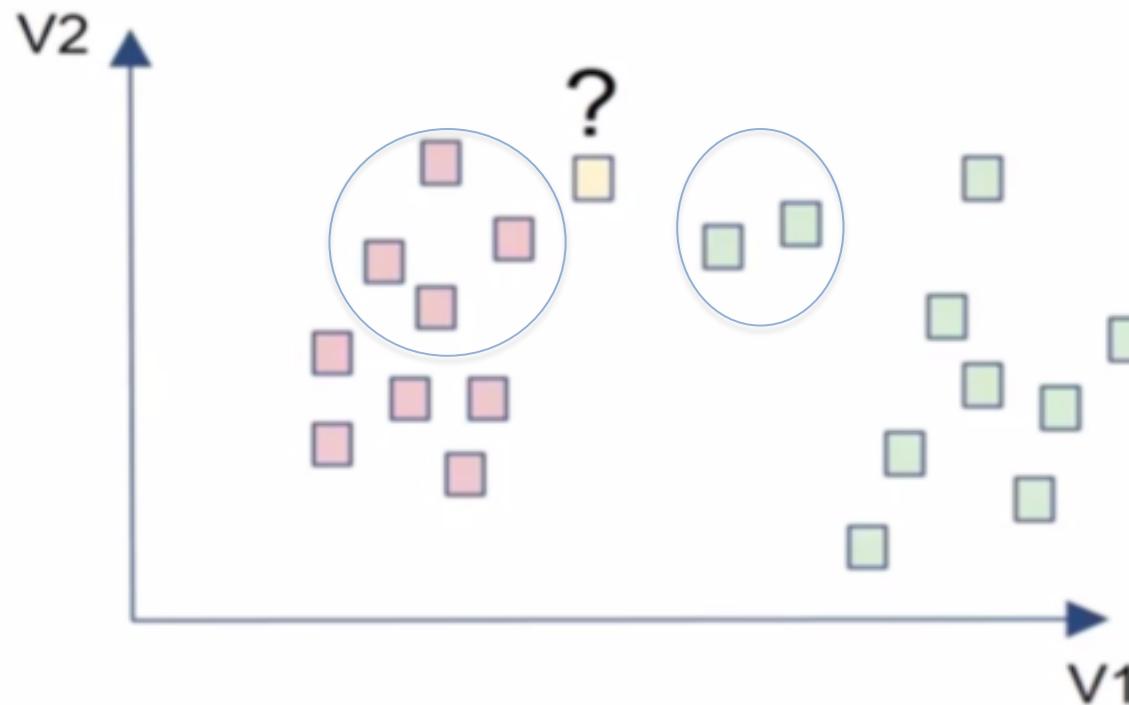
K vecinos cercanos

- Ejemplo: si K=6



K vecinos cercanos

- Ejemplo: si K=6
- La clasificación sería la clase de color rojo



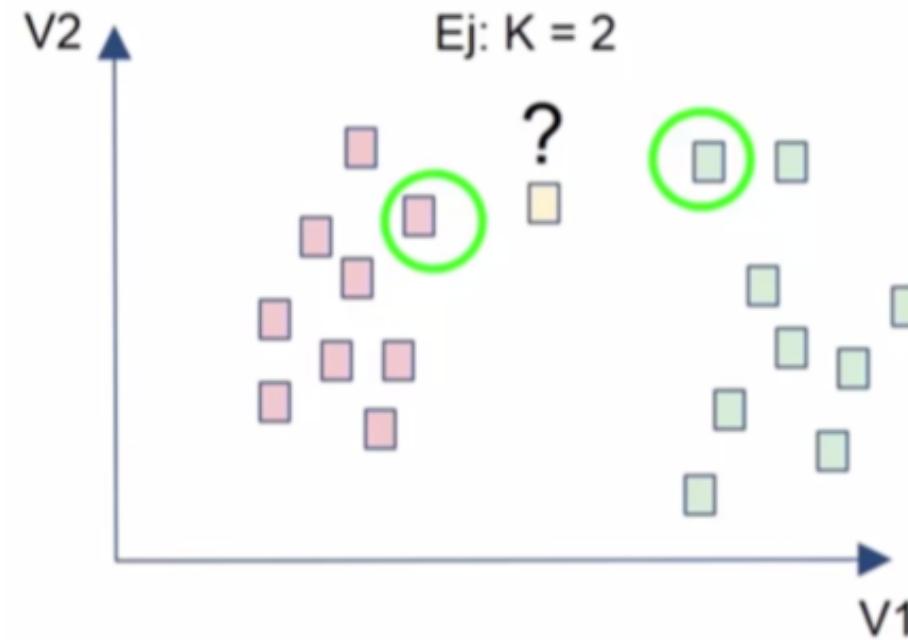
K vecinos cercanos

- Casos donde no es tan trivial encontrar una clase



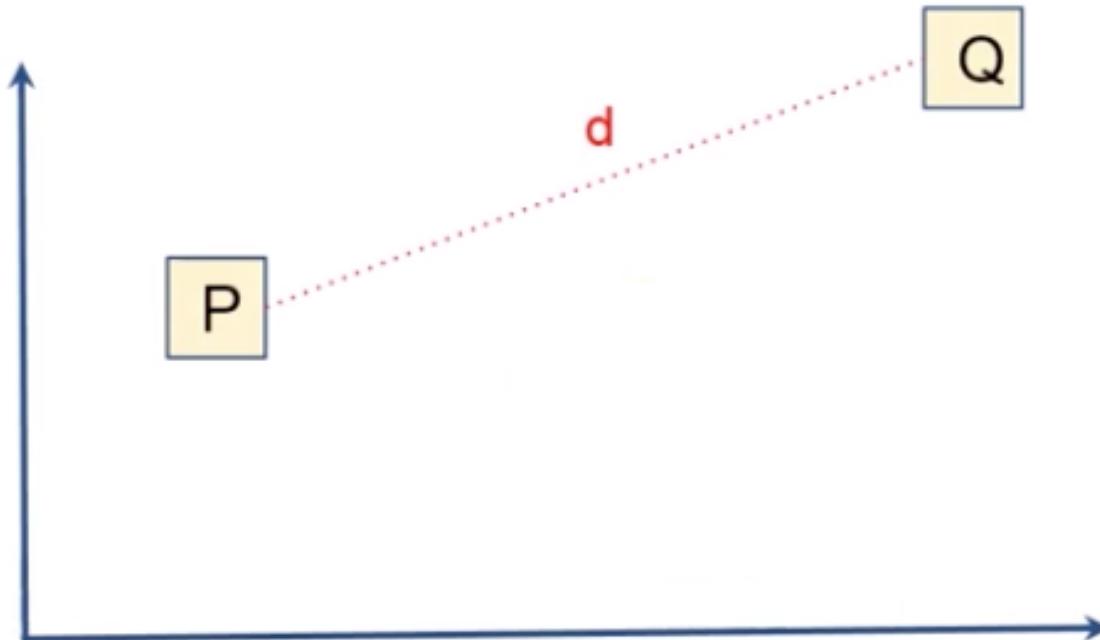
K vecinos cercanos

- Casos donde no es tan trivial encontrar una clase
- El problema es que cada uno pertenece a una clase distinta, por lo tanto no hay claridad sobre la clasificación que deberíamos proponer



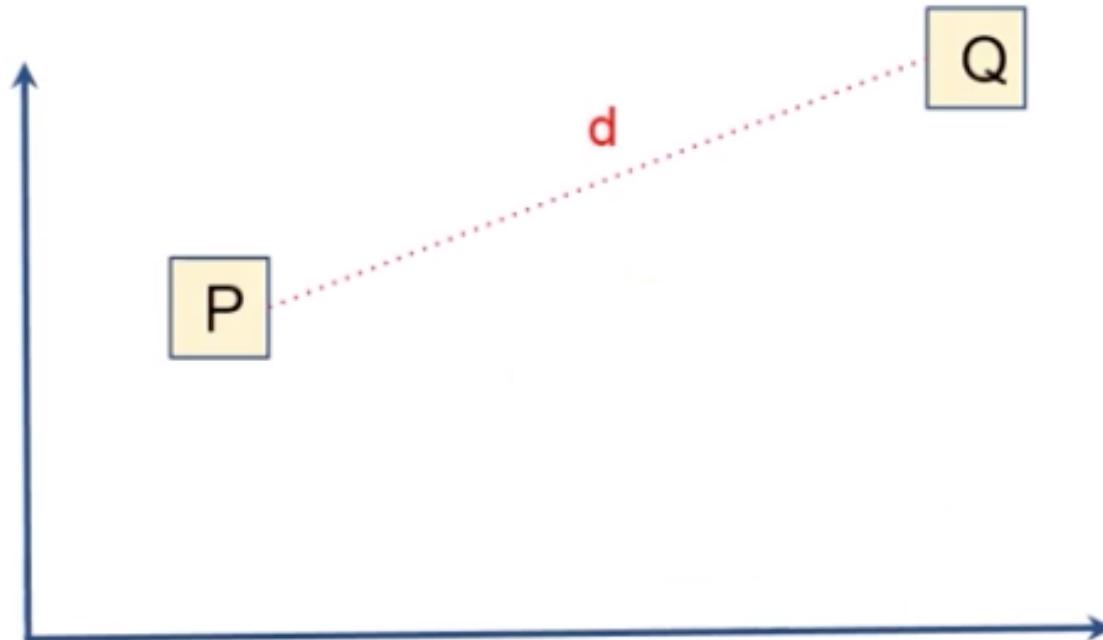
¿Cómo hacer comparaciones?

- Se explorarán distintas medidas de distancia.
- Dependiendo del tipo de variables que tenemos para describir los datos tendremos que cambiar la métrica a usar.



¿Cómo hacer comparaciones?

- Se explorarán distintas medidas de distancia.
- Dependiendo del tipo de variables que tenemos para describir los datos tendremos que cambiar la métrica a usar.



Ejemplo:

- ▶ datos son en general representados por un vector de descriptores.
- ▶ En general, típicamente los valores que pueden tomar las distintas columnas en este vector son de naturaleza muy distinta.

Id	Profesión	Rango Sueldo	Género	Monto Gasto Mensual	ubicación	...	Tipo Cliente
1	Ingeniero	500 - 1000	Masculino	50	(-27.98,33.45)		Preferencial

Ejemplo:

- ▶ ¿Cómo se podría calcular la siguiente distancia?
 - ▶ $d(\text{Juan}, \text{Maria})$

Nombre	Profesión	Rango Sueldo	Género	Monto Gasto Mensual	ubicación
Juan	Enfermera	500 - 1000	Masculino	30	(-27.98,33.45)
María	Músico	100 - 500	Femenino	10	(-35.65,22.16)

Ejemplo:

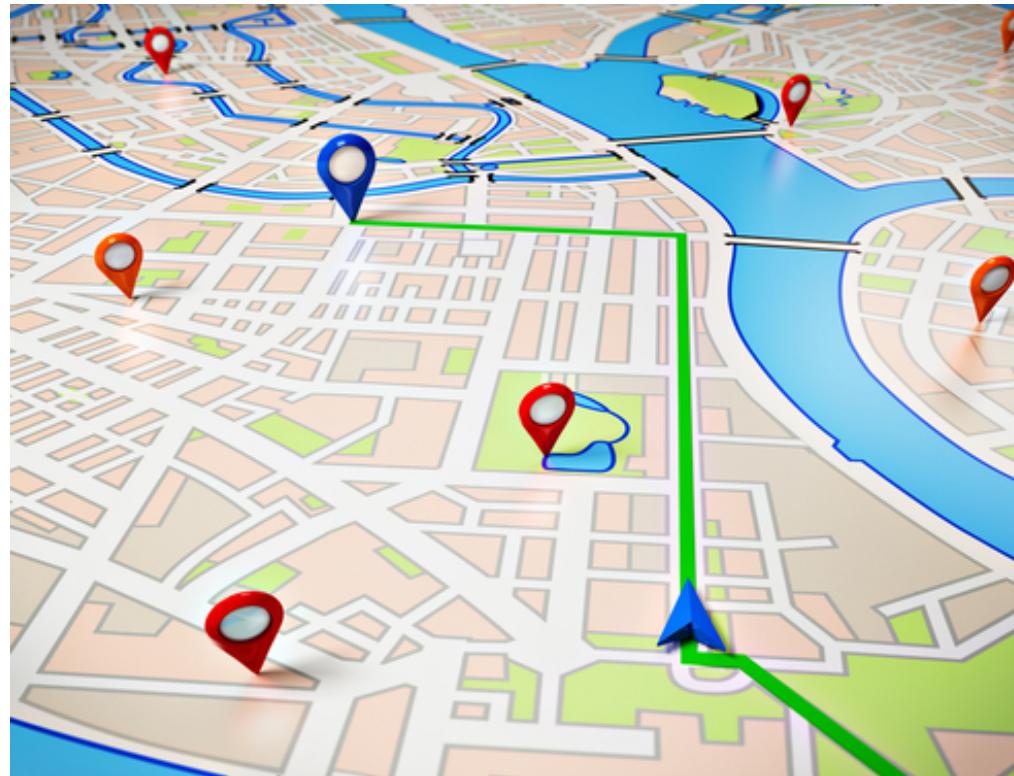
- ▶ ¿Cómo se podría calcular la siguiente distancia?
 - ▶ $d(\text{Juan}, \text{María})$

Nombre	Profesión	Rango Sueldo	Género	Monto Gasto Mensual	ubicación
Juan	Enfermera	500 - 1000	Masculino	30	(-27.98,33.45)
María	Músico	100 - 500	Femenino	10	(-35.65,22.16)

$$d(\text{Juan}, \text{María}) = \text{dif}(\text{Enfermera}; \text{Músico}) + \text{dif}(500/1000; 100/500) + \\ \text{dif}(\text{Femenino}; \text{Masculino}) + \text{dif}(30; 10) + \text{dif}((-27.98, \\ 33.45); (-35.65, 22.16))$$

Ejemplo:

- ▶ Distancia en mapa de la variable ubicación



Ejemplo:

- ▶ Distancia en mapa de la variable ubicación
 - ▶ Entre dos pares numéricos es común usar la distancia de Manhattan o Euclidiana.

(32.51,91.82)

(-23.99,24.25)

Ejemplo:

- ▶ Distancia en mapa de la variable ubicación
 - ▶ Entre dos pares numéricos es común usar la distancia de Manhattan o Euclidiana.

$$d(Ubicacion_{Cliente_1}, Ubicacion_{Cliente_2})$$



$$d((-23.99, 24.25), (32.51, 91.82))$$



$$\sqrt{(32.51 - (-23.99))^2 + (91.82 - 24.25)^2}$$



$$d(Ubicacion_{Cliente_1}, Ubicacion_{Cliente_2}) \approx 88.079$$

Distancia Euclídea

- ▶ Puede ser extensible a D dimensiones

$$d(P, Q) = \sqrt{(P_1 - Q_1)^2 + (P_2 - Q_2)^2 + \cdots + (P_D - Q_D)^2}$$

$$d(P, Q) = \sqrt{\sum_{i=1}^D (P_i - Q_i)^2}$$

Distancia de Manhattan

- ▶ Corresponde a la suma de los valores absolutos de las diferencias de cada coordenada ser extensible a D dimensiones

$$d(P, Q) = \sum_{i=1}^D |P_i - Q_i|$$

Distancia de Manhattan

- ▶ Ejemplo

$$d(P, Q) = \sum_{i=1}^D |P_i - Q_i|$$

$$d(Ubicacion_{Cliente_1}, Ubicacion_{Cliente_2})$$



$$d((-23.99, 24.25), (32.51, 91.82))$$



$$|32.51 - (-23.99)| + |91.82 - 24.25|$$



$$d(Ubicacion_{Cliente_1}, Ubicacion_{Cliente_2}) \approx 124.07$$

Otro tipo de variable

- ▶ Las variables que pueden tomar dos valores son conocidas como variables binarias, y generalmente las transformamos para que tomen los valores 0 o 1



Otro tipo de variable

- ▶ La distancia de Hamming es la más común para las variables binarias
- ▶ Retorna 1 si es que son distintos o 0 si es que son iguales.

$$\text{dist}(0;0) = 0$$

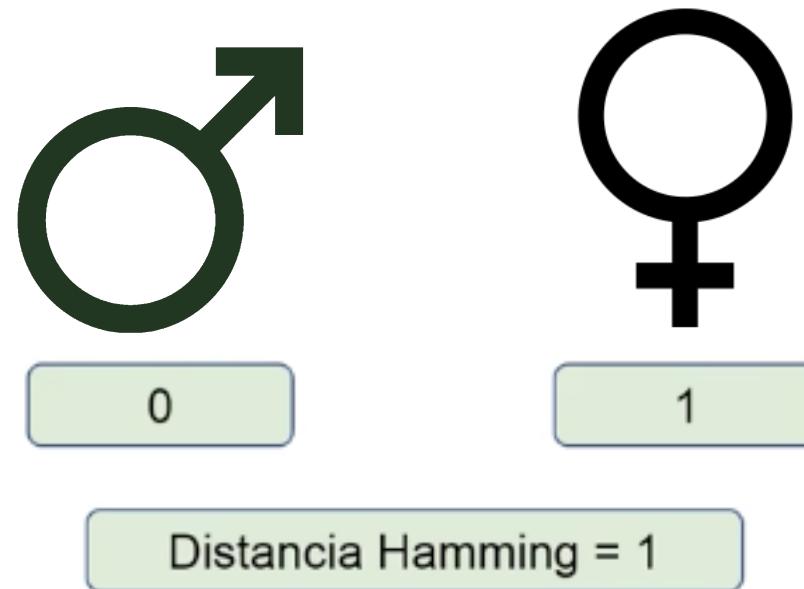
$$\text{dist}(0;1) = 1$$

$$\text{dist}(1;0) = 1$$

$$\text{dist}(1;1) = 0$$

Otro tipo de variable

- ▶ Ejemplo: $d(\text{hombre}, \text{mujer}) = 1$



Otro tipo de variable

- ▶ Ejemplo: $d(\text{mujer}, \text{mujer}) = 0$



1

1

Distancia Hamming = 0

Hamming

- ▶ Hamming para vectores se usa agrupando todas las variables y calculando la distancia de Hamming entre los conjuntos de variables

$$\begin{array}{c} \boxed{10\textcolor{red}{1}1101} \\ \text{y} \end{array} \quad \begin{array}{c} \boxed{10\textcolor{red}{0}0111} \\ = \quad \boxed{3} \end{array}$$

Hamming

- ▶ Ejemplo:

101

Es una **mujer** la cual **no** es
trabajadora independiente y **si**
tiene hijos

010

Es un **hombre** el cual **si** es
trabajador independiente y **no**
tiene hijos

Hamming

- ▶ Ejemplo: la distancia de los clientes sería 3
 $d(101,010) = 3$



101 010

Hamming

- ▶ Se utiliza mucho la normalización para usar valores pequeños.
- ▶ Sería dividir por la cantidad de bits del vector
 - ▶ $d(1011101, 1000111) = 3/7$

$$\boxed{10\textcolor{red}{1}1101} \text{ y } \boxed{10\textcolor{red}{0}01\textcolor{red}{1}1} = \boxed{3/7}$$

Hamming

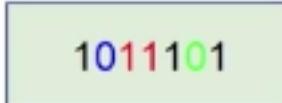
- ▶ La matriz de conteo permite visualizar la distancia entre 2 vectores

C_1  C_2 

$C_1 \setminus C_2$	0	1
0	1	1
1	2	3

Hamming

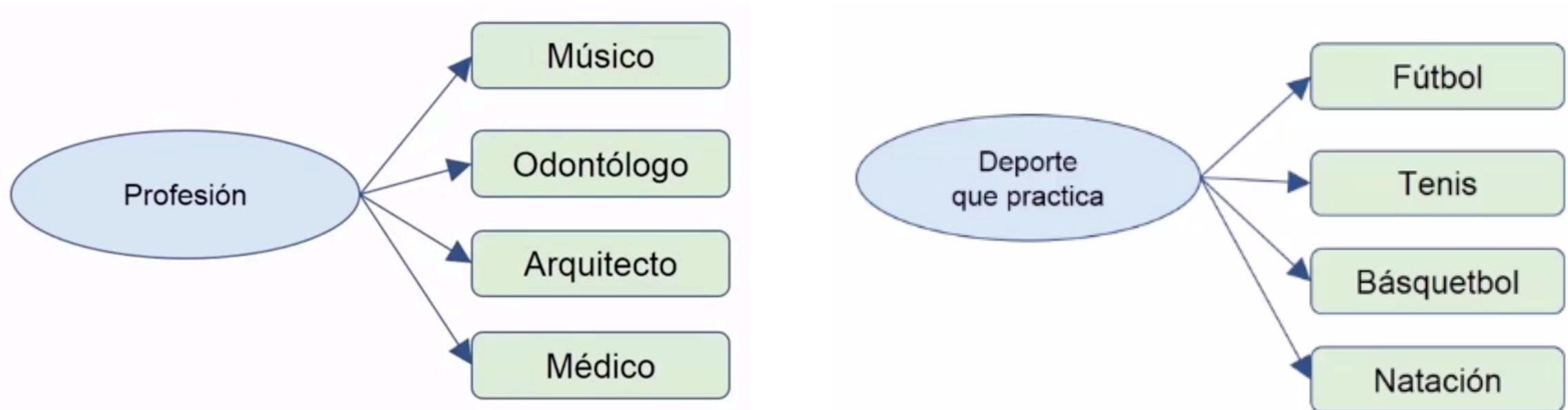
- ▶ La matriz de conteo permite visualizar la distancia entre 2 vectores

C_1  C_2 

$C_1 \setminus C_2$	0	1
0	1	1
1	2	3

Variables categóricas

- ▶ Cantidad fija de valores
- ▶ No tienen un orden definido



Variables categóricas

- ▶ Cálculo de distancia entre variables categóricas

Nombre	Profesión	Ciudad	Deporte
Pedro	Médico	Santiago	Fútbol
Sofía	Músico	Valparaíso	Fútbol

Variables categóricas

- ▶ Se cuenta el número de casos en que las variables son distintas dividido por el total de las variables categóricas.
- ▶ $d(\text{pedro}, \text{sofía}) = 2/3$

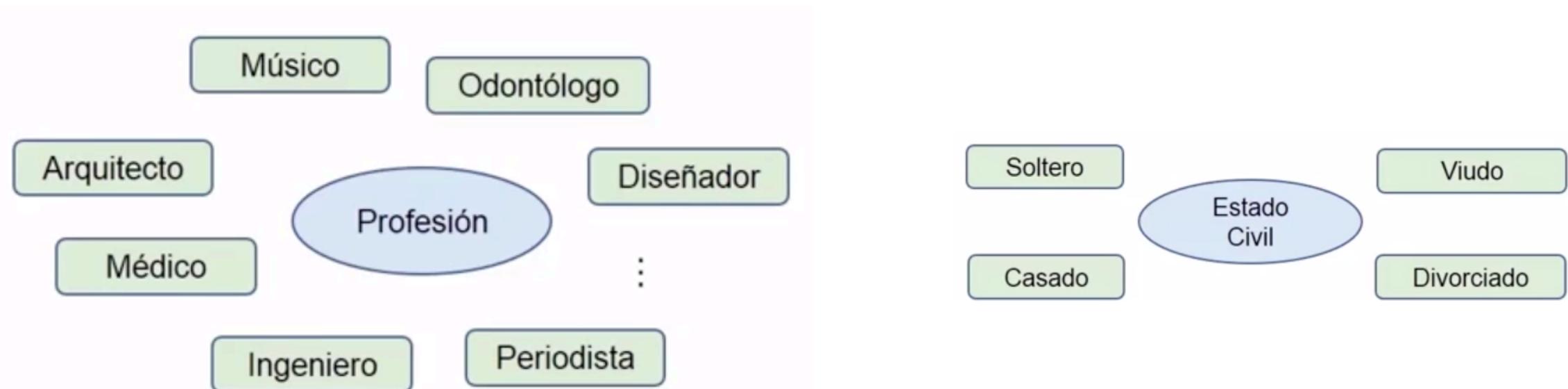
Nombre	Profesión	Ciudad	Deporte
Pedro	Médico	Santiago	Fútbol
Sofia	Músico	Valparaíso	Fútbol

The diagram illustrates the calculation of the distance between two individuals, Pedro and Sofia, across three categorical variables: Profession, City, and Sport. Red ovals highlight the differences between Pedro and Sofia in Profession (Médico vs Músico) and City (Santiago vs Valparaíso). A green oval highlights the similarity in Sport (Fútbol vs Fútbol). Below the table, red symbols ≠ and ≠ indicate the differences in Profession and City, while a green symbol = indicates the similarity in Sport.

≠ ≠ =

Variables categóricas: consideración

- Se debe considerar la cantidad de valores posibles que puede tomar una variable categórica porque mientras más valores la variable puede tomar serán menos probables que distintos objetos tengan el mismo valor en esa variable



Variables categóricas: consideración

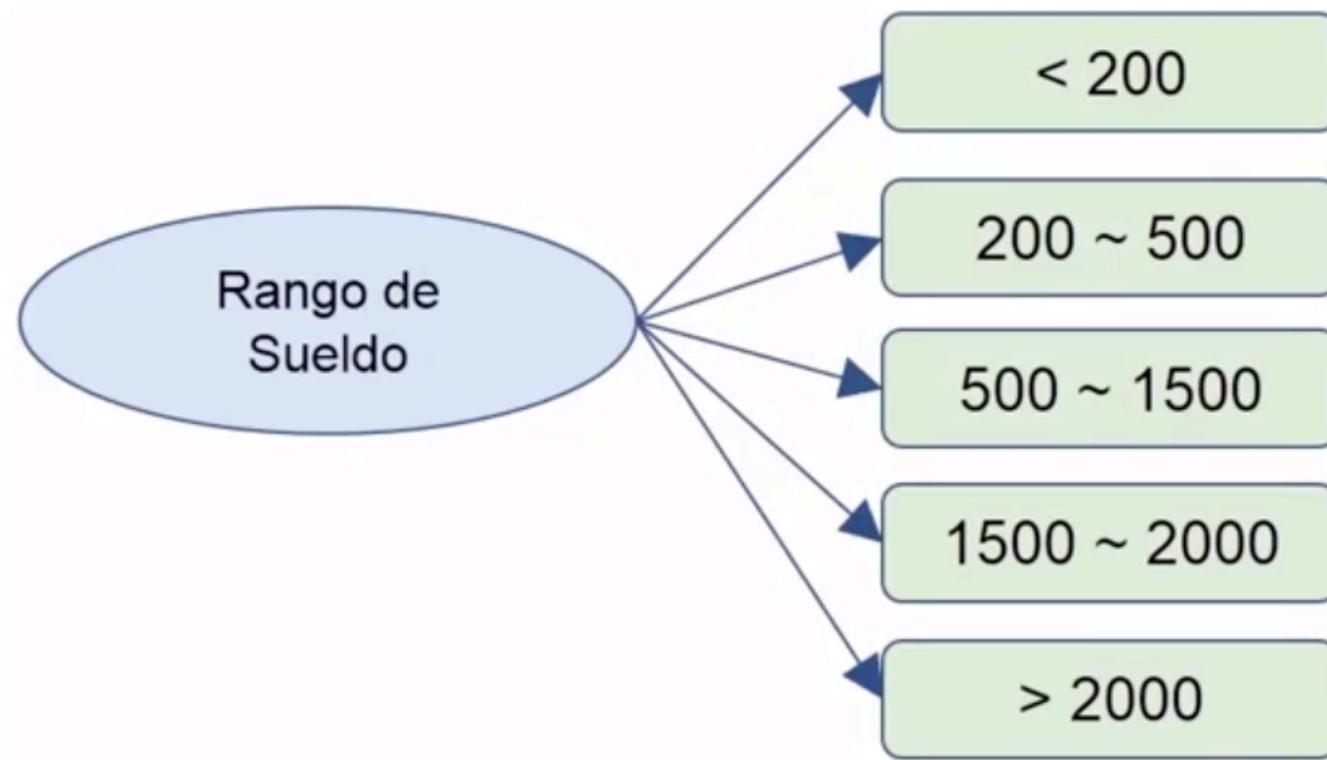
- ▶ Ejemplo: asumir que existen 100 posibles profesiones y 4 posibles estados civiles

Soltero Médico ; Casado Periodista

$$\text{dist}((\text{Soltero}; \text{Médico}) ; (\text{Casado}; \text{Periodista})) = (1/4)*(1) + \\ (1/100)*(1)$$

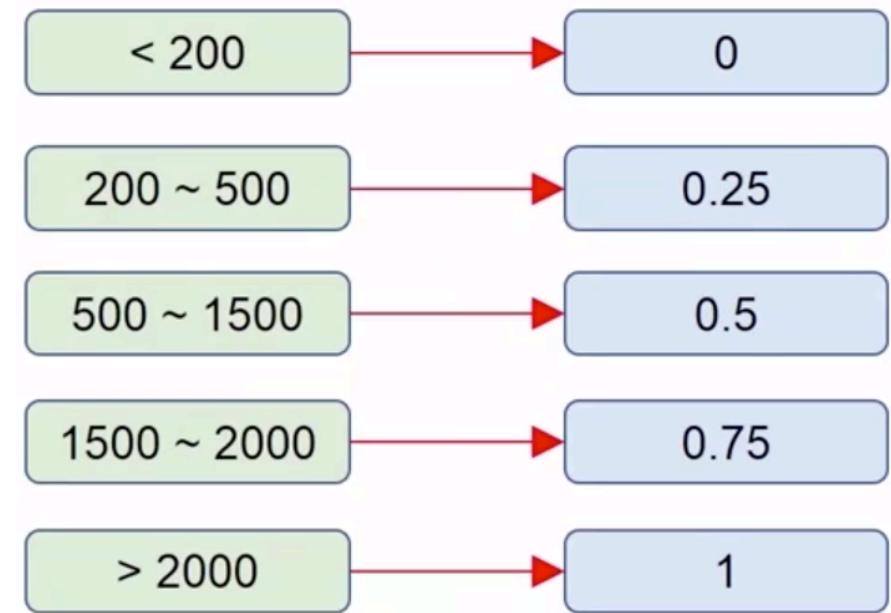
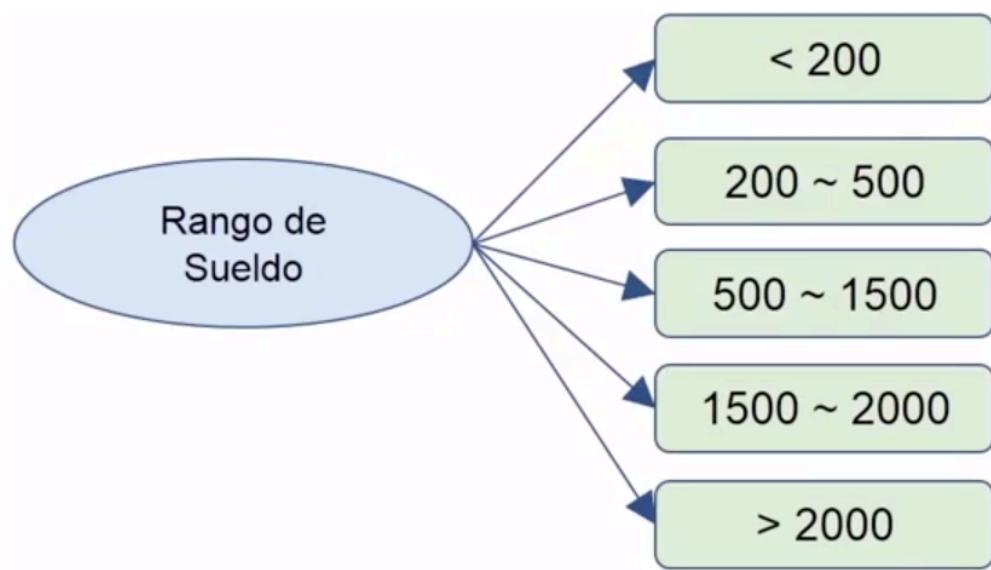
Variables ordinales

- ▶ Tienen un orden definido
- ▶ Tienen un cantidad fija de valores



Variables ordinales

- La forma más común es mapear los valores de mínimo a máximo entre 0 y 1



Ejemplo

Nombre	Posición en empresa	Rango Sueldo
Juan	Gerente	\$ 150
Maria	Analista	\$ 300

Posición en empresa:
{Analista (0), Sub-Gerente (0.5), Gerente (1)}

Rango de Sueldo:
{0-200 (0), 200-500 (0.25), 500-1500 (0.5),
1500-2000 (0.75), >2000 (1)}

Ejemplo

Nombre	Posición en empresa	Rango Sueldo
Juan	Gerente	\$ 150
Maria	Analista	\$ 300

A red circle highlights the "Posición en empresa" column. A red arrow points from this circle down to a second table below.

Juan	1	0
Maria	0	0.25

Ejemplo

Nombre	Posición en empresa	Rango Sueldo
Juan	1	0
Maria	0	0.25

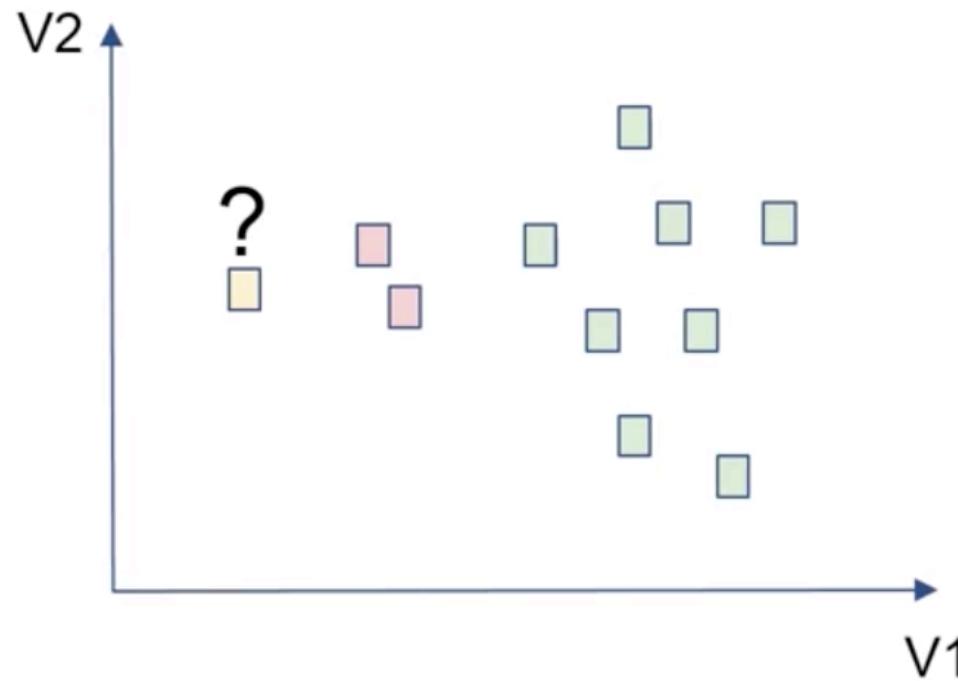
$$d(\text{Juan}, \text{María}) = \sqrt{(0 - 0.25)^2 + (1 - 0)^2} = 1.03$$

Variante del algoritmo de vecinos cercanos

- ▶ considerar la distancia hacia los vecinos a la hora de calcular la clase más votada
- ▶ Amplificar o disminuir ciertas variables al momento de calcular las distancias.

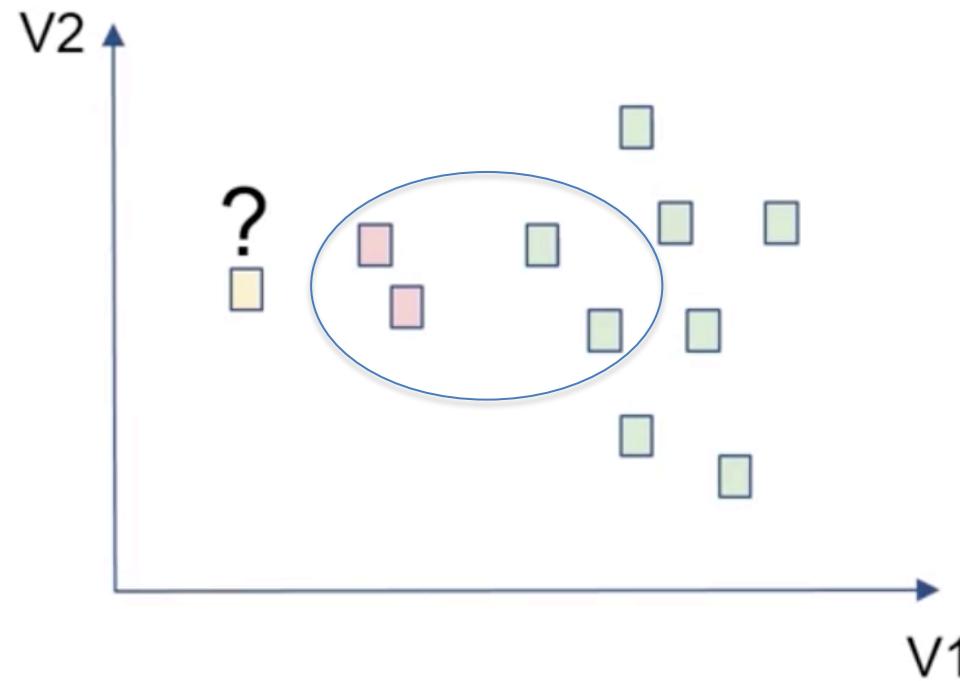
Variante del algoritmo de vecinos cercanos

- ▶ Amplificar o disminuir ciertas variables puede ser muy útil para dar énfasis a elementos más cercanos y menos énfasis a elemento más lejanos



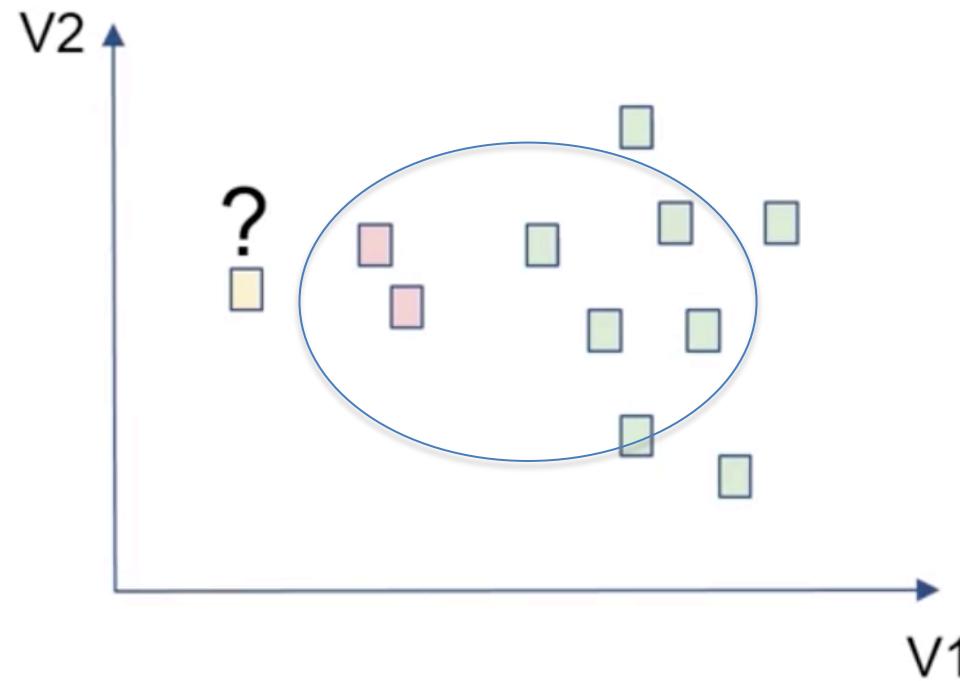
Variante del algoritmo de vecinos cercanos

- ▶ Ejemplo: si $k = 4$
 - ▶ K-nn tradicional marcaría empate, pero vemos que una clasificación correcta sería clase roja



Variante del algoritmo de vecinos cercanos

- ▶ Ejemplo: si $k = 7$
 - ▶ K-nn tradicional marcaría verde, pero vemos que una clasificación correcta sería clase roja



Variante del algoritmo de vecinos cercanos

- ▶ ¿Cómo se puede mejorar el algoritmo?
 - ▶ Usar pesos asociados a las distancias a la hora de calcular la clase más votada entre los vecinos
 - ▶ En general, estos pesos son inversamente proporcionales a la distancia, de tal forma de que si la distancia es más grande, el peso es más pequeño

Variante del algoritmo de vecinos cercanos

multiplicar cada voto por uno dividido por la distancia entre el vecino y el punto que se quiere clasificar.

$$w = \frac{1}{d}$$

$$w = \frac{1}{\sqrt{2\pi}} e^{\frac{-d^2}{2}}$$

decaimiento exponencial, de tal forma de que la penalización por distancia sea mayor mientras más cercano es el vecino, y menor mientras más lejano es el vecino

Variante del algoritmo de vecinos cercanos

- ▶ Finalmente cada vecino entrega su voto con respecto a la clase

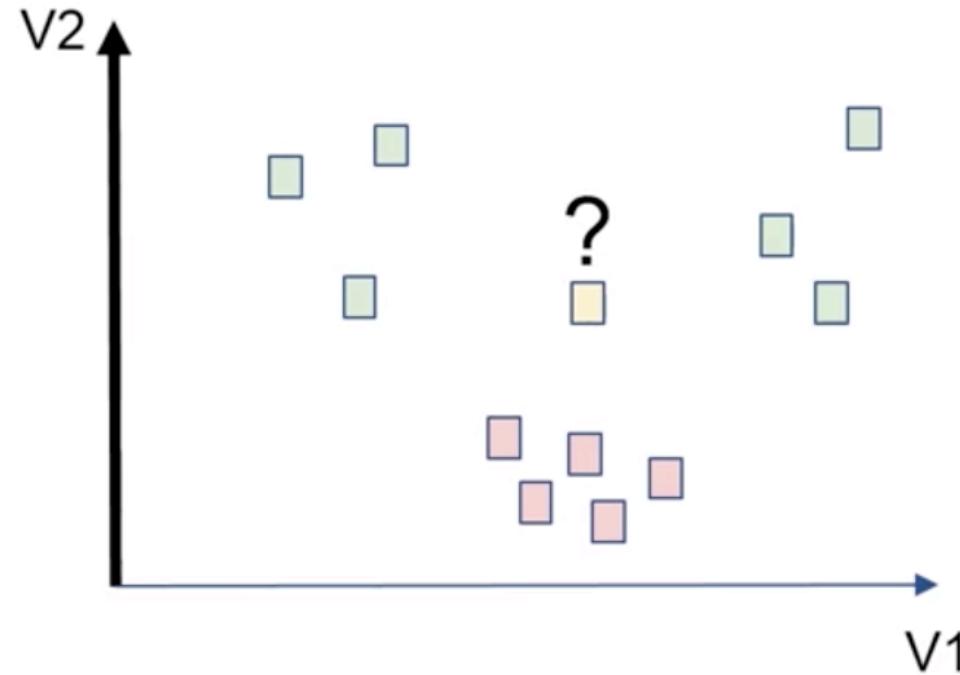
$$Clase(x_q) = \sum_{v=1}^K \hat{w}_v * Clase(x_v)$$

Donde:

$$\hat{w}_v = \frac{w_v}{\sum_{i=1}^k w_i}$$

Variante del algoritmo de vecinos cercanos

- ▶ Importancia de la variables a dimensiones



Variante del algoritmo de vecinos cercanos

- ▶ Poniendo los pesos en la distancia euclíadiana:

$$d(P, Q) = \sqrt{w_1 * (P_1 - Q_1)^2 + w_2 * (P_2 - Q_2)^2}$$

Y en D dimensiones:

$$d(P, Q) = \sqrt{\sum_{i=1}^D w_i * (P_i - Q_i)^2}$$

Ejemplo 1

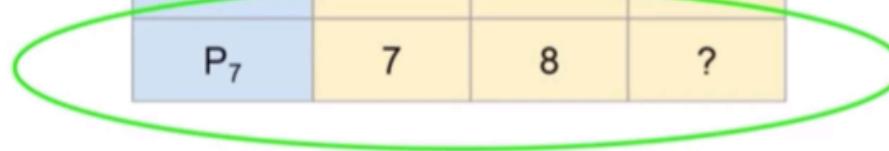
- ▶ Considere los siguientes registros

Puntos	X	Y	Clase
P_1	12	1	A
P_2	4	3	A
P_3	2	9	A
P_4	12	13	B
P_5	21	2	B
P_6	14	3	B

Ejemplo 1

- ▶ Determine a qué clase corresponde P7
- ▶ K = 1
- ▶ K = 3

Puntos	X	Y	Clase
P ₁	12	1	A
P ₂	4	3	A
P ₃	2	9	A
P ₄	12	13	B
P ₅	21	2	B
P ₆	14	3	B
P ₇	7	8	?



Ejemplo 2

- ▶ Considere los siguientes registros

Nombre	Antigüedad	Rango Sueldo	¿Hijos?	Género	Estado Civil	Lugar	Tipo Cliente
Juan	0	< 250	Si	Hombre	Soltero	Chile	Ocasional
Pedro	2	250 - 750	No	Hombre	Casado	Argentina	Frecuente
María	3	>750	Si	Mujer	Viudo	México	Ocasional

Ejemplo 2

- Determine a qué clase corresponde Ángela

Nombre	Antigüedad	Rango Sueldo	¿Hijos?	Género	Estado Civil	Lugar	Tipo Cliente
Juan	0	< 250	Si	Hombre	Soltero	Chile	Ocasional
Pedro	2	250 - 750	No	Hombre	Casado	Argentina	Frecuente
María	3	>750	Si	Mujer	Viudo	México	Ocasional
Angela	3	< 250	Si	Mujer	Casado	México	?

Ejemplo 2: solución

- ▶ Análisis de datos

Numérica

↓

Nombre	Antigüedad	Rango Sueldo	¿Hijos?	Género	Estado Civil	Lugar	Tipo Cliente
Juan	0	< 250	Si	Hombre	Soltero	Chile	Ocasional
Pedro	2	250 - 750	No	Hombre	Casado	Argentina	Frecuente
María	3	>750	Si	Mujer	Viudo	México	Ocasional

Ejemplo 2: solución

- ▶ Análisis de datos

Binarias



Nombre	Antigüedad	Rango Sueldo	¿Hijos?	Género	Estado Civil	Lugar	Tipo Cliente
Juan	0	< 250	Si	Hombre	Soltero	Chile	Ocasional
Pedro	2	250 - 750	No	Hombre	Casado	Argentina	Frecuente
Maria	3	>750	Si	Mujer	Viudo	México	Ocasional

Ejemplo 2: solución

- Análisis de datos

Categóricas

Nombre	Antigüedad	Rango Sueldo	¿Hijos?	Género	Estado Civil	Lugar	Tipo Cliente
Juan	0	< 250	Si	Hombre	Soltero	Chile	Ocasional
Pedro	2	250 - 750	No	Hombre	Casado	Argentina	Frecuente
María	3	>750	Si	Mujer	Viudo	México	Ocasional

Ejemplo 2: solución

- ▶ Análisis de datos

Clase
↓

Nombre	Antigüedad	Rango Sueldo	¿Hijos?	Género	Estado Civil	Lugar	Tipo Cliente
Juan	0	< 250	Si	Hombre	Soltero	Chile	Ocasional
Pedro	2	250 - 750	No	Hombre	Casado	Argentina	Frecuente
Maria	3	>750	Si	Mujer	Viudo	México	Ocasional

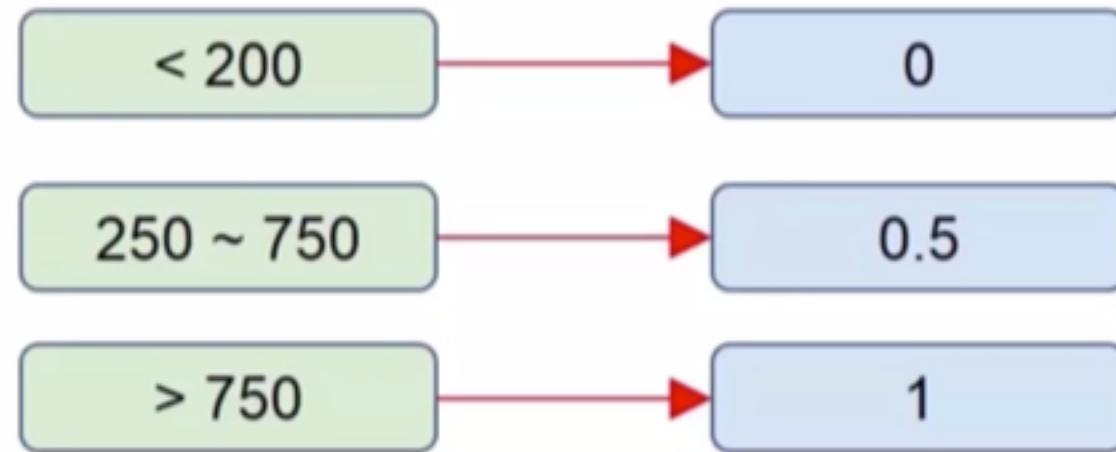
Ejemplo 2: solución

► Transformación

Nombre	Antigüedad	Rango Sueldo	¿Hijos?	Género	Estado Civil	Lugar	Tipo Cliente
Juan	0	< 250	1	0	Soltero	Chile	Ocasional
Pedro	2	250 - 750	0	0	Casado	Argentina	Frecuente
Maria	3	>750	1	1	Viudo	México	Ocasional
Angela	3	< 250	1	1	Casado	México	?

Ejemplo 2: solución

- ▶ Transformación: sueldo



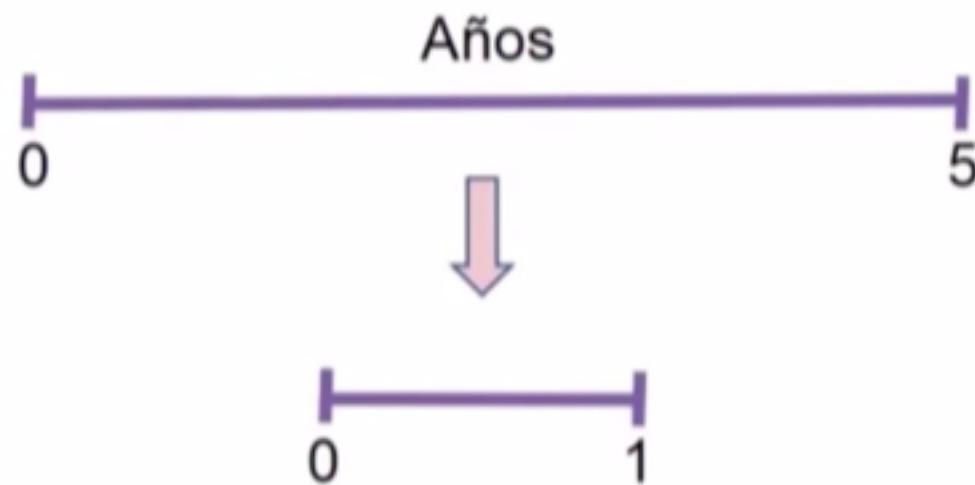
Ejemplo 2: solución

- ▶ Transformación: sueldo

Nombre	Antigüedad	Rango Sueldo	¿Hijos?	Género	Estado Civil	Lugar	Tipo Cliente
Juan	0	0	1	0	Soltero	Chile	Ocasional
Pedro	2	0.5	0	0	Casado	Argentina	Frecuente
Maria	3	1	1	1	Viudo	México	Ocasional
Angela	3	0	1	1	Casado	México	?

Ejemplo 2: solución

- ▶ Transformación: antiguedad (normalización)



Ejemplo 2: solución

- ▶ Normalización permite dar misma importancia a cada variable
- ▶ Ejemplo: esto no es deseable

$$d(P, Q) = \sqrt{(P_1 - Q_1)^2 + (P_2 - Q_2)^2 + \cdots + (P_D - Q_D)^2}$$

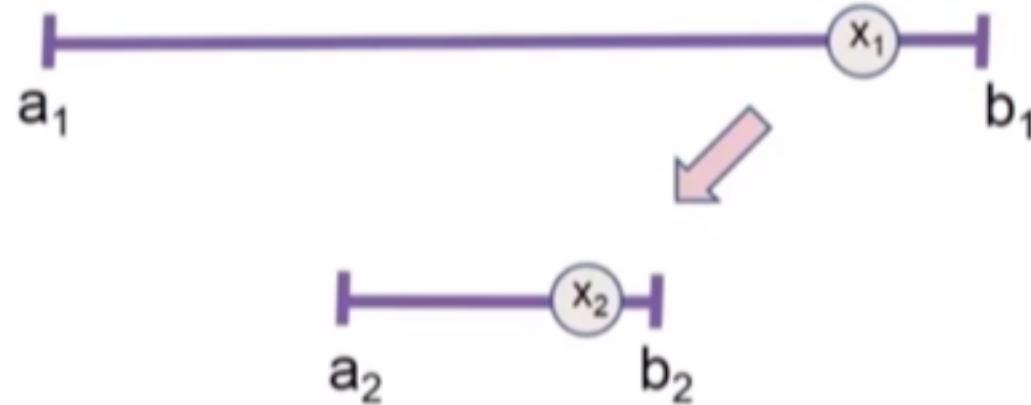
[0-1000] [0-1]

- ▶ Ejemplo: eso si es deseable

[0-1000]
↓
[0-1]

Ejemplo 2: solución

- ▶ Normalización : en general



$$x_2 = a_2 + \frac{x_1 - a_1}{b_1 - a_1} (b_2 - a_2)$$

Ejemplo 2: solución

- ▶ Normalización : en nuestro ejemplo

$$a_1 = 0, b_1 = 5$$

$$a_2 = 0, b_2 = 1$$

$$x_2 = a_2 + \frac{x_1 - a_1}{b_1 - a_1} (b_2 - a_2)$$

$$x_2 = 0 + \frac{x_1 - 0}{5 - 0} * (1 - 0) = \frac{x_1}{5}$$

Ejemplo 2: solución

- ▶ Normalización : en nuestro ejemplo

Nombre	Antigüedad	Rango Sueldo	¿Hijos?	Género	Estado Civil	Lugar	Tipo Cliente
Juan	0	0	1	0	Soltero	Chile	Ocasional
Pedro	0.4	0.5	0	0	Casado	Argentina	Frecuente
María	0.6	1	1	1	Viudo	México	Ocasional
Angela	0.6	0	1	1	Casado	México	?

Ejemplo 2: solución

- ▶ Estado civil y lugar

The diagram illustrates a decision tree for customer segmentation. The root node branches into two main categories: "4 estados civiles" (4 civil states) and "10 países posibles" (10 possible countries). These categories correspond to the "Estado Civil" and "Lugar" columns in the table below.

Nombre	Antigüedad	Rango Sueldo	¿Hijos?	Género	Estado Civil	Lugar	Tipo Cliente
Juan	0	< 250	Si	Hombre	Soltero	Chile	Ocasional
Pedro	2	250 - 750	No	Hombre	Casado	Argentina	Frecuente
Maria	3	>750	Si	Mujer	Viudo	México	Ocasional

Ejemplo 2: solución

- ▶ Cálculos: $d(\text{Pedro}, \text{Ángela})$

Nombre	Antigüedad	Rango Sueldo	¿Hijos?	Género	Estado Civil	Lugar	Tipo Cliente
Pedro	0.4	0.5	0	0	Casado	Argentina	Frecuente
Angela	0.6	0	1	1	Casado	México	?

Ejemplo 2: solución

Nombre	Antigüedad	Rango Sueldo	¿Hijos?	Género	Estado Civil	Lugar	Tipo Cliente
Pedro	0.4	0.5	0	0	Casado	Argentina	Frecuente
Angela	0.6	0	1	1	Casado	México	?



$$0.6 - 0.4$$

$$0.2$$



$$0.5 - 0$$

$$0.5$$



$$1$$



$$1$$



$$0 * 1/4$$



$$1 * 1/10$$

$d(\text{Pedro}, \text{Angela})$

$$\begin{aligned}
 &= \text{dif}(0.4; 0.6) + \text{dif}(0.5; 0) + \text{dif}(0; 1) + \text{dif}(0; 1) + \\
 &\quad \text{dif}(\text{Casado}; \text{Casado}) * 1/4 +
 \end{aligned}$$

$\text{dif}(\text{Argentina}; \text{México}) * 1/10$

$$= 0.2 + 0.5 + 1 + 1 + 0/4 + 1/10$$

$$= 2.8$$

Ejemplo 2: solución

- ▶ Resultados: $k = 1$

$$d(Juan, Angela) = 1.95$$

$$d(Pedro, Angela) = 2.8$$

$$d(María, Angela) = 1.25$$

- ▶ Matriz de distancia

	Juan	Pedro	María
Angela	1.95	2.8	1.25

Ejemplo 2: solución

- ▶ Resultados

$$d(Juan, Angela) = 1.95$$

$$d(Pedro, Angela) = 2.8$$

$$d(María, Angela) = 1.25$$

- ▶ Matriz de distancia

	Juan	Pedro	María
Angela	1.95	2.8	1.25

Menor distancia

Ejemplo 2: solución

- Matriz de distancia

	Juan	Pedro	María
Angela	1.95	2.8	1.25



Menor distancia

Nombre	Antigüedad	Rango Sueldo	¿Hijos?	Género	Estado Civil	Lugar	Tipo Cliente
Angela	0.6	0	1	1	Casado	México	Ocasional

REFERENCIAS

- ▶ Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
- ▶ Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
- ▶ Karim Pichara, curso minería de datos.
- ▶ Hand, D. J. (2006). Data Mining. *Encyclopedia of Environmetrics*, 2.