ALGORÍTMICA

El problema de Clasificación

1. Introducción al problema de Clasificación.
2. Ejemplo de Clasificador: k-NN.
3. Ejemplo de Clasificador: Naïve-Bayes.
4. Introducción al problema de reducción de datos.
5. Bases de Datos a utilizar

1. Introducción al problema de Clasificación

Disponemos de una muestra de objetos ya clasificados $w_i,\ldots,w_n$, representados en función de sus valores en una serie de atributos:

w_i tiene asociado el vector $(x_1(w_i),\ldots,x_c(w_i))$

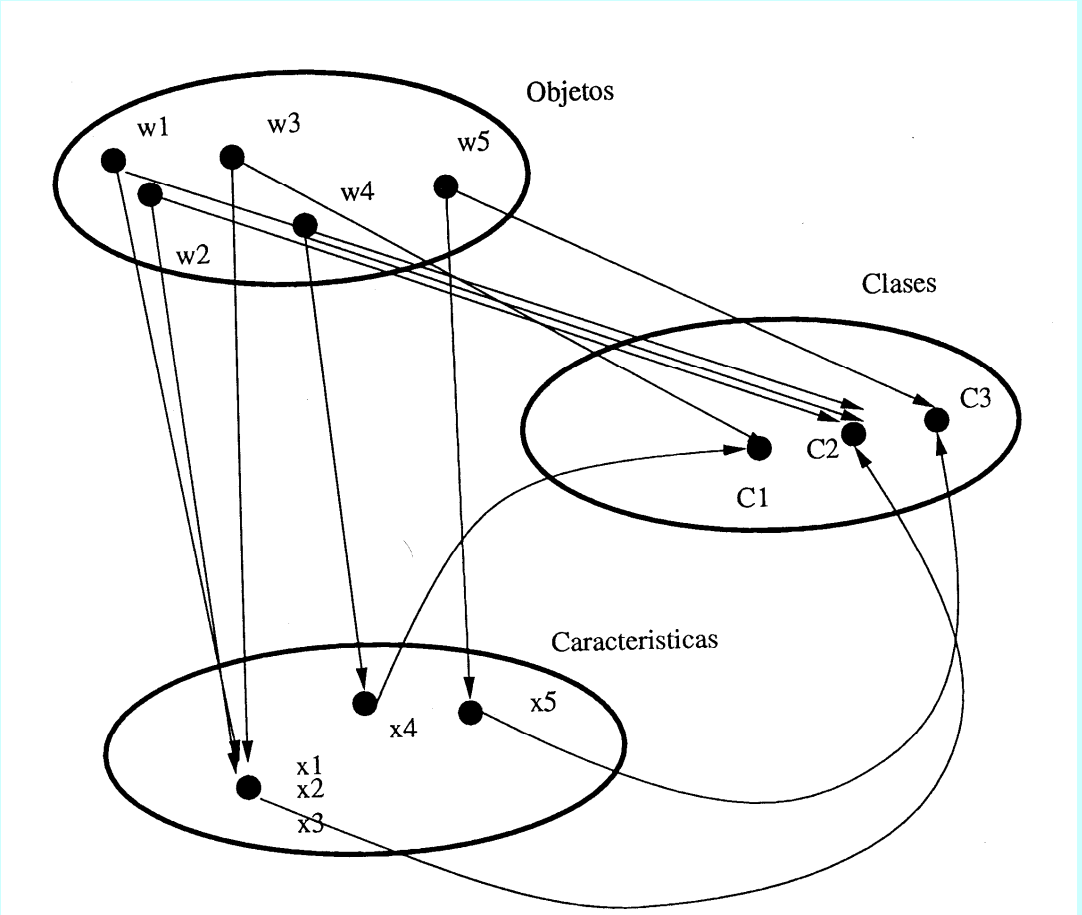
Cada objeto pertenece a una de las clases existentes $\{C_1,\ldots,C_M\}$.

OBJETIVO: Obtener un sistema que permita clasificar dichos objetos de modo automático.

$$\begin{array}{ccc} w_1 = (x_1(w_1),\mathrm{K}\,,x_n(w_1)) & \rightarrow & C_{i_1} \\ \mathrm{M} & & \mathrm{M} \\ w_k = (x_1(w_k),\mathrm{K}\,,x_n(w_k)) & \rightarrow & C_{i_k} \end{array} \qquad i_j \in \{1,\mathrm{K}\,,M\},\ \ j \in \{1,\mathrm{K}\,,k\}$$

1. Introducción al problema de Clasificación

El problema fundamental de la clasificación está directamente relacionado con la separablidad de las clases.



1. Introducción al problema de Clasificación

Concepto de aprendizaje supervisado en clasificación

Se conocen las clases existentes en el problema.

Se conoce la clase concreta a la que pertenece cada objeto del conjunto de datos.

Existen una gran cantidad de técnicas para el aprendizaje supervisado de Sistemas de Clasificación:

Técnicas estadísticas: k vecinos más cercanos, discriminadores bayesianos, etc…

Árboles de clasificación, Sistemas basados en reglas, Redes Neuronales, Máquinas de Soporte Vectorial, …

1. Introducción al problema de Clasificación

Ejemplo: Diseño de un Clasificador para la flor del *Iris*

- Problema simple muy conocido: *clasificación de lirios.*
- Tres clases de lirios: *setosa*, *versicolor* y *virginica*.
- Cuatro atributos: *longitud* y *anchura* de *pétalo* y *sépalo*, respectivamente.
- 150 ejemplos, 50 de cada clase.
- Disponible en http://www.ics.uci.edu/~mlearn/MLRepository.html

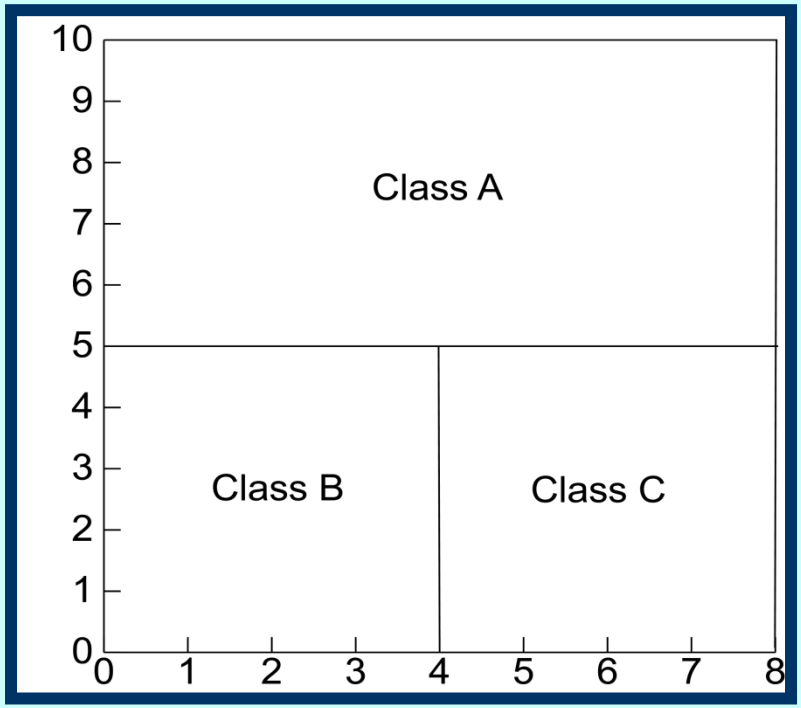
setosa

versicolor

virginica

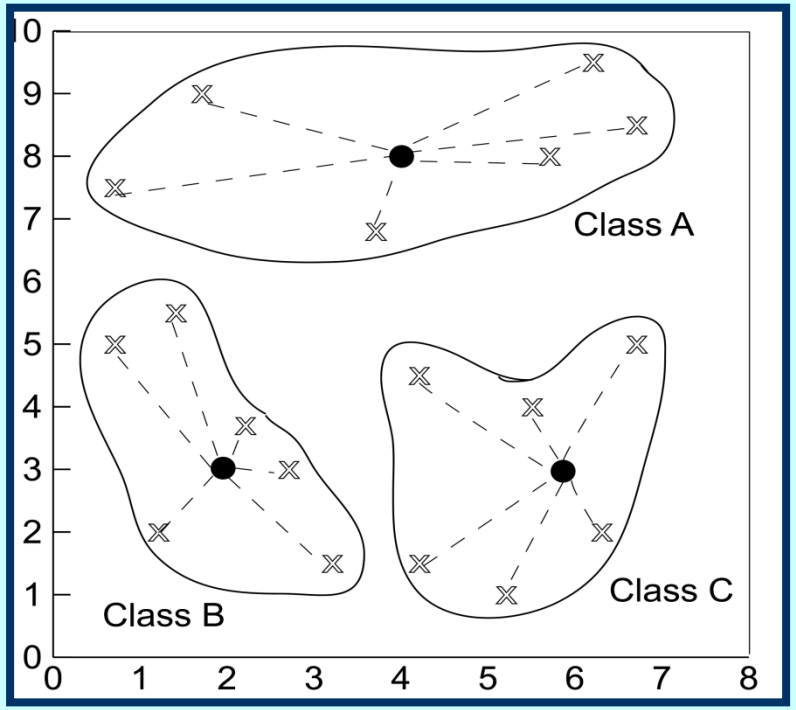
1. Introducción al problema de Clasificación

Ejemplos de clasificación sobre clases definidas: Basada en reglas y en distancias



Basado en Particiones

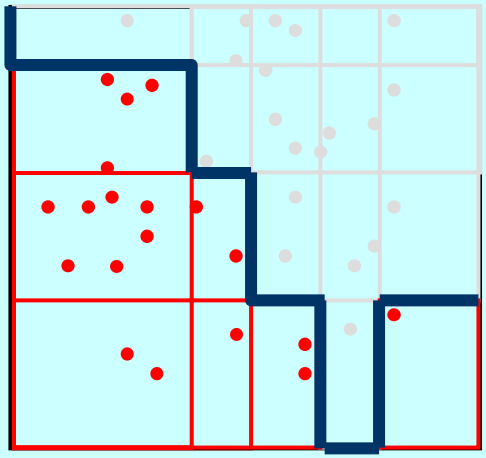
Basado en Distancias



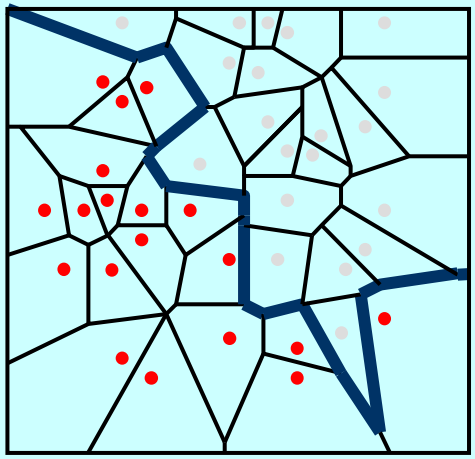
1. Introducción al problema de Clasificación

Ejemplos de clasificación sobre clases definidas

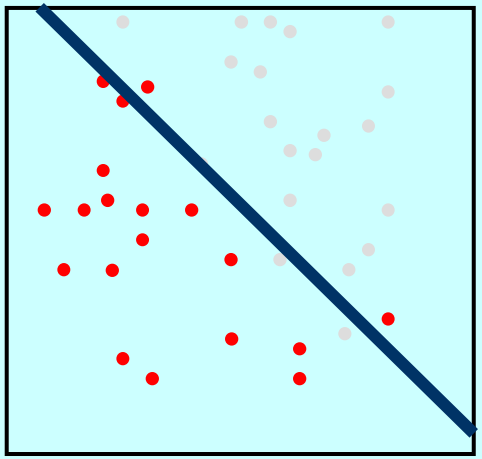
- **Reglas intervalares**



- **Basado en distancias**



- **Clasificador linear**



1. Introducción al problema de Clasificación

Para diseñar un clasificador, son necesarias dos tareas: *Aprendizaje* y *Validación*.

El conjunto de ejemplos se divide en dos subconjuntos:

- **Entrenamiento**: Utilizado para aprender el clasificador.
- **Test**: Se usa para validarlo. Se calcula el porcentaje de clasificación sobre los ejemplos de este conjunto (desconocidos en la tarea de aprendizaje) para conocer su poder de generalización.

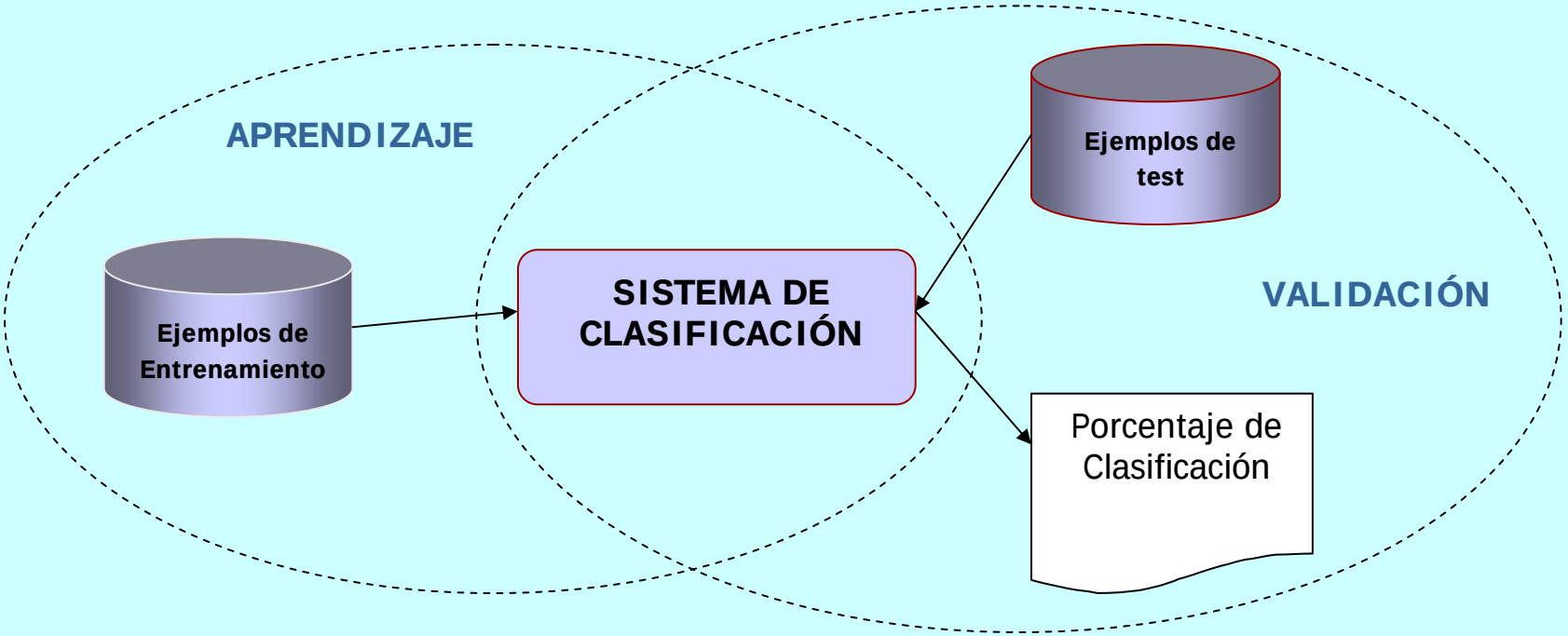
Para mayor seguridad, se suele hacer varias particiones entrenamiento-test.

Para cada partición, se diseña un clasificador distinto usando los ejemplos de entrenamiento y se valida con los de test.

Nosotros usaremos solo una partición (80% y 20%).

1. Introducción al problema de Clasificación

Esquema de aprendizaje en clasificación



2. Ejemplo de Clasificador: k-NN

k-NN (*k vecinos más cercanos*) es uno de los clasificadores más utilizados por su simplicidad.

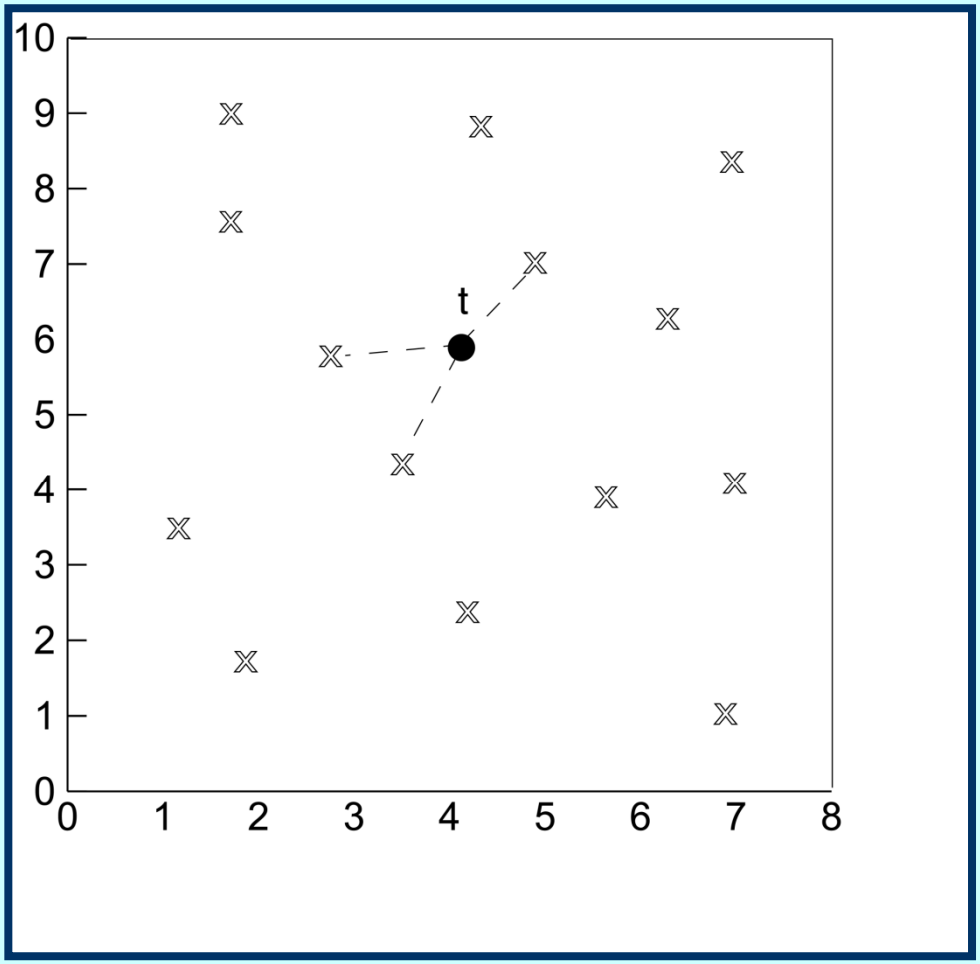
i. El proceso de aprendizaje de este clasificador consiste en almacenar una tabla con los ejemplos disponibles, junto a la clase asociada a cada uno de ellos.

ii. Ante un nuevo ejemplo a clasificar, se calcula su distancia (usaremos la Euclídea) con respecto a los *n* ejemplos existentes en la tabla, y se consideran los *k* más cercanos.

iii. El nuevo ejemplo se clasifica según la clase mayoritaria de los *k* ejemplos más cercanos.

iv. El caso más sencillo es cuando *k = 1 (1-NN)*.

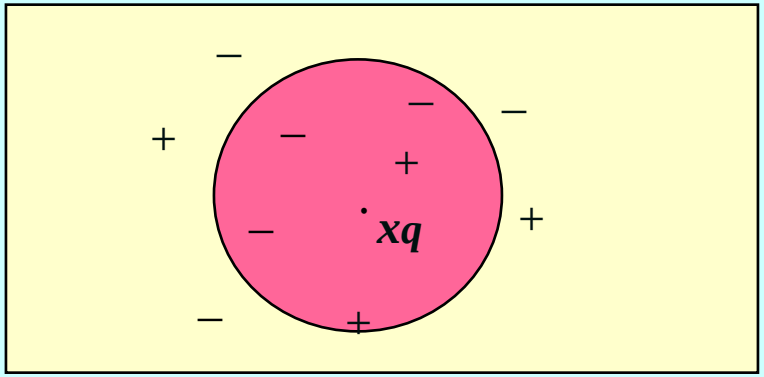
2. Ejemplo de Clasificador: k-NN

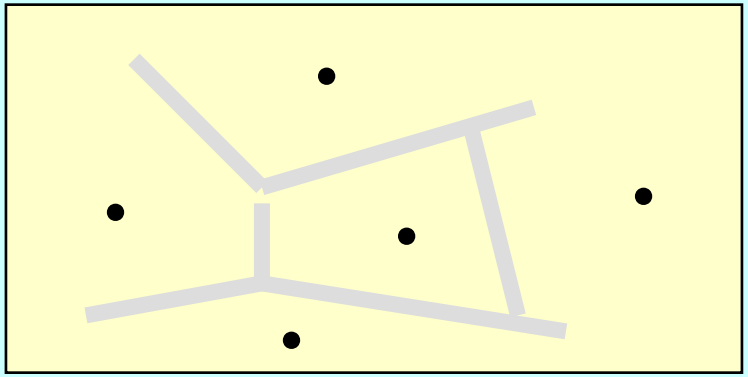


k = 3

2. Ejemplo de Clasificador: k-NN

- *k*-NN devuelve la clase más repetida de entre todos los k ejemplos de entrenamiento cercanos a *xq*.
- Diagrama de Voronoi: superficie de decisión inducida por 1-NN para un conjunto dado de ejemplos de entrenamiento.



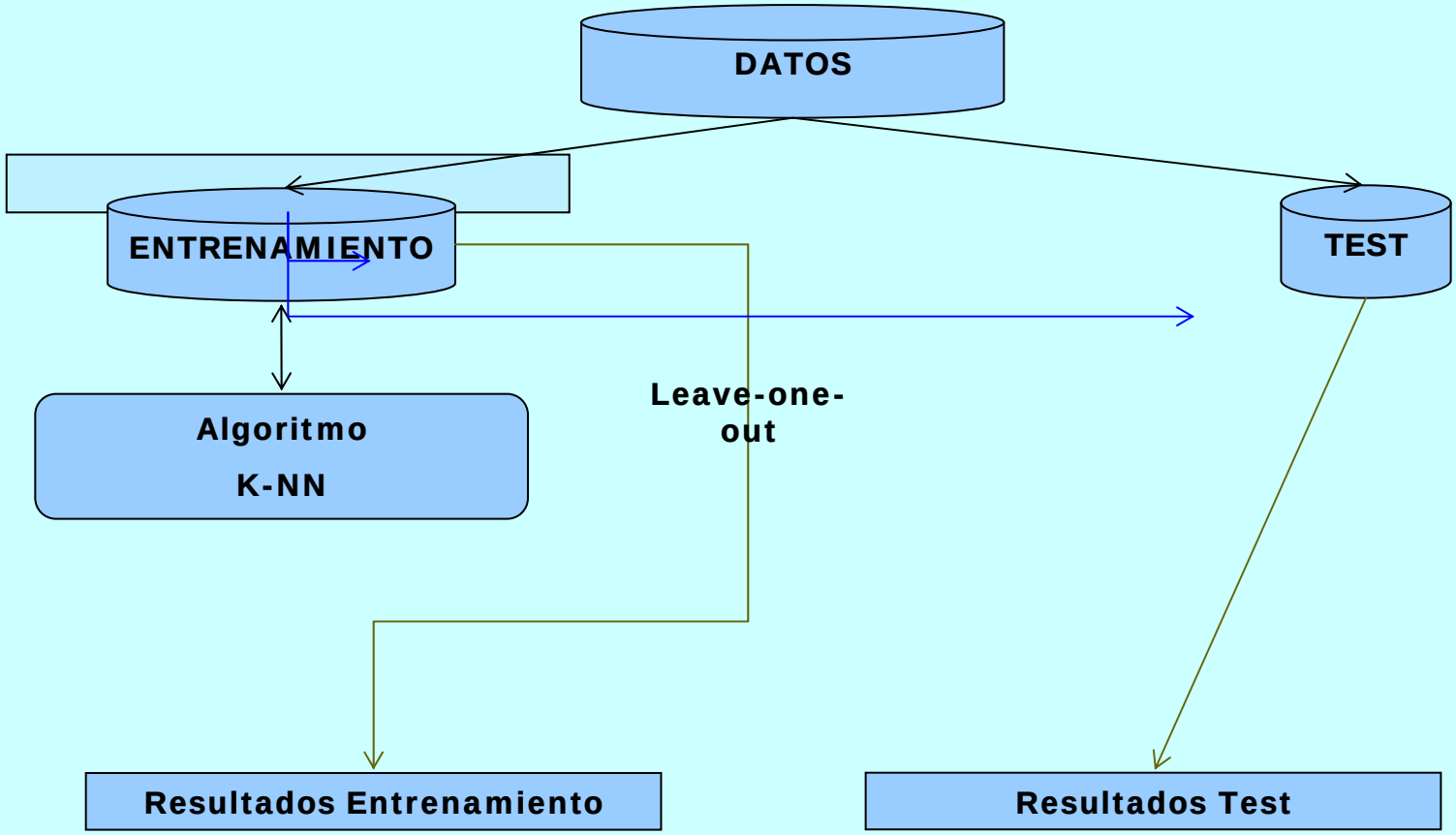


2. Ejemplo de Clasificador: k-NN

Cálculo del Porcentaje de Entrenamiento en 1-NN:

- En el algoritmo 1-NN, no es posible calcular el porcentaje de acierto sobre el conjunto de entrenamiento de un modo directo.
- Si intentásemos clasificar un ejemplo del conjunto de entrenamiento directamente con el clasificador 1-NN, el ejemplo más cercano sería siempre él mismo, con lo que se obtendría un 100% de acierto.
- Para remediar esto, se debe seguir el procedimiento denominado *dejar uno fuera ("leave one out").*
- Para clasificar cada ejemplo del conjunto de entrenamiento, se busca el ejemplo más cercano **sin considerar a él mismo**.
- Por lo demás, se opera igual: cada vez que la clase devuelta coincida con la clase real del ejemplo, se contabiliza un acierto.
- El porcentaje final de acierto es el número de aciertos entre el número total de ejemplos.

2. Ejemplo de Clasificador: k-NN



2. Ejemplo de Clasificador: k-NN

Dado el siguiente conjunto con 4 instancias, 3 atributos y 2 clases:

x_1: 0.4 0.8 0.2 positiva
x_2: 0.2 0.7 0.9 positiva
x_3: 0.9 0.8 0.9 negativa
x_4: 0.8 0.1 0.0 negativa

Queremos clasificar con 1-NN el siguiente ejemplo:

y_1: 0.7 0.2 0.1

Calculamos la distancia del ejemplo con todos los del conjunto:

$$d(x_1, y_1) = \sqrt{(0.4-0.7)^2 + (0.8-0.2)^2 + (0.2-0.1)^2} = 0.678$$
$$d(x_2, y_1) = \sqrt{(0.2-0.7)^2 + (0.7-0.2)^2 + (0.9-0.1)^2} = 1.068$$
$$d(x_3, y_1) = \sqrt{(0.9-0.7)^2 + (0.8-0.2)^2 + (0.9-0.1)^2} = 1.020$$
$$d(x_4, y_1) = \sqrt{(0.8-0.7)^2 + (0.1-0.2)^2 + (0.0-0.1)^2} = 0.173$$

Por tanto, el ejemplo se clasificará con respecto a la clase *negativa*.

IMPORTANTE: Los atributos deben estar normalizados [0,1] para no priorizarlos sobre otros.

3. Ejemplo de Clasificador: Naïve-Bayes

Naïve-Bayes es uno de los clasificadores más utilizados por su simplicidad y rapidez.

- El proceso de aprendizaje de este clasificador consiste en almacenar una tabla de contingencias con la información valor del atributo-clase.
- Se asume que los **atributos son independientes dada la clase**.
- Ante un nuevo ejemplo a clasificar, se calcula la probabilidad a priori de pertenecer a cada clase $P(C_k|w)$. El nuevo ejemplo se clasifica según la clase con **mayor probabilidad**.
- Al construir una tabla de contingencias, necesitamos que los atributos sean *nominales* o *discretos*.

3. Ejemplo de Clasificador: Naïve-Bayes

Cálculo de probabilidades en Naïve-Bayes:

$$P(C_i \mid w_k) = P(C_i)\prod_{j=1}^{c} P(x_j(w_k) = v_{jk} \mid C_i)$$

Dado el conjunto que utilizamos para calcular las contingencias, necesitamos 2 probabilidades diferentes.

Probabilidad de clase $P(C_i)$.

Es el cociente del número de ejemplos de clase C_i entre el número total de ejemplos.

Probabilidad de los atributos $P(x_j(w_k) = v_{jk} \mid C_i)$

Es una matriz 3D donde para cada clase C_i y atributo de entrada x_j, se cuenta el número de veces que aparecen cada uno de los valores v_{jk}.

Este recuento se normaliza usando Laplace: a cada valor v_{jk} se le incrementa 1 en el recuento (toda la matriz), y se suman todos los recuentos incrementados para una clase C_i y atributo de entrada x_j dados. Luego se divide cada recuento por esta última suma.

3. Ejemplo de Clasificador: Naïve-Bayes

Dado el siguiente conjunto con 4 instancias, 3 atributos y 2 clases:

Queremos clasificar con Naïve-Bayes el siguiente ejemplo:

y_1: 1 1 1

x_1: 1 2 1 positiva
x_2: 2 2 2 positiva
x_3: 1 1 2 negativa
x_4: 2 1 2 negativa

$$P(C_{pos}) = P(C_{neg}) = 0.5$$

Recuento inicial

Positiva	valor 1	valor 2
atr 1	1	1
atr 2	0	2
atr 3	1	1

Negativa	valor 1	valor 2
atr 1	1	1
atr 2	2	0
atr 3	0	2

3. Ejemplo de Clasificador: Naïve-Bayes

Dado el siguiente conjunto con 4 instancias, 3 atributos y 2 clases:

Queremos clasificar con Naïve-Bayes el siguiente ejemplo:

y_1: 1 1 1

x_1: 1 2 1 positiva
x_2: 2 2 2 positiva
x_3: 1 1 2 negativa
x_4: 2 1 2 negativa

$$P(C_{pos}) = P(C_{neg}) = 0.5$$

Corrección Laplace

Positiva	valor 1	valor 2
atr 1	2	2
atr 2	1	3
atr 3	2	2

Negativa	valor 1	valor 2
atr 1	2	2
atr 2	3	1
atr 3	1	3

3. Ejemplo de Clasificador: Naïve-Bayes

Dado el siguiente conjunto con 4 instancias, 3 atributos y 2 clases:

Queremos clasificar con Naïve-Bayes el siguiente ejemplo:

y_1: 1 1 1

x_1: 1 2 1 positiva
x_2: 2 2 2 positiva
x_3: 1 1 2 negativa
x_4: 2 1 2 negativa

$$P(C_{pos}) = P(C_{neg}) = 0.5$$

Normalización de potenciales

Positiva	valor 1	valor 2
atr 1	0.5	0.5
atr 2	0.25	0.75
atr 3	0.5	0.5

Negativa	valor 1	valor 2
atr 1	0.5	0.5
atr 2	0.75	0.25
atr 3	0.25	0.75

3. Ejemplo de Clasificador: Naïve-Bayes

Dado el siguiente conjunto con 4 instancias, 3 atributos y 2 clases:

x_1: 1 2 1 positiva
x_2: 2 2 2 positiva
x_3: 1 1 2 negativa
x_4: 2 1 2 negativa

Queremos clasificar con Naïve-Bayes el siguiente ejemplo:

y_1: 1 1 1

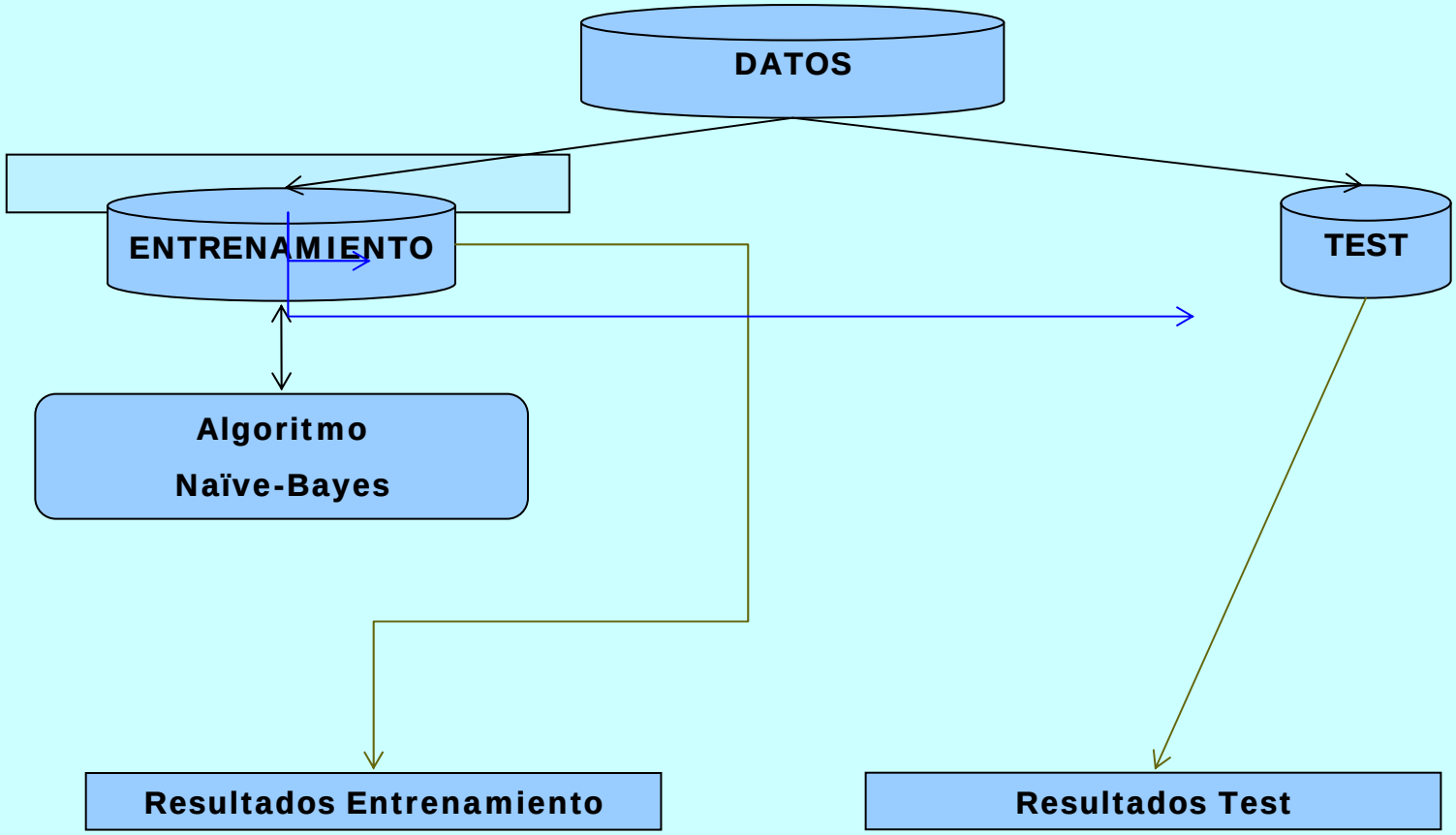
$$P(C_{pos}) = P(C_{neg}) = 0.5 \longrightarrow \begin{cases} P(C_{pos} \mid y_1) = 0.5 \cdot (0.5 \cdot 0.25 \cdot 0.5) = 0.03125 \\ P(C_{neg} \mid y_1) = 0.5 \cdot (0.5 \cdot 0.75 \cdot 0.25) = 0.046875 \end{cases}$$

Probabilidades usadas

Positiva	valor 1	valor 2
atr 1	0.5	0.5
atr 2	0.25	0.75
atr 3	0.5	0.5

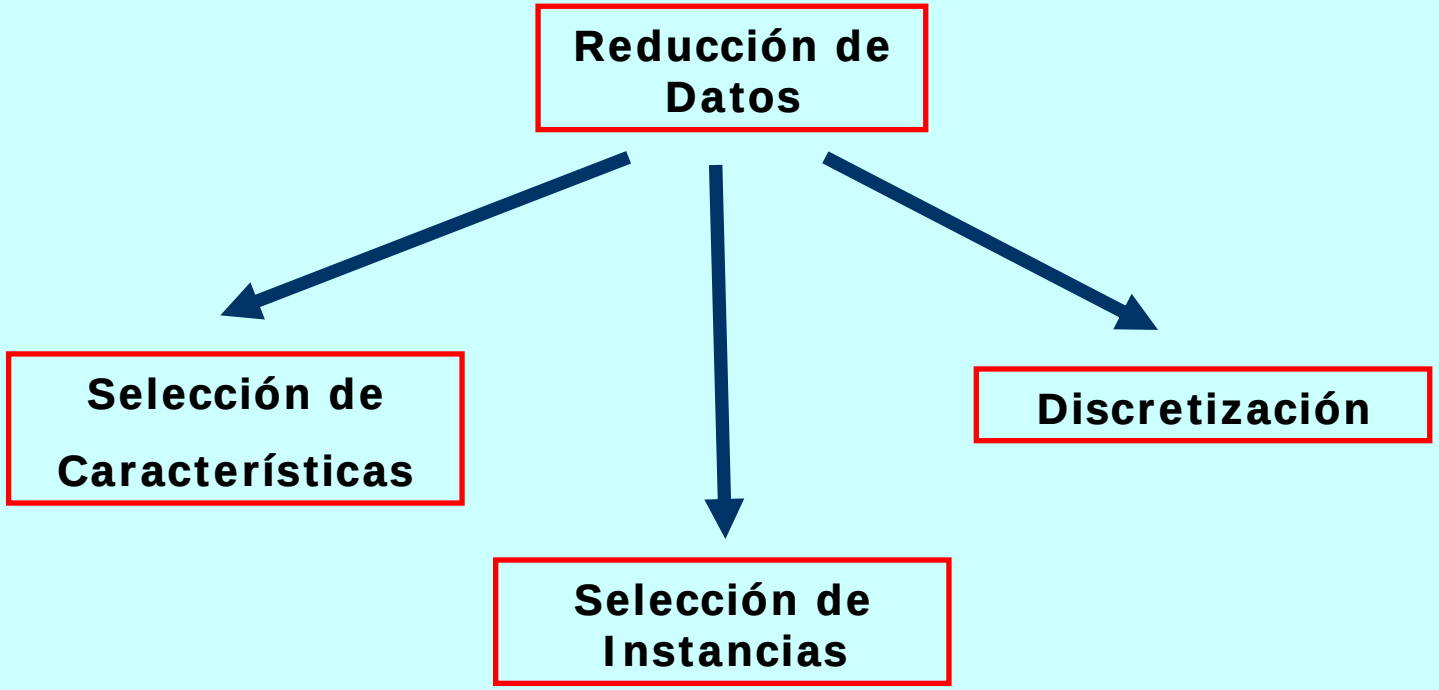
Negativa	valor 1	valor 2
atr 1	0.5	0.5
atr 2	0.75	0.25
atr 3	0.25	0.75

3. Ejemplo de Clasificador: Naïve-Bayes



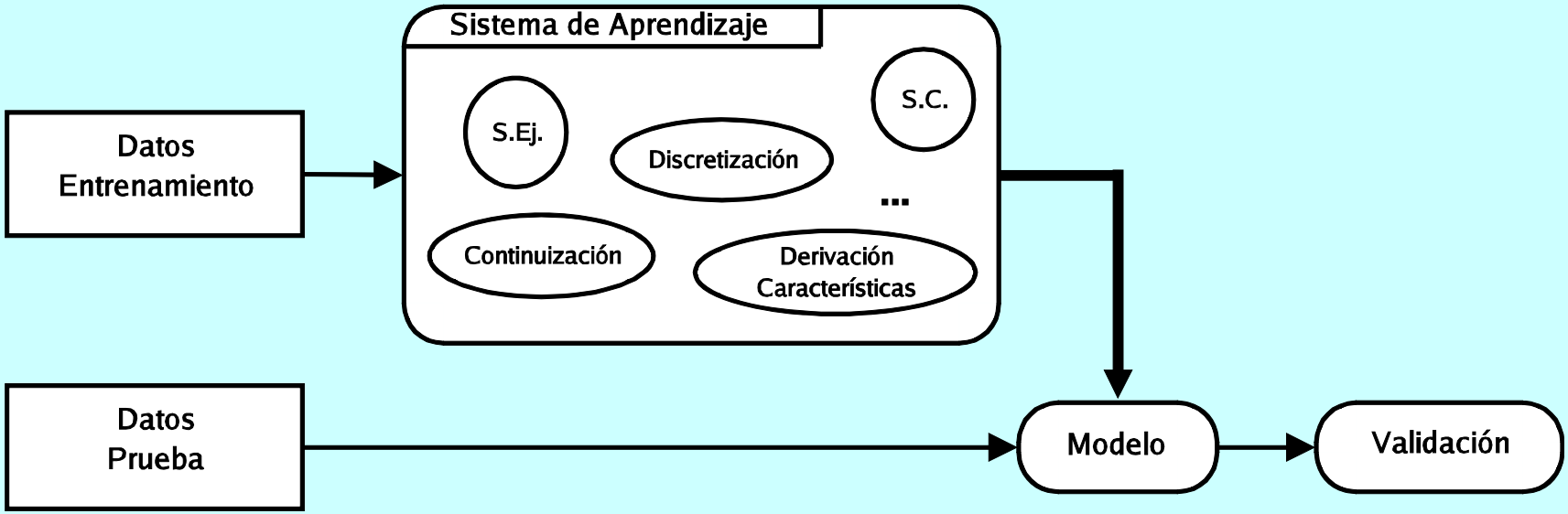
4. Introducción al problema de Reducción de Datos

A veces el problema de clasificación es demasiado complejo y necesita ser procesado previamente para ser tratable → **Reducción de datos**



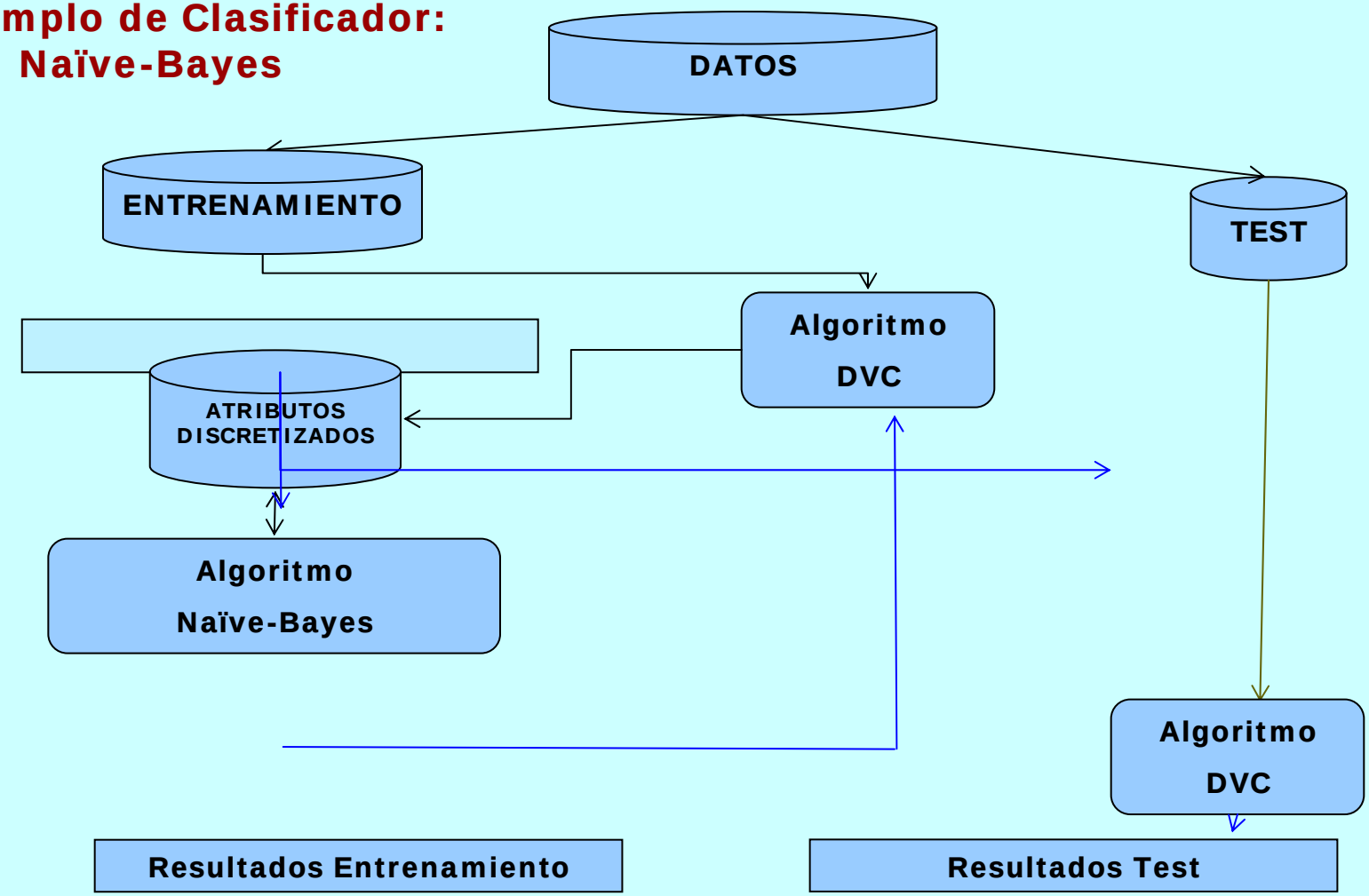
4. Introducción al problema de Reducción de Datos

Proceso de aprendizaje



4. Introducción al problema de Reducción de Datos

Ejemplo de Clasificador: Naïve-Bayes



5. Bases de Datos a utilizar

■ En la actualidad, hay muchas bases de datos que se utilizan como bancos de prueba (*benchmarks*) para comprobar el rendimiento de los algoritmos de clasificación

■ Un repositorio de bases de datos para aprendizaje automático muy conocido es el UCI

■ Es accesible en el Web en la siguiente dirección:

http://www.ics.uci.edu/~mlearn/MLRepository.html

5. Bases de Datos a utilizar

- A partir de este repositorio, se ha desarrollado un formato más completo para definir todas las cualidades de una base de datos en un único fichero
- Se trata del formato ARFF, utilizado en WEKA (http://www.cs.waikato.ac.nz/ml/weka/)
- Un fichero ARFF contiene dos partes:
  - **Datos de cabecera:** Líneas que comienzan por @
  - **Datos del problema:** Líneas de datos
- Los datos de cabecera contienen información acerca de los atributos: significado y tipo
- Los datos del problema son los ejemplos en sí. Cada línea corresponde con un ejemplo y los valores están separados por comas

5. Bases de Datos a utilizar (ejemplo fichero ARFF)

```
@relation iris
@attribute sepalLength real
@attribute sepalWidth real
@attribute petalLength real
@attribute petalWidth real
@attribute class {Iris-setosa, Iris-versicolor, Iris-virginica}
@data
5.1, 3.5, 1.4, 0.2, Iris-setosa
4.9, 3.0, 1.4, 0.2, Iris-setosa
7.0, 3.2, 4.7, 1.4, Iris-versicolor
6.0, 3.0, 4.8, 1.8, Iris-virginica
```

- 5 Atributos, los 4 primeros de tipo *real* y el último de tipo *nominal*
- La clase es el último atributo (atributo de salida) con los posibles valores definidos
- Los datos van a continuación de la directiva *@data*

5. Bases de Datos a utilizar

■ Junto a *Iris*, utilizaremos cinco bases de datos más: *Pima*, *Wine*, *Sonar*, *SpamBase y Ozone*

■ *Pima* es una base de datos con información sobre diabetes en mujeres indias de la región Pima. Se intenta comprobar si el paciente muestra síntomas de diabetes según la OMS.

■ Consta de 768 ejemplos.

■ Consta de 9 atributos (clase incluida).

■ Consta de 2 clases.

■ Los atributos indican, en este orden: Nº de embarazos, glucosa en plasma, presión diastólica, espesor de la piel del tríceps, cantidad de insulina en 2 horas, IMC, antecedentes familiares y edad.

■ Valores de la clase: tested_negative indica que el test no indicó diabetes (500 ejemplos), tested_positive indica que el test predice diabetes (268 ejemplos).

5. Bases de Datos a utilizar

■ Junto a *Iris*, utilizaremos cinco bases de datos más: *Pima*, *Wine*, *Sonar*, *SpamBase y Ozone*.

■ *Wine* es una base de datos con información química sobre vinos de viñedos diferentes de la misma región en Italia.

- ■ Consta de 178 ejemplos.
- ■ Consta de 14 atributos (clase incluida).
- ■ Consta de 3 clases.
- ■ Atributos: Resultados del análisis químico (alcohol encontrado, ácido málico, magnesio, fenoles, etc.).
- ■Las clases indican tres tipos de vinos: tipo 1 (59 ejemplos), tipo 2 (71 ejemplos) y tipo 3 (48 ejemplos).

Tipo 1

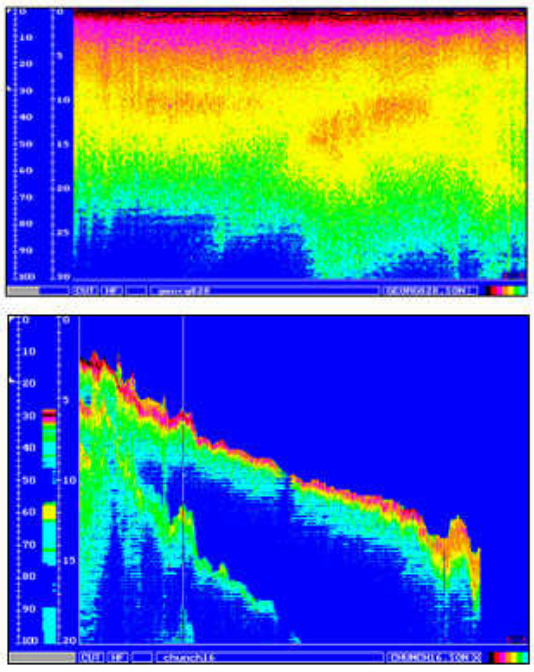
Tipo 2

Tipo 3

5. Bases de Datos a utilizar

■ Junto a *Iris*, utilizaremos cinco bases de datos más: *Pima*, *Wine*, *Sonar*, *SpamBase y Ozone*.

■ *Sonar* es una base de datos sobre detección de materiales mediante señales de sónar, discriminando entre objetos metálicos y rocas.



- Consta de 208 ejemplos.
- Consta de 61 atributos (clase incluida).
- Consta de 2 clases.
- Atributos: Cada atributo representa la energía dentro de una banda de frecuencia en concreto. Están representados en el rango [0.0,1.0].
- Hay 111 ejemplos referentes a señales obtenidas a partir de cilindros metálicos a partir de diversos ángulos y condiciones y 97 referentes a rocas bajo condiciones similares.

5. Bases de Datos a utilizar

■ Junto a *Iris*, utilizaremos cinco bases de datos más: *Pima*, *Wine*, *Sonar*, *SpamBase y Ozone*.

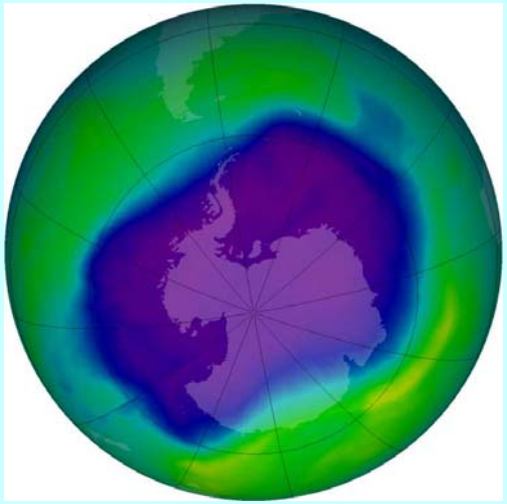
■ *SpamBase* es una base de datos que versa sobre la detección de correo electrónico de tipo SPAM, frente a correo seguro.

- Consta de 460 ejemplos.
- Consta de 58 atributos (clase incluida).
- Consta de 2 clases.
- Atributos: 48 atributos sobre la frecuencia de una palabra concreta en el correo, 6 atributos sobre la frecuencia de un carácter, 1 atributo sobre el número de caracteres seguidos en mayúsculas, etc.
- Las clases claramente discriminan entre SPAM y correo seguro.

5. Bases de Datos a utilizar

■ Junto a *Iris*, utilizaremos cinco bases de datos más: *Pima*, *Wine*, *Sonar*, *SpamBase y Ozone*.

■ *Ozone* es una base de datos para la detección del nivel de ozono según las mediciones realizadas a lo largo del tiempo.



- Consta de 185 ejemplos.
- Consta de 74 atributos (clase incluida).
- Consta de 2 clases.
- Atributos: Predicción del nivel de ozono local, Temperatura, Radiación Solar, Velocidad del viento…
- Las clases simplemente indican si el día será normal o tendrá una alta concentración de ozono.