5.1.- Introducción a problemas de clasificación

5.2.- Enfoque basado en la Teoría de la Decisión

5.3.- Técnicas básicas de clasificación

5.3.1.- Clasificadores bayesianos

5.3.2.- Clasificadores basados en distribuciones normales

5.3.3.- Clasificador k-vecinos más cercanos

5.4.- Desarrollo de sistema de reconocimiento de objetos

5.4.1.- Análisis y pre-procesamiento de datos

5.4.2.- Selección de características

5.4.3.- Evaluación de modelos

FILTRO AUTOMÁTICO DE CORREO ELECTRÓNICO ANTISPAM

Objetivo: diseñar un sistema inteligente que etiquete los correos electrónicos que llegan en dos categorías, «SPAM» o «LEGÍTIMO»

→ ¿ PROBLEMA DE CLASIFICACIÓN O REGRESIÓN ?

CLASIFICACIÓN: se predice una categoría (no el valor de una variable numérica – ejemplo: cuál va a ser el precio de un artículo, o el número de reservas que se harán en mayo en un hotel – PROBLEMA DE REGRESIÓN).

→ **CLASIFICACIÓN SUPERVISADA (datos históricos etiquetados)**

PROCESO DE PREDICCIÓN:

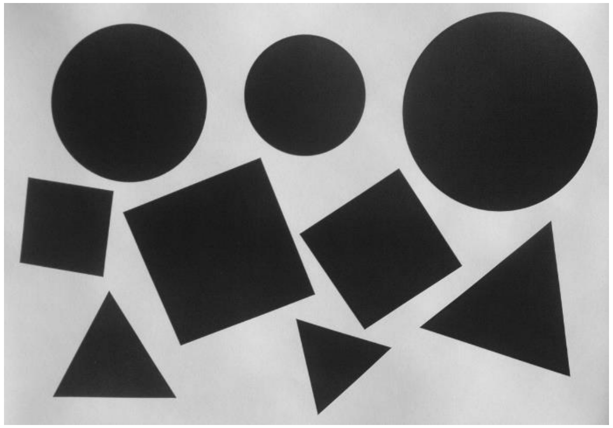
- **1.- FUENTE DE DATOS:** hay que disponer de datos etiquetados, un conjunto de correos electrónicos etiquetados como «spam» o «legítimo» por los usuarios.

- **2.- DEFINIR Y EXTRAER ATRIBUTOS DE LOS DATOS: establecer y obtener características o propiedades de los datos que permitan discriminar correos de ambas categorías.** Ejemplo: dirección IP, dominio de la cuenta, contenido de ciertas palabras, tipo de archivos adjuntos, lugar de origen, tipo de caracteres o alfabeto utilizado etc…

- **3.- FASE DE APRENDIZAJE – ENTRENAMIENTO DEL MODELO: detección de patrones que tienen los correos marcados con ambas etiquetas.** Ejemplo: un correo es «spam» si viene de determinadas direcciones IP, y además contiene ciertas palabras, o presenta archivos adjuntos de ciertos tipos o tamaños, etc.

- **4.- FASE DE APLICACIÓN DEL MODELO – INFERENCIA: una vez entrenado el modelo, se utiliza para clasificar los correos nuevos, no marcados por usuarios, en función de sus características, como «spam» o «legítimo»**

EJEMPLO DE PROBLEMA DE CLASIFICACÍÓN



"Dado un objeto en una imagen, reconocer su forma geométrica"

Conocimiento a priori: el objeto únicamente puede presentar forma circular, cuadrada o triangular

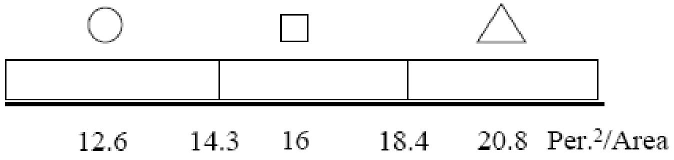
Planteamiento: cálculo de atributos o descriptores matemáticos que sean representativos de cada forma geométrica

- Ejemplo de descriptores matemáticos:
  - Área
  - Perímetro
  - Relación perímetro^2 y área

Análisis teórico de la idoneidad del descriptor

	Área	Perímetro	Perímetro^2 / Área
O	πr^2	$2\pi r$	**12.56**
□	l^2	$4l$	**16**
Δ	$\sqrt{3}\,l^2/4$	$3l$	**20.8**

- Clasificador:
  - Regla de decisión
    - Si per.^{2} / área < 14.3 entonces objeto = círculo.
    - Si 14.3 < per.^{2} / área < 18.4 entonces objeto = cuadrado
    - Si 18.4 < per.^{2} / área entonces objeto = triángulo



EJEMPLO DE PROB LEMA DE CLASIFICACÍÓN

Dado un objeto en una imagen, reconocer su forma geométrica

FUENTE DE DATOS: IMAGEN

ADQUISICIÓN DE LA IMAGEN

⬇

PREPROCESAMIENTO

⬇

EXTRACCIÓN DE REGIONES DE INTERÉS: SEGMENTACIÓN

⬇

DESCRIPCIÓN MATEMÁTICA Y RECONOCIMIENTO DE OBJETOS

1. **Calcular perímetro a partir del borde previamente detectado del objeto.**
2. **Calcular área a partir de la segmentación previa del objeto.**
3. **Calcular descriptor: perímetro^2 / área,**
4. **Aplicar el clasificador diseñado en base al conocimiento a priori del problema.**

- Clasificador basado en Regla de decisión:
  - Si per.^{2} / área < 14.3 entonces objeto = círculo.
  - Si $14.3 <$ per.^{2} / área < 18.4 entonces objeto = cuadrado
  - Si $18.4 <$ per.^{2} / área entonces objeto = triángulo

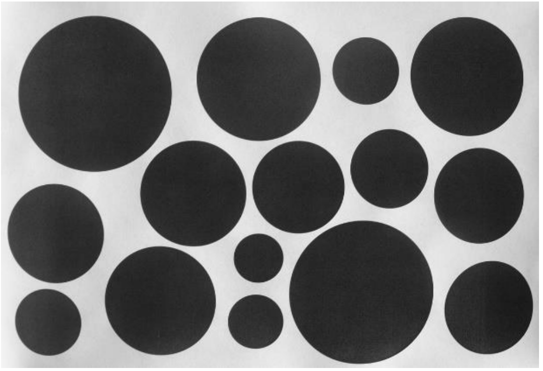
ATRIBUTOS Y CLASIFICADOR: PLANTEAMIENTO TEÓRICO

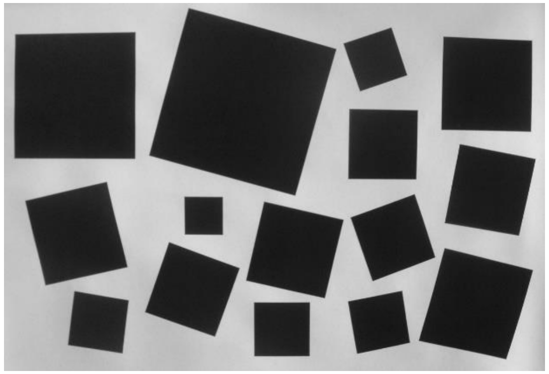
SELECCIÓN DE ATRIBUTOS Y ENTRENAMIENTO DEL CLASIFICADOR

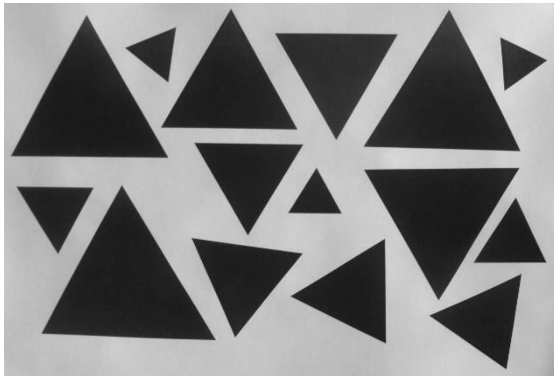
A PARTIR DE DATOS SIMILARES Y REPRESENTATIVOS DE LA SITUACIÓN REAL DE FUNCIONAMIENTO DEL CLASIFICADOR

EJEMPLO ANTERIOR: dado un objeto en una imagen, reconocer su forma geométrica

DISPONIBILIDAD DE IMÁGENES DE PRUEBA







EJEMPLO DE IMÁGENES DE PRUEBA PARA LA SELECCIÓN DE ATRIBUTOS Y DISEÑO DEL CLASIFICADOR

- ❖ **OBTENCIÓN DE DATOS: DATOS DE ATRIBUTOS O DESCRIPTORES MATEMÁTICOS «PROMETEDORES» SOBRE MUESTRAS DE FORMA GEOMÉTRICA ES CONOCIDA**
- ❖ **SELECCIÓN DE DESCRIPTORES ADECUADOS**
- ❖ **ELECCIÓN DE ESTRATEGIA DE CLASIFICACIÓN Y DISEÑO DE CLASIFICADOR**

PROBLEMA DE CLASIFICACÍÓN: TERMINOLOGÍA Y DEFINICIONES

- **Clases:** salidas del modelo de predicción, categorías o grupos representativos en los que se quieren clasificar los datos.

- **Instancia, ejemplo o registro (instance, sample, record):** cada una de las muestras disponibles para entrenar/evaluar un modelo (en los ejemplos anteriores, cada uno de los correos electrónicos; cada objeto cuadrado, circular o triangular).

- **Característica, atributo, propiedad o campo (feature, attribute, property, field):** cada instancia se describe por medio de un conjunto de atributos. Al conjunto de atributos que definen una instancia se le denomina vector de atributos o características (en los ejemplos anteriores, cada una de las medidas que se utilizan para describir un correo electrónico; relación perímetro^2/area).

- **Espacio de características (feature space):** espacio definido por cada uno de los atributos que componen el vector de atributos. En este espacio, cada instancia se representa mediante un punto cuyas coordenadas están definidas por los valores que tienen los atributos de dicha instancia.

- **Conjunto de datos (dataset):** el conjunto de datos está formado por instancias; cada instancia se compone de los valores de los atributos que conforman el vector de atributos. Además, en aprendizaje supervisado, cada instancia está etiquetada con la codificación asignada a la clase a la que pertenece.

    - **Conjunto de entrenamiento (patrones de entrenamiento):** subconjunto de datos utilizados en la fase de aprendizaje para el diseño y entrenamiento del modelo (en ocasiones este conjunto se subdivide en entrenamiento + validación).

    - **Conjunto de test:** subconjunto de datos utilizados en la evaluación del modelo entrenado.

EJEMPLO: CONCESIÓN DE CRÉDITOS BANCARIOS

- Concesión de créditos bancarios
  - Un banco por Internet desea obtener reglas para predecir qué personas de las que solicitan un crédito no van a devolverlo.
  - La entidad bancaria cuenta con una gran base de datos correspondientes a los créditos concedidos (o no) a otros clientes con anterioridad.
  - Instancias (de la base de datos del banco):
    - Atributos de entrada: años del crédito, cuantía del crédito, tiene cuentas morosas, tiene casa propia
    - Clase: si/no
  - Modelo que se podría aprender:
    - **SI** (cuentas-morosas > 0) **ENTONCES** Devuelve-crédito = no
    - **SI** (cuentas-morosas = 0) **Y** ((salario > 2500) **O** (años > 10)) **ENTONCES** devuelve-crédito = si

EJEMPLO: CONCESIÓN DE CRÉDITOS BANCARIOS

Instancia de test

Años	Euros	Salario	Casa propia	Cuentas morosas	Crédito
10	50000	3000	Si	0	??

T = Conjunto de instancias de entrenamiento (o ejemplos, datos, patrones, …)

Años	Euros	Salario	Casa propia	Cuentas morosas	Crédito
15	60000	2200	Si	2	No
2	30000	3500	Si	0	Si
9	9000	1700	Si	1	No
15	18000	1900	No	0	Si
10	24000	2100	No	0	No
	…	…	…	…	…

Debido a esta columna, la tarea es supervisada

x: atributos (o características, predictores, variables independientes, variables de entrada, …)

y: clase (o etiqueta, atributo de salida, variable dependiente, variable respuesta, …)

Algoritmo

Modelo

IF CM >0 **THEN** NO

IF CM =0 **Y** S>2500 **THEN** SI

…

Crédito = Si

EJEMPLO

Sistema que produce objetos en madera para su posterior decoración. Estos objetos deben clasificarse como pertenecientes a la clase Peras o Manzanas:

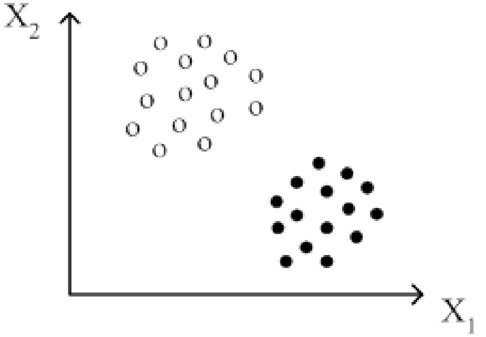
→ **Clases: clase Manzanas (C_m) y de la clase Peras (C_p)**

→ **Atributos empleados para caracterizar cada objeto: compacticidad y excentricidad.**

Vector de atributos o características (describen cada instancia, cada objeto): $\vec{x} = (x_1 , x_2) \Rightarrow \begin{cases} x_1 \equiv \text{compacticidad} \\ x_2 \equiv \text{excentricidad} \end{cases}$

→ **Patrones de entrenamiento:**

Se tienen disponibles para crear el modelo, 15 objetos «pera» y 15 objetos «manzana». De cada uno de ellos (de cada instancia), se obtienen sus atributos (valores de x1 y x2), que pueden ser representados en el espacio de características (espacio de dos dimensiones definidos por x1 y x2).

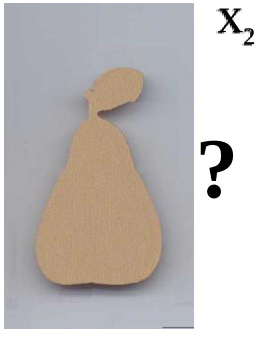


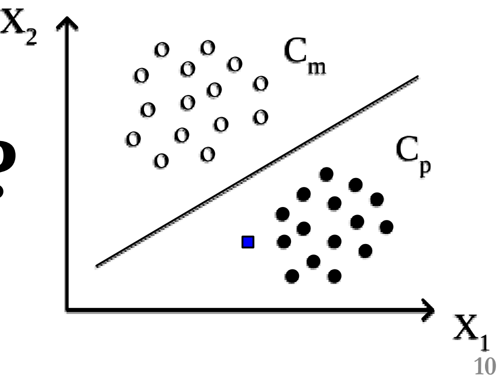
→ **Planteamiento del modelo de clasificación:**

Si los atributos son representativos de cada clase, los puntos de las muestras de cada clase tienden a agruparse en regiones diferenciadas del espacio de características → Planteamiento: división del espacio de características en regiones correspondientes a las distintas clases bajo consideración.

→ **Entrenamiento del modelo (fase de aprendizaje): diseño de una función (en el ejemplo: línea recta – clasificador lineal) que divide el espacio de características en regiones que corresponden a cada una de las clases del problema.**

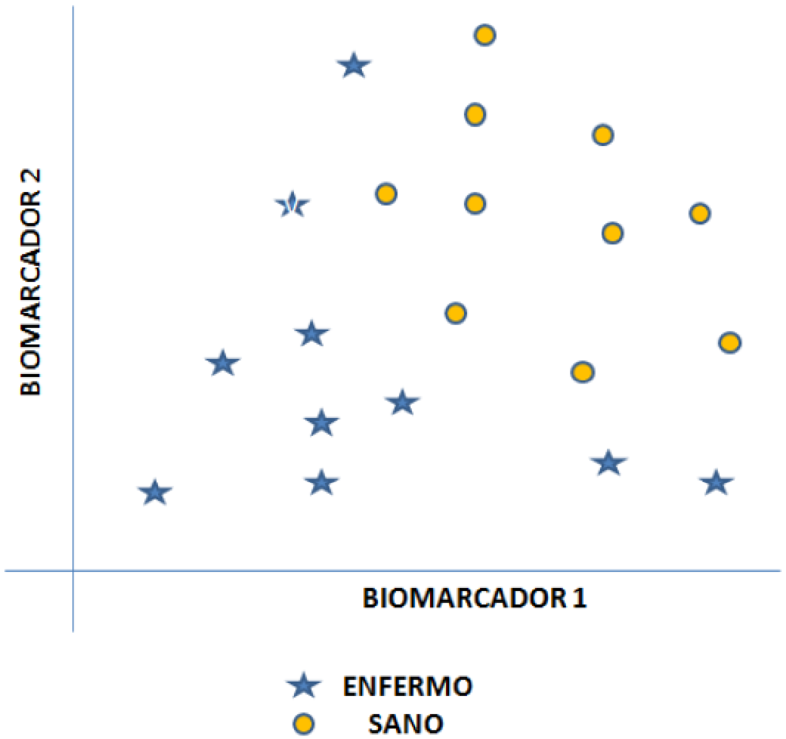
→ **Aplicación del modelo (fase de inferencia): de un objeto de clase desconocida, se mide (x_1 , x_2) y se evalúa el modelo que decidirá la clase dependiendo de la región donde se encuentre.**





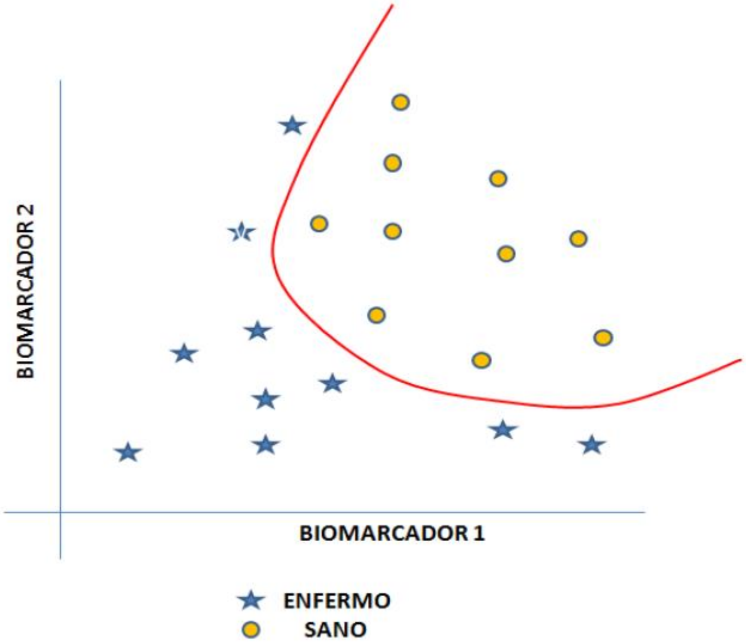
EJEMPLO

Sistema de diagnóstico



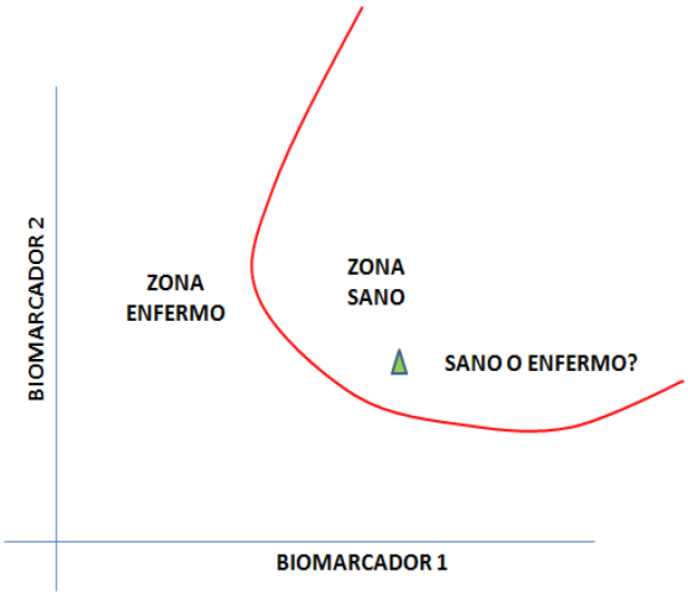
EJEMPLO

Sistema de diagnóstico

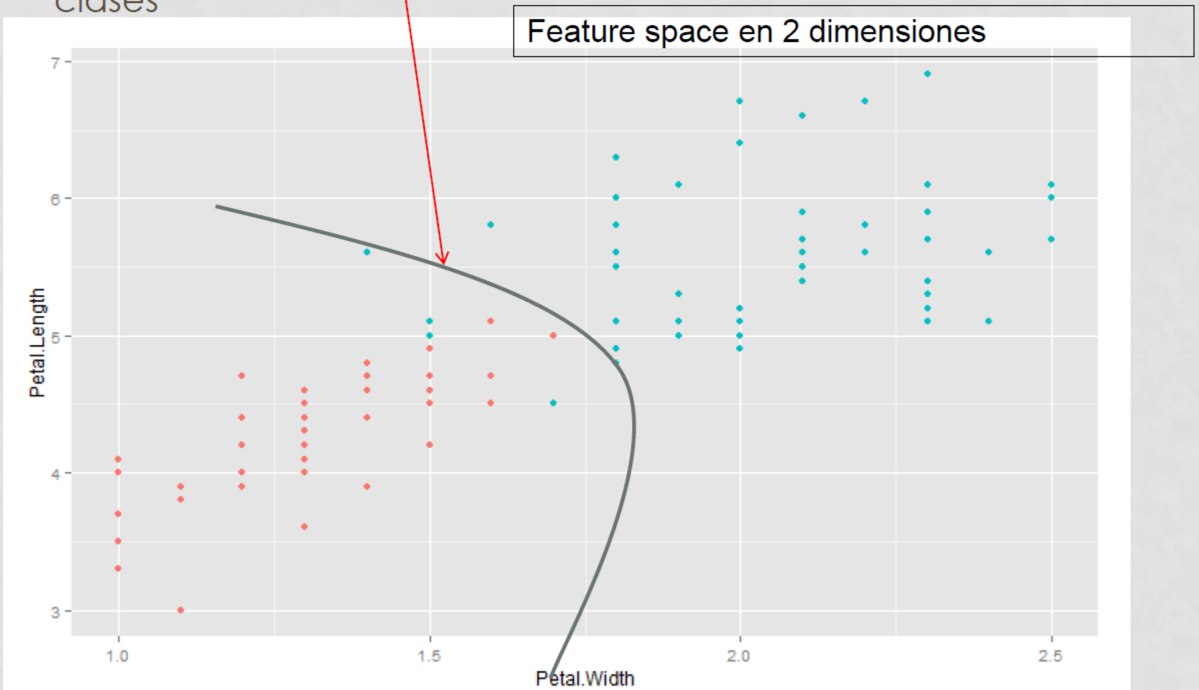


EJEMPLO

Sistema de diagnóstico

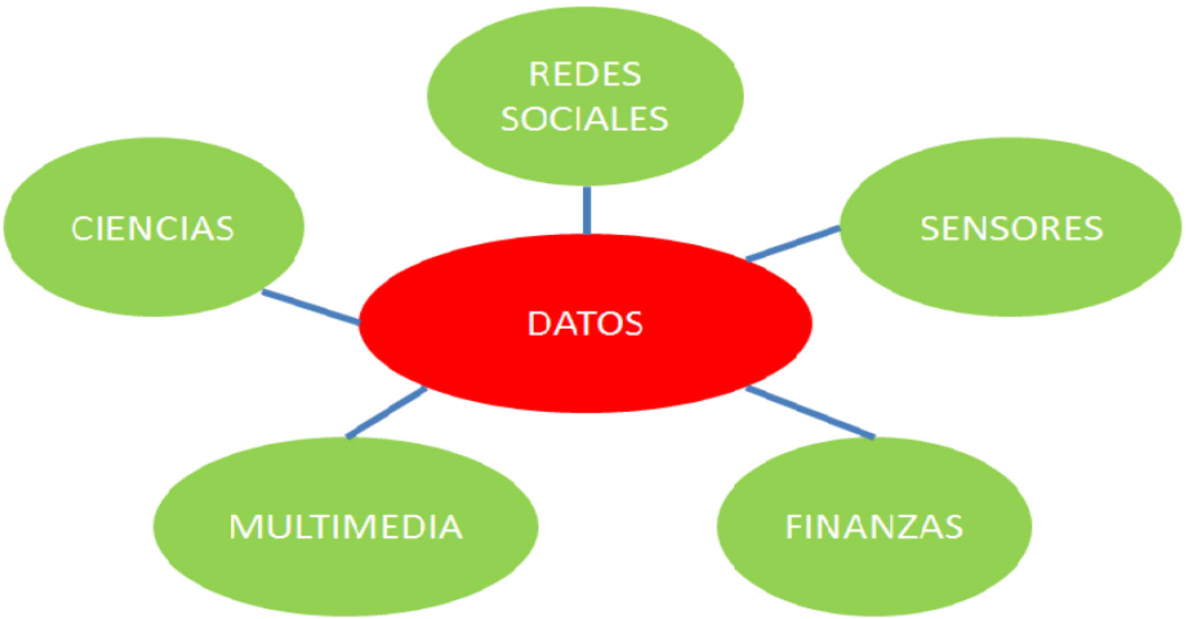


- Ejemplo: clasificar plantas en dos clases ("versicolor" / roja vs. "virginica" / azul)
- 2 atributos = (Petal.Width, Petal.Length) = 2 dimensiones
- Clasificación = encontrar una función **g: X → Y** frontera entre las clases



CONSIDERACIONES: Fuente de datos

- **Diversidad en la naturaleza de los datos**



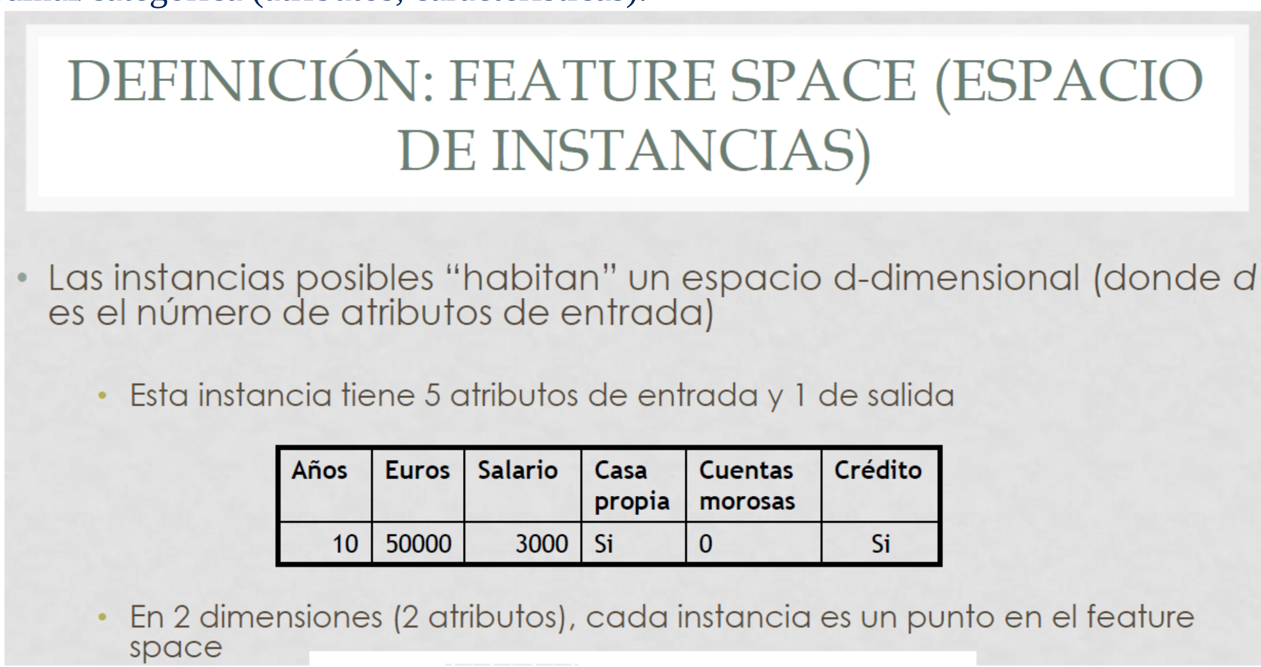
CONSIDERACIONES: Planteamiento Metodológico Común

- **Clasificación Supervisada:**

1. ***Definición de grupos representativos o clases:*** **posibles tipos de objetos que, según el conocimiento a priori del problema, se espera puedan aparecer.**

2. ***Creación de conjunto de datos***

   2.1- EXTRACCIÓN DE ATRIBUTOS DE CADA INSTANCIA DISPONIBLE: de las muestras disponibles de cada clase del problema, se extraen o calculan propiedades de naturaleza cuantitativa/ordinal/categórica (atributos, características).



CONSIDERACIONES: Planteamiento Metodológico Común

- **Clasificación Supervisada:**

1. ***Definición de grupos representativos o clases:*** **posibles tipos de objetos que, según el conocimiento a priori del problema, se espera puedan aparecer.**

2. ***Creación de conjunto de datos*:**

   2.1- EXTRACCIÓN DE ATRIBUTOS DE CADA INSTANCIA DISPONIBLE: de las muestras disponibles de cada clase del problema, se extraen o calculan propiedades de naturaleza cuantitativa/ordinal/categórica (atributos, características).

- Tipos de atributos:
  - Nominales / categóricos: verde, rojo, amarillo
  - Ordinales: frío, templado, caliente
  - Reales / enteros: 1.3, 7.9, 10.798, …
- $Y = \{c_1, c_2, \ldots, c_k\}$ son las clases.
  - Si k=2, problema de clasificación binaria: cáncer / no-cáncer
  - Si K>2, problema de clasificación multi-clase: peligroso / normal / inofensivo

CONSIDERACIONES: Planteamiento Metodológico Común

- **Clasificación Supervisada:**

1. ***Definición de grupos representativos o clases:*** **posibles tipos de objetos que, según el conocimiento a priori del problema, se espera puedan aparecer.**

2. ***Creación de conjunto de datos*:**

   2.1- EXTRACCIÓN DE ATRIBUTOS DE CADA INSTANCIA DISPONIBLE: de las muestras disponibles de cada clase del problema, se extraen o calculan propiedades de naturaleza cuantitativa/ordinal/categórica (atributos, características).

   2.2- SELECCIÓN DE ATRIBUTOS - DEFINICIÓN DE UN VECTOR DE ATRIBUTOS: de todos los atributos, se seleccionan y extraen aquellos que describirán finalmente a las muestras involucradas en el problema de clasificación. Este conjunto de datos de clase conocida, constituye el conjunto de datos de entrenamiento que se utilizará para diseñar (conjunto de entrenamiento) y evaluar (conjunto de test) el modelo de clasificación.

   ⇒ Deben ser, **idealmente**, discriminantes de las diferentes clases de interés e invariantes a todas sus posibles versiones (p. ej. cambios en posición, tamaño, orientación, intensidad de color, timbre o velocidad de la voz, etc.).

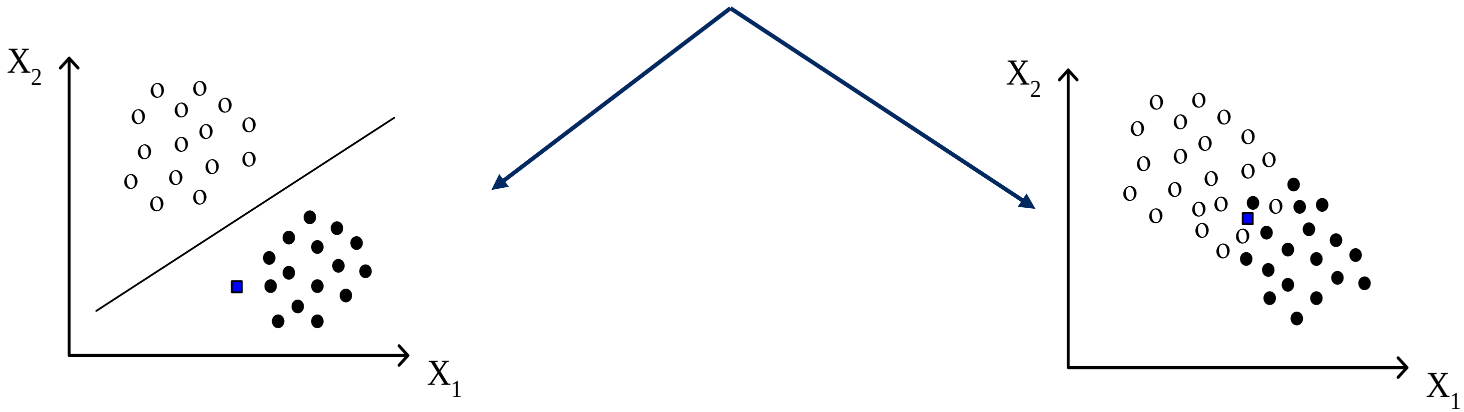
   ⇒ Un conjunto de propiedades de mala calidad produce un solapamiento de clases y por tanto una gran probabilidad de error en la clasificación.

¡¡¡ IMPORTANCIA DE SELECCIÓN DE CARACTERÍSTICAS ADECUADAS !!!

*** OBJETIVO: Proporcionar la mayor separabilidad posible entre las muestras de las clases**

⇒ Ejemplo: dos atributos y dos clases.

⇒ Si tras calcular las dos características en las instancias disponibles de cada clase:



- Las características separan bien las dos clases.

- Los valores de las características de los objetos presentan cierta tolerancia o están entre cierto rango.

- Clasificador sencillo.

- Aunque se observa una cierta separación entre clases, crear una frontera de decisión se hace más complejo.

- Solución: buscar nuevas características o hacer uso de un enfoque estadístico del problema.

CONCLUSIÓN: cuando el conjunto de propiedades obtenidas de las instancias es suficientemente discriminante, la complejidad del clasificador se reduce sensiblemente. En caso contrario, un diseño correcto del clasificador contribuiría, al menos, a disminuir la proporción de errores.

CONSIDERACIONES: Planteamiento Metodológico Común

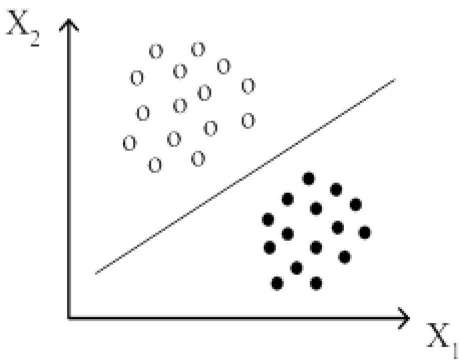
Requisitos de atributos

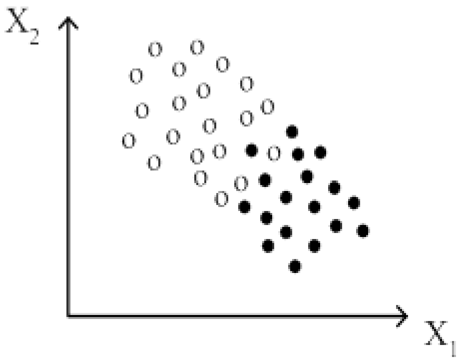
➢ **Poder de discriminación: las características a escoger deben tomar valores distintos para muestras que pertenecen a clases diferentes.**

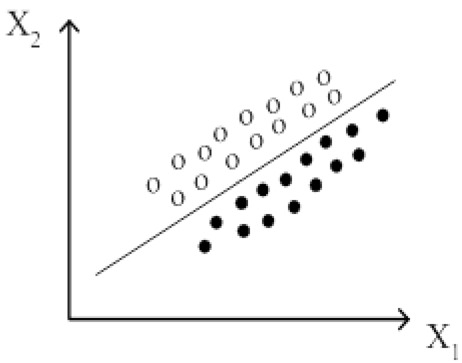
- ❑ **Sensibles: deben reflejar diferencias para muestras «similares» de diferentes clases.**

➢ **Representatividad: las características han de tomar valores similares para muestras de la misma clase.**

➢ **Número de atributos: debe ser el más pequeño posible ya que la complejidad del modelo aumenta rápidamente con la dimensión del vector de atributos.**







¡¡¡ CUIDADO: los atributos pueden no tener individualmente ningún poder de discriminación, pero sí cuando son considerados de forma conjunta !!!

CONSIDERACIONES: Planteamiento Metodológico Común

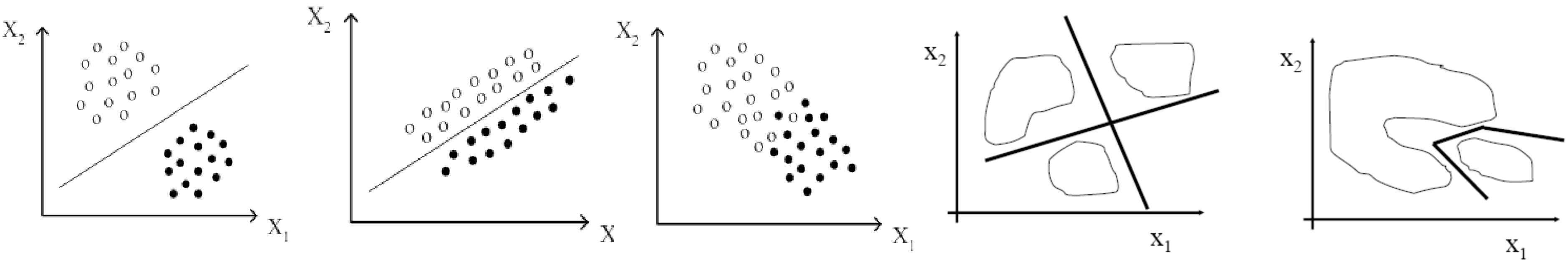
- **Clasificación Supervisada:**

1. ***Definición de grupos representativos o clases***

2. ***Creación de conjunto de datos***

3. ***Clasificación***

***3.1. Fase de aprendizaje:** diseño y entrenamiento del modelo de clasificación*

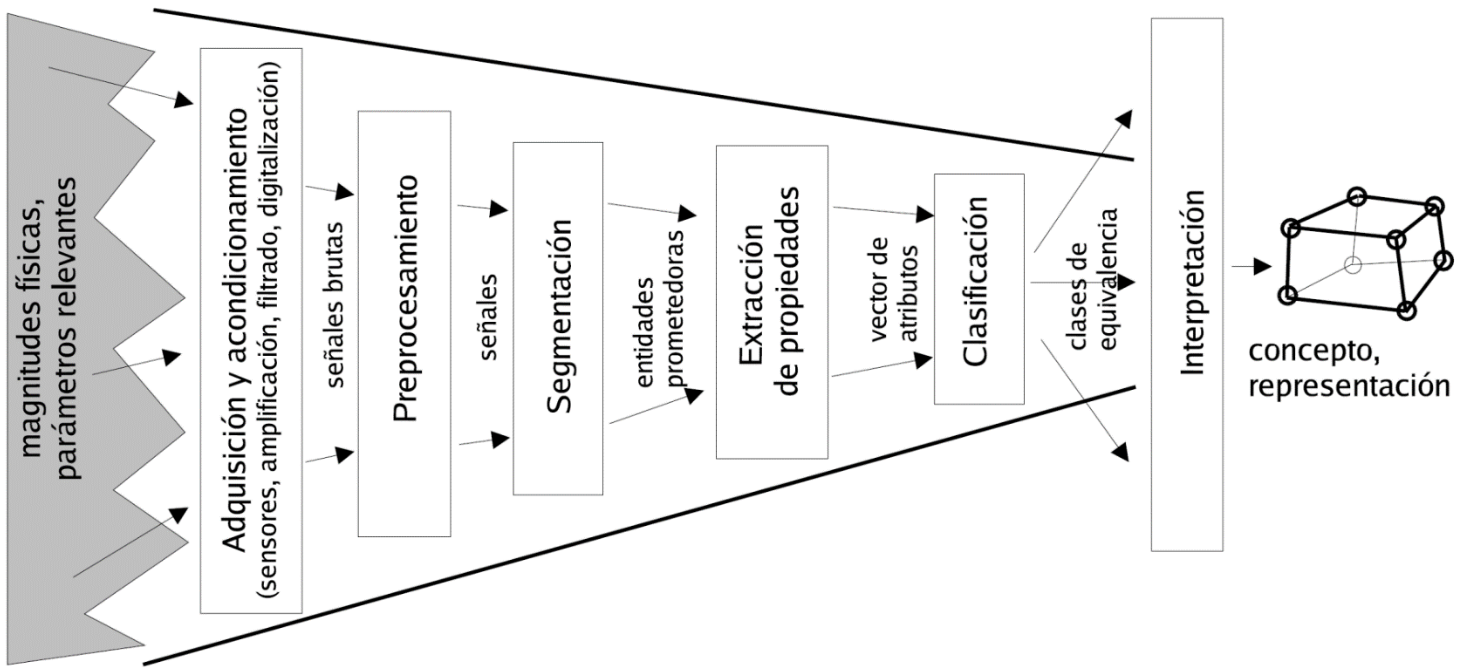


***3.2. Fase de inferencia:** aplicación del modelo entrenado para predecir la clase de una nueva muestra caracterizada por su vector de atributos*

CONSIDERACIONES: Planteamiento Metodológico Común

***Técnicas de Reconocimiento de Patrones* (*Pattern Recognition*):** Técnicas encaminadas a detectar patrones o regularidades existentes en un conjunto de datos con el objetivo de clasificarlos dentro de un conjunto de categorías de interés.

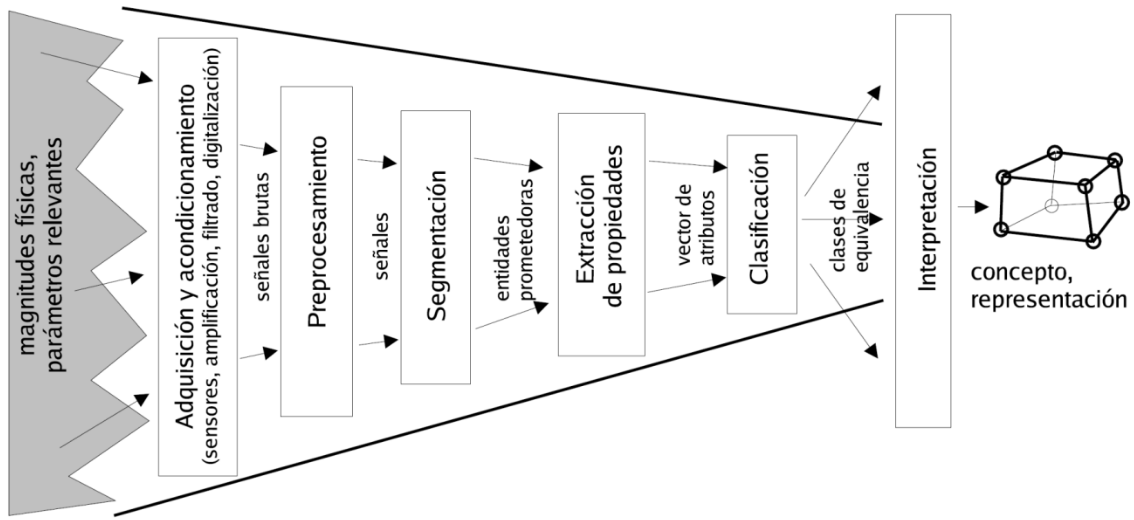
- **Descompone la etapa de aprendizaje y aplicación de un problema de clasificación en una serie de etapas bien definidas, independientemente de la naturaleza de la fuente de datos del proceso.**



- **Dependiendo de la naturaleza de los datos, para la extracción del conjunto de datos pueden ser necesarias la aplicación de nuevas etapas de adquisición y pre-procesamiento de la información, segmentación de los elementos de interés.**

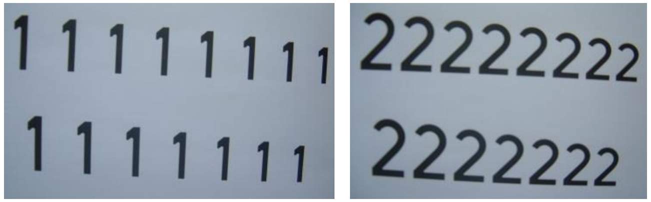
- **Ejemplo: reconocimiento de caracteres de matrícula a partir de imágenes**

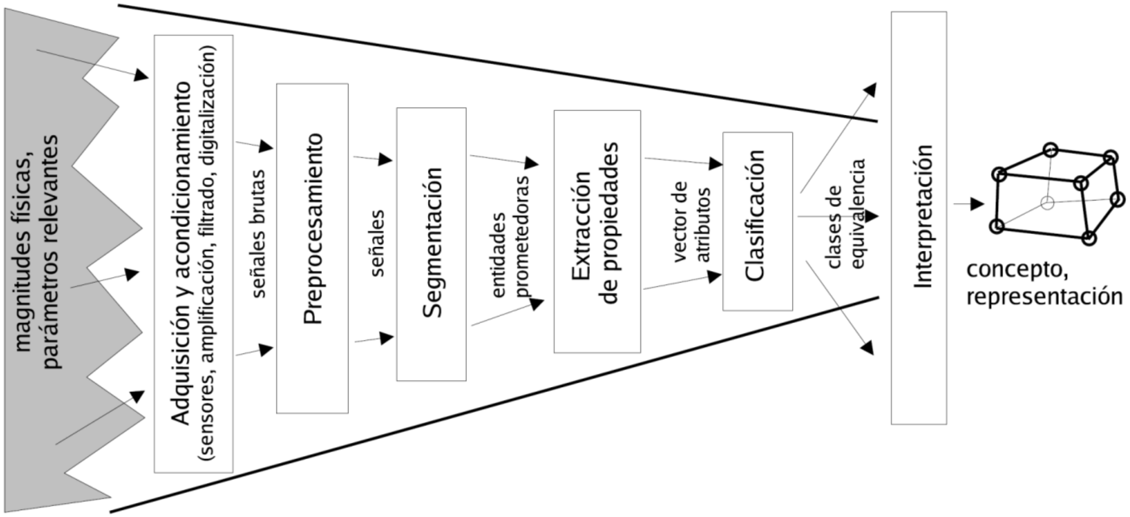
ADQUISICIÓN, PREPROCESAMIENTO, SEGMENTACIÓN

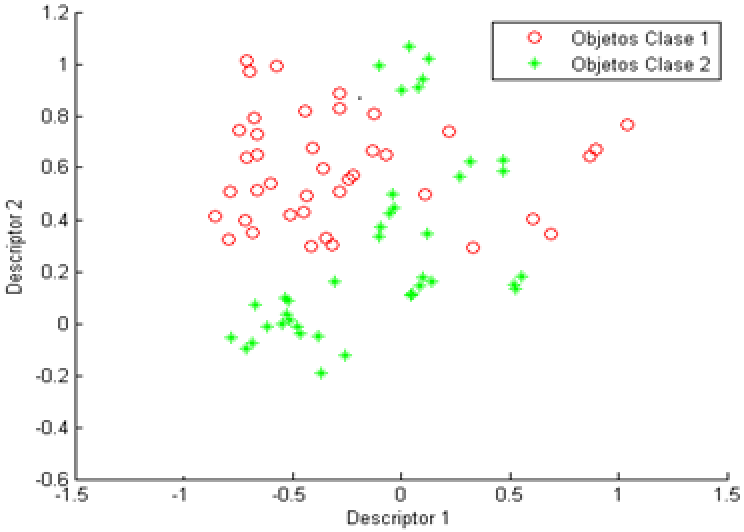


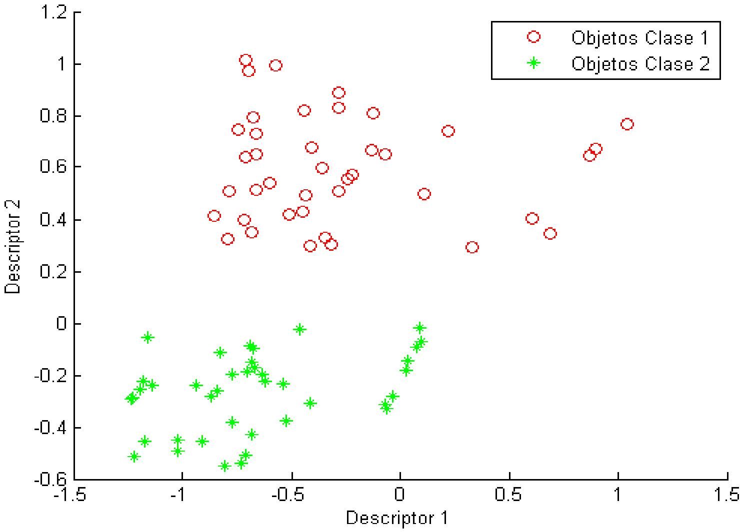
➢ **Ejemplo: reconocimiento de caracteres de matrícula a partir de imágenes**

EXTRACCIÓN DE PROPIEDADES Y CLASIFICACIÓN



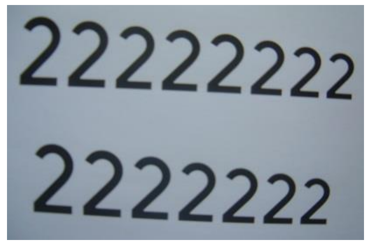


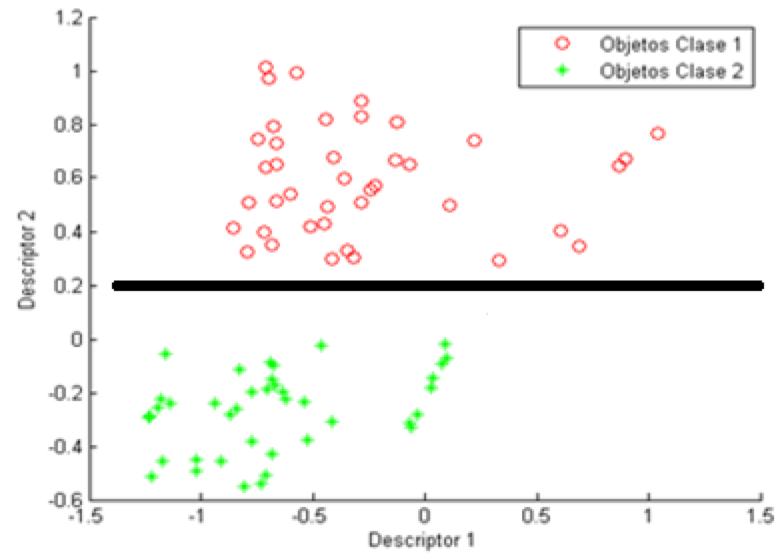




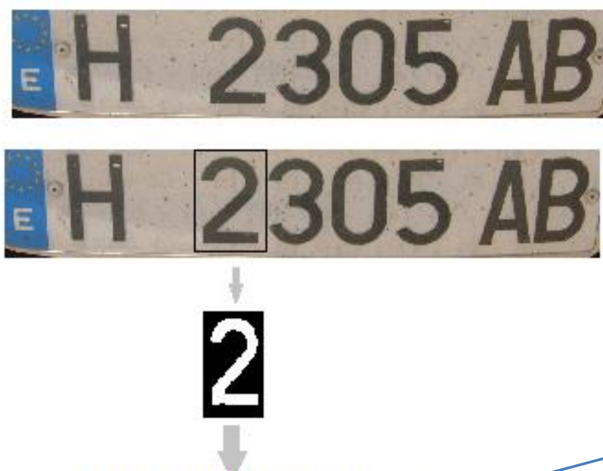
➢ **Ejemplo: reconocimiento de caracteres de matrícula a partir de imágenes**

DISEÑO DEL CLASIFICADOR

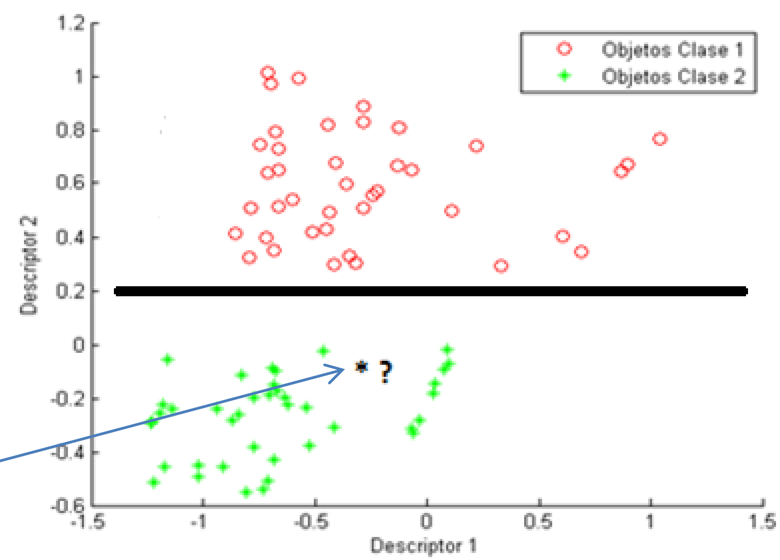




RECONOCIMIENTO: APLICACIÓN DEL CLASIFICADOR



Medida de Descriptor 1 y 2



EN ESTE BLOQUE DE LA ASIGNATURA ABORDAMOS LA ÚLTIMA ETAPA DEL PROCESO DE VISIÓN ARTIFICIAL PARA RECONOCER OBJETOS EN IMÁGENES COMO UN PROBLEMA DE CLASIFICACIÓN SUPERVISADA

1. ***Definición de grupos representativos o clases:*** **posibles tipos de objetos que, según el conocimiento a priori del problema, se espera puedan aparecer.**

2. ***Creación de conjunto de datos*:**

   2.1- Extracción de atributos de cada instancia disponible

   2.2- Selección de atributos - definición de un vector de atributos

3. ***Clasificación***

   3.1. Fase de aprendizaje: diseño y entrenamiento del modelo de clasificación

   3.2. Fase de inferencia: aplicación del modelo entrenado para predecir la clase de una nueva muestra caracterizada por su vector de atributos

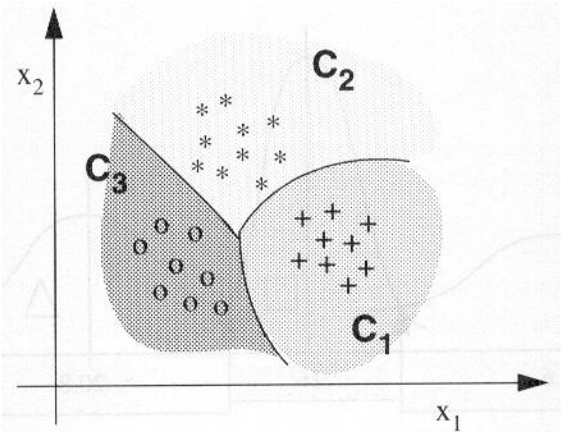
PROBLEMA DE CLASIFICACIÓN: ENFOQUE BASADO EN LA TEORÍA DE LA DECISIÓN

- **Problema de clasificación:**

  - Planteamiento matemático bien definido: Teoría de la Decisión, enfoque probabilístico-estadístico, enfoque basado en la optimización de funciones discriminantes.

  - Consiste en dividir el espacio de características en regiones o subespacios representativos de cada clase considerada en el problema.

  - Implica la definición de funciones de decisión que particionan el espacio de características.



Partición del espacio de características x_1-x_2 en 3 regiones correspondientes a 3 clases

➢ **La clasificación se formula en base a unas *funciones* denominadas *de decisión* o *discriminantes* que son evaluadas para decidir la clase de una muestra «desconocida» descrita mediante su vector de atributos.**

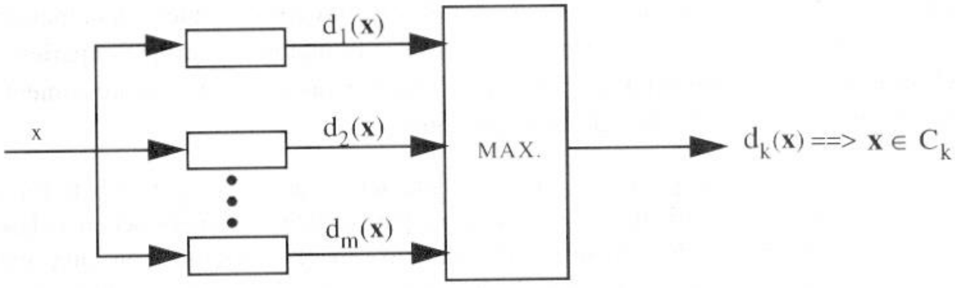


Diagrama esquemático de un sistema de clasificación:

- El vector de características x se evalúa según m funciones de decisión.

- El objeto caracterizado por x se asigna a la clase C_k cuya función de decisión sea máxima.

Funciones de decisión lineales:

- **CASO BI-DIMENSIONAL (vector de atributos compuesto por 2 características x_1 y x_2)**

- **2 clases, C_1 y C_2**

$\Rightarrow$ **Las dos poblaciones de patrones pueden separarse mediante una recta:**

$$w_1x_1 + w_2x_2 + w_3 = 0 \Leftrightarrow \begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix} = 0 \Leftrightarrow W^T X = 0$$

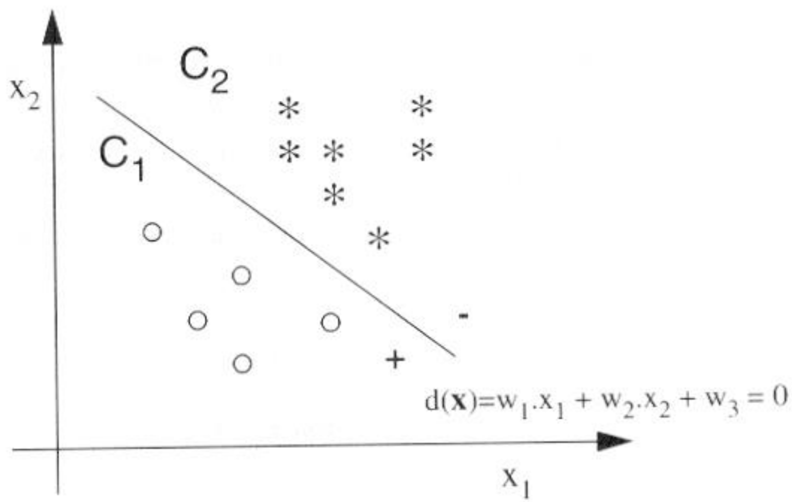
$$W = \begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix}^T \equiv \text{Vector de pesos} \;\; ; \;\; X = \begin{bmatrix} x_1 & x_2 & 1 \end{bmatrix}^T \equiv \text{Vector de características}$$

$\Rightarrow$ **Función de decisión lineal:**

$$d(X) = w_1x_1 + w_2x_2 + w_3 = W^T X$$

Para un vector de caracterísitcas de clasificación desconocida:

$$d(X) \begin{cases} > 0 \Rightarrow X \in C_1 \\ < 0 \Rightarrow X \in C_2 \\ = 0 \Rightarrow X \in \text{ frontera de separación} \end{cases}$$



Funciones de decisión lineales:

- **CASO BI-DIMENSIONAL (vector de atributos compuesto por 2 características x_1 y x_2)**
- **2 clases, C_1 y C_2**

$\Rightarrow$ **Las dos poblaciones de patrones pueden separarse mediante una recta:**

$$w_1x_1 + w_2x_2 + w_3 = 0 \iff \begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix} = 0 \iff W^T X = 0$$

$$W = \begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix}^T \equiv \text{Vector de pesos} \;\; ; \;\; X = \begin{bmatrix} x_1 & x_2 & 1 \end{bmatrix}^T \equiv \text{Vector de características}$$

- **CASO n-DIMENSIONAL**
- **2 clases, C_1 y C_2**

$$d(X) = w_1x_1 + w_2x_2 + \ldots + w_nx_n + w_{n+1} = W^T X \begin{cases} > 0 \Rightarrow X \in C_1 \\ < 0 \Rightarrow X \in C_2 \\ = 0 \Rightarrow X \in \text{ frontera de separación} \end{cases}$$

$$W = \begin{bmatrix} w_1 & w_2 & \ldots & w_n & w_{n+1} \end{bmatrix}^T \equiv \text{Vector de pesos} \;\; ; \;\; X = \begin{bmatrix} x_1 & x_2 & \ldots & x_n & 1 \end{bmatrix}^T \equiv \text{Vector de características}$$

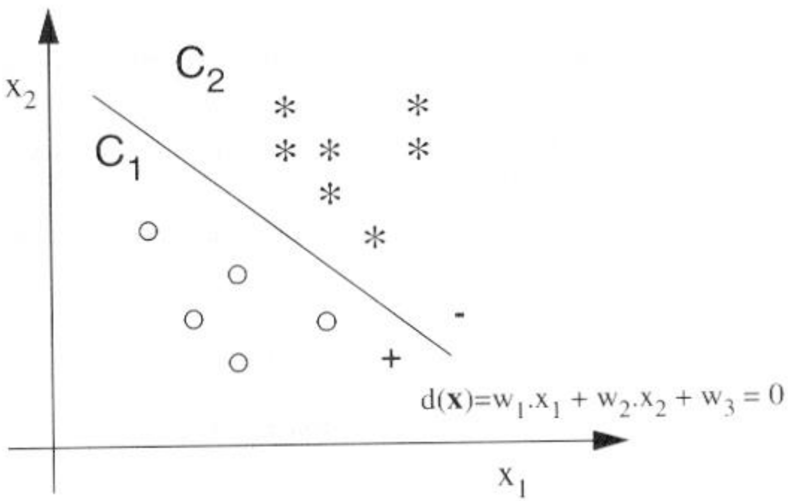
Funciones de decisión lineales:

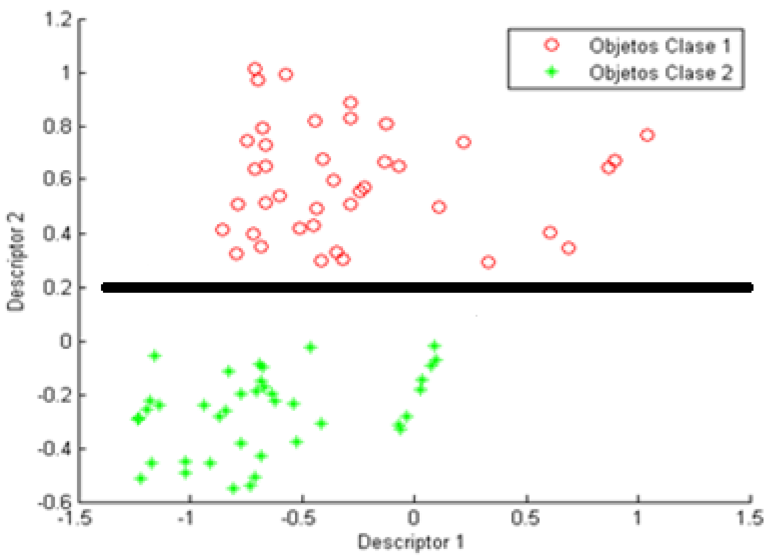
- **EJEMPLO CASO BI-DIMENSIONAL (vector de atributos compuesto por 2 características x_1 y x_2)**
- **2 clases, C_1 y C_2**

$$d(X) = w_1 x_1 + w_2 x_2 + w_3 = W^T X$$

Para un vector de caracterísitcas de clasificación desconocida:

$$d(X) \begin{cases} > 0 \Rightarrow X \in C_1 \\ < 0 \Rightarrow X \in C_2 \\ = 0 \Rightarrow X \in \text{ frontera de separación} \end{cases}$$





$$x_2 = 0,2 \rightarrow \textit{Frontera: Separación (FS): } x_2 - 0,2 = 0$$

$$d(x_1, x_2) = x_2 - 0,2$$

$$d(x_1, x_2) = \begin{cases} > 0 \rightarrow X \in C_1 \\ < 0 \rightarrow X \in C_2 \\ = 0 \rightarrow X \epsilon\, FS \end{cases}$$

Funciones de decisión lineales:

CASO GENERALIZADO

- **CASO n-DIMENSIONAL (vector de atributos compuesto por n características: x_1 , x_2 , … , x_n**
- ***m* clases, C_1 , C_2 , …, C_m**

Distintas posibilidades de clasificación:

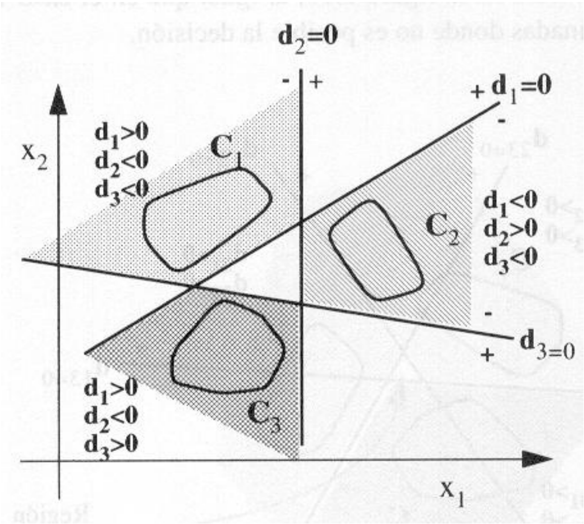
⇒ CASO 1: se consideran *m* funciones de decisión con la propiedad:

$$d_i(X) = w_{i1}x_1 + w_{i2}x_2 + \ldots + w_{in}x_n + w_{in+1} = W_i^T X \begin{cases} > 0 \text{ si } X \in C_1 \\ < 0 \text{ en otro caso} \end{cases}$$

Vector de pesos asociado a la función de decisión i-ésima:

$$W_i = \left[w_{i1} \;\; w_{i2} \; \ldots \; w_{in} \;\; w_{in+1} \right]^T$$

Vector de características: $X = \left[x_1 \;\; x_2 \; \ldots \; x_n \;\; 1 \right]^T$



Funciones de decisión lineales:

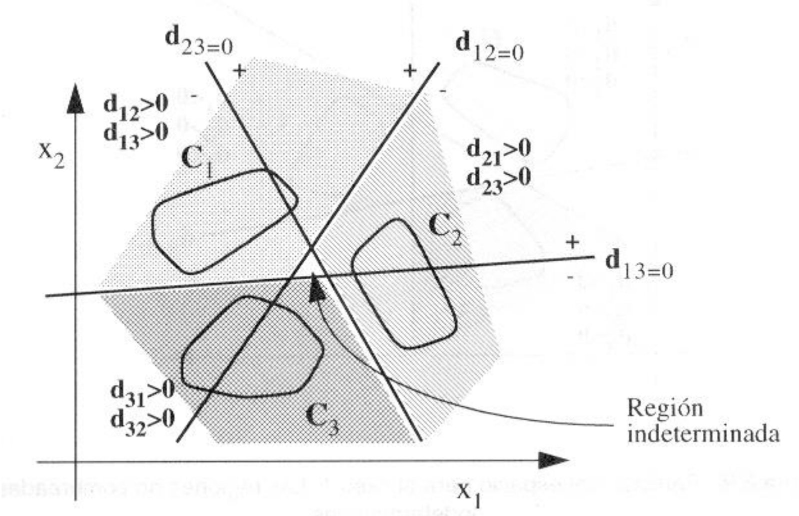
CASO GENERALIZADO

- **CASO n-DIMENSIONAL (vector de atributos compuesto por n características: $x_1 , x_2 , \ldots , x_n$**
- ***m* clases, $C_1 , C_2 , \ldots, C_m$**

Distintas posibilidades de clasificación:

⇒ CASO 2: se consideran *m (m-1) / 2* funciones de decisión (combinaciones de m clases tomadas de dos en dos).

$$d_{ij}(X) = W_{ij}^T X \; > 0 \text{ si } X \in C_i \quad \forall j \neq i$$



Funciones de decisión lineales:

CASO GENERALIZADO

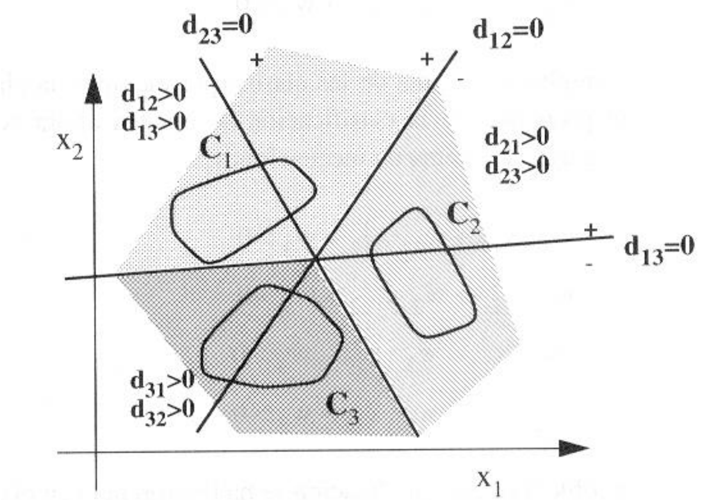
- **CASO n-DIMENSIONAL (vector de atributos compuesto por n características: x_1 , x_2 , … , x_n**
- ***m* clases, C_1 , C_2 , …, C_m**

Distintas posibilidades de clasificación:

⇒ CASO 3: se consideran *m* funciones de decisión (tantas como clases) de la forma:

$$d_k(X) = W_k^T X \text{ con } k = 1, 2, \ldots, m,$$

Cumpliéndos que si $X \in C_i \;\Rightarrow\; d_i(X) > d_j(X) \quad \forall j \neq i$



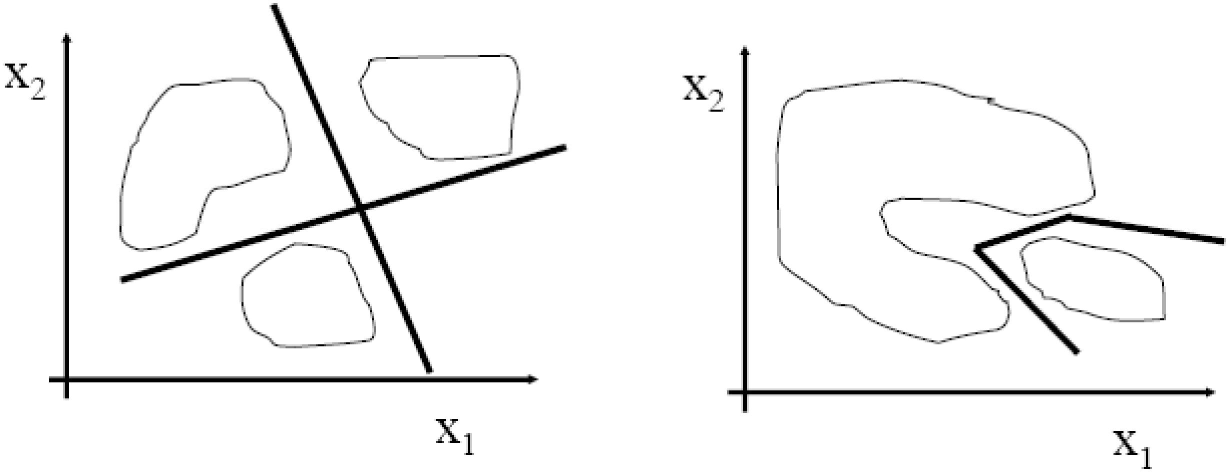
¿ Y SI LAS CLASES NO SON LINEALMENTE SEPARABLES ?

Funciones de decisión generalizadas:

➢ **Las clases cuya envolvente conexa no se corte son separables:**

⇒ Clases de patrones linealmente separables: si pueden separarse mediante funciones de decisión lineales.

⇒ Clases no linealmente separables: se incluyen situaciones en las que para separar las clases hay que recurrir a fronteras lineales a trozos.



Clases linealmente separables y clases separables no linealmente

<u>Funciones de decisión generalizadas:</u>

➢ **Objetivo: generalizar el concepto de función de decisión lineal, tratando funciones de decisión en principio complejas como si fueran lineales.**

$$d(X) = w_1 f_1(X) + w_2 f_2(X) + \ldots + w_k f_k(X) + w_{k+1} \quad ; \quad k \geq n \quad ; \quad X = [x_1 \;\; x_2 \;\ldots\; x_n \;\; 1]^T$$

$$d(X) = [w_1 \;\; w_2 \;\ldots\; w_{k+1}] \begin{bmatrix} f_1(X) \\ f_2(X) \\ \\ f_k(X) \\ 1 \end{bmatrix} = W^T X^* \quad ; \quad W \equiv [w_1 \;\; w_2 \;\ldots\; w_{k+1}]^T \quad ; \quad X^* = \begin{bmatrix} f_1(X) \\ f_2(X) \\ \\ f_k(X) \\ 1 \end{bmatrix}$$

⇒Permite la representación de una gran variedad de funciones de decisión, dependiendo de la elección de las {$f_i(X)$}, y el número de términos usados en la expansión.

⇒ En el espacio transformado X^* (transformación del vector de atributos original X en X^* mediante las funciones {$f_i(X)$}), se diseñan funciones de decisión lineales:

→ **Ventaja: simplificación del problema de separación de clases que requieren funciones de decisión no lineales (permite aplicar funciones lineales mediante la transformación del espacio de características).**

→ **Inconveniente: aumento de la dimensionalidad del problema ($k \geq n$).**

Funciones de decisión generalizadas: *ejemplo caso cuadrático (polinomio de grado 2) bidimensional (dos atributos)*

$$\text{Para el caso bidimensional:} \quad X = [x_1 \ \ x_2 \ \ 1]^T$$

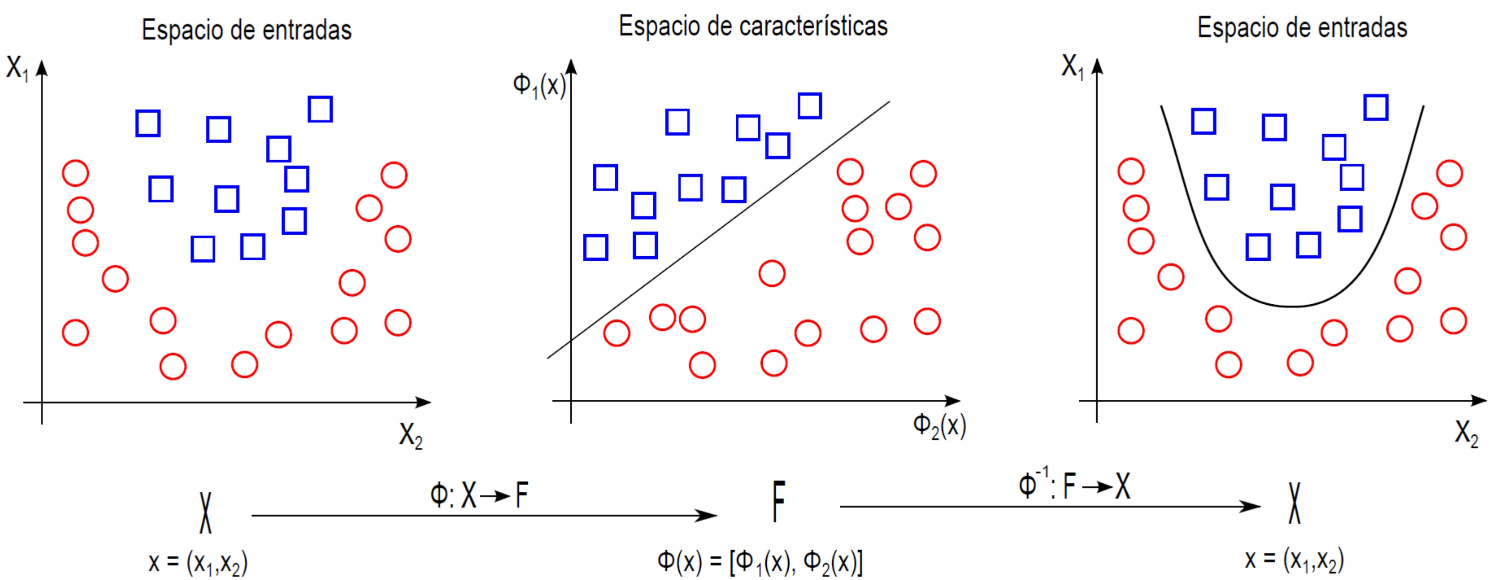
$$\Downarrow$$

$$d(X) = w_{11}x_1^{\ 2} + w_{12}x_1x_2 + w_{22}x_2^{\ 2} + w_1x_1 + w_2x_2 + w_3$$

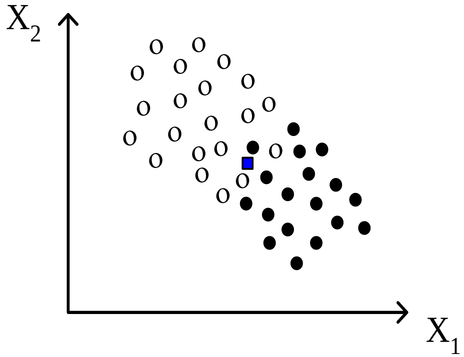
$$\Downarrow$$

$$\text{En forma lineal:}$$

$$d(X) = [w_{11} \ \ w_{12} \ \ w_{22} \ \ w_1 \ \ w_2 \ \ w_3] \begin{bmatrix} x_1^{\ 2} \\ x_1x_2 \\ x_2^{\ 2} \\ x_1 \\ x_2 \end{bmatrix} = W^T X^* \quad ; \quad W \equiv [w_{11} \ \ w_{12} \ \ w_{22} \ \ w_1 \ \ w_2 \ \ w_3]^T \quad ; \quad X^* = \begin{bmatrix} x_1^{\ 2} \\ x_1x_2 \\ x_2^{\ 2} \\ x_1 \\ x_2 \end{bmatrix}$$



Clasificadores estadísticos (paramétricos): Clasificador Bayesiano:



➢ **Asocian una instancia a la clase cuya probabilidad de pertenencia sea mayor.**

⇒ **Las funciones de decisión se diseñan a partir de las distribuciones estadísticas que caracterizan las distintas clases que se consideran en la clasificación.**

➢ **Teoría de decisión bayesiana aplicada a problemas de clasificación. Basada en dos suposiciones:**

1. **El problema de decisión se puede describir en términos probabilísticos:**

   ⇒ **Dado un conjunto de datos, *D*, para cada instancia descrita por *X*, decidir la mejor hipótesis *h* del conjunto de hipótesis *H***

   - ❑ ***X*: vector de atributos que describe una instancia del conjunto de datos *D***
   - ❑ **H: clases del problema**

   ⇒ **La mejor hipótesis es la hipótesis más probable.**

2. **Asume que las probabilidades implicadas en el problema son conocidas.**

- **PROBABILIDADES INVOLUCRADAS**

  - **Probabilidades a Priori (incondicional):**

    - ⇒ **Probabilidad a priori de la hipótesis *h, P(h)*: probabilidad de que la hipótesis *h* sea cierta**

      - ❑ **Refleja el conocimiento que tenemos sobre las oportunidades de que la hipótesis *h* sea cierta antes de recibir ninguna observación.**

      - ❑ **Si no tenemos ningún conocimiento a priori, se le podría asignar la misma probabilidad a todas las hipótesis**

    - ⇒ **Probabilidad a priori de la observación *X, P(X)*: probabilidad de que recibamos una instancia descrita por *X***

      - ❑ **Refleja la probabilidad de recibir la observación *X*, cuando no tenemos ninguna idea sobre cuál es la hipótesis real.**

  - **Probabilidades a Posteriori (condicional):**

    - ⇒ **Probabilidad a posteriori de la observación X, *P(X | h)*: probabilidad de observar el dato X cuando se cumple la hipótesis h.**

    - ⇒ **Probabilidad a posteriori de la hipótesis *h, P(h / X)*: probabilidad de que se cumpla la hipótesis *h* habiéndose recibido una muestra descrita por *X*.**

➢ TEOREMA DE BAYES APLICADO A PROBLEMAS DE CLASIFICACIÓN

⇒ **Probabilidad a priori de la hipótesis *h*, *P(h)*: probabilidad de que la hipótesis *h* sea cierta**

⇒ **Probabilidad a priori de la observación *X*, *P(X)*: probabilidad de que recibamos una instancia descrita por *X***

⇒ **Probabilidad a posteriori de la observación *X*, *P(X|h)*: probabilidad de observar la observación *X* cuando se cumple la hipótesis *h*.**

⇒ **Probabilidad a posteriori de la hipótesis *h*, *P(h / X)*: probabilidad de que se cumpla la hipótesis *h* habiéndose recibido una muestra descrita por *X*.**

$$P(h|X) = \frac{P(X|h)P(h)}{P(X)}$$

$$P(X) = \sum_{i=1}^{Numero\ de\ Clases} \left(P(X|hi)P(hi)\right)$$

Ejemplo: Moneda trucada

Espacio de hipótesis: *{cara, cruz} (Clases del problema)*

Conocimiento a priori: *las probabilidades de que al lanzar la moneda salga cara o cruz son 0.2 y 0,8 respectivamente*

Espacio de observaciones: {brillo, mate} . *De cada muestra (hecho de lanzar una moneda) se recibe una observación, que es el tipo de acabado de la moneda (atributo X utilizado), brillo o mate.*

Objetivo: *al lanzar una moneda, y tras recibir la observación de su atributo (si es brillo o mate), hay que generar la mejor hipótesis (si ha salido cara o cruz)*

Procedimiento: *disponer de un conjunto de observaciones, de lanzamientos de monedas con ambos tipos de acabado, y anotar si ha salido cara o cruz.*

➢ TEOREMA DE BAYES ENFOCADO A CLASIFICACIÓN: EJEMPLO MONEDA TRUCADA

⇒ **Probabilidad a priori de la hipótesis *h*, *P(h)*: probabilidad de que la hipótesis *h* sea cierta**

Conocimiento a priori: *las probabilidades de que al lanzar la moneda salga cara o cruz son 0.2 y 0,8 respectivamente*

$$P(cara) = 0.2 \;\; ; \;\; P(cruz) = 0.8$$

⇒ **Probabilidad a posteriori de la observación *X*, *P(X|h)*: probabilidad de observar la observación *X* cuando se cumple la hipótesis *h*.**

Supongamos que del conjunto de observaciones realizadas se obtiene la siguiente información:

Probabilidad de recibir la observación «BRILLO» en los lanzamientos que han salido «CARA»: $P(brillo \mid cara) = 0.9$

Probabilidad de recibir la observación «MATE» en los lanzamientos que han salido «CARA»: $P(mate \mid cara) = 0.1$

Probabilidad de recibir la observación «BRILLO» en los lanzamientos que han salido «CRUZ»: $P(brillo \mid cruz) = 0.6$

Probabilidad de recibir la observación «MATE» en los lanzamientos que han salido «CRUZ»: $P(mate \mid cruz) = 0.4$

⇒ **Probabilidad a priori de la observación *X*, *P(X)*: probabilidad de que recibamos una instancia descrita por *X***

Probabilidad de recibir la observación «BRILLO»: $P(brillo) = P(brillo \mid cara) * P(cara) + P(brillo \mid cruz) * P(cruz) = 0.66$

Probabilidad de recibir la observación «MATE»: $P(mate) = P(mate \mid cara) * P(cara) + P(mate \mid cruz) * P(cruz) = 0.34$

➢ TEOREMA DE BAYES ENFOCADO A CLASIFICACIÓN: EJEMPLO MONEDA TRUCADA

⇒ **Probabilidad a posteriori de la hipótesis *h*, *P(h / X)*: probabilidad de que se cumpla la hipótesis *h* habiéndose recibido una muestra descrita por *X*.**

$$P(h|X) = \frac{P(X|h)P(h)}{P(X)}$$

Tiro la moneda y recibo observación «BRILLO»:

Probabilidad de que el lanzamiento salga «CARA» habiendo recibido como observación «BRILLO»:

$$P(cara|brillo) = \frac{P(brillo|cara)P(cara)}{P(brillo)} = 0,273$$

Probabilidad de que el lanzamiento salga «CRUZ» habiendo recibido como observación «BRILLO»:

$$P(cruz|brillo) = \frac{P(brillo|cruz)P(cruz)}{P(brillo)} = 0,727$$

La hipótesis más probable es que, tras recibir BRILLO el lanzamiento salga CRUZ

Tiro la moneda y recibo observación «MATE»:

Probabilidad de que el lanzamiento salga «CARA» habiendo recibido como observación «MATE»:

$$P(cara|mate) = \frac{P(mate|cara)P(cara)}{P(mate)} = 0,059$$

Probabilidad de que el lanzamiento salga «CRUZ» habiendo recibido como observación «MATE»:

$$P(cruz|mate) = \frac{P(mate|cruz)P(cruz)}{P(mate)} = 0,941$$

La hipótesis más probable es que, tras recibir MATE, el lanzamiento salga CRUZ

➢ TEOREMA DE BAYES ENFOCADO A CLASIFICACIÓN: EJEMPLO MONEDA TRUCADA

⇒ **Probabilidad a posteriori de la hipótesis *h*, *P(h / X)*: probabilidad de que se cumpla la hipótesis *h* habiéndose recibido una muestra descrita por *X*.**

$$P(h|X) = \frac{P(X|h)P(h)}{P(X)}$$

CONCLUSIONES:

Se ha diseñado una función de decisión para cada hipótesis (salir cara o cruz).

Esta función de decisión evalúa la probabilidad de ser cierta la hipótesis dada una determinada observación

Se predice que la hipótesis del suceso es aquella más probable para la observación medida

Dado que el denominador de la expresión es común en las funciones de decisión de cada hipótesis, la hipótesis más probable es aquella que, para la observación medida, presenta el numerador más alto: decisor máximo a posteriori

Decisor Máximo a Posteriori:

$$h_{MAP} = arg \max_{h \in H} P(h|X) = arg \max_{h \in H} \left(\frac{P(X|h)P(h)}{P(X)}\right) = arg \max_{h \in H} \left(P(X|h)P(h)\right)$$

Ejemplo: tiro la moneda y obtengo brillo:

$$h_{MAP} = arg \max_{cara,\, cruz} \left(P(brillo|cara)P(cara)\, ,P(brillo|cruz)P(cruz)\right) = arg \max_{cara,\, cruz} (0.18\, , 0.48) = \text{CRUZ}$$

➢ TEOREMA DE BAYES: EJEMPLO MONEDA SIN TRUCAR

⇒ **Probabilidad a posteriori de la hipótesis *h*, *P(h / X)*: probabilidad de que se cumpla la hipótesis *h* habiéndose recibido una muestra descrita por *X*.**

$$P(h|X) = \frac{P(X|h)P(h)}{P(X)}$$

Si por conocimiento a priori: *las probabilidades de salir cara o cruz son las mismas: P(cara) = 0.5 ; P(cruz) = 0.5*

HIPÓTESIS EQUIPROBABLES A PRIORI

↓

Al evaluar $P(h|X)$ para cada hipótesis y seleccionar la hipótesis que presente el valor más alto, P(X) y P(h) no influyen

↓

Decisor de Máxima Verosimilitud (ML, maximum likelihood)

Decisor Máxima Verosimilitud:

$$h_{ML} = arg \max_{h \in H} P(h|X) = arg \max_{h \in H} \left(\frac{P(X|h)P(h)}{P(X)}\right) = arg \max_{h \in H} \left(P(X|h)\right)$$

Ejemplo: tiro la moneda y obtengo brillo:

$$h_{MAP} = arg \max_{cara,\, cruz} \left(P(brillo|cara)\,, P(brillo|cruz)\right) = arg \max_{cara,\, cruz} (0.9\,, 0.6) = \text{CARA}$$

➢ TEOREMA DE BAYES ENFOCADO A CLASIFICACIÓN: EJEMPLO – JUGAR AL TENIS

EJERCICIO PROPUESTO: CONTESTAR DE FORMA RAZONADA A LAS SIGUIENTES PREGUNTAS UTILIZANDO UN SISTEMA DE CLASIFICACIÓN BAYESIANO DECISOR MÁXIMO A POSTERIORI.

¿ Se podrá jugar al tenis con un día soleado?

¿ Se podrá jugar al tenis con un día frío?

Day	outlook	temperature	humidity	windy	play
D1	sunny	hot	high	weak	no
D2	sunny	hot	high	strong	no
D3	overcast	hot	high	weak	yes
D4	rainy	mild	high	weak	yes
D5	rainy	cool	normal	weak	yes
D6	rainy	cool	normal	strong	no
D7	overcast	cool	normal	strong	yes
D8	sunny	mild	high	weak	no
D9	sunny	cool	normal	weak	yes
D10	rainy	mild	normal	weak	yes
D11	sunny	mild	normal	strong	yes
D12	overcast	mild	high	strong	yes
D13	overcast	hot	normal	weak	yes
D14	rainy	mild	high	strong	no

- **TEOREMA DE BAYES ENFOCADO A CLASIFICACIÓN: EJEMPLO – JUGAR AL TENIS**

UN PASO MÁS:

¿ Se podrá jugar al tenis con un día soleado, frío, de humedad alta y viento fuerte ?

Day	outlook	temperature	humidity	windy	play
D1	sunny	hot	high	weak	no
D2	sunny	hot	high	strong	no
D3	overcast	hot	high	weak	yes
D4	rainy	mild	high	weak	yes
D5	rainy	cool	normal	weak	yes
D6	rainy	cool	normal	strong	no
D7	overcast	cool	normal	strong	yes
D8	sunny	mild	high	weak	no
D9	sunny	cool	normal	weak	yes
D10	rainy	mild	normal	weak	yes
D11	sunny	mild	normal	strong	yes
D12	overcast	mild	high	strong	yes
D13	overcast	hot	normal	weak	yes
D14	rainy	mild	high	strong	no

➢ TEOREMA DE BAYES ENFOCADO A CLASIFICACIÓN: EJEMPLO – JUGAR AL TENIS

Clases del problema: $C = \{C_1, C_2\}$ = *{ sí, no }*

Vector de atributos: $X = [\, x1, x2, x3, x4 \,]^t$ = *[estado del cielo, temperatura, humedad, viento]*^{t}

Se dispone de 14 observaciones previas: *datos de la tabla*

Objetivo: *por ejemplo,* ***saber si se podrá jugar al tenis con un día soleado, frío, de humedad alta y viento fuerte.***

Day	outlook	temperature	humidity	windy	play
D1	sunny	hot	high	weak	no
D2	sunny	hot	high	strong	no
D3	overcast	hot	high	weak	yes
D4	rainy	mild	high	weak	yes
D5	rainy	cool	normal	weak	yes
D6	rainy	cool	normal	strong	no
D7	overcast	cool	normal	strong	yes
D8	sunny	mild	high	weak	no
D9	sunny	cool	normal	weak	yes
D10	rainy	mild	normal	weak	yes
D11	sunny	mild	normal	strong	yes
D12	overcast	mild	high	strong	yes
D13	overcast	hot	normal	weak	yes
D14	rainy	mild	high	strong	no

➢ TEOREMA DE BAYES ENFOCADO A CLASIFICACIÓN: EJEMPLO – JUGAR AL TENIS

¿ Se podrá jugar al tenis con un día soleado, frío, de humedad alta y viento fuerte ?

⇒ **Hay que decidir qué clase («sí», «no») es más probable que suceda para una instancia dada por un vector de atributo de valores <*sunny, cool, high, strong*>**

Clasificador Bayesiano – Decisor Máximo a Posteriori

$$C_{MAP} = arg \max_{i=1,2} P(C_i|X) = arg \max_{i=1,2} \left(\frac{P(X|C_i)P(C_i)}{P(X)}\right) = arg \max_{i=1,2} \left(P(X|C_i)P(C_i)\right)$$

$$Xoi = \langle sunny, cool, high, strong \rangle$$

$$D_{yes} = \left(P(Xoi \mid yes) * P(yes)\right) \;; D_{no} = \left(P(Xoi \mid no) * P(no)\right)$$

⇒ **El problema implica determinar las siguientes probabilidades:**

❖ **Probabilidades a priori de las clases del problema , *$P(C_i)$, i = 1, 2***

$$P(C_1) = P(yes) = \frac{9}{14} = 0.64 \quad ; P(C_2) = P(no) = \frac{5}{14} = 0.36$$

❖ ***Probabilidad de la observación Xoi en los datos disponibles de la clase «yes»: P(Xoi | yes) = ???***

❖ ***Probabilidad de la observación Xoi en los datos disponibles de la clase «no»: P(Xoi | no) = ???***

➢ *NAIVE* BAYES – CLASIFICADOR BAYESIANO *INGENUO*

⇒ **ASUME INDEPENDENCIA LINEAL ENTRE LOS ATRIBUTOS: las variables que componen el vector de atributos son estadísticamente independientes, no están correlados:**

$$X = [x_1, x_2, \dots, x_n] \qquad P(X|C_i) = P([x_1, x_2, \dots, x_n]|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

EJEMPLO ANTERIOR: *P(Xoi | yes)* = ??? ; *P(Xoi | no)* = ???

$$X = [estado\ del\ cielo, temperatura, humedad, viento]\ ;\ Xoi = \langle\, sunny\ , cool\ , high\ , strong \,\rangle$$

Asumiendo independencia lineal de atributos

$$P(Xoi\mid yes) = P(sunny\mid yes)*P(cool\mid yes)*P(high\mid yes)*P(strong\mid yes) = \frac{2}{9}*\frac{3}{9}*\frac{3}{9}*\frac{3}{9} = 0.0082$$

$$P(Xoi\mid no) = P(sunny\mid no)*P(cool\mid no)*P(high\mid no)*P(strong\mid no) = \frac{3}{5}*\frac{1}{5}*\frac{4}{5}*\frac{3}{5} = 0.0576$$

Funciones de decisión:

$$D_{yes} = \big(P(Xoi\mid yes)\ *\ P(yes)\big) = 0.005\ \ ; D_{no} = \big(P(Xoi\mid no)\ *\ P(no)\big) = 0{,}021 \rightarrow C_{NB}(X_{OI}) = NO$$

⇒ **Clasificador NAIVE BAYES**

$$C_{MAP} = arg \max_{i=1,2} P(C_i|X) = arg \max_{i=1,2} \left(\frac{P(X|C_i)P(C_i)}{P(X)}\right) = arg \max_{i=1,2} \big(P(X|C_i)P(C_i)\big)$$

$$C_{NB} = arg \max_{i=1,2} \left(\prod_{k=1}^{n} P(x_k|C_i) * P(C_i)\right)$$

- **CONSIDERACIONES**

⇒ **Si el ejemplo a clasificar, no tiene el valor de algún atributo:** se omite dicho atributo

⇒ **Si hay faltas en la muestra de entrenamiento de un atributo:** esas instancias no cuentan en la estimación de probabilidades de ese atributo

⇒ **Si alguna clase no presenta ningún valor posible de un determinado atributo (muestras poco representativas):**

- **Problema:** no es posible evaluar la probabilidad de la clase cuando la instancia a clasificar presente ese valor del atributo - saldría siempre 0, independientemente de los valores del resto de los atributos de la instancia

Ejemplo: si en el conjunto de entrenamiento no hay ninguna instancia con temperatura fría «cool» en la clase «no», siempre se podría jugar al tenis en un día «frío»

$$X = [estado\ del\ cielo, temperatura, humedad, viento]\ ;\ Xoi = \langle sunny, cool, high, strong \rangle$$

$$P(Xoi \mid no) = P(sunny \mid no) * P(cool \mid no) * P(high \mid no) * P(strong \mid no) = \frac{3}{5} * \frac{0}{5} * \frac{4}{5} * \frac{3}{5} = 0$$

$$D_{yes} = \big(P(Xoi \mid yes) * P(yes)\big) = 0{,}005\ ;\ D_{no} = \big(P(Xoi \mid no) * P(no)\big) = 0 \rightarrow C_{NB}(X_{OI}) = SI$$

- **Probabilidades nulas o muy bajas, por ausencia en el conjunto de entrenamiento de algunos valores de atributos en algunas categorías.**

- **Solución:** se supone que hay K ejemplos virtuales en cada clase por cada posible valor de cada atributo – K suele tener valores bajos (1-5).

$$P(x_i = V_j | C_k) = \frac{(\text{número de ejemplos de la clase } C_k \text{ con } x_i = V_j) + K}{(\text{número de ejemplos de la clase } C_k) + K * (\text{número de posibles valores del atributo } x_i)}$$

EN LOS EJEMPLOS ANTERIORES:

⇒ **PROBABILIDAD DE LAS CLASES DEL PROBLEMA:**

$$C = \{C_1, C_2, \ldots, C_M\} \;\; ; P(C_i) = \frac{N_i}{N} \; donde \; N = \sum_{i=1}^{M} N_i \;\; ; \;\; N_i: número \; de \; instancias \; de \; la \; clase \; C_i$$

⇒ **PROBABILIDAD A POSTERIORI DE QUE UN ATRIBUTO DADO TENGA UN DETERMINADO VALOR EN LAS INSTANCIAS DE UNA DETERMINADA CLASE:**

$$X = [x_1, x_2, \ldots, x_n] \;\; ; \;\; P\left(x_i = V_j \middle| C_k\right) = \frac{número \; de \; ejemplos \; de \; la \; clase \; C_k \; con \; x_i = V_j}{número \; de \; ejemplos \; de \; la \; clase \; C_k}$$

⇓

¡ ATRIBUTOS DISCRETOS (NOMINALES O CATEGÓRICOS) !

¿ QUÉ OCURRE SI LOS ATRUBITUOS TIENEN NATURALEZA CONTINUA ?

❖ **Al igual que en el caso anterior, tomar decisiones a partir de las observaciones disponibles**

⇒ **Se supone que los valores de los atributos de las muestras de las distintas clases siguen una distribuciones de probabilidad conocida**

⇒ **Estimación de parámetros de las distribuciones de probabilidad conocidas que permiten calcular probabilidades de aparición de valores que pueden ser desconocidas: aprendizaje paramétrico**

- **Para cada clase, se hacen pasar una cierta cantidad de muestras por el sistema, se miden los valores numéricos de las características y los valores representativos de su función de distribución (si por ejemplo distribución normal, valor nominal o media y la tolerancia o desviación respecto a ese valor, desviación típica).**

SI ATRIBUTOS CONTINUOS:

- **Asumir que sus valores siguen una distribución Normal o Gaussiana N(μ , σ):** $f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

- **Para los datos del atributo de la clase en cuestión, se calcula la media y desviación típica:** μ_{ik} , σ_{ik}

- **Asumir como valor representativo de la probabilidad de aparición de un determinado valor del atributo para la clase en cuestión, el valor de la función de densidad de probabilidad para ese valor del atributo:**

$$X = [x_1, x_2, \dots, x_n] \; ; \; P\left(x_i = V_j \middle| C_k\right) \rightarrow f\left(x_i = V_j \middle| C_k\right) = \frac{1}{\sqrt{2\pi}\;\sigma_{ik}} e^{-\frac{\left(V_j-\mu_{ik}\right)^2}{2\sigma_{ik}{}^2}}$$

Ejemplo: *¿es posible jugar al tenis en un día soleado, frío, viento fuerte y de humedad 89?*

Ejemplo	Vista	Temperatura	Humedad	Viento	Jugar
1	Soleado	Alta (85)	Alta (85)	No	No
2	Soleado	Alta (80)	Alta (90)	Sí	No
3	Nublado	Alta (83)	Alta (86)	No	Sí
4	Lluvioso	Media (70)	Alta (96)	No	Sí
5	Lluvioso	Baja (68)	Normal (80)	No	Sí
6	Lluvioso	Baja (65)	Normal (70)	Sí	No
7	Nublado	Baja (64)	Normal (65)	Sí	Sí
8	Soleado	Media (72)	Alta (95)	No	No
9	Soleado	Baja (69)	Normal (70)	No	Sí
10	Lluvioso	Media (75)	Normal (80)	No	Sí
11	Soleado	Media (75)	Normal (70)	Sí	Sí
12	Nublado	Media (72)	Alta (90)	Sí	Sí
13	Nublado	Alta (81)	Normal (75)	No	Sí
14	Lluvioso	Media (71)	Alta (91)	Sí	No

$$Xoi = \langle\, sunny\,, cool\,, 89, strong \,\rangle$$

$$D_{yes} = \big(P(Xoi \mid yes) \ * \ P(yes)\big)$$

$$D_{no} = \big(P(Xoi \mid no) \ * \ P(no)\big)$$

$$P(C_1) = P(yes) = \frac{9}{14} = 0.64$$

$$P(C_2) = P(no) = \frac{5}{14} = 0.36$$

$$P(Xoi \mid yes) = P(sunny \mid yes) \ * \\ P(cool \mid yes) * P(89 \mid yes) * P(strong \mid yes)$$

$$P(hum = 89 | yes) \rightarrow$$

$$\rightarrow f(hum = 89 | yes) = \frac{1}{\sqrt{2\pi}\ \sigma_{hum,yes}} e^{-\frac{\left(89-\mu_{hum,yes}\right)^2}{2\sigma_{hum,yes}{}^2}}$$

De la misma forma, se calcularía $P(Xoi \mid no)$

Clasificadores estadísticos (paramétricos): Clasificador Bayesiano - RESUMEN:

Ejemplo: Clases: tuercas y tornillos, Característica: diámetro → ¿Probabilidad de que un objeto cuyo diámetro es 3mm sea un tornillo?

- **Procedimiento:**

1. **Determinar la probabilidad a posteriori del vector de atributos dado por X para cada clase C_i , $p(X|C_i)$:**

   ⇒ **A partir del conjunto de observaciones disponibles**

Ejemplo:

- **A partir de todas las muestras de tornillos disponibles: determinar la probabilidad de que el diámetro de un tornillo sea de 3 mm.**
- **A partir de todas las muestras de tuercas disponibles: determinar la probabilidad de que el diámetro de una tuerca sea de 3 mm.**

❖ **Si no hay más factores, regla de decisión:** ***para una X dada decidir la clase C_i tal que $p(X|C_i)$ sea máxima.***

2. **Determinar la probabilidad a priori de que se presente un elemento de la clase C_i , $P(C_i)$:**

⇒ **Probabilidad a priori de que un patrón cualquiera sea de la clase C_i.**

⇒ **Justificación: hay más objetos de una clases que de otras (ejemplo: se fabrican 3 tornillos por cada tuerca).**

Clasificadores estadísticos (paramétricos): Clasificador Bayesiano - RESUMEN:

Ejemplo: Clases: tuercas y tornillos, Característica: diámetro → ¿Probabilidad de que un objeto cuyo diámetro es 3mm sea un tornillo?

- **Procedimiento:**

3. Determinar, para cada clase, la probabilidad a posteriori de asignación del patrón X a la clase Ci, p(Ci|X):

⇒ **Regla de Bayes:** $p(C_i|X) = \dfrac{p(X|C_i)P(C_i)}{p(X)}$; $p(X) = \sum_i p(X|C_i)P(C_i)$

$p(X) \equiv$ Probabilidad a priori de que se presente un elemento con vector de características igual a X (función de probabilidad total de que se dé el patrón X)

4. Calcular una función de decisión para cada clase basada en la probabilidad del punto 3

Funciones de decisión válidas: $d_i(X) = \dfrac{p(X|C_i)P(C_i)}{p(X)}$; $d_i(X) = p(X|C_i)P(C_i)$; $\underline{d_i(X) = \ln\left[p(X|C_i)\right] + \ln\left[P(C_i)\right]}$

⇒ **Forma habitual de construir un clasificador: tomar tantas funciones de decisión como clases y evaluarlas para el patrón dado X de la muestra cuya clase se desconoce.**

5. Criterio de decisión: decidir que pertenece a la clase cuya función de decisión sea máxima

Asignar X a la clase C_j que verifique $p(X|C_j)P(C_j) > p(X|C_i)P(C_i)$

- *Clasificadores basados en distribuciones normales*

- **Clasificadores basados en la función de decisión obtenida para el clasificador bayesiano particularizada**

$$d_i(X) = \ln\left[p\left(X|C_i\right)\right] + \ln\left[P\left(C_i\right)\right]$$

para el caso de que las funciones de probabilidad de la característica X para cada clase C_i , *p(X|C_i)*, sean funciones gaussianas

$\Rightarrow$ **Clasificador Natural o de mínima distancia Euclídea:**

$\rightarrow$ **Basados en medir la distancia de un patrón de característicos X respecto a un vector prototipo de cada clase y seleccionar la clase que produzca la distancia más pequeña (clasificador mediante el vecino más próximo).**

$\Rightarrow$ **Clasificador de mínima distancia de Mahalanobis:**

$\rightarrow$ **Utiliza como función discriminante la distancia de Mahalanobis.**

$\rightarrow$ **Distancia de Mahalanobis: distancia Euclídea normalizada por la dispersión de los datos.**

- **Conceptos básicos → Función de distribución normal o Gaussiana para un vector n-dimensional:**

Caso unidimensional: $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right]$ Media $\equiv m = E[x] = \int_{-\infty}^{\infty} x p(x) dx$; Varianza $\equiv \sigma^2 = E[(x-m)^2]$

Caso n-dimensional $\rightarrow X = [x_1 \ \ x_2 \ \ \ldots \ \ x_n]^T$: $p(X) = \frac{1}{(2\pi)^{n/2} |C|^{1/2}} \exp\left[-\frac{1}{2}(X-M)^T C^{-1} (X-M)\right]$ donde:

- Vector de Medias $\equiv M = E[X] = [m_1 \ \ m_2 \ \ \ldots \ \ m_n]^T$ con $m_i = E[x_i]$ (información del punto donde se agrupan los datos)

- Matriz de covarianzas $\equiv C = E[(X-M)(X-M)^T] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$

con $\sigma_i^2 = E\left[(x_i - m_i)^2\right]$ y $\sigma_{ij} = \sigma_{ji} = E\left[(x_i - m_i)(x_j - m_j)\right]$ (matriz simétrica) ($|C| \equiv$ determinante).

Si las características x_i de X son estadísticamente independientes: $\sigma_{ij} = 0 \ \ \forall i \neq j$ (matriz diagonal)

Distancia de Mahalanobis $\equiv D_M \rightarrow D_M{}^2 = (X - M)^T C^{-1} (X - M)$
– Distancia de un valor del vector X respecto al vector de medias, normalizada por la dispersión de los datos.

- **Clasificadores basados en distribuciones normales: caso general**

⇒ **La función de probabilidad a priori de la característica X para cada clase C_i , *p(X|C_i)*, se ajusta a una función gaussiana.**

⇒ **Para cada clase C_i , a partir de un conjunto de muestras conocidas, se determina su vector de medias M^i y su matriz de covarianzas C^i.**

Para un vector características $X=\left[x_1 \;\; x_2 \;\; \ldots \;\; x_n\right]^T : p\left(X|C_i\right)=\frac{1}{\left(2\pi\right)^{n/2}\left|C^i\right|^{1/2}}\exp\left[-\frac{1}{2}\left(X-M^i\right)^T\left(C^i\right)^{-1}\left(X-M^i\right)\right]$

Vector de medias de la clase $C_i \equiv M^i=\left[m_1^i \;\; m_2^i \;\; \ldots \;\; m_n^i\right]^T$

Matriz de covarianzas de la clase $C_i \equiv \quad C^i=\begin{bmatrix} \sigma^i{}_{11} & \sigma^i{}_{12} & \cdots & \sigma^i{}_{1n} \\ \sigma^i{}_{21} & \sigma^i{}_{22} & \cdots & \sigma^i{}_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma^i{}_{n1} & \sigma^i{}_{n2} & \cdots & \sigma^i{}_{nn} \end{bmatrix} \;;\; \sigma_{jk}{}^i=\sigma_{kj}{}^i=E\left[\left(x_j{}^i-m_j{}^i\right)\left(x_k{}^i-m_k{}^i\right)\right]$

⇒ **La función de decisión d_i para la clase C_i:** $\quad d_i\left(X\right)=\ln\left[p\left(X|C_i\right)\right]+\ln\left[P\left(C_i\right)\right]=$

$$=-\frac{1}{2}\left(X-M^i\right)^T\left(C^i\right)^{-1}\left(X-M^i\right)-\frac{n}{2}\log\left(2\pi\right)-\frac{1}{2}\log\left|C^i\right|+\ln\left[P\left(C_i\right)\right]$$

➢ **Clasificadores basados en distribuciones normales: *Clasificador de Mínima Distancia de Mahalanobis***

Condiciones:

1. **Las matrices de covarianzas de las clases son iguales: $C^1 = C^2 = \ldots = C^n = C$**

❖ **Suponemos que esta matriz de covarianza C no es diagonal, $\sigma_{ij} \neq 0$: las variables de las características elegidas x_i no son independientes, existiendo una correlación entre ellas.**

2. **Las probabilidades a priori de las clases son iguales: $P(C_1) = P(C_2) = \ldots = P(C_n)$**

⇒ Teniendo en cuenta las 2 condiciones los términos subrayados son constantes para todas las clases, por lo que pueden ser eliminados en las funciones de decisión:

$$d_i(X) = -\frac{1}{2}(X - M^i)^T (C^i)^{-1} (X - M^i) \underline{- \frac{n}{2}\log(2\pi)} \underline{- \frac{1}{2}\log|C^i|} + \underline{\ln[P(C_i)]}$$

$$\Downarrow$$

$$d_i(X) = -\frac{1}{2}(X - M^i)^T (C)^{-1} (X - M^i) = -\frac{1}{2}(D_M^i)^2$$

$$\Downarrow$$

$$d_i(X) = -(D_M^i)^2$$

⇒ Este clasificador asigna un patrón X a aquella clase C_i cuyo vector promedio M^i esté a una distancia de Mahalanobis (da una medida de la distancia al centro, media, de las clases) mínima.

- **Clasificadores basados en distribuciones normales:** ***Clasificador de Mínima Distancia Euclídea***

Condiciones:

1. **Las matrices de covarianzas de las clases son iguales: $C^1 = C^2 = \dots = C^n = C$**

2. **Las variables de las características elegidas x_i son estadísticamente independientes, no están correladas, siendo, además las varianzas de cada variable x_i son iguales: $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_n^2 = \sigma^2$.**

   $\Rightarrow$ **Matriz de covarianza diagonal, $\sigma_{ij} = 0$, $i \neq j$.**

$$C^i = C = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \sigma^2 I$$

3. **Las probabilidades a priori de las clases son iguales: $P(C_1) = P(C_2) = \dots = P(C_n)$.**

Ejemplo de clasificadores estadísticos (paramétricos): - Clasificadores basados en distribuciones normales

- **Clasificadores basados en distribuciones normales: *Clasificador de Mínima Distancia Euclídea***

⇒ **Teniendo en cuenta las condiciones y eliminando los términos constantes para todas las clases, las funciones de decisión para cada clase:**

$$\text{Funciones de decisión} \equiv d_i(X) = -\frac{1}{2}(X - M^i)^T (C^i)^{-1} (X - M^i) \underline{- \frac{n}{2}\log(2\pi)} \underline{- \frac{1}{2}\log|C^i|} + \underline{\ln[P(C_i)]}$$

$$\Downarrow$$

$$d_i(X) = -\frac{1}{2}(X - M^i)^T (C)^{-1} (X - M^i) = -\frac{1}{2\sigma^2}(X - M^i)^T (X - M^i)$$

$$\Downarrow$$

$$d_i(X) = -(X - M^i)^T (X - M^i) = -\sum_{f=1}^{n}(x_f - m_f^i)^2 = -(D_E^i)^2$$

$$\left\{ \text{Vector características} \equiv X = [x_1 \;\; x_2 \;\; \ldots \;\; x_n]^T \;\; ; \;\; \text{Vector de medias de la clase } C_i \equiv M^i = [m_1^i \;\; m_2^i \;\; \ldots \;\; m_n^i]^T \right\}$$

⇒ **Este clasificador asigna un patrón X a aquella clase C_i cuyo vector promedio M^i esté a una distancia Euclídea mínima.**

⇒ **El criterio es, por tanto, medir la distancia a cada prototipo de las clases y elegir aquel cuya distancia sea más pequeña – Clasificador mediante el vecino más próximo.**

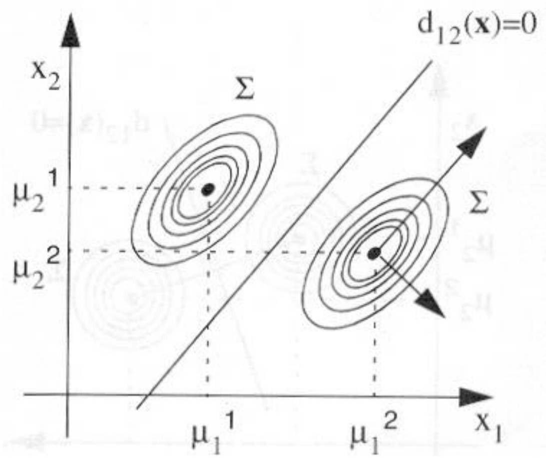
- **Clasificadores basados en distribuciones normales: *observaciones***

Clasificador de Mínima Distancia de Mahalanobis:

⇒ **El lugar geométrico de aquellos puntos equidistantes del centro (vector de medias) son hiperelipsoides.**

Clasificador de Mínima Distancia Euclídea

⇒ **El lugar geométrico de aquellos puntos equidistantes del centro (vector de medias) son hiperesferas.**



Clasificador basado en el criterio de mínima distancia de Mahalanobis

Para ambos clasificadores:

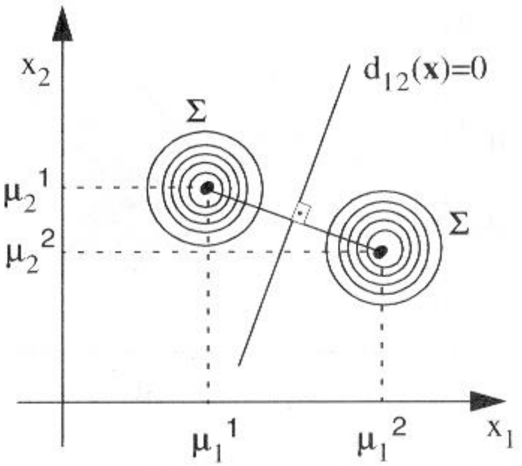
⇒ **Funciones de decisión cuadráticas y fronteras entre clases lineales.**

⇒ **Fronteras de decisión y criterio de clasificación:**

Frontera de las regiones de decisión entre las clases C_i y C_j:

$$d_{ij}(X) = d_i(X) - d_j(X) = 0 \quad \text{(aquellos puntos que satisfacen } d_i(X) = d_j(X) \text{)}$$

$$\text{Criterio de clasificación: Si } X \in C_i \;\Rightarrow\; d_i(X) > d_j(X) \;\; \forall j \neq i \;\;\equiv\;\; d_{ij}(X) > 0 \;\; \forall j \neq i$$



Clasificador basado en el criterio de mínima distancia Euclidea

RESUMEN (I): CLASIFICADORES BASADOS EN DISTRIBUCIONES NORMALES

$$d_i(X) = - (X - M^i)^T (C^i)^{-1} (X - M^i) - \log|C^i| + \ln[P(C_i)]$$

MÍNIMA DISTANCIA MAHALANOBIS:

- Matrices de covarianzas aproximadamente iguales (las clases presentan la misma dispersión en sus datos)
- Probabilidades a priori de cada clase iguales

Posibles diseños:

1. Diseñar una función de decisión por cada clase **(clasificador no lineal cuadrático)**. Elegir la clase que presenta una función máxima (en este caso, se puede utilizar la matriz de covarianza de la clase en cuestión):

$$\text{Funciones de decisión} \equiv d_i(X) = - (X - M^i)^T (C^i)^{-1} (X - M^i) = -(D_M^i)^2$$

$$X \in C_i \Rightarrow d_i(X) > d_j(X) \quad \forall j \neq i$$

2. Diseñar una función de decisión por cada dos clases para separar sus muestras (tantas funciones de decisión como combinaciones de m clases tomadas de dos haya). Estas funciones son lineales **(clasificador lineal)** si se asume una única matriz de covarianzas para todas las clases.

$$d_i(X) = - (X - M^i)^T (C)^{-1} (X - M^i) = - (D_M^i)^2 \qquad \textbf{con } C = \frac{1}{N_{TOTAL}} \sum_{i=1}^{m\ clases} N_i C_i$$

$$d_{ij}(X) = d_i(X) - d_j(X): \quad X \in C_i \Rightarrow d_i(X) > d_j(X) \quad \forall j \neq i \quad \equiv \quad d_{ij}(X) > 0 \quad \forall j \neq i$$

RESUMEN (II): CLASIFICADORES BASADOS EN DISTRIBUCIONES NORMALES

MÍNIMA DISTANCIA EUCLIDEA:

- Matrices de covarianzas aproximadamente iguales (las clases presentan la misma dispersión en sus datos)
- Los atributos son estadísticas independientes y además sus varianzas son iguales
- Probabilidades a priori de cada clase iguales

Posibles diseños:

1. Diseñar una función de decisión por cada clase **(clasificador no lineal cuadrático)**. Elegir la clase que presenta una función máxima.

$$d_i(X) = -\left(X - M^i\right)^T \left(X - M^i\right) = -\left(D_E^i\right)^2$$

$$X \in C_i \;\Rightarrow\; d_i(X) > d_j(X) \quad \forall j \neq i$$

2. Diseñar una función de decisión por cada dos clases para separar sus muestras **(clasificador lineal)**

$$d_{ij}(X) = d_i(X) - d_j(X) \qquad X \in C_i \;\Rightarrow\; d_i(X) > d_j(X) \quad \forall j \neq i \quad \equiv \quad d_{ij}(X) > 0 \quad \forall j \neq i$$

SI NO SE CUMPLEN LAS SUPOSICIONES ANTERIORES, CONSIDEAR LAS FUNCIONES DE DECISIÓN PLANTEADAS SIN ELIMINAR TÉRMINOS:

$$d_i(X) = -\left(X - M^i\right)^T \left(C^i\right)^{-1} \left(X - M^i\right) - \log\left|C^i\right| + \ln\left[P(C_i)\right]$$

$$X \in C_i \;\Rightarrow\; d_i(X) > d_j(X) \quad \forall j \neq i$$

- **Ejemplo 1: Diseñar un Clasificador de Mínima Distancia Mahalanobis suponiendo 3 clases de patrones, cada uno de ellos representados por 2 características, y los siguientes datos:**

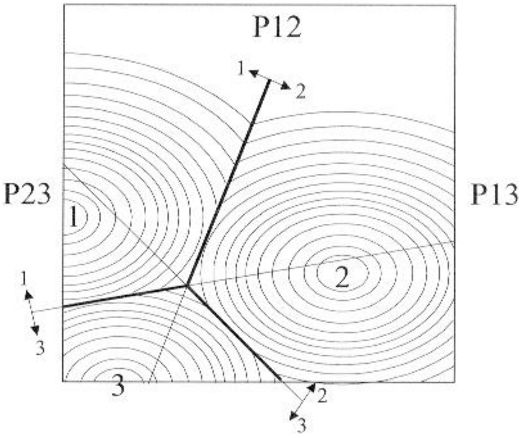
$$\mathbf{X}=\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \;;\; \mathbf{P}(C_1)=\mathbf{P}(C_2)=\mathbf{P}(C_3) \;;\; \mathbf{C}^1=\mathbf{C}^2=\mathbf{C}^3=\mathbf{C}=\begin{bmatrix} 1/2 & 0 \\ 0 & 1/4 \end{bmatrix} \left\{\Rightarrow \mathbf{C}^{-1}=\begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}\right\} ;$$

$$\mathbf{M}^1=\begin{bmatrix} 0 \\ 3 \end{bmatrix} ,\; \mathbf{M}^2=\begin{bmatrix} 5 \\ 2 \end{bmatrix} ,\; \mathbf{M}^3=\begin{bmatrix} 1 \\ 0 \end{bmatrix} \left\{\Rightarrow \text{Vector de medias de la clase } C_i \equiv \mathbf{M}^i=\begin{bmatrix} m_1^i & m_2^i & \ldots & m_n^i \end{bmatrix}^T\right\}$$

Funciones de decisión de cada clase: $d_i(X)=-\left(D_M^i\right)^2=-\left(X-M^i\right)^T(C)^{-1}\left(X-M^i\right)$

$$d_1(X)=-\begin{bmatrix} x_1-0 & x_2-3 \end{bmatrix}\begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}\begin{bmatrix} x_1-0 \\ x_2-3 \end{bmatrix}=-\begin{bmatrix} 2x_1 & 4x_2-12 \end{bmatrix}\begin{bmatrix} x_1-0 \\ x_2-3 \end{bmatrix}=-2x_1-4(x_2-3)^2$$

$$d_2(X)=-2(x_1-5)^2-4(x_2-2)^2 \quad ; \quad d_3(X)=-2(x_1-1)^2-4x_2^2$$



Fronteras de decisión: $d_{12}=d_1-d_2=-20x_1+8x_2+30=0$

$d_{12}\equiv -10x_1+4x_2+15=0 \quad ; \quad d_{23}\equiv x_1+x_2-4=0 \quad ; \quad d_{13}\equiv -2x_1+12x_2-17=0$

Criterio de clasificación: $X \in C_1$ si $d_{12}>0$ y $d_{13}>0$; $X \in C_2$ si $d_{12}<0$ y $d_{23}>0$; $X \in C_3$ si $d_{13}<0$ y $d_{23}<0$

- **Ejemplo 2: Teniendo en cuenta la muestra de la tabla, diseñar un Clasificador de Mínima Distancia Mahalanobis suponiendo que las dos clases tienen la misma matriz de covarianzas (para su estimación considerar conjuntamente los patrones de ambas clases):**

CLASE 1

PATRON	1	2	3	4
x_1	2	3	3	4
x_2	1	2	3	2

CLASE 2

PATRON	1	2	3
x_1	6	5	7
x_2	1	2	3

$$\text{Vector de medias: } M^i = \begin{bmatrix} m_1{}^i \\ m_2{}^i \end{bmatrix} = \begin{bmatrix} E\left[x_1{}^i\right] \\ E\left[x_2{}^i\right] \end{bmatrix} = \begin{bmatrix} \frac{1}{n_i}\sum_{k=1}^{n_i} x_{1k}{}^i \\ \frac{1}{n_i}\sum_{k=1}^{n_i} x_{2k}{}^i \end{bmatrix} \Rightarrow \begin{cases} M^1 = \begin{bmatrix} m_1{}^1 \\ m_2{}^1 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \\ M^2 = \begin{bmatrix} m_1{}^2 \\ m_2{}^2 \end{bmatrix} = \begin{bmatrix} 6 \\ 2 \end{bmatrix} \end{cases}$$

$$\text{Matriz de covarianzas: } C = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \quad \sigma_{jk}{}^i = E\left[\left(x_j{}^i - m_j{}^i\right)\left(x_k{}^i - m_k{}^i\right)\right] \quad j,k = 1,2$$

$$C = \frac{1}{n_1+n_2}\left[\sum_{f=1}^{n_1}\left(X_f{}^1 - M^1\right)\left(X_f{}^1 - M^1\right)^T + \sum_{f=1}^{n_2}\left(X_f{}^2 - M^2\right)\left(X_f{}^2 - M^2\right)^T\right] = \frac{1}{7}[\sum_{f=1}^{4}\left(\begin{bmatrix} x_{1f}{}^1 \\ x_{2f}{}^1 \end{bmatrix} - \begin{bmatrix} m_1{}^1 \\ m_2{}^1 \end{bmatrix}\right)\left(\begin{bmatrix} x_{1f}{}^1 \\ x_{2f}{}^1 \end{bmatrix} - \begin{bmatrix} m_1{}^1 \\ m_2{}^1 \end{bmatrix}\right)^T +$$

$$+\sum_{f=1}^{3}\left(\begin{bmatrix} x_{1f}{}^2 \\ x_{2f}{}^2 \end{bmatrix} - \begin{bmatrix} m_1{}^2 \\ m_2{}^2 \end{bmatrix}\right)\left(\begin{bmatrix} x_{1f}{}^2 \\ x_{2f}{}^2 \end{bmatrix} - \begin{bmatrix} m_1{}^2 \\ m_2{}^2 \end{bmatrix}\right)^T] = \frac{1}{7}\left[\sum_{f=1}^{4}\left(\begin{bmatrix} x_{1f}{}^1 - 3 \\ x_{2f}{}^1 - 2 \end{bmatrix}\begin{bmatrix} x_{1f}{}^1 - 3 & x_{2f}{}^1 - 2 \end{bmatrix}\right) + \sum_{f=1}^{3}\left(\begin{bmatrix} x_{1f}{}^2 - 6 \\ x_{2f}{}^2 - 2 \end{bmatrix}\begin{bmatrix} x_{1f}{}^2 - 6 & x_{2f}{}^2 - 2 \end{bmatrix}\right)\right] =$$

$$= \frac{1}{7}\left[\sum_{f=1}^{4}\begin{bmatrix} \left(x_{1f}{}^1 - 3\right)^2 & \left(x_{1f}{}^1 - 3\right)\left(x_{2f}{}^1 - 2\right) \\ \left(x_{2f}{}^1 - 2\right)\left(x_{1f}{}^1 - 3\right) & \left(x_{2f}{}^1 - 2\right)^2 \end{bmatrix} + \sum_{f=1}^{3}\begin{bmatrix} \left(x_{1f}{}^2 - 6\right)^2 & \left(x_{1f}{}^2 - 6\right)\left(x_{2f}{}^2 - 2\right) \\ \left(x_{2f}{}^2 - 2\right)\left(x_{1f}{}^2 - 6\right) & \left(x_{2f}{}^2 - 2\right)^2 \end{bmatrix}\right] = \rightarrow$$

- **Ejemplo 2:**

CLASE 1

PATRON	1	2	3	4
x_1	2	3	3	4
x_2	1	2	3	2

CLASE 2

PATRON	1	2	3
x_1	6	5	7
x_2	1	2	3

$$C = \frac{1}{7}\left[\sum_{f=1}^{4}\begin{bmatrix} \left(x_{1f}^{\ 1}-3\right)^2 & \left(x_{1f}^{\ 1}-3\right)\left(x_{2f}^{\ 1}-2\right) \\ \left(x_{2f}^{\ 1}-2\right)\left(x_{1f}^{\ 1}-3\right) & \left(x_{2f}^{\ 1}-2\right)^2 \end{bmatrix} + \sum_{f=1}^{3}\begin{bmatrix} \left(x_{1f}^{\ 2}-6\right)^2 & \left(x_{1f}^{\ 2}-6\right)\left(x_{2f}^{\ 2}-2\right) \\ \left(x_{2f}^{\ 2}-2\right)\left(x_{1f}^{\ 2}-6\right) & \left(x_{2f}^{\ 2}-2\right)^2 \end{bmatrix}\right] =$$

$$= \frac{1}{7}\left\{\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\right\} = \frac{1}{7}\begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix} \quad \Rightarrow \quad C^{-1} = \frac{7}{6}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

Funciones de decisión de cada clase: $d_i(X) = -\left(D_M^i\right)^2 = -\left(X - M^i\right)^T (C)^{-1}\left(X - M^i\right)$

$$d_1(X) = -\frac{7}{6}\begin{bmatrix} x_1-3 \\ x_2-2 \end{bmatrix}^T \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} x_1-3 \\ x_2-2 \end{bmatrix} = -\frac{7}{3}\left(x_1^{\ 2} + x_2^{\ 2} - 4x_1 - x_2 - x_1x_2 + 7\right)$$

$$d_2(X) = -\frac{7}{3}\left(x_1^{\ 2} + x_2^{\ 2} - 10x_1 + 2x_2 - x_1x_2 + 28\right)$$

- **Ejemplo 2:**

CLASE 1

PATRON	1	2	3	4
x_1	2	3	3	4
x_2	1	2	3	2

CLASE 2

PATRON	1	2	3
x_1	6	5	7
x_2	1	2	3

Frontera de decisión

$$\Downarrow$$

$$d_{12}(X) = d_1(X) - d_2(X) = 0$$

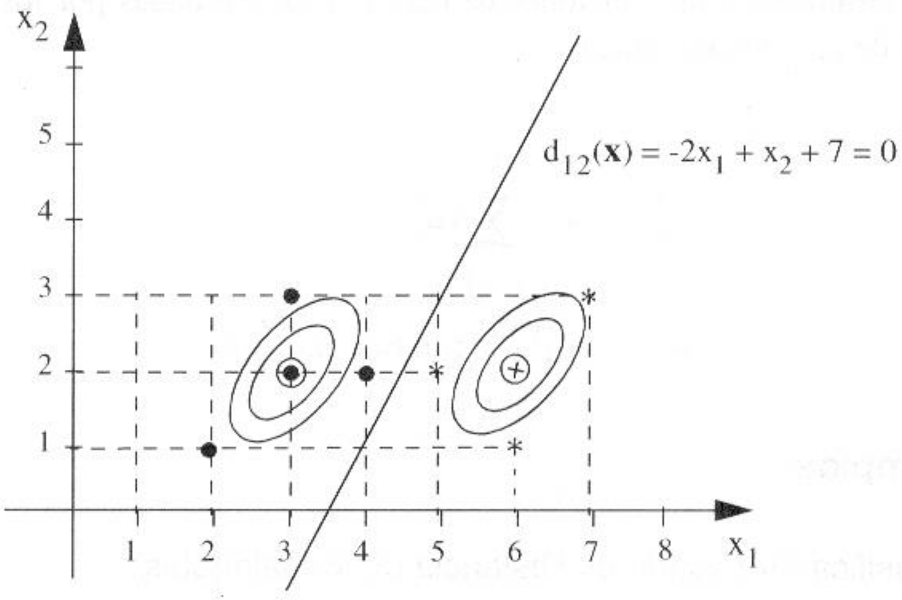
$$\Downarrow$$

$$d_{12}(X) \equiv -2x_1 + x_2 + 7 = 0$$

Criterio de clasificación

$$\Downarrow$$

$$X \in C_1 \ \text{si} \ d_{12}(X) \geq 0$$
$$X \in C_2 \ \text{si} \ d_{12}(X) < 0$$



Muestra controlada en el espacio de características. Se dibujan los vectores promedios, las elipses de iso-distancias y la frontera resultante de aplicar el criterio de Mínima Distancia de Mahalanobis.

- **Ejemplo 3: Teniendo en cuenta la muestra de la tabla, construir un Clasificador de Mínima Distancia Euclídea.**

CLASE 1

PATRON	1	2	3	4	5	6	7	8	9	10
x_1	1	2	2	2	2	3	3	4	5	1
x_2	3	1	2	3	4	2	3	3	2	2

CLASE 2

PATRON	1	2	3	4	5	6	7	8	9	10
x_1	4	5	5	4	6	6	6	7	4	8
x_2	5	5	6	7	5	6	7	6	6	7

$$\text{Vector de medias: } M^i = \begin{bmatrix} m_1{}^i \\ m_2{}^i \end{bmatrix} = \begin{bmatrix} E\left[x_1{}^i\right] \\ E\left[x_2{}^i\right] \end{bmatrix} = \begin{bmatrix} \frac{1}{n_i}\sum_{k=1}^{n_i} x_{1k}{}^i \\ \frac{1}{n_i}\sum_{k=1}^{n_i} x_{2k}{}^i \end{bmatrix} \Rightarrow M^1 = \begin{bmatrix} m_1{}^1 \\ m_2{}^1 \end{bmatrix} = \begin{bmatrix} 2.5 \\ 2.5 \end{bmatrix} \;;\; M^2 = \begin{bmatrix} m_1{}^2 \\ m_2{}^2 \end{bmatrix} = \begin{bmatrix} 5.5 \\ 6 \end{bmatrix}$$

- **Ejemplo 3: Teniendo en cuenta la muestra de la tabla, construir un Clasificador de Mínima Distancia Euclídea.**

$$\text{Funciones de decisión } \equiv d_i(X) = -\left(X - M^i\right)^T \left(X - M^i\right) = -\sum_{f=1}^{n}\left(x_f - m_f^i\right)^2 = -\left(D_E^i\right)^2$$

$$d_1(X) = -\left[x_1 - 2.5 \quad x_2 - 2.5\right]\begin{bmatrix} x_1 - 2.5 \\ x_2 - 2.5 \end{bmatrix} = -\left[\left(x_1 - 2.5\right)^2 + \left(x_2 - 2.5\right)^2\right] = -x_1^2 - x_2^2 + 5x_1 + 5x_2 - 12.5$$

$$d_2(X) = -\left[\left(x_1 - 5.5\right)^2 + \left(x_2 - 6\right)^2\right] = -x_1^2 - x_2^2 + 11x_1 + 12x_2 - 66.25$$

Frontera de decisión

$$\Downarrow$$

$$d_{12}(X) = d_1(X) - d_2(X) = 0$$

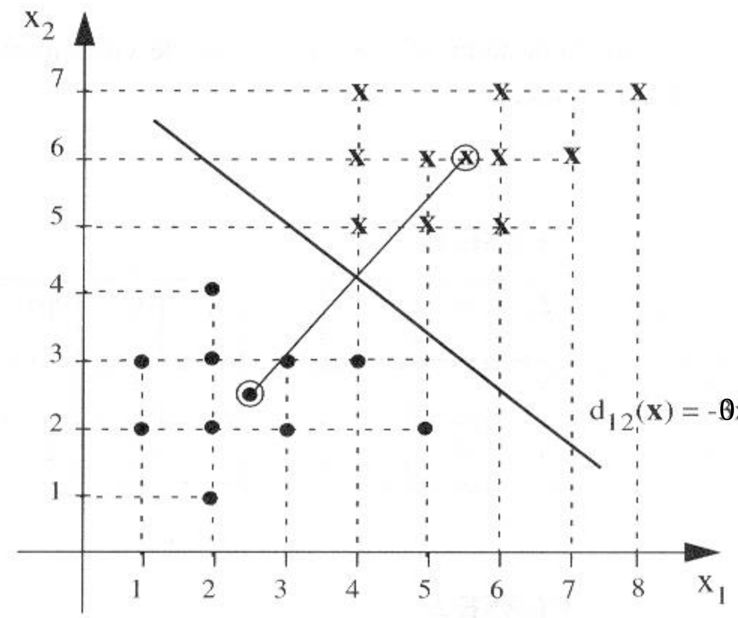
$$\Downarrow$$

$$d_{12}(X) \equiv -6x_1 - 7x_2 + 53.75 = 0$$

Criterio de clasificación

$$\Downarrow$$

$$X \in C_1 \text{ si } d_{12}(X) \geq 0$$
$$X \in C_2 \text{ si } d_{12}(X) < 0$$



5.1.- Introducción a problemas de clasificación

5.2.- Enfoque basado en la Teoría de la Decisión

5.3.- Técnicas básicas de clasificación

5.3.1.- Clasificadores bayesianos

5.3.2.- Clasificadores basados en distribuciones normales

5.3.3.- Clasificador k-vecinos más cercanos

5.4.- Desarrollo de sistema de reconocimiento de objetos

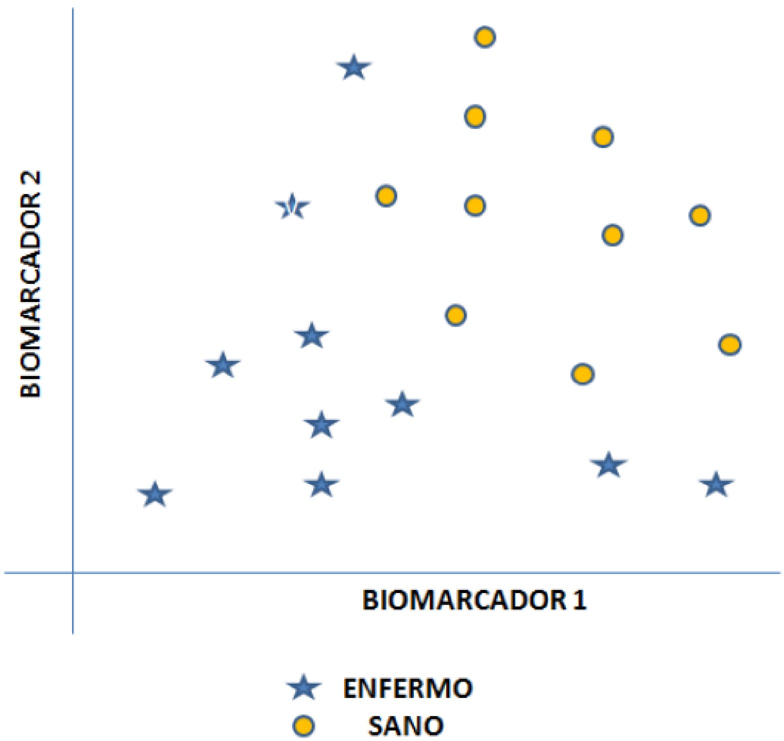
5.4.1.- Análisis y pre-procesamiento de datos

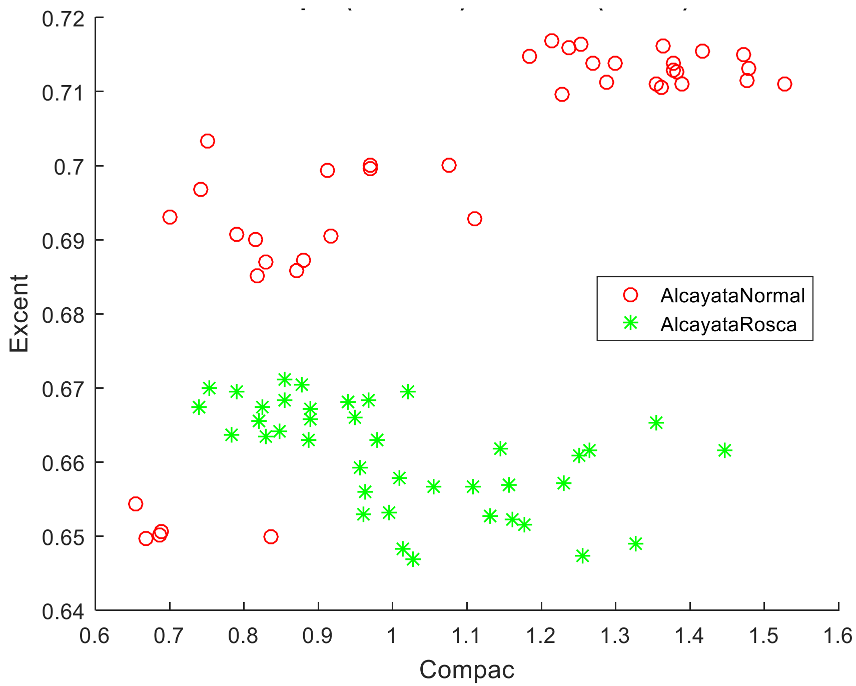
5.4.2.- Selección de características

5.4.3.- Evaluación de modelos

CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

- **EJEMPLOS:**



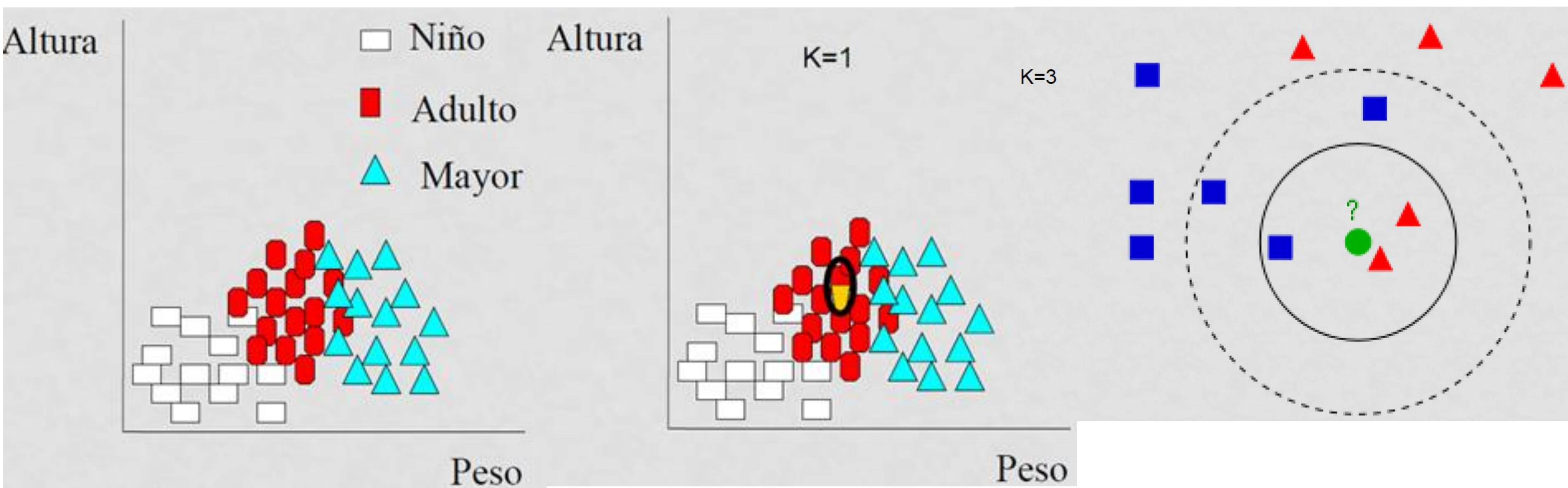


CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

IDEA GENERAL:

1. **Dado un conjunto de muestras prototipo, descrito por los correspondientes valores del vector de atributos, de clase conocida (patrones de entrenamiento) y dado una nueva muestra cuya clase es desconocida (patrón de test), se busca entre el conjunto de prototipos los "K" más parecidos al nuevo objeto.**

2. **La clase predicha para la instancia de test es la clase más numerosa de las clases a las que pertenecen las "K" muestras más cercanas.**



CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

- **"K MÁS PARECIDOS"**

- **MEDIDAS DE SIMILITUD (SEMEJANZA) / DISIMILITUD (DESEMEJANZA) (*SIMILARITY / DISSIMILARITY*):**

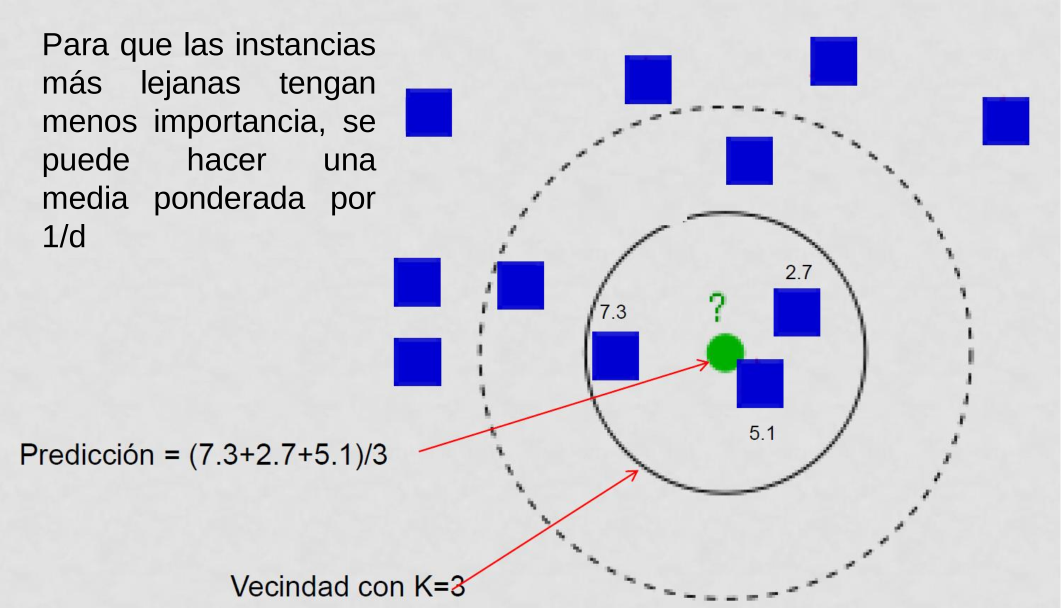
  1. **SIMILITUD (Ejemplo: correlación, coseno)**
     - → Medida numérica del parecido de dos muestras de datos
     - → Valores más altos indican muestras más parecidas
     - → Generalmente su rango de variación se sitúa entre 0 y 1
  2. **DISIMILITUD (Ejemplo: distancia)**
     - → Medida numérica de cómo de diferentes son dos muestras de datos
     - → Valores más bajos indican muestras más parecidas
     - → El mínimo valor es generalmente 0, el valor máximo puede diferir dependiendo de la métrica

CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

CARACTERÍSTICAS GENERALES

- **KNN para clasificación:** predice la clase de un instancia como la clase mayoritaria de entre los *k* vecinos mas cercanos de entre los datos de entrenamiento.

  → **Si clases continuas (regresión):** se elige la media de las clases de los k vecinos más cercanos



- **Algoritmo "perezoso" (*lazy*):**
  - K-NN no genera un modelo fruto del aprendizaje con datos de entrenamiento, sino que el aprendizaje sucede en el mismo momento en el que se prueban los datos de test.

    → Durante el entrenamiento, sólo "guardan" las instancias, no se construye ningún modelo.

    → La clasificación se hace cuando llega la instancia de test.

- **Es no paramétrico:** no se hacen suposiciones sobre la distribución que siguen los datos (como, por ejemplo, hacen clasificadores basados en distribuciones normales); el mejor modelo de los datos son los propios datos.

- **Es local:** asume que la clase de un dato depende sólo de los *k* vecinos mas cercanos (no se construye un modelo global)

CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

SELECCIÓN DEL VALOR DE K

- **K: hiperparámetro de este modelo de clasificación**

→ **Parámetros de un modelo:** son las variables que se estiman durante el proceso de entrenamiento con los conjuntos de datos (coeficientes en una regresión lineal, pesos en una red neuronal, vectores de soporte en una máquina de vectores de soporte).

→ **Hiperparámetros de un modelo:** son parámetros que se configuran antes del entrenamiento del modelo y no forman parte del modelo como tal; generalmente, sus valores óptimos no se conocen a priori, para establecerlos se deben utilizar reglas genéricas, valores que han funcionado en problemas similares o ajustarlos mediante prueba y error (mediante validación cruzada o, si es demasiado costoso en tiempo, utilizando un único conjunto de validación).

- **K = 1:** las instancias que son ruido (o solape entre clases) tienen mucha influencia
- **K > 1:** se consideran mas vecinos y las instancias de ruido pierden influencia
- **K muy altos:** se pierde la idea de localidad.

  → Para evitar que los vecinos lejanos tengan mucha influencia, se puede hacer que cada vecino vote de manera inversamente proporcional a la distancia (así, cuanto más lejos, tiene menos peso en la votación)

CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

LIMITACIONES / SOLUCIÓN

- **Muy sensible a los atributos irrelevantes**

  → **Fase de pre-proceso (selección de atributos)**

- **Muy sensible al aumento de la dimensión del problema (maldición de la dimensionalidad)**

  → **Fase de pre-proceso (selección de atributos)**

- **Lento, si hay muchos datos de entrenamiento (en almacenamiento y en tiempo)**

  → **Fase de pre-proceso (selección de instancias): eliminación de instancias superfluas**

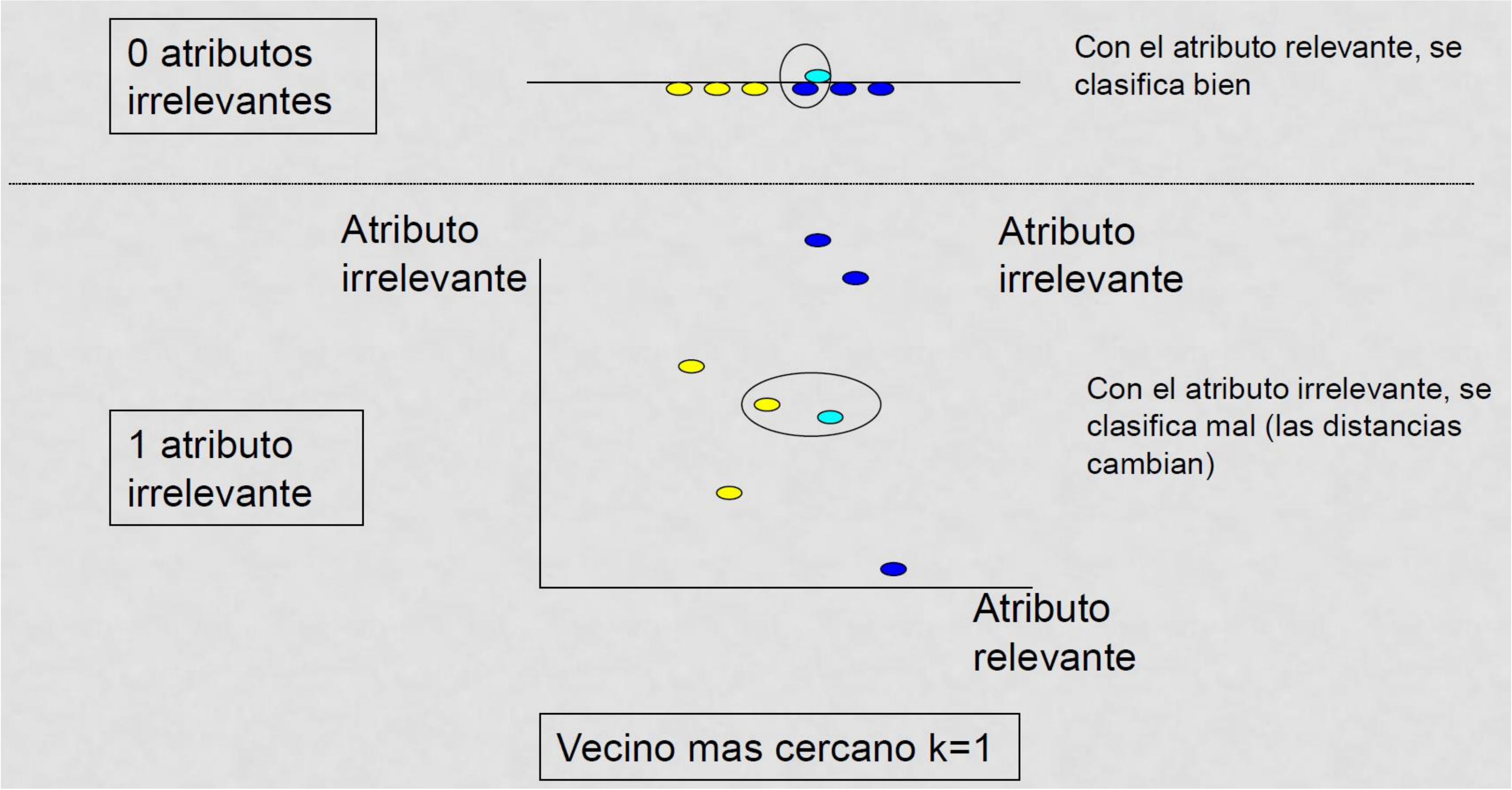
- **Muy sensible al ruido o instancias "engañosas"**

  → **Ajuste del hiperparámetro del número de vecinos (K)**

  → **Fase de pre-proceso (selección de instancias): eliminación de instancias "engañosas" o de solape entre clases**

CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

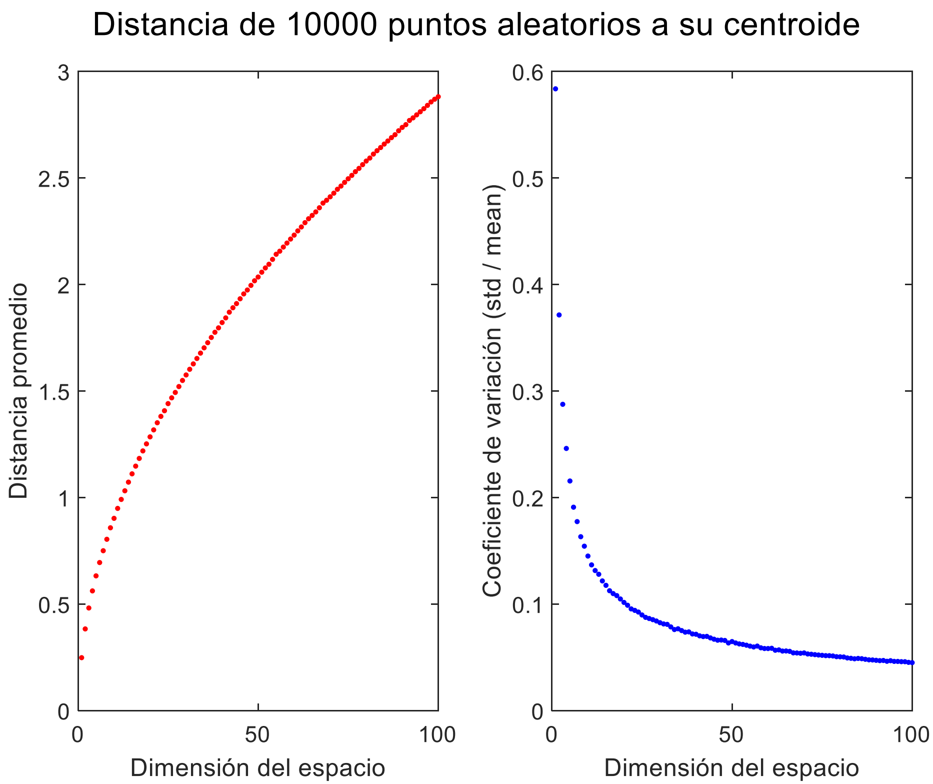
LIMITACIÓN: Sensible a atributos irrelevantes



CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

LIMITACIÓN: sensible al aumento de la dimensión del problema (maldición de la dimensionalidad)

- **La maldición de la dimensión:**
  - La distancia media entre los datos aumenta con el número de dimensiones
  - La variabilidad de la distancia disminuye exponencialmente con el número de dimensiones. Este es el verdadero problema:
    - → Cuando hay un gran número de atributos, los datos están todos a casi la misma distancia. Es decir, no hay variabilidad entre sus distancias y es más difícil saber qué puntos están más cerca de otros.



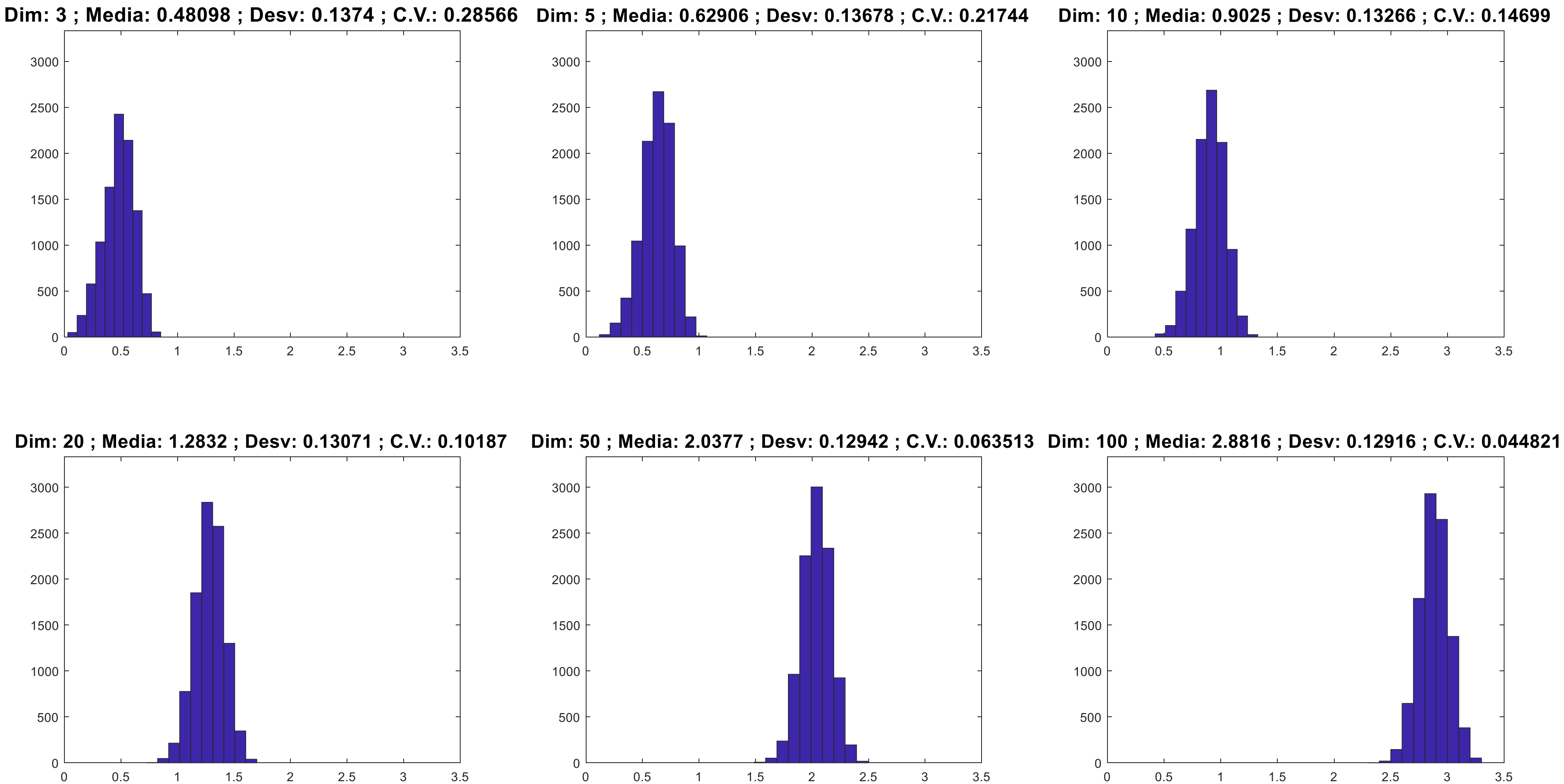
- **Solución: Aplicar selección de atributos o aumentar (exponencialmente) la cantidad de datos**

→ A más dimensiones, más datos se necesitan para llenar el espacio. Cuando el número de dimensiones es muy alto, el espacio está casi vacío. El aumento de datos debe ser exponencial con el número de dimensiones, para contrarrestar el efecto exponencial de la pérdida de variabilidad de las distancias.

CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

Ejemplo Maldición de la Dimensionalidad

Histograma - Distancia de 10000 puntos aleatorios a su centroide



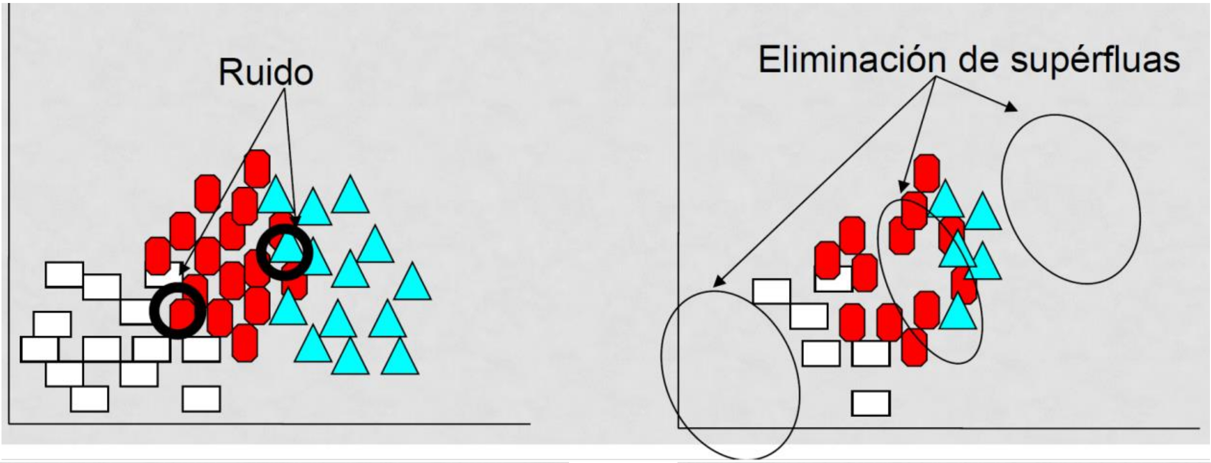
CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

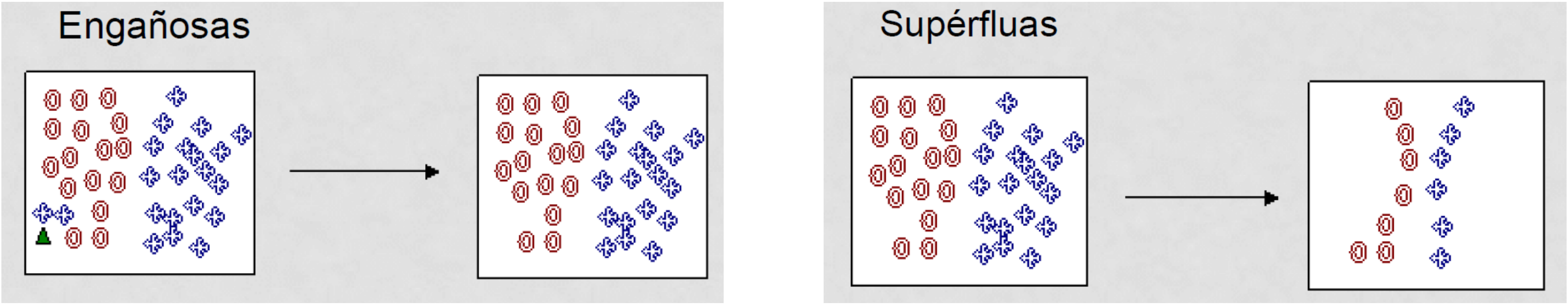
LIMITACIÓN: sensible a instancias "ruidosas" y lento en caso de conjuntos de datos muy grandes

SOLUCIÓN: SELECCIÓN DE INSTANCIAS EN UNA FASE DE PREPROCESO

→ **ELIMINACIÓN DE INSTANCIAS "RUIDOSAS":** instancias engañosas o de solape entre clases que confunden al clasificador – si se eliminan mejorará el porcentaje de aciertos en la clasificación.

→ **ELIMINACIÓN DE INSTANCIAS SUPÉRFLUAS:** instancias que no aumentan el porcentaje de aciertos en la clasificación – si se eliminan se decrementará el tiempo de clasificación.





CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

SELECCIÓN DE INSTANCIAS EN UNA FASE DE PREPROCESO

TÉCNICAS DE ELIMINACIÓN DE INSTANCIAS RUIDOSAS: ALGORITMOS DE EDICIÓN (*EDITING*)

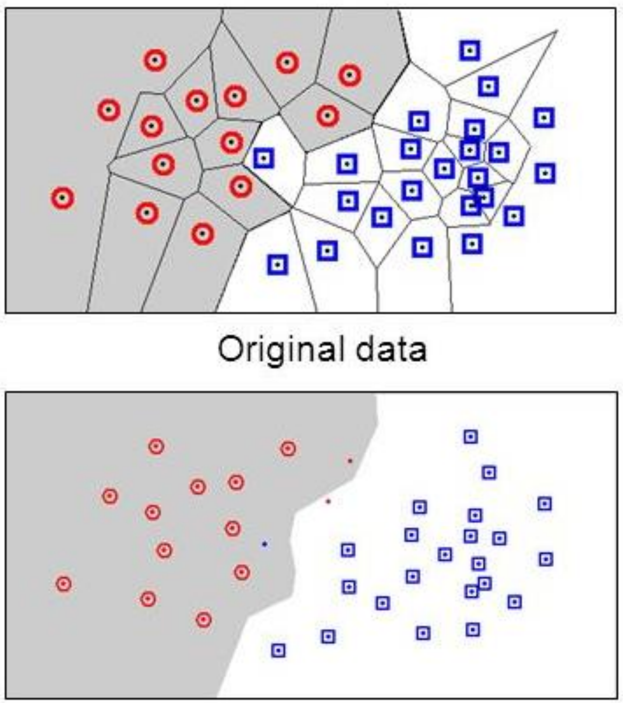
- ***WILSON EDITING*:** Elimina una instancia $\mathbf{x_i}$ si es clasificada incorrectamente por sus **k** vecinos

- ***REPEATED WILSON EDITING*:** repite *Wilson Editing* hasta que no que se eliminen instancias

(Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data sets. IEEE Trans. on Systems, Man and Cybernetics 2, 408–421, 1972)

En los ejemplos de los datos originales, se muestran las superficies de decisión inducidas por la aplicación de un 1-NN (diagrama de Vonoroi)

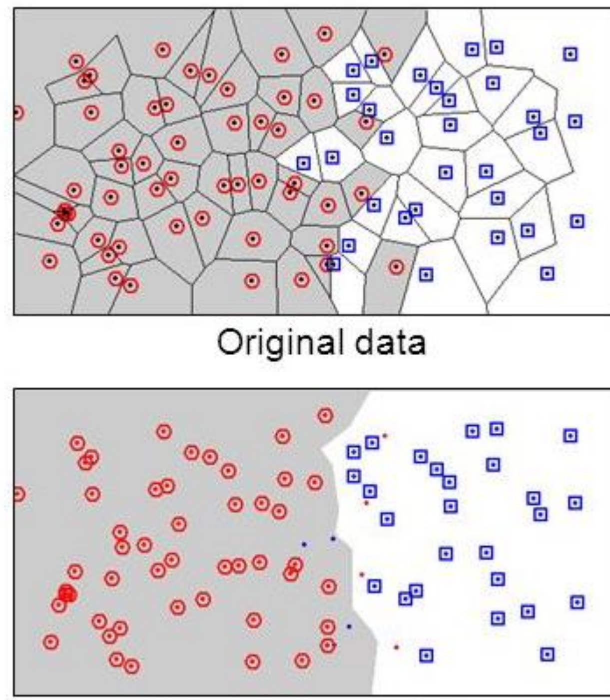
Example



Original data

Wilson editing with k=7

Overlapping classes



Original data

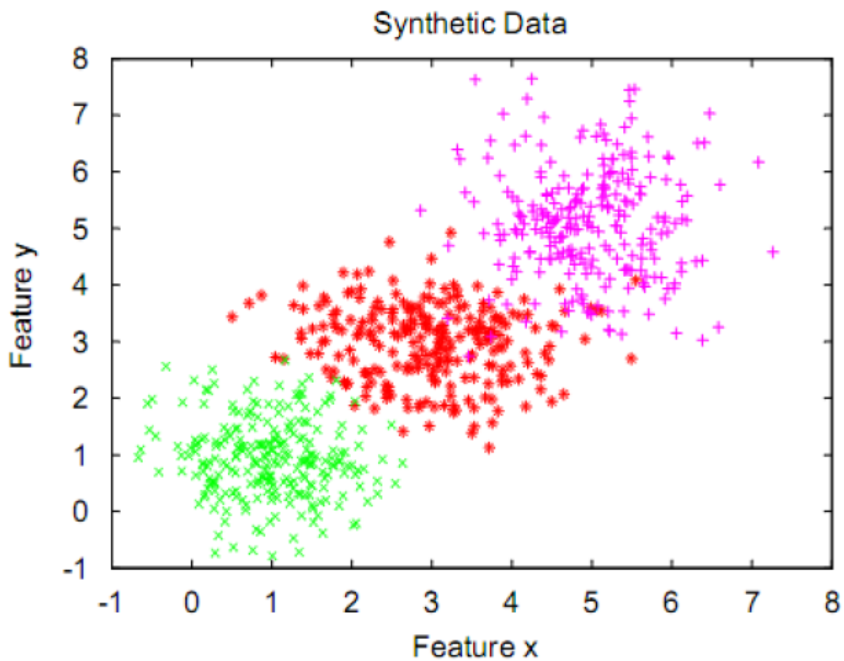
Wilson editing with k=7

CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

SELECCIÓN DE INSTANCIAS EN UNA FASE DE PREPROCESO

TÉCNICAS DE ELIMINACIÓN DE INSTANCIAS RUIDOSAS: ALGORITMOS DE EDICIÓN (*EDITING*)

- ***WILSON EDITING*:** otro ejemplo



- **Elimina pocos datos; funciona bien si no hay demasiado ruido**
  - → Elimina bien excepciones, ruido aislado, en el interior de una clase
  - → Si hay mucho ruido, las instancias con ruido clasificarán bien a otras instancias con ruido.
  - → Elimina puntos en las fronteras (suaviza las fronteras)

CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

SELECCIÓN DE INSTANCIAS EN UNA FASE DE PREPROCESO

TÉCNICAS DE ELIMINACIÓN DE INSTANCIAS SUPERFLUAS: ALGORITMOS DE CONDENSACIÓN (*CONDENSED NEAREST NEIGHBOURS – CNN*)

- **CNN – ALGORITMO DE HART** (Hart, P. E. The Condensed Nearest Neighbor Rule". IEEE Transactions on Information Theory 18, 515–516, 1968)

→ **Objetivo:** reducir el número de instancias, eliminando las superfluas.

→ **Idea general:** generar un "almacén" de instancias críticas; se recorren las instancias del conjunto de datos original y, si esa instancia se clasifica bien con las que ya hay en el "almacén", no pasa a formar parte de él – sólo se mueven al "almacén" aquellas instancias que no se clasifican bien con las instancias ya almacenadas.

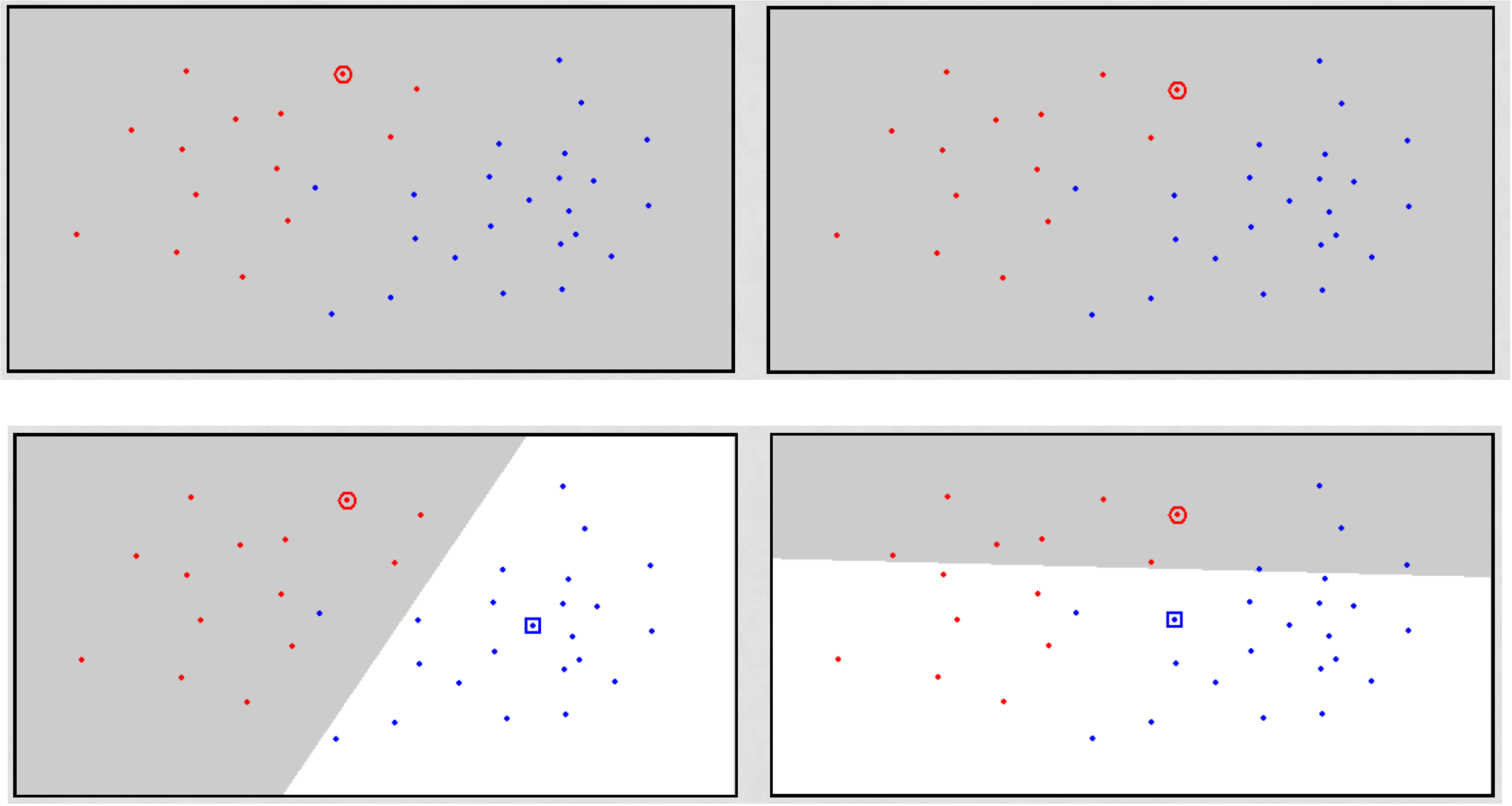
1. Inicializar el "almacén" incorporando una instancia x_1 del conjunto original (se mueve) .
2. Incorporar al "almacén" otra instancia x_i que NO se clasifique bien con las instancias del "almacén".
3. Repetir 2 hasta que no se muevan más instancias del conjunto de instancias restantes de fuera al "almacén".
4. Utilizar para clasificar el conjunto de instancias del "almacén" en lugar del original.

CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

SELECCIÓN DE INSTANCIAS EN UNA FASE DE PREPROCESO

TÉCNICAS DE ELIMINACIÓN DE INSTANCIAS SUPERFLUAS: ALGORITMOS DE CONDENSACIÓN (*CONDENSED NEAREST NEIGHBOURS – CNN HART ALGORITHM*)

→ **Limitación:** el resultado final depende de la inicialización y orden en el que toman las instancias. Ejemplo:

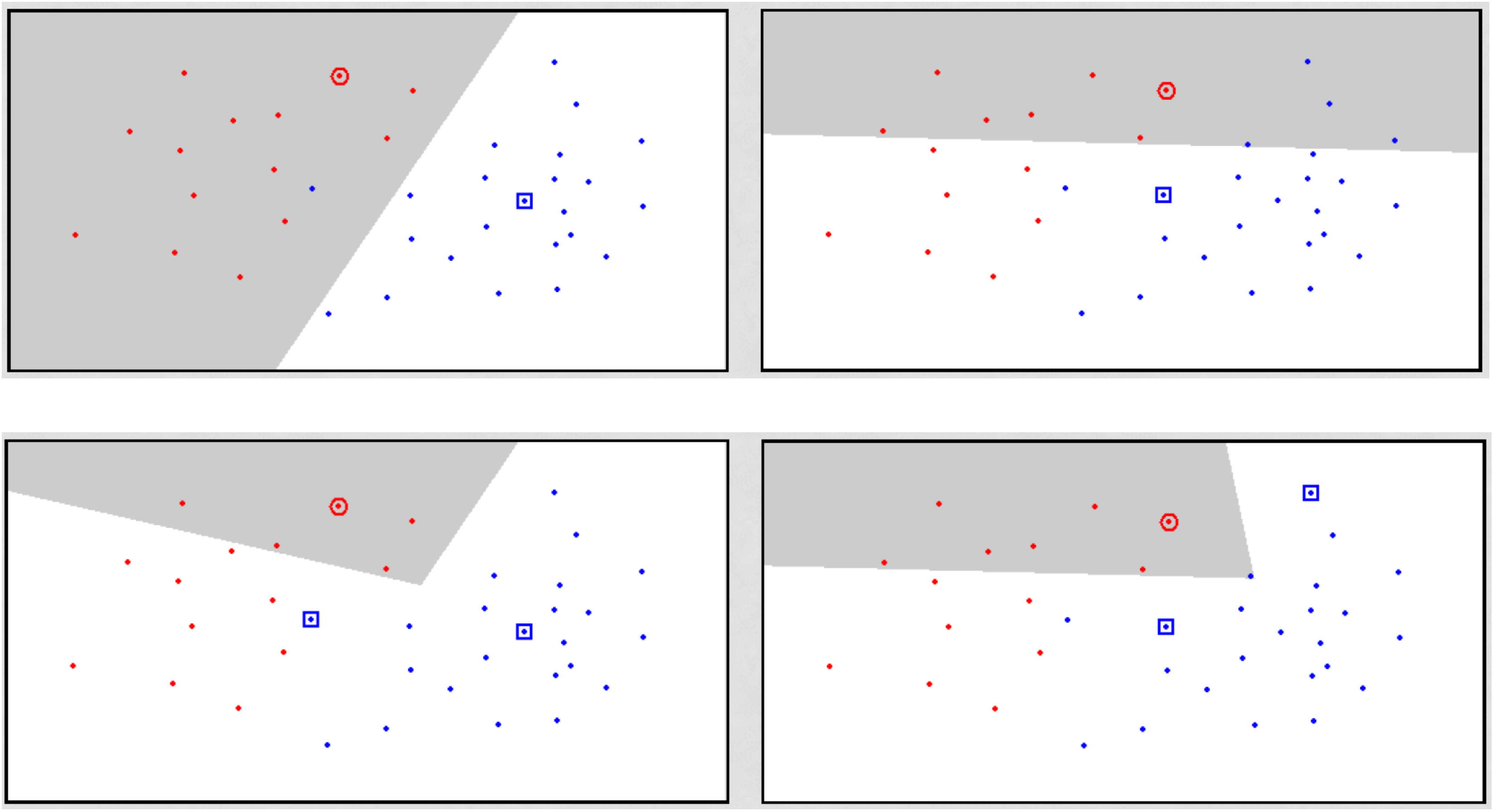


CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

SELECCIÓN DE INSTANCIAS EN UNA FASE DE PREPROCESO

TÉCNICAS DE ELIMINACIÓN DE INSTANCIAS SUPERFLUAS: ALGORITMOS DE CONDENSACIÓN (*CONDENSED NEAREST NEIGHBOURS – CNN HART ALGORITHM*)

→ **Limitación:** el resultado final depende de la inicialización y orden en el que toman las instancias. Ejemplo:

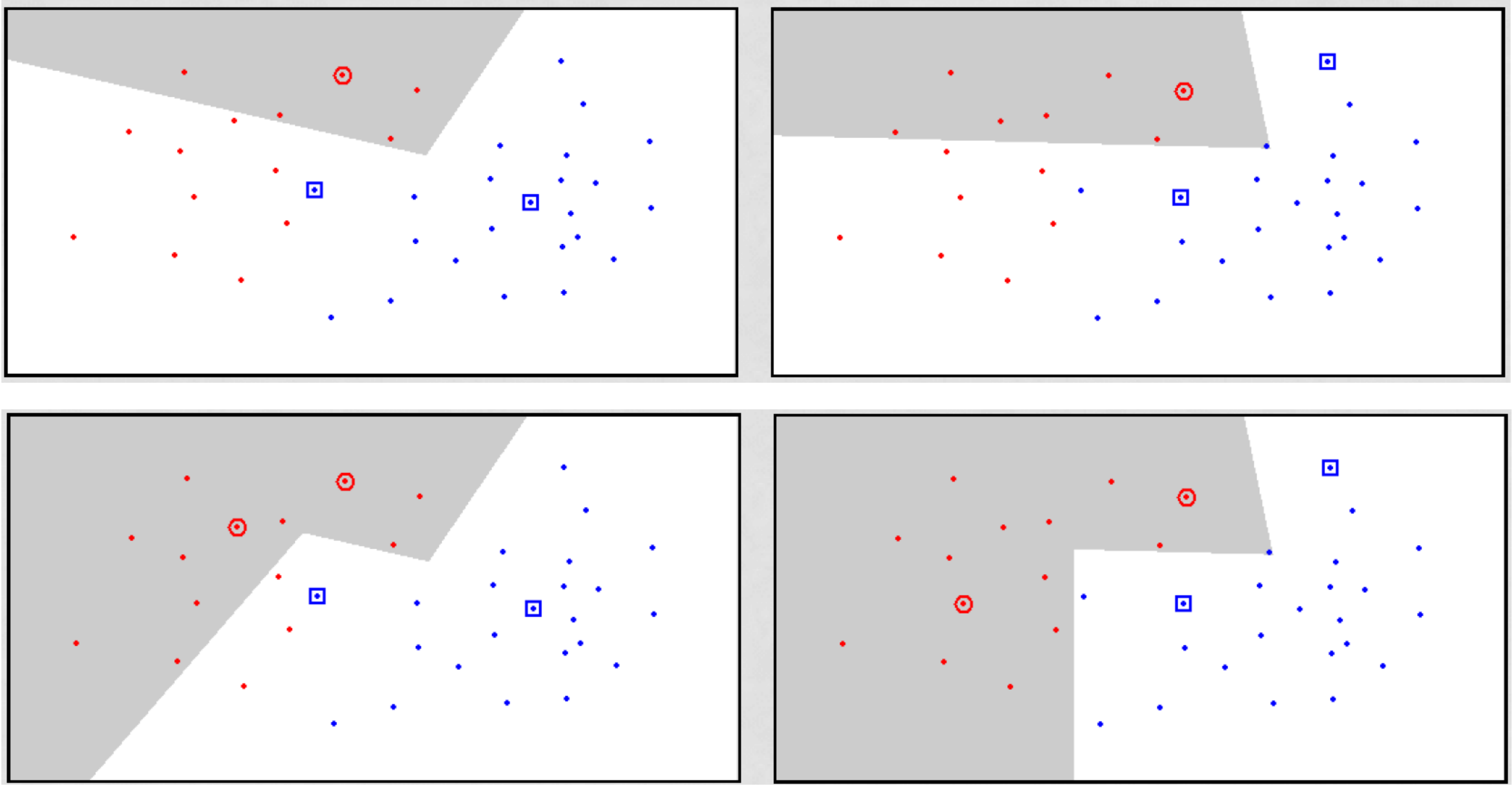


CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

SELECCIÓN DE INSTANCIAS EN UNA FASE DE PREPROCESO

TÉCNICAS DE ELIMINACIÓN DE INSTANCIAS SUPERFLUAS: ALGORITMOS DE CONDENSACIÓN (*CONDENSED NEAREST NEIGHBOURS – CNN HART ALGORITHM*)

→ **Limitación:** el resultado final depende de la inicialización y orden en el que toman las instancias. Ejemplo:

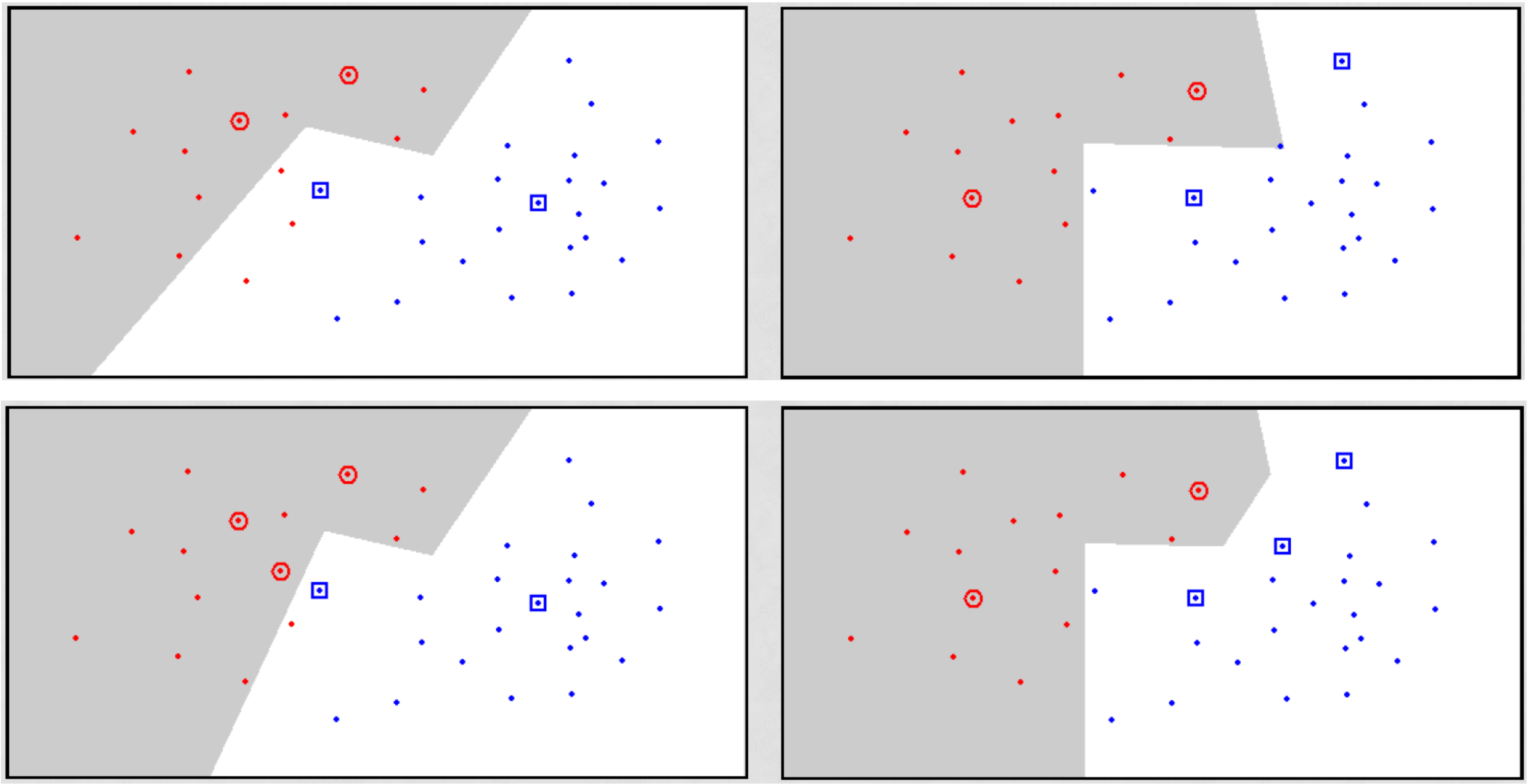


CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

SELECCIÓN DE INSTANCIAS EN UNA FASE DE PREPROCESO

TÉCNICAS DE ELIMINACIÓN DE INSTANCIAS SUPERFLUAS: ALGORITMOS DE CONDENSACIÓN (*CONDENSED NEAREST NEIGHBOURS – CNN HART ALGORITHM*)

→ **Limitación:** el resultado final depende de la inicialización y orden en el que toman las instancias. Ejemplo:

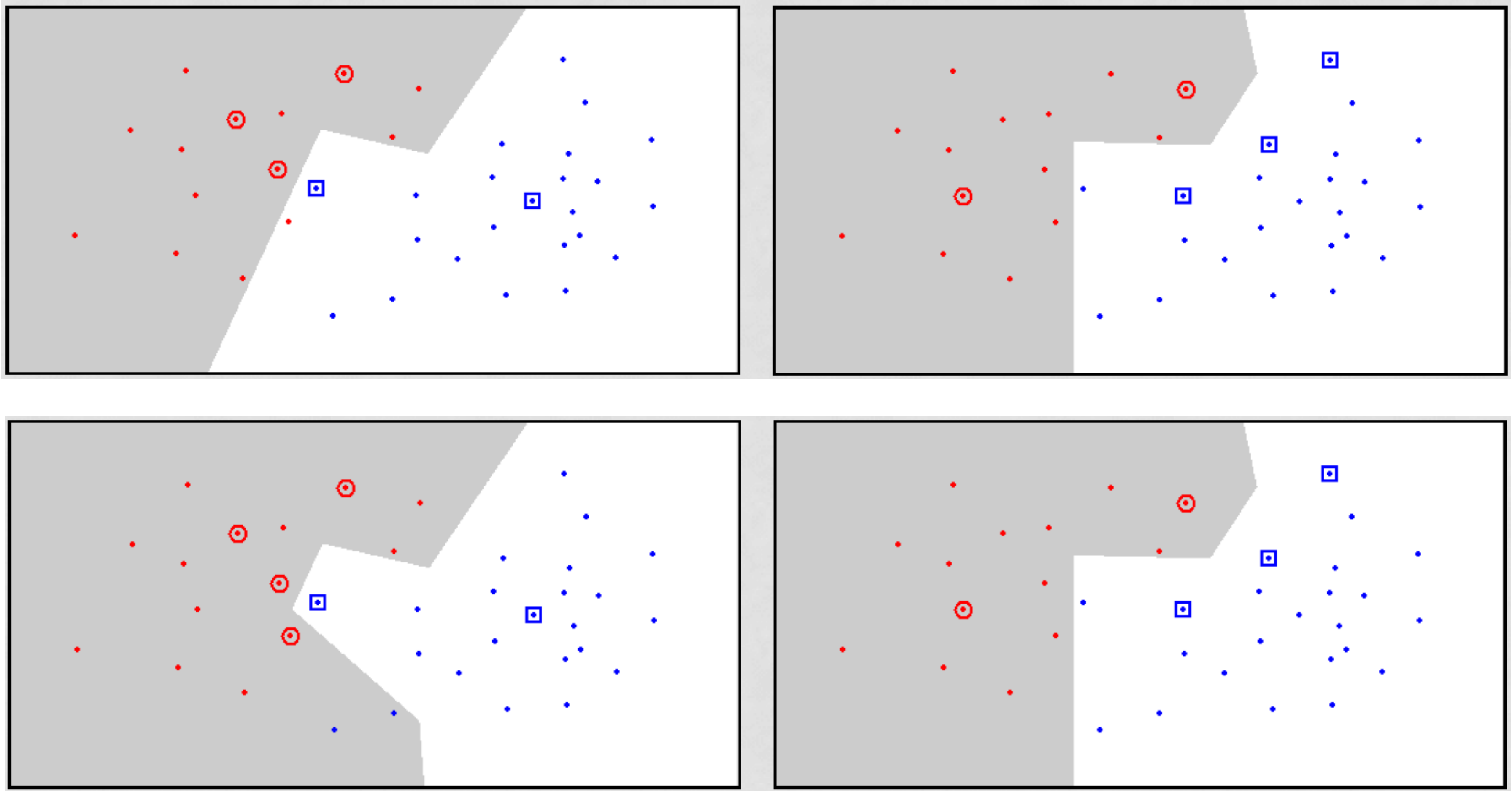


CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

SELECCIÓN DE INSTANCIAS EN UNA FASE DE PREPROCESO

TÉCNICAS DE ELIMINACIÓN DE INSTANCIAS SUPERFLUAS: ALGORITMOS DE CONDENSACIÓN (*CONDENSED NEAREST NEIGHBOURS – CNN HART ALGORITHM*)

→ **Limitación:** el resultado final depende de la inicialización y orden en el que toman las instancias. Ejemplo:

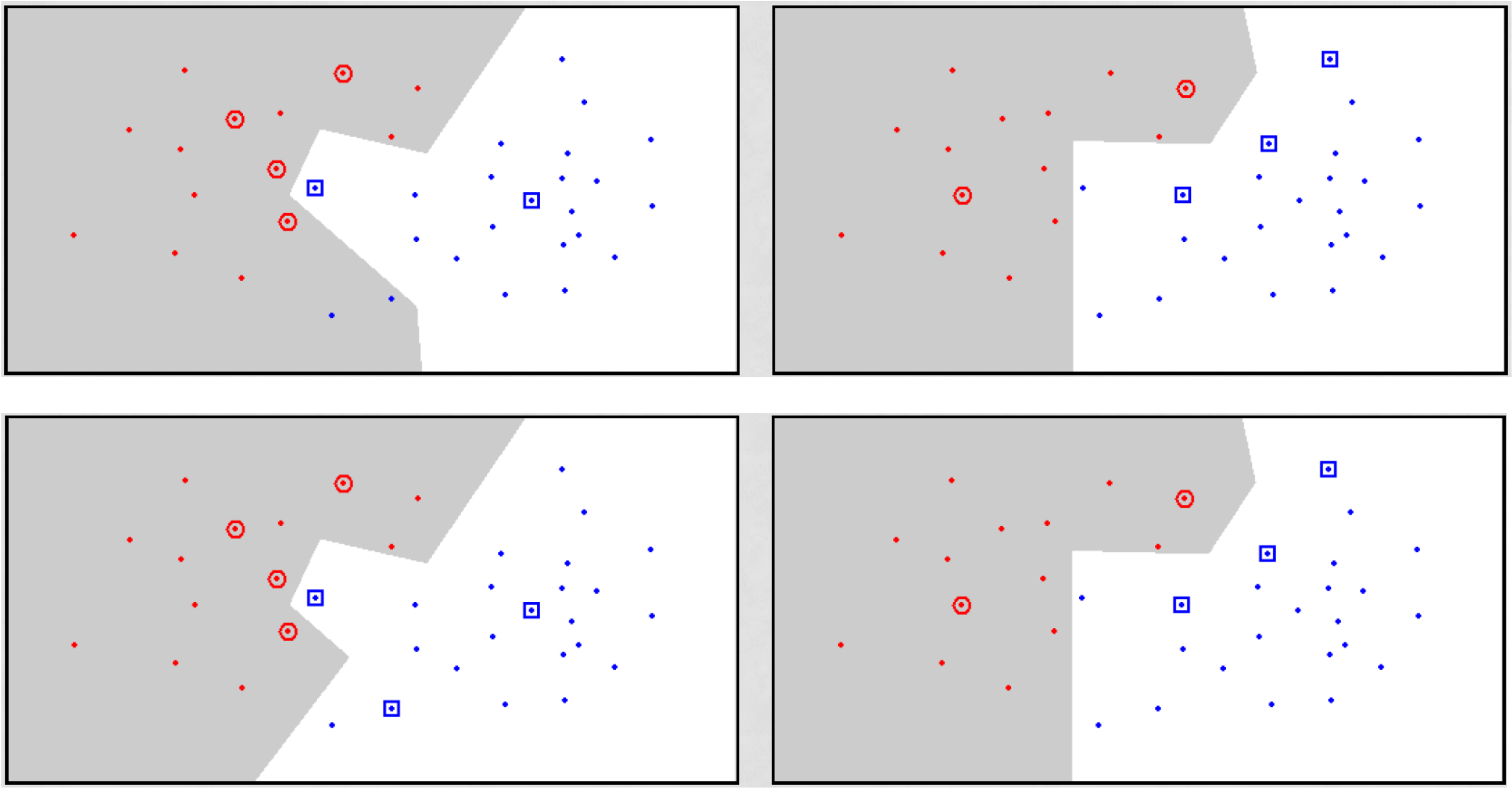


CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

SELECCIÓN DE INSTANCIAS EN UNA FASE DE PREPROCESO

TÉCNICAS DE ELIMINACIÓN DE INSTANCIAS SUPERFLUAS: ALGORITMOS DE CONDENSACIÓN (*CONDENSED NEAREST NEIGHBOURS – CNN HART ALGORITHM*)

→ **Limitación:** el resultado final depende de la inicialización y orden en el que toman las instancias. Ejemplo:



CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

SELECCIÓN DE INSTANCIAS EN UNA FASE DE PREPROCESO

TÉCNICAS DE ELIMINACIÓN DE INSTANCIAS SUPERFLUAS: ALGORITMOS DE CONDENSACIÓN (*CONDENSED NEAREST NEIGHBOURS – CNN HART ALGORITHM*)

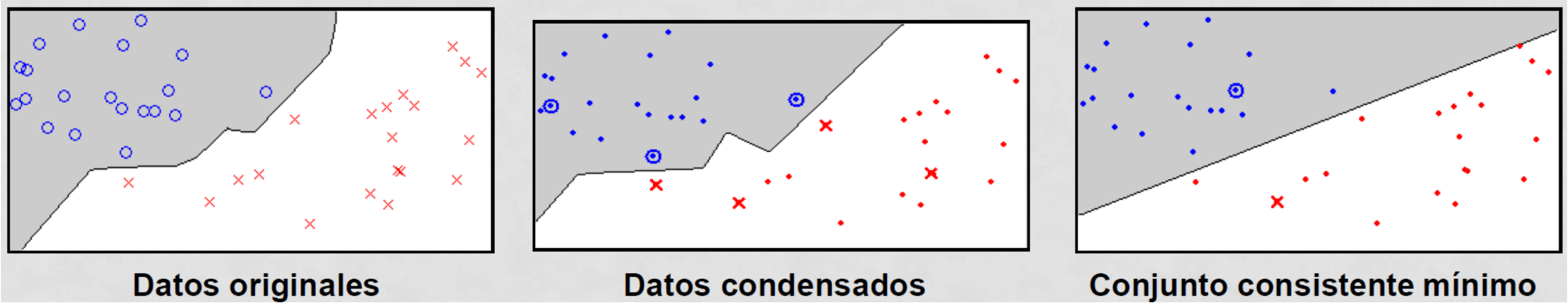
- **CNN – ALGORITMO DE HART - Resumiendo:**

→ **Elimina todas aquellas instancias no críticas para la clasificación (reduce mucho la necesidad de almacenamiento).**

→ **Tiende a conservar aquellas instancias con ruido (puesto que son mal clasificadas por las instancias en el "almacén").**

→ **Depende mucho del orden en el que se toman las instancias**

→ **No hay garantía de que se genera un conjunto de datos consistente mínimo**

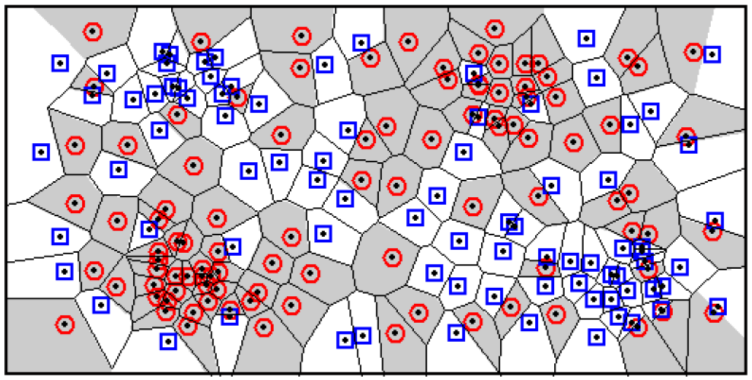


CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

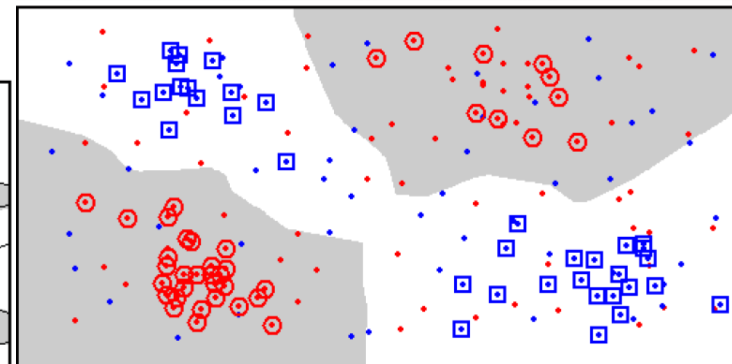
ALGORITMOS DE EDICIÓN Y DE CONDENSACIÓN: Resumiendo

- **EDICIÓN:** elimina el ruido y suaviza las fronteras, pero mantiene la mayor parte de los datos (mejora la capacidad de generalización pero no mejora la eficiencia)
- **CONDENSACIÓN:** elimina gran cantidad de datos superfluos, pero mantiene los datos con ruido (se mantienen aquellos datos clasificados mal por los demás datos (y el ruido tiene siempre esta propiedad)

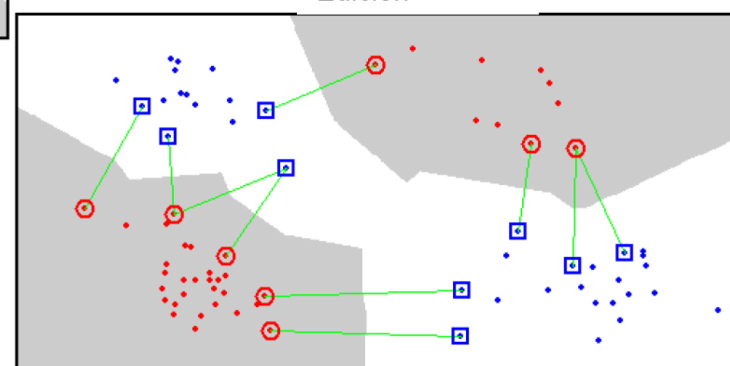
ALGORITMOS HÍBRIDOS: 1.- EDITAR ; 2.- CONDENSAR



Original



Edición



Condensación

Ejemplos de algoritmos híbridos avanzados:

RT3:
D. Randall Wilson, Tony R. Martinez: Instance Pruning Techniques. ICML 403-411, 1997.

Iterative case filtering (ICF):
Henry Brighton, Chris Mellish: Advances in Instance Selection for Instance-Based Learning Algorithms. Data Min. Knowl. Discov. 6(2): 153-172, 2002.

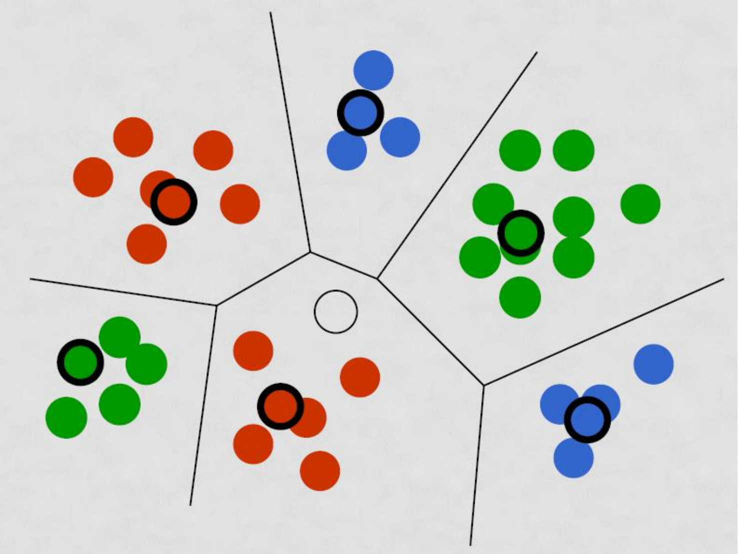
CLASIFICADOR KNN – "K-VECINOS MÁS PRÓXIMOS" (K – NEAREST NEIGHBOURS)

CLASIFICACIÓN BASADA EN PROTOTIPOS:

→ Cada prototipo tiene una etiqueta

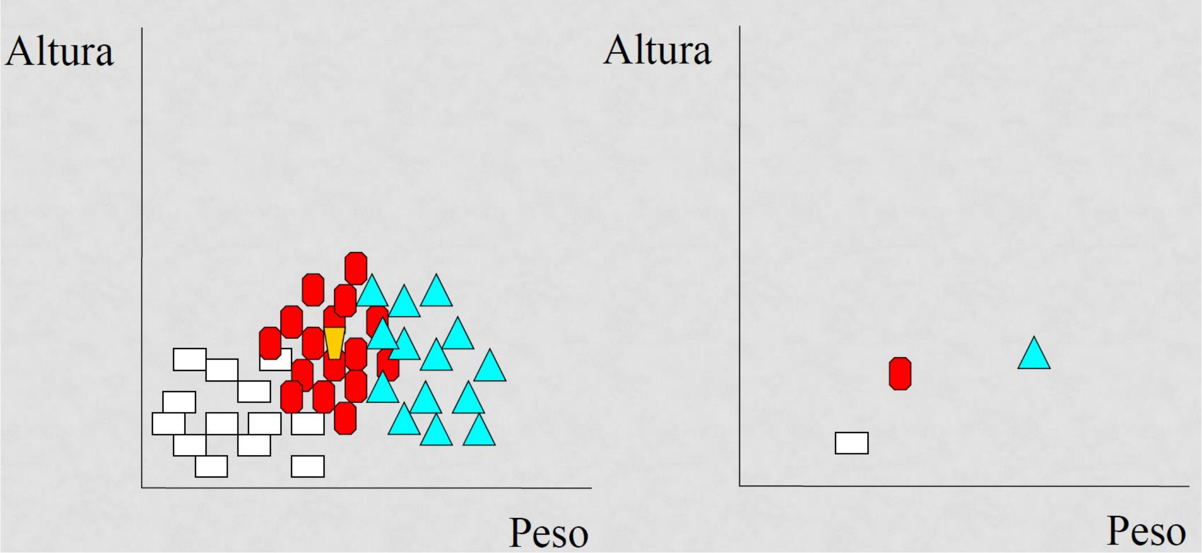
→ Se clasifica según la clase del prototipo más cercano (o según sus regiones de Voronoi)

→ Mejora la eficiencia en espacio (sólo se guardan unos pocos prototipos) y en tiempo (se computan muchas menos distancias cuando llega el dato de test)



→ Algoritmo de posicionamiento de prototipos: LVQ - Learning Vector Quantization

Kohonen T. Learning Vector Quantization. In: Self-Organizing Maps. Springer Series in Information Sciences, vol 30. Springer, Berlin, Heidelberg (1995 ; 2001)



CLASIFICADOR KNN – CUESTIONES DE IMPLEMENTACIÓN

PLANTEAMIENTO DE CLASIFICACIÓN: *m* clases, vector de atributos *p*-dimensional definido, *n* instancias de entrenamiento → predecir la clase de una instancia de test dada por su vector de atributos

$$\boldsymbol{X_{TRAIN}} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \ ; \ \boldsymbol{Y_{TRAIN}} = \begin{bmatrix} C_z \\ \vdots \\ C_z \end{bmatrix} \ \boldsymbol{z} = 1, 2, \ldots, \text{ó } m \ ; \ \boldsymbol{X_{TEST}} = \begin{bmatrix} \boldsymbol{X_{j1}} & \ldots \boldsymbol{X_{jp}} \end{bmatrix} \ ; \boldsymbol{Y_{TEST}} = ?$$

- **CLASIFICADOR KNN:** Decidir la clase más numerosa de las K instancias más parecidas del conjunto de muestras de entrenamiento a la instancia de test

- **HAY QUE CALCULAR CUÁLES SON LAS K INSTANCIAS MÁS PARECIDAS DEL CONJUNTO DE ENTRANAMIENTO:** cuantificando la similitud (o disimilitud) de cada instancia del conjunto de entrenamiento a la instancia de test.

  → MEDIDAS DE SIMILITUD (SEMEJANZA) / DISIMILITUD (DESEMEJANZA) ENTRE INSTANCIAS DESCRITAS POR UN VECTOR DE ATRIBUTOS:

  1. SIMILITUD (correlación, coseno)

  2. DISIMILITUD **(medidas de distancia)**

CLASIFICADOR KNN – CUESTIONES DE IMPLEMENTACIÓN

KNN: CLASIFICACIÓN BASADA EN MEDIR DISTANCIAS ENTRE INSTANCIAS DESCRITAS POR UN VECTOR DE ATRIBUTOS (DISTANCIAS ENTRE PUNTOS EN EL ESPACIO DE ATRIBUTOS)

- **¿ CÓMO MEDIR DISTANCIAS ENTRE INSTANCIAS DESCRITAS POR ATRIBUTOS DE DISTINTA NATURALEZA (NOMINAL, ORDINAL, NUMÉRICA) ?**
- **¿ QUÉ FUNCIÓN DE DISTANCIA ES LA ADECUADA ?**
- **¿ CÓMO MEDIR DISTANCIAS ENTRE INSTANCIAS DESCRITAS POR ATRIBUTOS QUE PUEDAN TENER RANGOS DE VARIACIÓN Y/O DISPERSIÓN MUY DIFERENTES ?**

A CONTINUACIÓN, VAMOS A:

- **DEFINIR TIPOS DE ATRIBUTOS Y FUNCIONES DE DISTANCIAS SEGÚN SU TIPOLOGÍA:**
  - → **DISTANCIA PARA ATRIBUTOS NUMÉRICOS, ORDINALES Y NOMIMALES**
- **DEFINIR ESTANDARIZACIÓN DE ATRIBUTOS NUMÉRICOS PARA QUE LOS ATRIBUTOS TENGAN EL MISMO PESO INDEPENDIENTEMENTE DE SU RANGO DE VARIACIÓN Y/O DISPERSIÓN**
- **DEFINIR MEDIDA DE SIMILITD BASADA EN EL COSENO DEL ÁNGULO QUE FORMAN VECTORES DE OBSERVACIONES DE DATOS**

TIPOS DE ATRIBUTOS SEGÚN LA ESCALA DE MEDIDA

- **ATRIBUTO NUMÉRICO (indica cantidad – valor real o entero)**

1. DE INTERVALO

→ Los valores tienen orden, valores más altos indican más cantidad de la propiedad que se cuantifica.

→ Escala de medida basada en unidades de igual tamaño.

→ El punto o valor cero no indica la ausencia de la propiedad, no es el origen absoluto de la escala. De esta forma, podemos referenciar los valores por medio de la suma o resta de intervalos, pero no tiene sentido la proporción o razón entre ellos.

→ Ejemplo: temperatura en ºC (por ejemplo, respecto a 4º y 8º, podemos decir que 8º es una temperatura 4 unidades mayor que 4º, pero no es el doble de temperatura – notar que el cero absoluto de temperatura es 0Kelvin = -273ºC).

2. DE RAZÓN

→ Los datos tienen todas las propiedades de los datos de intervalo, pero ahora, la proporción entre ellos tiene sentido.

→ De esta forma, el valor cero de la escala indica la ausencia de la propiedad a medir..

→ Ejemplo: temperatura en K (ahora 8K si es el doble de temperatura que 4K), conteos, salario, tiempo en realizar una tarea, medidas de altura, longitud...

TIPOS DE ATRIBUTOS SEGÚN LA ESCALA DE MEDIDA

- **NOMINAL O CATEGÓRICO (indica categorías, estados o nombres de cosas, sin importar el orden)**
  - → En esta escala carecen de sentido el orden de las etiquetas, así como la comparación y las operaciones aritméticas. La única finalidad de este tipo de datos es clasificar a las observaciones.
  - → Ejemplos: color del pelo, género, estado civil…
  - ❖ **ATRIBUTO BINARIO: caso particular de atributo nominal que únicamente tiene dos estados (0 y 1)**
    - → **Atributo binario simétrico:** los dos estados tienen el mismo peso o importancia (no importa qué estado se codifique con 0 o 1). Ejemplo: género.
    - → **Atributo binario asimétrico:** uno de los estados es más relevante (por convención, se le asigna el valor 1). Ejemplo: test médico (positivo vs negativo).

- **ORDINAL (indica categorías, estados o nombres de cosas en las que el orden tiene significado)**
  - → Muestran las propiedades de los datos nominales, pero además tiene sentido el orden (o jerarquía)
  - → En esta variable sigue sin tener sentido las operaciones aritméticas.
  - → Los valores indican un orden (ranking) pero la magnitud entre ellos no es conocida.
  - → Ejemplos: tamaño (pequeño, mediano, grande), nivel de satisfacción (en escala de 1 a 5)…

MEDIDA DE DISTANCIA Y PROCESAMIENTO DE DATOS SEGÚN LA NATURALEZA DEL ATRIBUTO

- **NOTACIÓN:** sean $\mathbf{X_i}$ y $\mathbf{X_j}$ dos muestras de un conjunto de datos $\mathbf{X}$ → Evaluación de la distancia entre esas muestras (distancia de dos puntos en el espacio *p*-dimensional definido por las variables que componen el vector de atributos de las muestras)

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \; ; \; \boldsymbol{X_i} = \left(\boldsymbol{x_{i1}, x_{i2}, \dots, x_{ip}}\right) \; ; \; \boldsymbol{X_j} = \left(\boldsymbol{x_{j1}, x_{j2}, \dots, x_{jp}}\right) \; ; \; \boldsymbol{d}\left(\boldsymbol{X_i}, \boldsymbol{X_j}\right) = ?$$

ATRIBUTOS NUMÉRICOS: MÉTRICA DE DISTANCIA

DISTANCIA DE MINKOWKI (Norma L_h):

$$\boldsymbol{d}\left(\boldsymbol{X_i}, \boldsymbol{X_j}\right) = \sqrt[h]{\left(\left|\boldsymbol{x_{i1} - x_{j1}}\right|^{\boldsymbol{h}} + \left|\boldsymbol{x_{i2} - x_{j2}}\right|^{\boldsymbol{h}} + \dots + \left|\boldsymbol{x_{ip} - x_{jp}}\right|^{\boldsymbol{h}}\right)} \;, \boldsymbol{q\ entero\ positivo}$$

- **Es una métrica de distancia: D es una métrica de distancia si dados 3 puntos p_1, p_2 y p_3, se verifica:**
  1. $\mathrm{D}(\mathrm{p_1}, \mathrm{p_2}) \geq 0 \quad \left[\mathrm{D}(\mathrm{p_1}, \mathrm{p_2}) = 0 \text{ si } \mathrm{p_1} = \mathrm{p_2}\right]$ - Definida positiva
  2. $\mathrm{D}(\mathrm{p_1}, \mathrm{p_2}) = \mathrm{D}(\mathrm{p_2}, \mathrm{p_1})$ - Simetría
  3. $\mathrm{D}(\mathrm{p_1}, \mathrm{p_3}) \leq \mathrm{D}(\mathrm{p_1}, \mathrm{p_2}) + \mathrm{D}(\mathrm{p_2}, \mathrm{p_3})$ - Desigualdad triangular

MEDIDA DE DISTANCIA Y PROCESAMIENTO DE DATOS SEGÚN LA NATURALEZA DEL ATRIBUTO

- **ATRIBUTOS NUMÉRICOS: MÉTRICAS DE DISTANCIA**

DISTANCIA DE MINKOWKI:

$$d(X_i, X_j) = \sqrt[h]{\left(|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \ldots + |x_{ip} - x_{jp}|^h\right)}$$

→ **DISTANCIA DE MANHATAN (*CITY BLOCK*, Norma L_1): h = 1**

$$d(X_i, X_j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ip} - x_{jp}|$$

- **DISTANCIA HAMMING** - caso particular cuando los atributos son binarios (número de bits diferentes en dos vectores binarios)

→ **DISTANCIA EUCLIDEA (Norma L_2): h = 2**

$$d(X_i, X_j) = \sqrt{\left(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2\right)}$$

→ **DISTANCIA DE CHEBYSHEV (Norma *suprema* L_∞ , expresa la diferencia máxima en cualquiera de las componentes o atributos del vector): h → ∞**

$$d(X_i, X_j) = \lim_{h\to\infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h\right)^{1/h} = \max_f |x_{if} - x_{jf}|^h$$

MEDIDA DE DISTANCIA Y PROCESAMIENTO DE DATOS SEGÚN LA NATURALEZA DEL ATRIBUTO

❑ ATRIBUTOS NUMÉRICOS: MÉTRICAS DE DISTANCIA - EJEMPLO

→ **DISTANCIA DE MANHATAN:** $d(X_i, X_j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ip} - x_{jp}|$

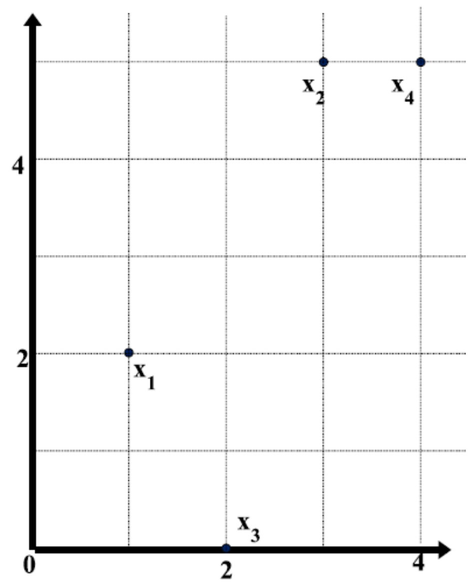
→ **DISTANCIA EUCLIDEA:** $d(X_i, X_j) = \sqrt{\left(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2\right)}$

→ **DISTANCIA DE CHEBYSHEV (*SUPREMUM*):**

$$d(X_i, X_j) = \max_f |x_{if} - x_{jf}|^h$$

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Ejemplo: 4 puntos descritos por dos atributos



Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

MEDIDA DE DISTANCIA Y PROCESAMIENTO DE DATOS SEGÚN LA NATURALEZA DEL ATRIBUTO

❑ **ATRIBUTOS NUMÉRICOS DE INTERVALO: TRATAMIENTO PREVIO (dependiendo de la aplicación)**

→ **NECESIDAD DE ESTANDARIZACIÓN.** **JUSTIFICACIÓN:** vectores de atributos compuestos por variables de distinta naturaleza, rangos de variación y dispersión diferentes

→ **La estandarización permite que todos los atributos tengan el mismo peso en el cálculo de distancias:** los valores de cada atributo se transforman en variables sin unidad variando entorno a su valor medio representativo en unidades relativas a la desviación que presentan (el valor medio y la desviación de cada atributo se calculan en el conjunto de datos disponibles del atributo en cuestión).

$$m_f = \frac{1}{n}\sum_{i=1}^{n} x_{if}$$

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \; ; \quad \begin{matrix} \textit{Mean}: \; M = (m_1, m_2, \ldots, m_p) \\ \textit{Standard Deviation}: \; STD = (\sigma_1, \sigma_2, \ldots, \sigma_p) \\ \textit{Mean Absolute Deviation}: MAD\; (s_1, s_2, \ldots, s_p) \end{matrix} \; ; \quad \sigma_f = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{if} - m_f\right)^2} \; ; \; f = 1, \ldots, p$$

❖ **PROCEDIMIENTO ESTANDARIZACIÓN: obtención de medida estandarizada de cada atributo o Z-Score:**

$$s_f = \frac{1}{n}\sum_{i=1}^{n}\left|x_{if} - m_f\right|$$

→ **Opción 1: expresar los valores en unidades de la desviación estándar** $x_{if} \rightarrow z_{if} = \dfrac{x_{if} - m_f}{\sigma_f}$

→ **Opción 2: expresar los valores en unidades de la desviación absoluta media** $x_{if} \rightarrow z_{if} = \dfrac{x_{if} - m_f}{s_f}$

❖ Esta opción tiende a ser más robusta (reduce el efecto de posibles *outliers*, que permanecen detectables - desviación no cuadrada)-

❖ **OTRA OPCIÓN: sólo normalizar y emplear distancia de Mahalanobis:** $x_{if} \rightarrow x^*_{if} = \dfrac{x_{if} - min_f}{max_f - min_f} \in [0, 1]$

MEDIDA DE DISTANCIA Y PROCESAMIENTO DE DATOS SEGÚN LA NATURALEZA DEL ATRIBUTO

❑ ATRIBUTOS ORDINALES: TRATAMIENTO PREVIO PARA TRATARLOS COMO NUMÉRICOS

→ **PROCEDIMIENTO:**

$$N^{\underline{o}}\ Valores\ Atributos\ (M_1, M_2, \dots, M_p)$$

Notación: supongamos que un atributo cualquiera x_f tiene M_f posibles valores.

1. Establecer una escala de valores numérica para cada atributo, un ranking: $Ranking\_x_f = \{1, 2, \dots, M_f\}$
2. Reemplazar los valores de las muestras de cada atributo por su ranking: $x_{if} \rightarrow r_{if} \in \{1, 2, \dots, M_f\}$
3. Mapear el rango de cada atributo a [0, 1]: $r_{if} \rightarrow z_{if} = \dfrac{r_{if\_}1}{M_f - 1} \in [0, 1]$
4. Computar la medidas de distancias utilizando estos valores (como atributo numérico de intervalo).

❑ ATRIBUTOS NUMÉRICOS DE RAZÓN: TRATAMIENTO PREVIO – OPCIONES:

Justificación: atributos cuyas medidas presentan una escala de variación no lineal (tratarlos directamente como si fueran atributos numéricos de intervalo no es buena idea, su escala puede estar distorsionada)**:**

→ Variables de rápido crecimiento/decrecimiento → distribución de valores asimétrica

❖ **APLICAR TRANSFORMACIÓN LOGARÍTMICA**

→ Escala aproximadamente exponencial (ejemplo: crecimiento de una bacteria)

❖ **CONSIDERARLOS COMO ATRIBUTOS ORDINALES CONTINUOS Y TRATAR SUS RANKINGS COMO VARIABLE NUMÉRICA DE INTERVALO.**

MEDIDA DE DISTANCIA Y PROCESAMIENTO DE DATOS SEGÚN LA NATURALEZA DEL ATRIBUTO

❑ ATRIBUTOS NOMINALES: MEDIDA DE DISTANCIA

Dado un espacio definido por $X = (x_1, x_2, \dots, x_p)$ *con* x_f *nominal* $\forall f$ *de 1 a p. Determinar la distancia entre dos puntos de este espacio:* $X_i = (x_{i1}, x_{i2}, \dots, x_{ip})$; $X_j = (x_{j1}, x_{j2}, \dots, x_{jp}) \rightarrow d(X_i, X_j) = ?$

1. ATRIBUTOS BINARIOS: $X = (x_1, x_2, \dots, x_p)$ *con* $x_f \in \{0,1\}$, *atributos binarios* $\forall f$ *de 1 a p*

→ Distancia para atributos simétricos:

$$d(X_i, X_J) = \frac{\textit{Número de discrepancias}}{\textit{Número de atributos}} = \frac{M_{10} + M_{01}}{M_{11} + M_{10} + M_{01} + M_{00}}$$

Tabla de contingencia para datos binarios

		X_j 1	X_j 0	Total
X_i	1	M_{11}	M_{10}	$M_{11}+M_{10}$
	0	M_{01}	M_{00}	$M_{01}+M_{00}$
	Total	$M_{11}+M_{01}$	$M_{10}+M_{00}$	p

→ Distancia para atributos asimétricos (1 – Coeficiente de *Jaccard*):

$$d(X_i, X_J) = \frac{\textit{Número de discrepancias}}{\textit{Número de atributos sin coincidencias en } 00} = \frac{M_{10} + M_{01}}{M_{11} + M_{10} + M_{01}}$$

Coeficiente de *Jaccard (medida de similitud)*:

$$SimJaccard(X_i, X_J) = \frac{\textit{Número de coincidencias } 11}{\textit{Número de atributos sin coincidencias en } 00}$$

2. ATRIBUTOS NOMINALES DE MÁS DE DOS ESTADOS:

→ Opción 1: $$d(X_i, X_J) = \frac{\textit{Nº de discrepancias}}{\textit{Nº de atributos}} = \frac{\textit{Nº de atributos} - \textit{Nº de coincidencias}}{\textit{Número de atributos}} = \frac{p - m}{p}$$

→ Opción 2: transformar los datos a forma binaria, creando un atributo binario por cada uno de los estados categóricos posibles de cada atributo

MEDIDA DE DISTANCIA ENTRE MUESTRAS DE DATOS DEFINIDOS POR ATRIBUTOS MIXTOS

- Un conjunto de datos puede estar definido por atributos de los siguientes tipos: numérico (de intervalo y de razón), ordinal, nominal binario (simétrico o asimétrico) y nominal de más de dos estados.

- Para medir la distancia entre dos puntos en ese espacio de atributos mixtos: promediar de forma ponderada las distancias de cada atributo individual

MEDIDA DE DISTANCIA PARA VECTORES DE ATRIBUTOS MIXTOS: UNA OPCIÓN

$$X_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \;;\; X_j = (x_{j1}, x_{j2}, \dots, x_{jp}) \;\rightarrow\; d(X_i, X_j) = \frac{\sum_{f=1}^{p} \delta_f * d(x_{if}, x_{jf})}{\sum_{f=1}^{p} \delta_f}$$

Donde:

- Si x_f es nominal: $d(x_{if}, x_{jf}) = 0$ si $x_{if} = x_{jf}$ y $d(x_{if}, x_{jf}) = 1$ si $x_{if} \neq x_{jf}$

- Si x_f es numérico: evaluar distancia estandarizada $sd(x_{if}, x_{jf})$ (es habitual considerar distancia Euclidea dividida por la desviación estándar del atributo)

- Si x_f es ordinal (o numérico de razón):

$$\boldsymbol{x_{if}, x_{jf}} \rightarrow \boldsymbol{r_{if}, r_{jf}} \in \{\boldsymbol{1, 2, \dots, M_f}\} \rightarrow \boldsymbol{z_{if}} = \frac{\boldsymbol{r_{if-1}}}{\boldsymbol{M_f - 1}}, \boldsymbol{z_{jf}} = \frac{\boldsymbol{r_{jf-1}}}{\boldsymbol{M_f - 1}} \Rightarrow \text{evaluar distancia estandarizada } d(z_{if}, z_{jf})$$

- δ_f: pesos de ponderación a la distancia de cada atributo

MEDIDA DE SIMILITUD POR COSENO

- Hasta ahora, se han medido distancias entre dos observaciones de datos que se han tratado como puntos en el espacio de atributos (MEDIDA DE DISIMILITUD: distancia entre puntos)

- En este apartado se introduce una medida de similitud basada en tratar las observaciones como vectores en ese espacio de atributos: MEDIDA DE SIMILITUD POR COSENO (medida basada en el ángulo que forman esos vectores)

- **MEDIDA DE SIMILITUD POR COSENO DEL ÁNGULO QUE FORMAN LOS VECTORES ASOCIADOS A DOS OBSERVACIONES DE DATOS**

$$V_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \; ; \; V_j = (x_{j1}, x_{j2}, \dots, x_{jp}) \;\rightarrow cos(V_i, V_j) = (V_i \bullet V_j)/(\|V_i\| * \|V_j\|)$$

donde:

$$\|V_i\| = \sqrt{(V_i \bullet V_i)} \; ; \; \|V_j\| = \sqrt{(V_j \bullet V_j)}$$

- $\bullet$: denota el producto escalar
- $\|\;\;\|$: denota el módulo o longitud del vector

→ Ejemplo de aplicación: vectores de atributos definidos en términos de frecuencia de aparición o conteo de determinada información o magnitud

MEDIDA DE SIMILITUD POR COSENO: EJEMPLO

- Un documento puede ser descrito por miles de atributos, cada uno registrando la frecuencia de aparición en el documento de una determinada palabra (palabras clave) o frase

	team	coach	baseball	soccer	penalty	score	win	loss
Doc1	5	0	0	2	0	0	2	0
Doc2	3	0	2	1	0	0	3	0
Doc3	0	7	0	1	0	0	3	0
Doc4	0	1	0	1	2	2	0	3

→ Encontrar la similitud entre los documentos 1 y 2 $cos(V_i, V_j) = (V_i \bullet V_j)/(\|V_i\| * \|V_j\|)$

$d_1 =$ (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)
$d_2 =$ (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

$d_1 \bullet d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25$
$||d_1|| = (5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$
$||d_2|| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5} = (17)^{0.5} = 4.12$
$\cos(d_1, d_2) = 0.94$