

Ingeniería de Conocimiento: Tipos de Tarea de Analítica de Datos

Jose Aguilar

- Privacidad
- Seguridad
- Decisiones basado en datos incompletos
- Decisiones con datos inexactos
- Usando sólo los datos que apoyan nuestras decisiones
- Llegar a la conclusión errónea de los datos: por ejemplo, los precios de las acciones

Ciencias de los Datos

toma de decisiones basada en datos(DDD)

se refiere a la práctica de basar las decisiones en el análisis de los datos, en lugar de únicamente en la intuición.

- Por ejemplo, un vendedor podría seleccionar los anuncios publicitarios basados puramente en su larga experiencia en el campo y su ojo para lo que va a funcionar.

O,

- podría basar su selección en el análisis de los datos con respecto a cómo los consumidores reaccionan a diferentes anuncios.

También podría utilizar una combinación de estos enfoques. DDD no es una práctica de todo o nada, y diferentes empresas se dedican a DDD a mayor o menor grado.

Analizando los Datos

~ ~ Los métodos exploratorios

Este método implica a menudo una gran cantidad de cálculo de promedios y porcentajes, y la visualización de la información en un gráfico. Aunque los métodos exploratorios pueden proporcionar muchas piezas de información, no puede responder a preguntas específicas o hacer declaraciones definitivas acerca de un problema.

~ ~ Los métodos de confirmación

Este método se utiliza para la conclusión de los resultados de encuestas y la información estadística al responder a preguntas específicas. Por ejemplo, utilizando un método de confirmación, un estadístico puede decir "Los precios del petróleo que salen de Arabia Saudita han ido en aumento, y aumentarán los precios."

Ambos métodos se deben utilizar ampliamente para analizar los resultados

Los métodos cuantitativos y cualitativos producen diferentes tipos de datos

- Los datos cuantitativos produce valores numéricos
- Los datos cualitativos produce narrativas

Sin embargo, para los datos cuantitativos y cualitativos, las mismas estrategias analíticas se utilizan para la interpretación de los datos

Categorías de AD

- Narrativa (e.g. leyes)
- Descriptiva (e.g. ciencias sociales)
- Estadística/matemática (pura/aplicada)
- Audio-Optico (e.g. telecomunicación)
- Otros

La mayoría de análisis, sin duda, adoptan las primeras tres.

La segunda y tercera son más popular y las ciencias aplicadas y puras, y sociales

¿Qué es la AD?

- **Métodos Descriptivos**

Encontrar patrones interpretable que describen los datos.

- **Métodos de Predicción**

Utilizar algunas variables para predecir los valores desconocidos o futuros de otras variables.

MODELOS!!!

Descriptivo

Predictivo

Prescriptivo

Preguntas

Qué paso?
Qué está pasando?
Cuál es el problema?
Qué acciones son necesarias?

Por qué esta pasando?
Qué se producira?
Por qué se producira?

Qué debería hacerse?
Por qué debería hacerse?
Qué pasa si se intenta eso?

Enablers

- Reportes
- Dashboards
- Data Warehousing
- Alertas

- Data Mining
- Text Mining
- Web/Media Mining
- Forecasting

- Optimización
- Simulación
- Modelos de Decisión

Resultados

Bien definidos los problemas y oportunidades

Proyección de los futuros estados y condiciones

Mejores posibles decisiones y transacciones

Las estrategias analíticas básicas:

Describiendo

Factorización

Agrupación

Comparando

Clasificación

encontrar puntos comunes

encontrar covarianza

Descartar alternativas



Las estrategias analíticas básicas.

- Clasificación
- Pronóstico (Predicción)
- Asociación (reglas)
- Agrupación o segmentación (Clustering)

Tipos de aplicaciones de la AD

- Clasificación [**predictivo** y descriptivo]
- Clustering [**descriptivo**]
- Descubrimiento de Regla Asociación [**descriptivo**]
 - Análisis de dependencia de datos
 - correlación y causalidad
- Descubrimiento Patrones Secuenciales [**descriptivo**]
 - Análisis de series de tiempo, asociaciones secuenciales
- Regresión [**predictivo**]
- Detección de Tendencia y Desviaciones [**predictivo**]
- Filtros Colaborativos [**predictivo** y descriptivo]
- Resumir [**descriptivo**]
- Descripción de Conceptos [**descriptivo**]
 - Descripción de características
 - Descripción de su identidad discriminante



Clasificación

Examinar las características de un nuevo objeto y **asignarle** una clase o categoría de acuerdo a un conjunto de tales objetos previamente clasificados.

- Ejemplos:
 - Clasificar los estudiantes en categorías según sus rendimiento: bajo, medio y alto
 - Detectar los estados operacionales de un sistema: con falla, seguro, inactivo.



Pronóstico

Predecir un valor futuro con base a valores pasados

Prognosis

Predicción

- Ejemplos:
 - Predecir cuánto efectivo requerirá un cajero automático en un fin de semana



Asociación

Determinar cosas u objetos que van juntos

- Ejemplo:
 - Determinar que productos se adquieren conjuntamente en un supermercado



Agrupación o segmentación

Dividir una población en un número de grupos más homogéneos

- No depende de clases pre-definidas a diferencia de la clasificación
- Ejemplo:
 - Dividir la base de clientes de acuerdo con los hábitos de consumo
 - Establecer los grupos de estudiante según sus estilos de aprendizaje

Clasificación

Clasificación

Email: Spam / No es Spam?

Transacciones en línea: Fraudulento (Si / No)?

Tumor: Maligno / Benigno ?

$$y \in \{0, 1\}$$

0: “Clase negativa” (tumor benigno)

1: “Clase positiva” (tumor maligno)

Clasificación

Obtener una función o modelo que determine la clase de un objeto basado en las características de sus atributos.

- Para generar dicho modelo o función, es necesario definir un conjunto de datos de entrenamiento, compuesto por objetos que ya tienen su clase asignada, también denominados ejemplos etiquetados.
- El modelo o función es creado analizando las relaciones entre los atributos de los objetos en el conjunto de entrenamiento y las clases.
- Mientras mas variedad de escenarios se presenten en el conjunto de entrenamiento, mas se enriquece el modelo de clasificación (mejores resultados en la clasificación de nuevas entradas no etiquetadas).

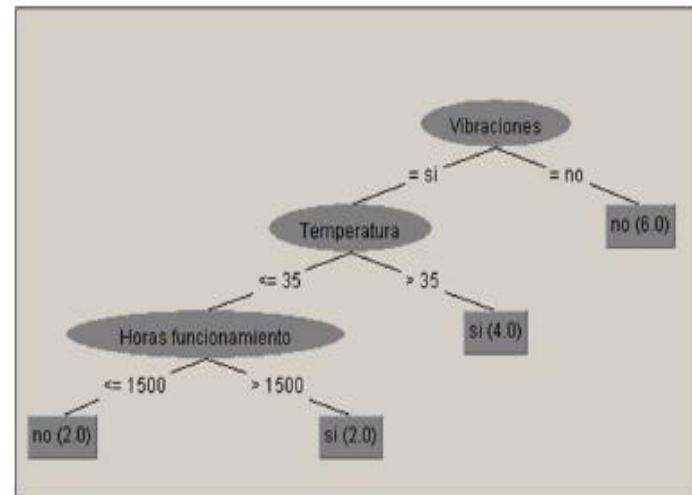
Clasificación

Ante nuevos valores de unas variables independientes(variables de entrada), el modelo obtenido permite prever la clase correspondiente de la variable dependiente(de salida).

Por ejemplo:

Árbol de decisión para predecir el fallo de una máquina según la temperatura, horas de funcionamiento, si hay vibraciones y días desde la ultima revisión.

No.	Temperatura Numeric	Vibraciones Nominal	Horas funcionamiento Numeric	Días desde revisión Numeric	FALLO Nominal
1	55.0	si	500.0	55.0	si
2	23.0	no	30.0	17.0	no
3	45.0	no	1500.0	72.0	no
4	47.0	no	650.0	43.0	no
5	32.0	si	700.0	58.0	no
6	35.0	si	2500.0	93.0	si
7	50.0	si	150.0	21.0	si
8	53.0	si	550.0	50.0	si
9	21.0	no	35.0	12.0	no
10	47.0	no	1200.0	75.0	no
11	43.0	no	750.0	51.0	no
12	35.0	si	680.0	63.0	no
13	30.0	si	2300.0	87.0	si
14	52.0	si	180.0	23.0	si



Clasificación suave

Los modelos de clasificación suave, además de clasificar la clase, determinan el grado o probabilidad de certeza de cada una de las clasificaciones.

Por ejemplo:

Clasificador Naive Bayes del modelo de predicción del fallo de una maquina indicando el grado de probabilidad de fallo o no fallo de cada ejemplo.

Naive Bayes Classifier		inst#,	actual,	predicted,	error,	probability distribution
Class si: Prior probability = 0.44		1	1:si	1:si		*0.967 0.033
		2	2:no	2:no		0.028 *0.972
Temperatura: Normal Distribution. Mean = 45.8485 StandardDev = 9.6788 WeightSum = 6 Precision = 3.050505050505051		3	2:no	2:no		0.028 *0.972
Vibraciones: Discrete Estimator. Counts = 7 1 (Total = 8)		4	2:no	2:no		0.028 *0.972
Horas funcionamiento: Normal Distribution. Mean = 1045 StandardDev = 954.7382 WeightSum = 6 Precision = 190.0		5	2:no	2:no		0.492 *0.508
Dias desde revisión: Normal Distribution. Mean = 55.0385 StandardDev = 28.1919 WeightSum = 6 Precision = 6.230769230769231		6	1:si	2:no	+	0.492 *0.508
		7	1:si	1:si		*0.967 0.033
		8	1:si	1:si		*0.967 0.033
Class no: Prior probability = 0.56		9	2:no	2:no		0.028 *0.972
		10	2:no	2:no		0.028 *0.972
Temperatura: Normal Distribution. Mean = 36.3182 StandardDev = 10.1047 WeightSum = 8 Precision = 3.090909090909091		11	2:no	2:no		0.028 *0.972
Vibraciones: Discrete Estimator. Counts = 3 7 (Total = 10)		12	2:no	2:no		0.492 *0.508
Horas funcionamiento: Normal Distribution. Mean = 688.75 StandardDev = 483.8243 WeightSum = 8 Precision = 190.0		13	1:si	2:no	+	0.492 *0.508
Dias desde revisión: Normal Distribution. Mean = 49.0673 StandardDev = 21.9048 WeightSum = 8 Precision = 6.230769230769231		14	1:si	1:si		*0.967 0.033

Clasificación

categoría

categoría

continuo

clase

ID	Reemb	Edo Civil	pago Impuest	Enga ña
1	Si	Soltero	125K	No
2	No	Casado	100K	No
3	No	Soltero	70K	No
4	Si	Casado	120K	No
5	No	Divorc.	95K	Si
6	No	Casado	60K	No
7	Si	Divorciad	220K	No
8	No	Soltero	85K	Si
9	No	Casado	75K	No
10	No	Soltero	90K	Si

nominales
numérico
s

Reemb	Edo. Civil	pago Impuest	Enga ña
No	Soltero	75K	?
Si	Casado	50K	?
No	Casado	150K	?
Si	Divorciad	90K	?
No	Solero	40K	?
No	Casado	80K	?



Clasificación

Detección de Fraude

Objetivo: Predecir casos fraudulentos en las transacciones con tarjetas de crédito.

Enfoque:

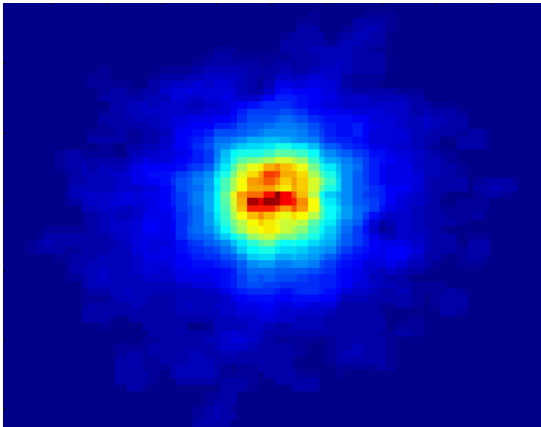
- Utilizar las transacciones con tarjetas de crédito y la información sobre la cuenta del titular como atributos.
 - ¿Cuándo compra un cliente?, ¿qué compra?, ¿con qué frecuencia paga a tiempo?, etc.
- Etiquetar transacciones pasadas: fraudulentas o correctas. Esto forma el atributo de clase.
- Aprender un modelo para las clases de transacciones.
- Utilice este modelo para detectar futuros fraudes mediante la observación de las transacciones de las tarjetas de crédito en una cuenta.

Clasificación

Clasificar cada imagen como una estrella (y su estado de formación) o una imagen no estelar (galaxia) (no-estelar)

<http://aps.umn.edu>

Temprano



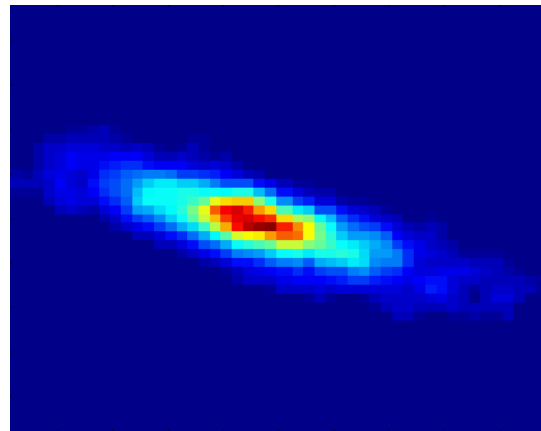
Clases:

- Estado de Formación

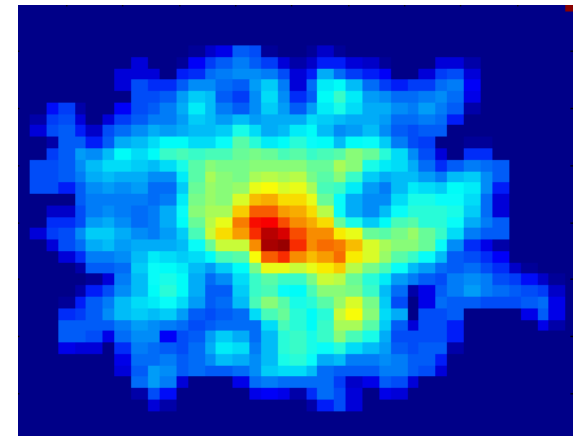
Atributos:

- Caract. Imagen,
- Caract. Ondas, luz, etc.

Intermedio



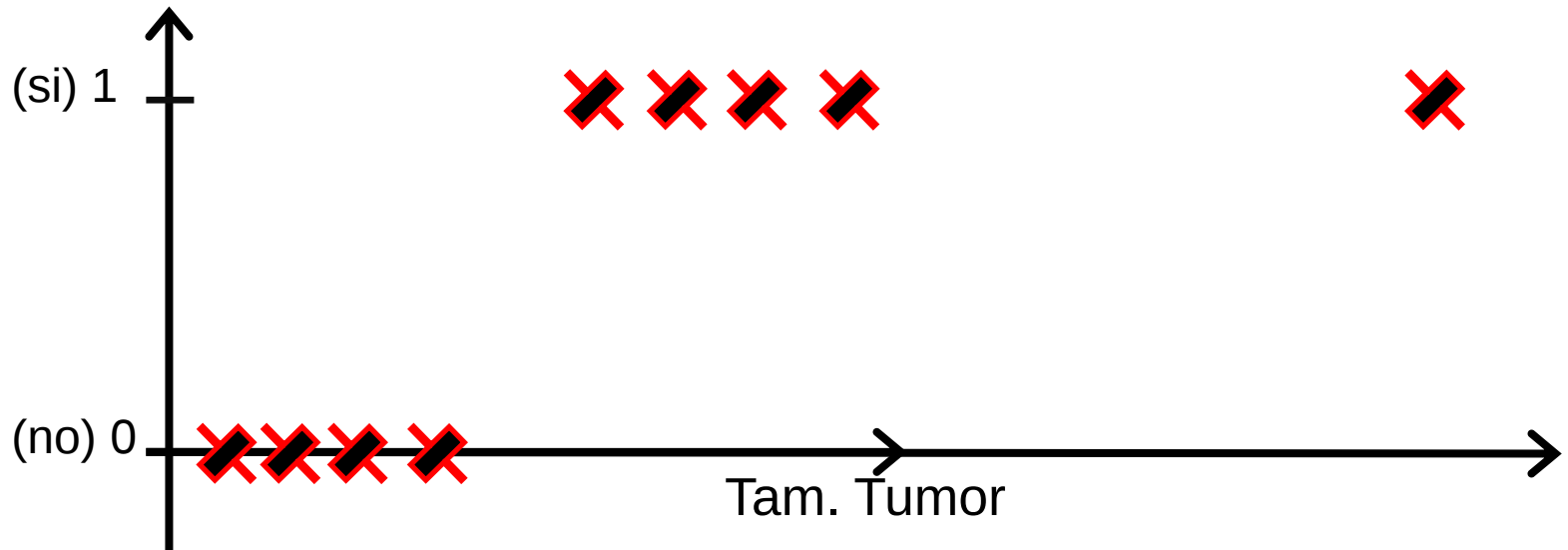
Tarde



Tam. datos:

- 72 millones estrellas, 20 millones galaxias
- BD Imagen: 150 GB

Clasificación



Umbral clasificador

$$h_{\theta}(x) = 0.5$$

Si $h_{\theta}(x) \geq 0.5$, predice “y = 1”

Si $h_{\theta}(x) < 0.5$, predice “y = 0”

Algoritmos de Clasificación

- **Basados en análisis estadísticos (Clasificación Bayesiana):** Gaussian Naive Bayes, Bernoulli Naive Bayes, La regresión y sus variantes: regresión lineal, regresión logística, isotonic regresión, entre otros, Procesos gaussianos, Redes bayesianas
- **Basados en Árboles de decisión:** J48, CART, C4.5, ADtree, randomTree, REPTree,
- **Basados en reglas:** ZeroR, M5Rule, ConjunctiveRule, PART
- **Basados en distancia**
- **Basados en redes neuronales:** perceptron simple y el perceptron multicapa
- **Híbridos**

Métricas para evaluar un algoritmo de clasificación

Clasificación	Clasificados positivos	Clasificados negativos
Pos	Verdaderos positivos (tp)	Falsos negativos (fn)
Neg	Falsos positivos (fp)	Verdaderos negativos (tn)

Métrica	Formula
Tasa de error	$\frac{\sum_{i=1}^n \frac{tp_i + tn_i}{tp_i + tn_i + fp_i + fn_i}}{n}$
Precisión	$\frac{\sum_{i=1}^n tp_i}{\sum_{i=1}^n (tp_i + fp_i)}$
Recall _μ	$\frac{\sum_{i=1}^n tp_i}{\sum_{i=1}^n (tp_i + fn_i)}$

Modelos de Agrupamiento (Segmentación)

Agrupamiento (Clustering)

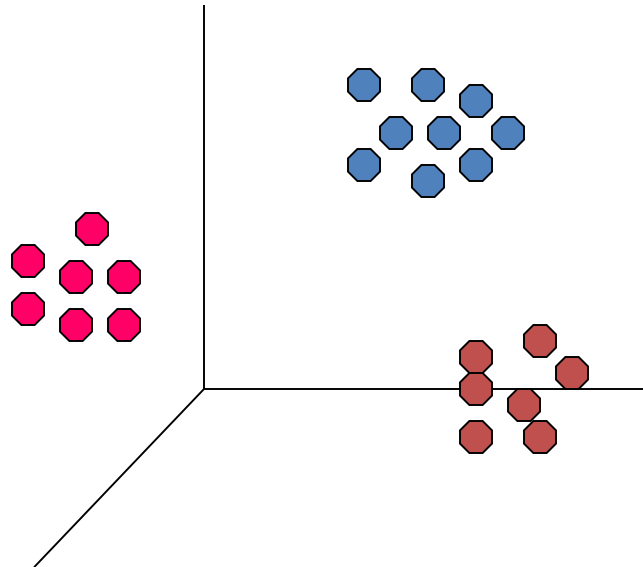
Dado un conjunto de datos, cada una con un conjunto de atributos, y una medida de similitud entre ellos, **encontrar grupos** de tal manera que:

- Los puntos de datos en un clúster son los **más similares** entre sí.
- Los puntos de datos en grupos separados son **menos similares** entre sí.

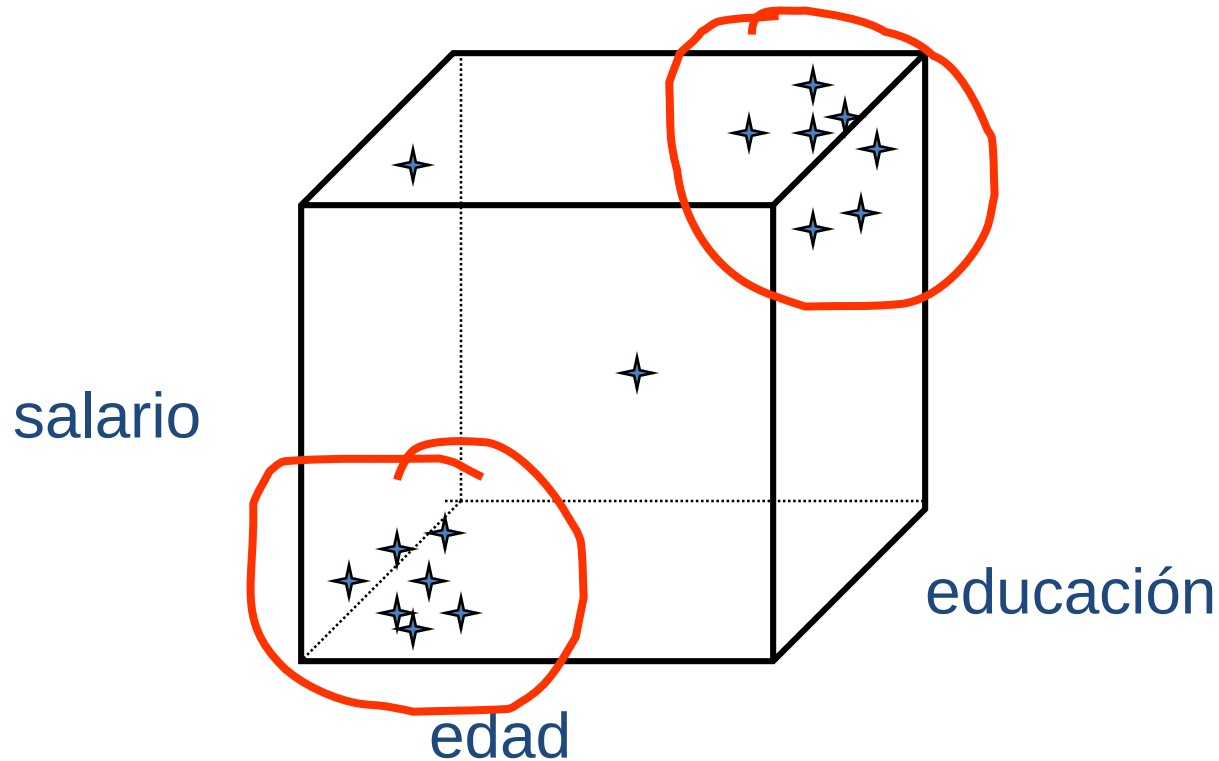
Agrupamiento (Clustering)

Distancias Intracluster
son minimizadas

Distancias Intercluster
son maximizadas

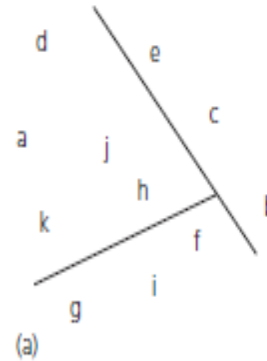


Agrupamiento (Clustering)



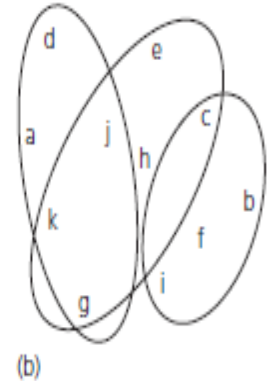
Clusters

Técnica muy usadas
son inferir un **árbol de
decisión** o un **conjunto
de reglas** que asignan a
cada instancia **al grupo**
al que pertenece

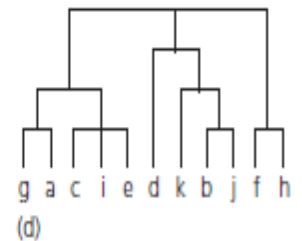


	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
f	0.1	0.4	0.5
g	0.7	0.2	0.1
h	0.5	0.4	0.1

(c)



(b)



(d)

No hay etiquetas de por medio

Ejemplo de Clustering

Agrupación de documento:

- **Objetivo:** encontrar grupos de documentos que son similares entre sí sobre la base de los términos importantes que aparecen en ellos.
- **Enfoque:** Identificar términos que aparecen con frecuencia en cada documento. Formar una medida de similitud basada en las frecuencias de los diferentes términos.

Agrupación de documento

- 3204 Artículos de un periódico.
- **Medida Similitud:** ¿Cuántas palabras son comunes en estos documentos (después de algún tipo de filtrado de palabras).

<i>Categoría</i>	<i>Total Articulos</i>	<i>Grupos</i>
<i>Financiero</i>	555	36
<i>Extranjero</i>	341	20
<i>Nacional</i>	273	6
<i>Ciudad</i>	943	76
<i>Deportes</i>	738	73
<i>Entretenimiento</i>	354	28

Tipos de clustering

- **Clustering particional**

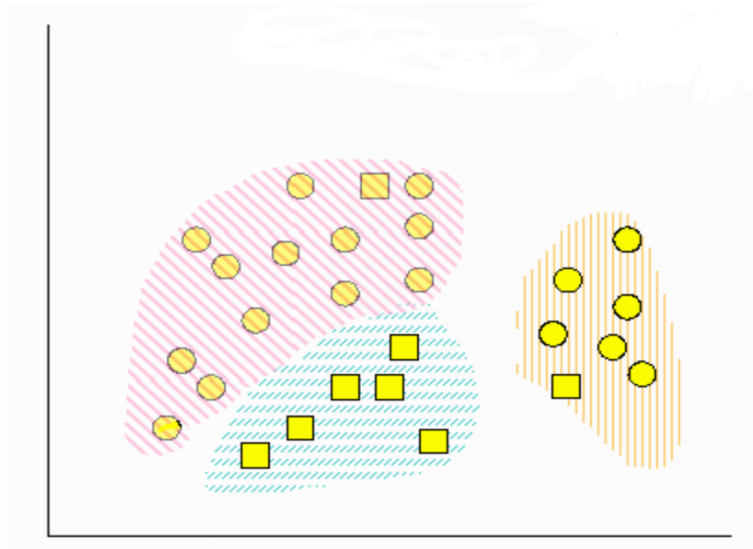
- Partición de los objetos en grupos o clusters. Todos los objetos pertenecen a alguno de los k clusters, los cuales son disjuntos. **Problema => elección de k**

- **Clustering ascendente jerárquico**

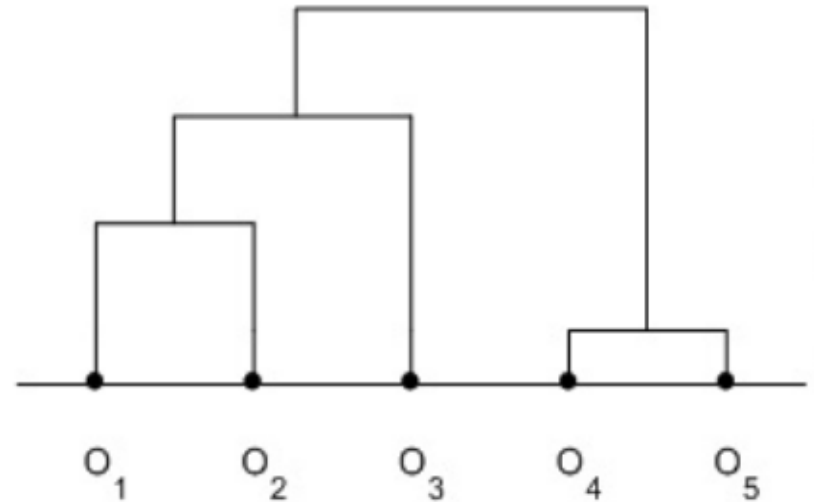
- Crear un dendograma, es decir, crear un conjunto de agrupaciones anidadas hasta construir **un árbol jerárquico**

Clusterización

Dados unos **datos sin etiquetar**, el objetivo es encontrar grupos naturales de instancias



a) Particional



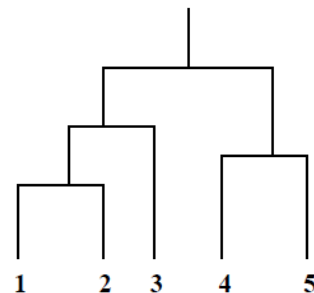
b) Jerárquico

Tipos de clustering

Métodos Jerárquicos

- **Los métodos aglomerativos:** Comienzan con la creación de un cluster para cada uno de los ejemplos individuales, seguidamente se van mezclando en pares los clusters más cercanos hasta que queden K clusters.
- **Los métodos divisivos:** Comienza con la creación de uno o dos cluster que contienen todos los ejemplos, seguidamente se va dividiendo hasta que se crearon K clusters

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



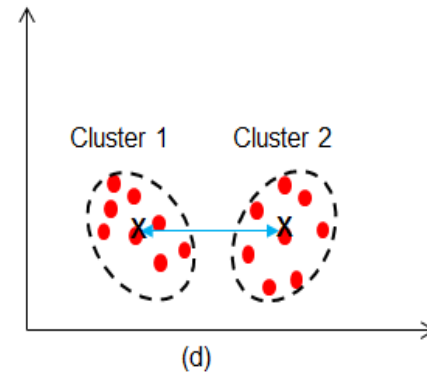
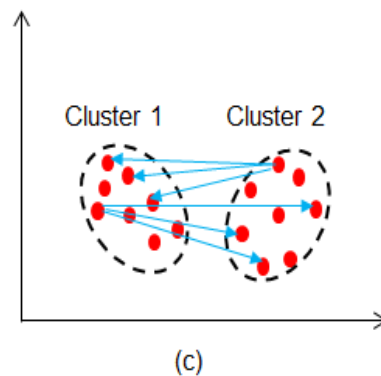
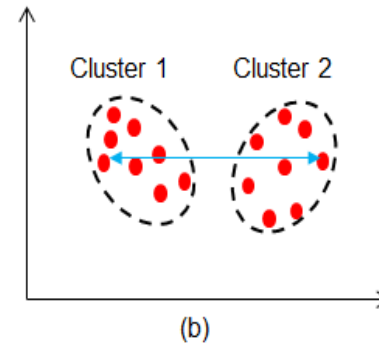
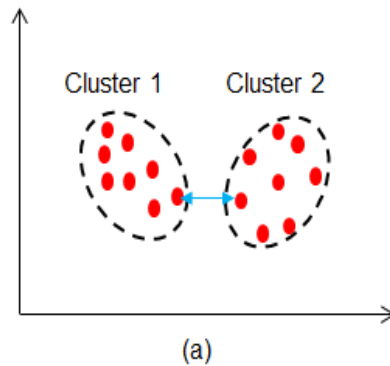
Tipos de clustering: basados en distancia

(a) Distancia mínima

(b) Distancia máxima

(c) Distancia de promedio del grupo

(d) Distancia con respecto al centroide



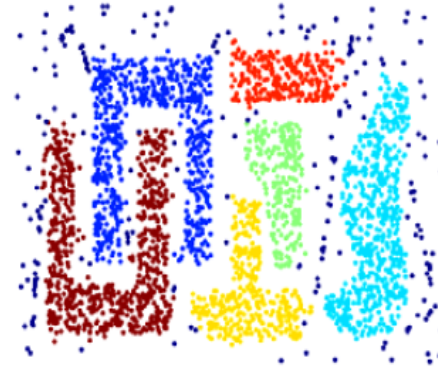
Tipos de clustering

Basados en densidad

Los algoritmos basados en densidad, tratan de formar agrupaciones en áreas con altas densidades de ejemplos



a) Datos originales



b) Datos después de clusterin

Tipos de clustering

Algoritmo de DBSCAN

```
current_cluster_label  $\leftarrow$  1
for all core points do
  if the core point has no cluster label then
    current_cluster_label  $\leftarrow$  current_cluster_label + 1
    Label the current core point with cluster label current_cluster_label
  end if
  for all points in the Eps-neighborhood, except  $i^{th}$  the point itself do
    if the point does not have a cluster label then
      Label the point with cluster label current_cluster_label
    end if
  end for
end for
```

1. Comienza eliminando los puntos de ruido, es decir que no tienen una densidad mayor a un umbral en un radio previamente definido,
2. seguidamente se procede a realizar el agrupamiento de los puntos

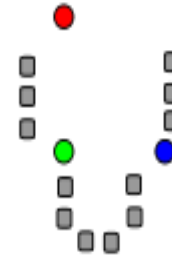
Tipos de clustering

Basados en prototipos

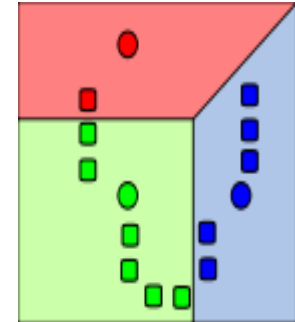
- Un cluster se define por un conjunto de objetos, donde cada objeto esta más cerca (o es más similar) al **prototipo** que define al cluster donde fue asignado que a cualquier otro prototipo de otro cluster existente.
- El prototipo que define al cluster normalmente **denota sus características principales**, por ejemplo, para atributos solo numéricos el prototipo del cluster a menudo se representa como el centroide.

Algoritmo K-medias

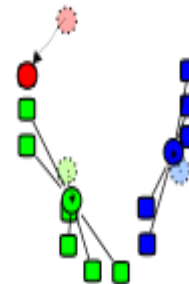
es un método de agrupamiento, que tiene como objetivo la partición de un conjunto (n) en k grupos en el que **cada ejemplo pertenece al grupo más cercano a la media**



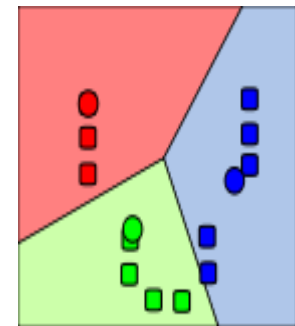
1) k centroides iniciales generados aleatoriamente (en este caso $k=3$)



2) k grupos son generados asociándole el punto



3) El centroide de cada uno de los k grupos se recalcula



4) Pasos 2 y 3 se repiten hasta que se logre la convergencia.

Algoritmo K-medias

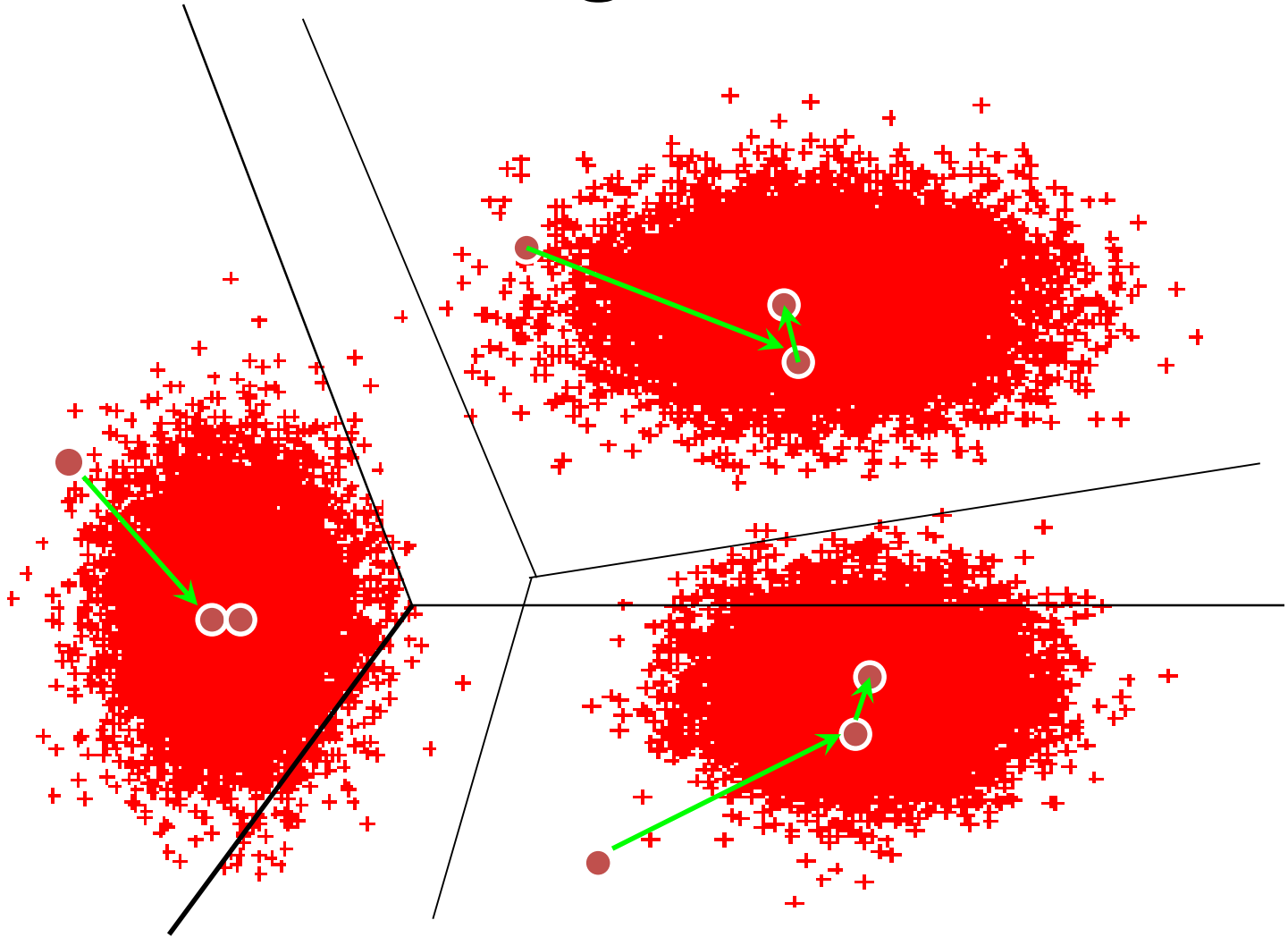
1. **Método más utilizado** de clustering particional
2. La idea es situar **los prototipos o centros en el espacio**, de forma que los datos pertenecientes al mismo prototipo tengan características similares
3. Los datos se **asignan a cada centro según la menor distancia**, normalmente usando la distancia euclídea
4. Una vez introducidos todos los datos, se **desplazan los prototipos hasta el centro de masas** de su nuevo conjunto, esto se repite hasta que no se desplazan más.

Clusterización: Algoritmo K-Medias

- Seleccionar centroides aleatorios
- Asignar cada objeto al grupo cuyo centroide sea el más cercano al objeto.
- Cuando todos los objetos hayan sido asignados, recalcular la posición de los k centroides.
- Repetir los pasos 2 y 3 hasta que los centroides no varíen

Distancia Euclídea $\delta^2_E (X_i, X_j) = || X_i - X_j ||^2$

Clusterización: Algoritmo K-Medias



Corrida en frio K-means

Input:

- K (number of clusters)
- Trainingset $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$

Randomly initialize K cluster centroids $\mu_1, \mu_2, \dots, \mu_K \in \mathbb{R}^n$

Repeat {

 for $i = 1$ to m

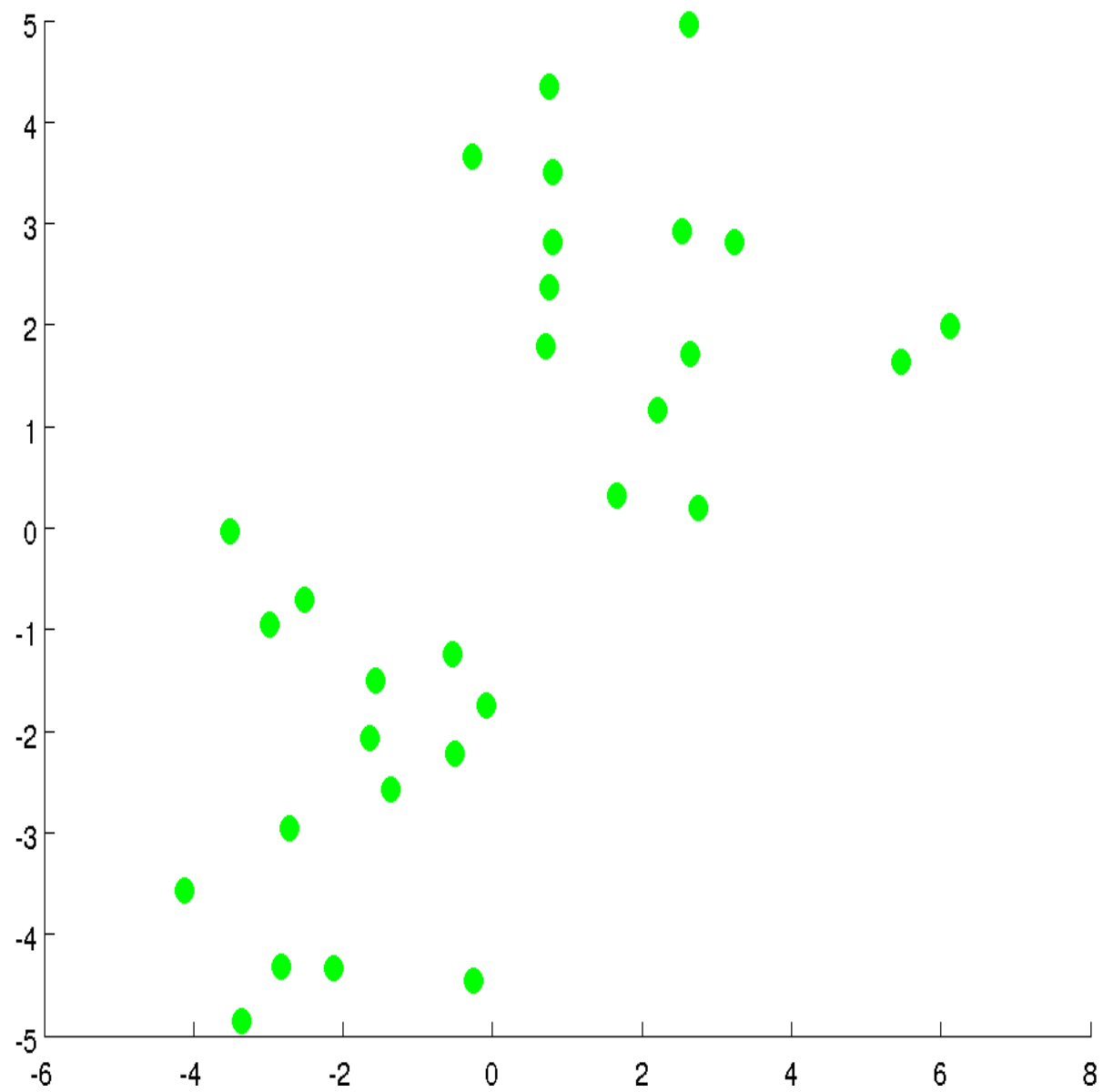
$c^{(i)} :=$ index (from 1 to K) of cluster centroid
 closest to $x^{(i)}$

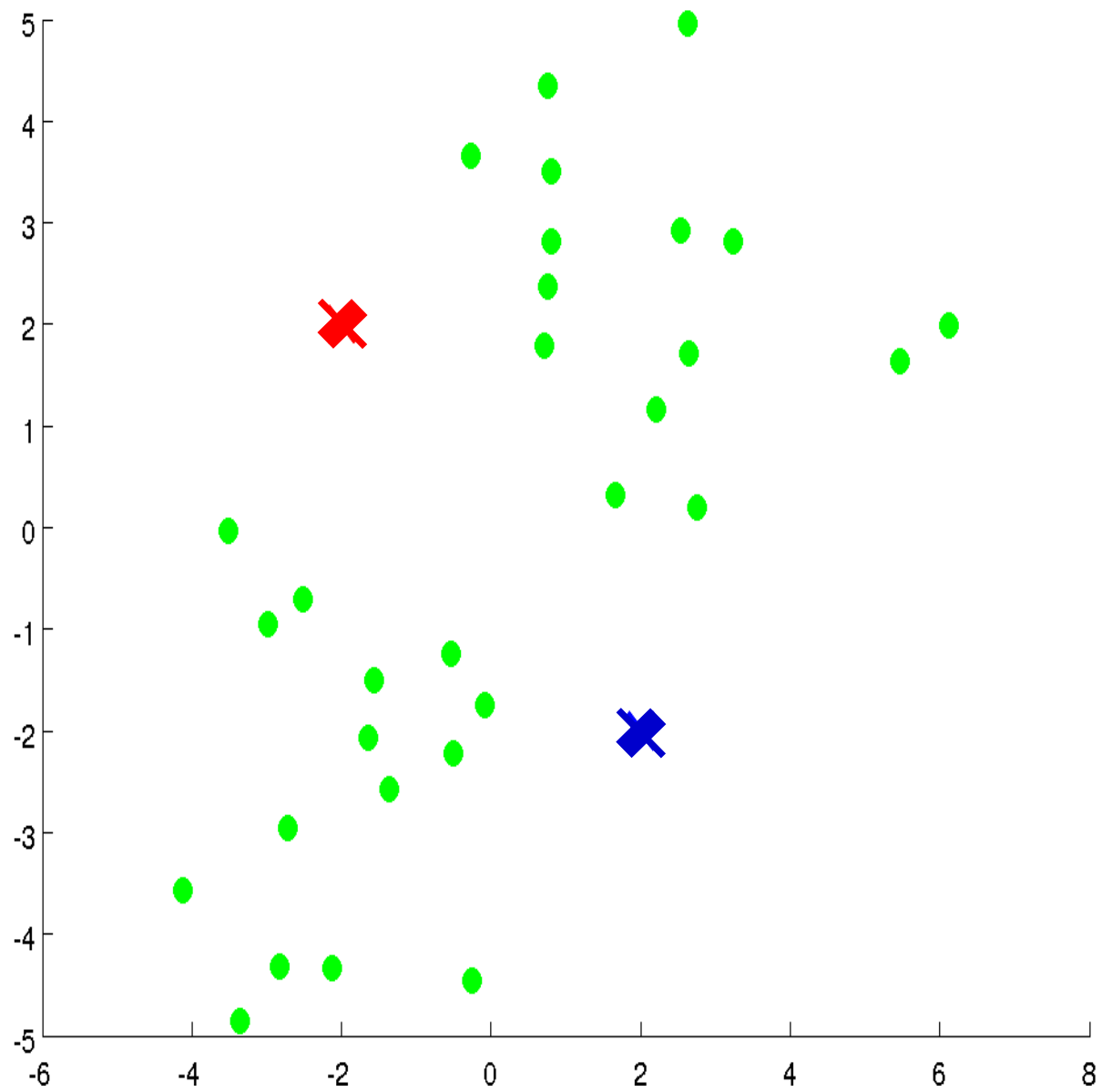
 for $k = 1$ to K

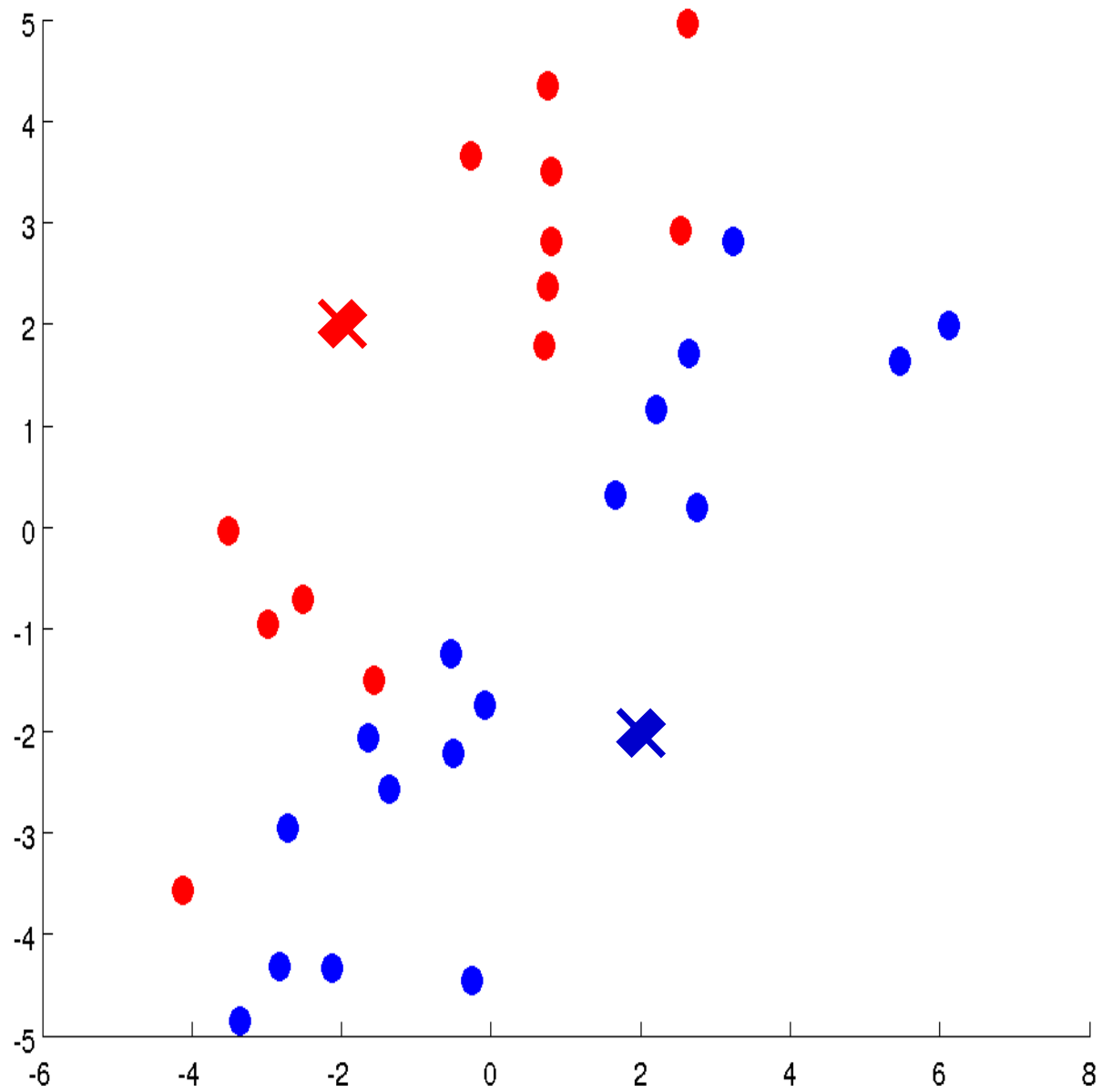
$\mu_k :=$ average (mean) of points assigned to

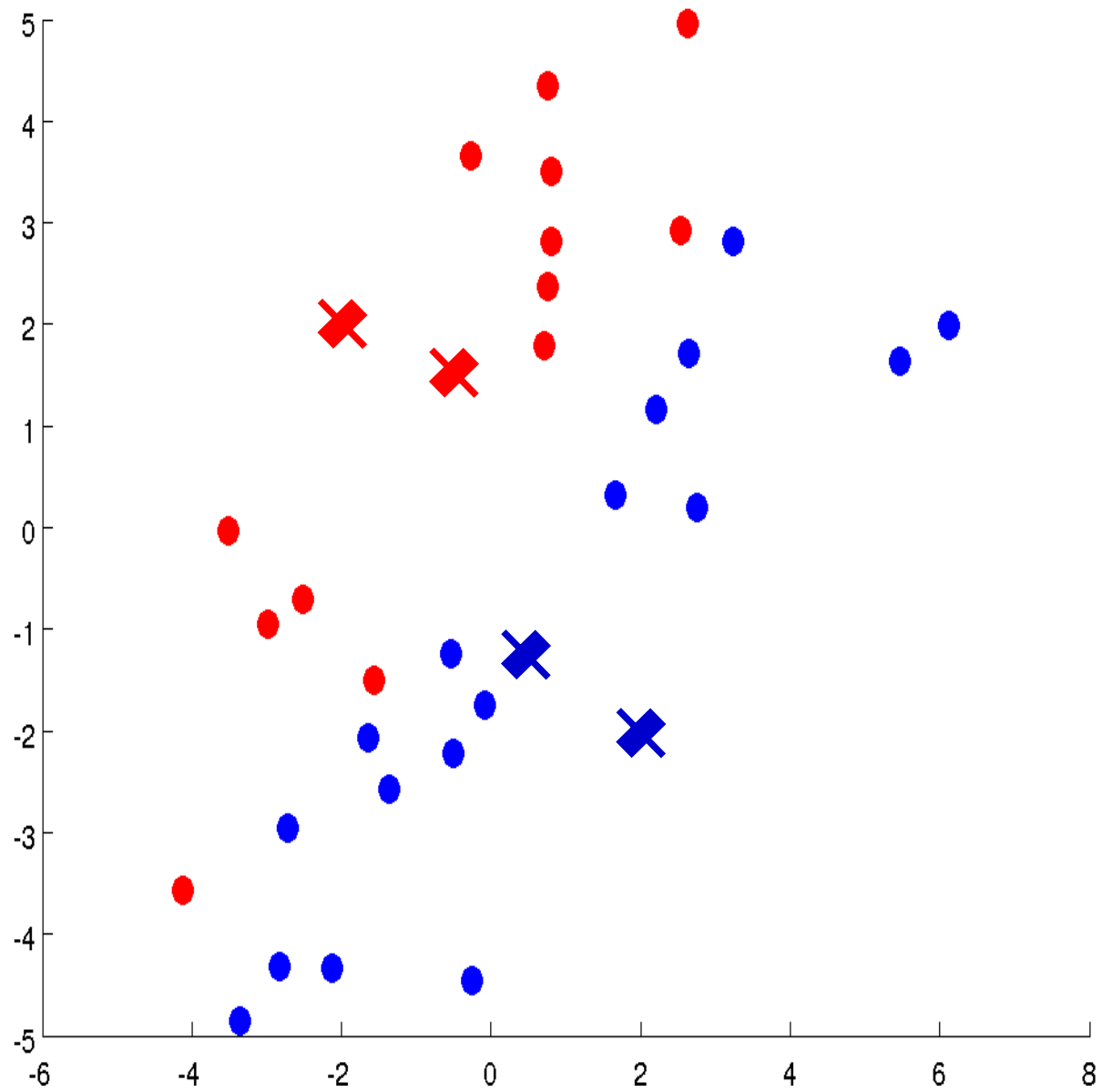
cluster

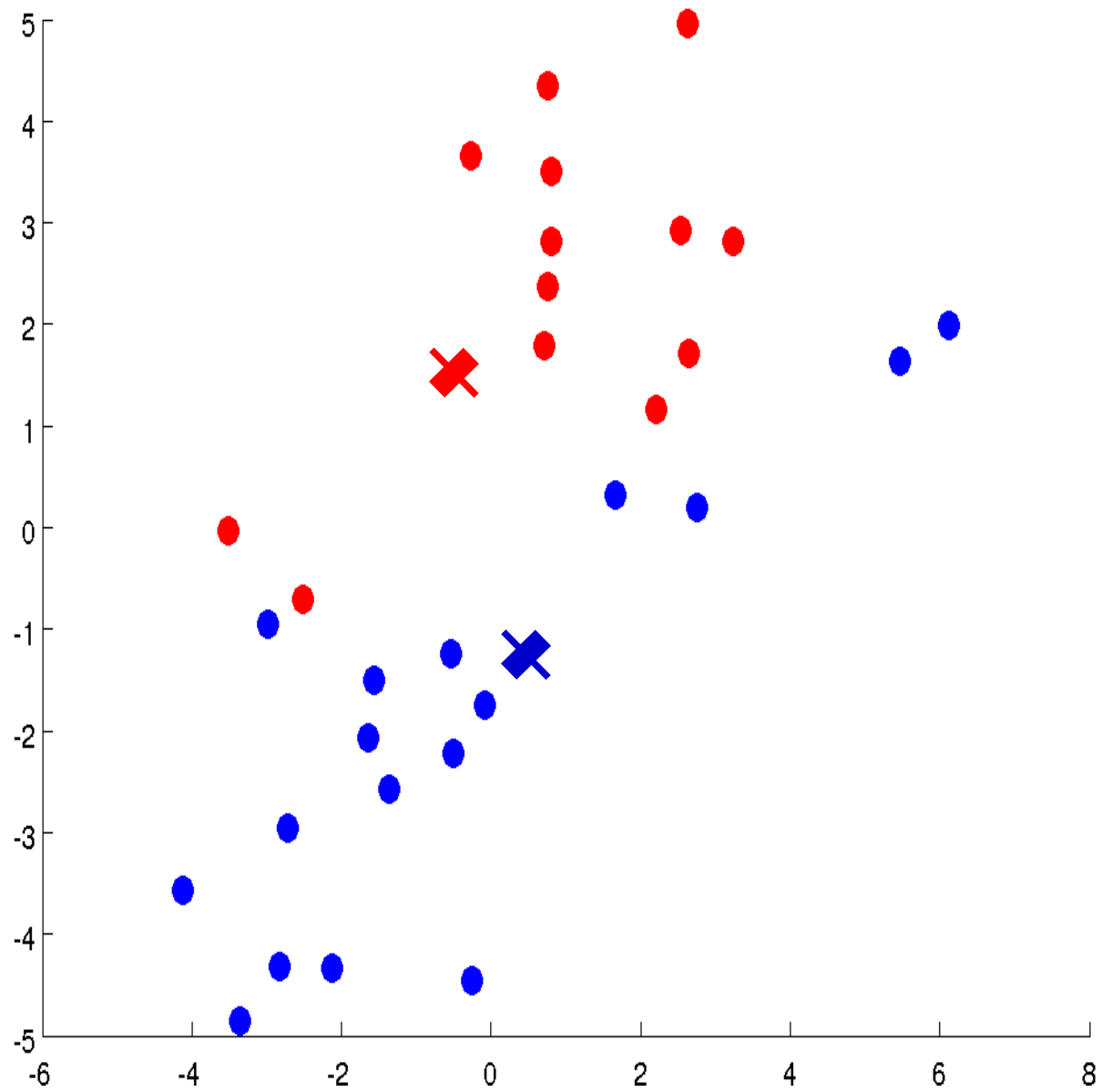
 } until convergence criteria is met

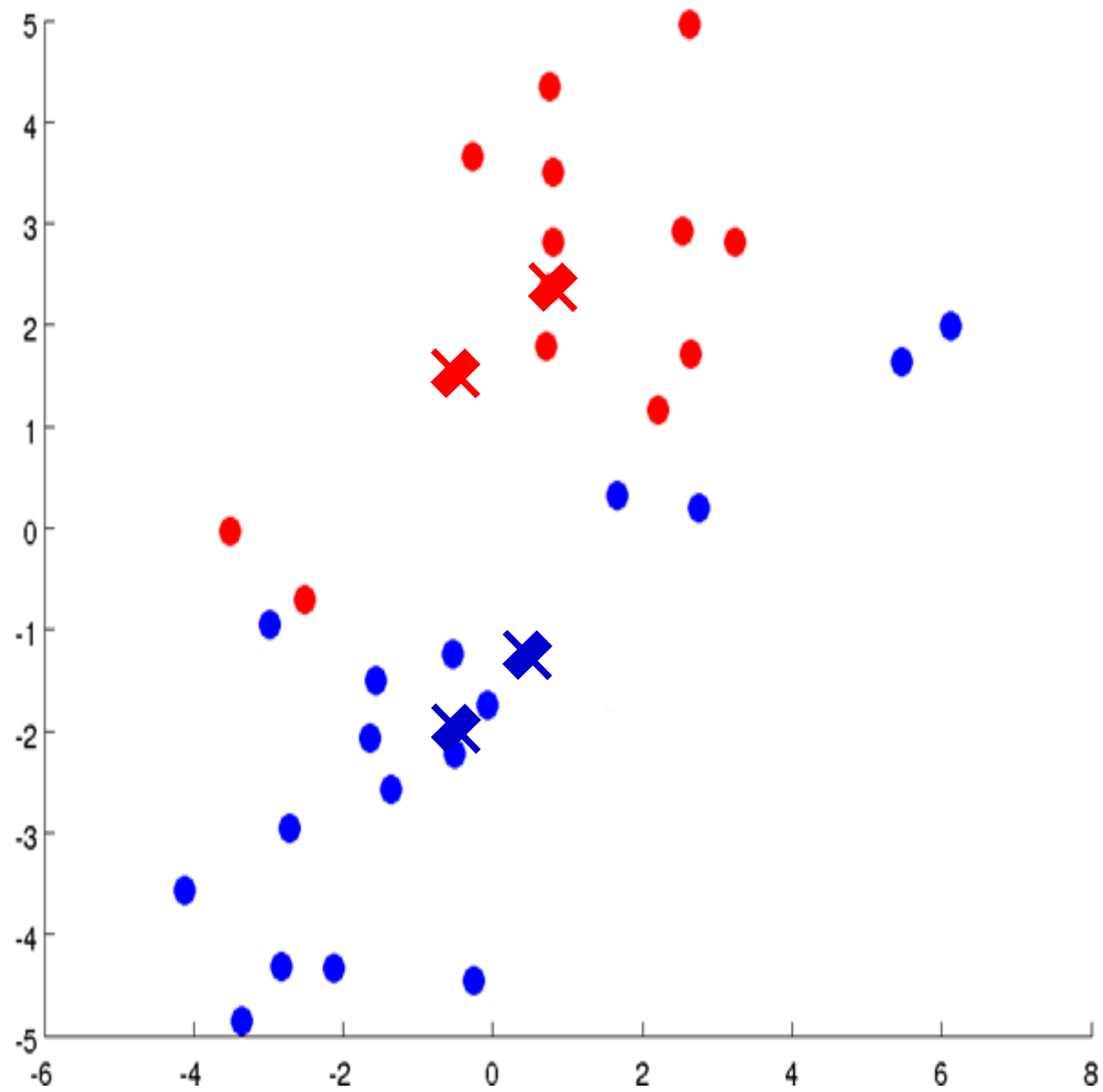


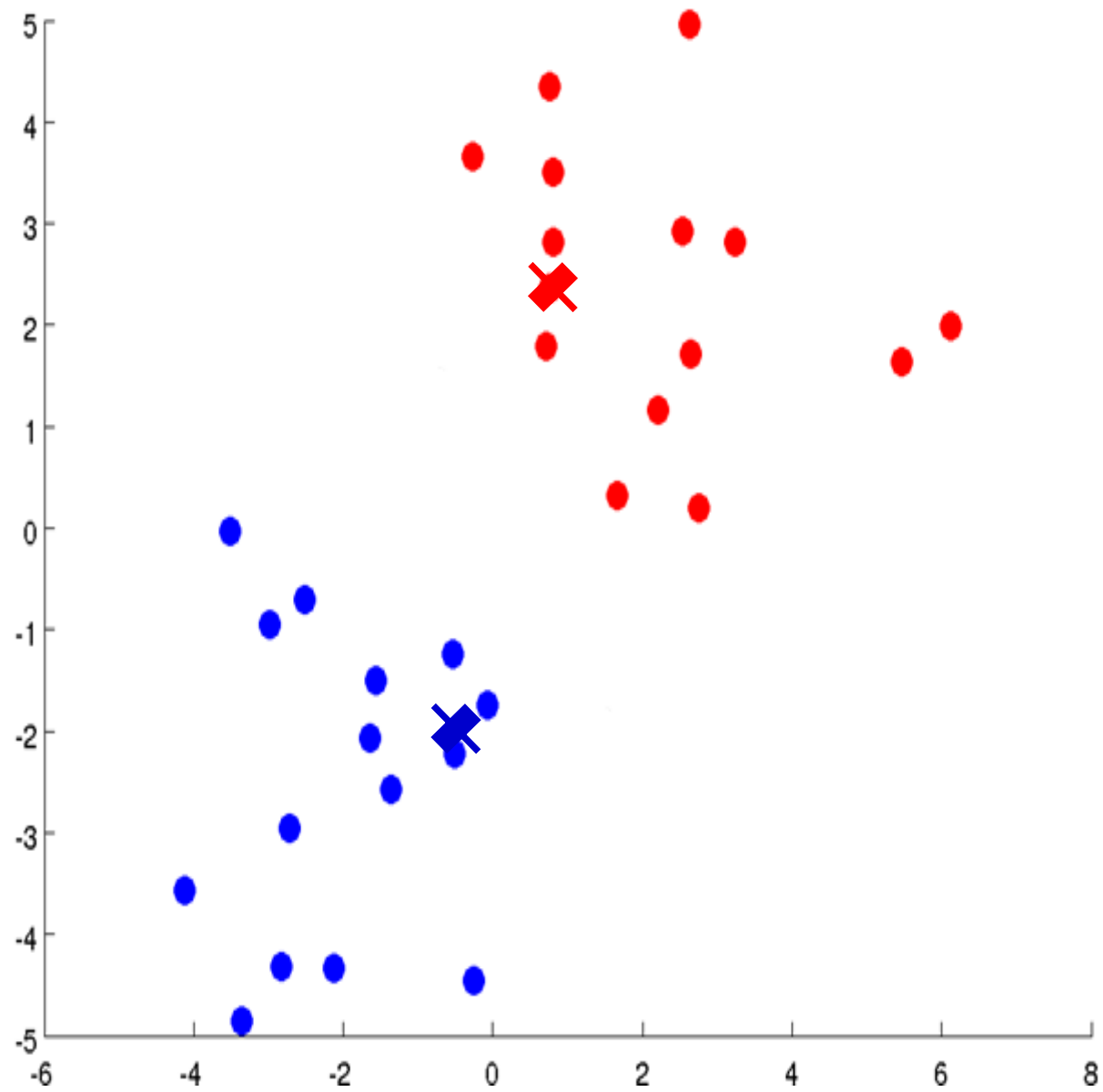


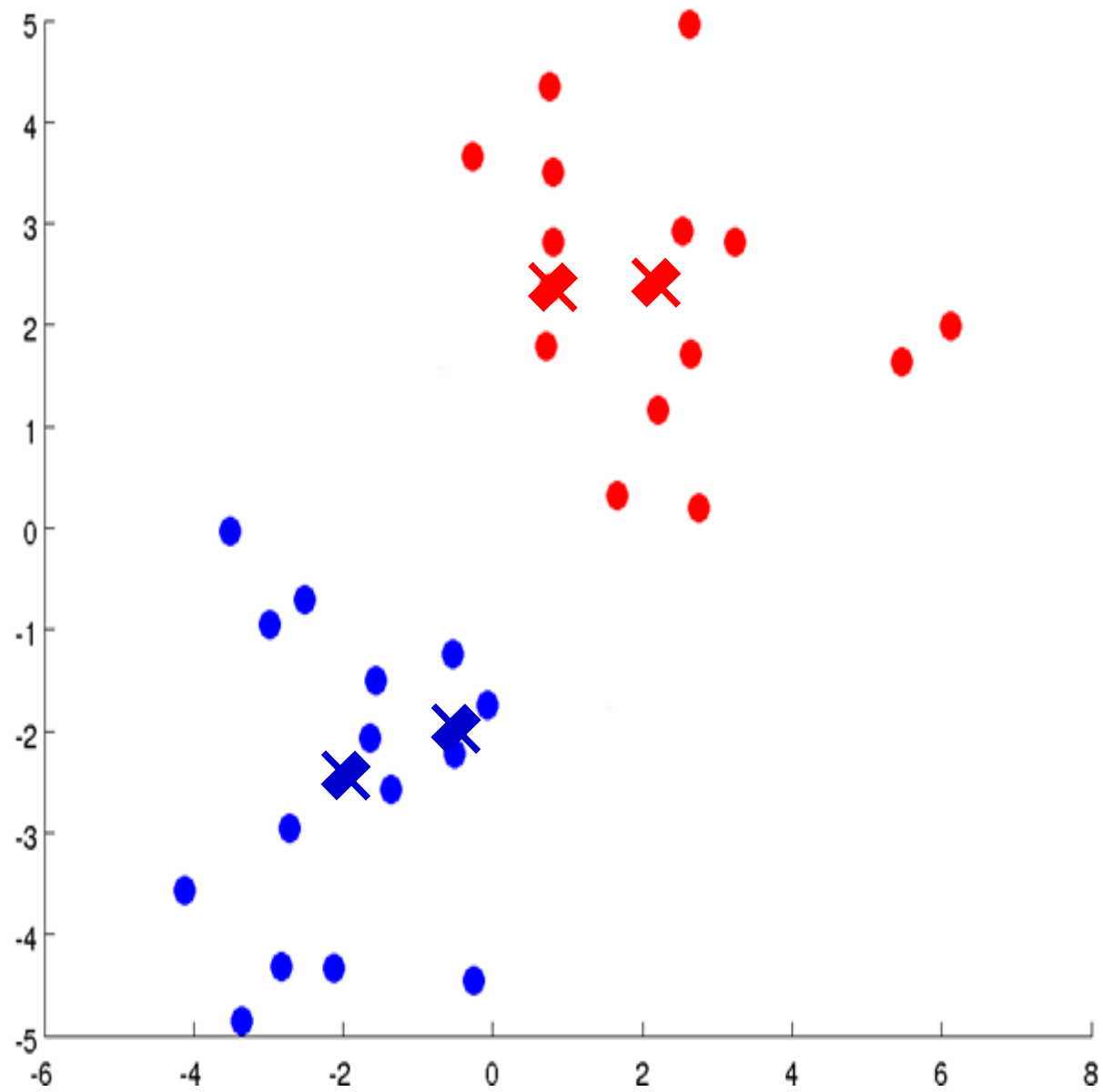


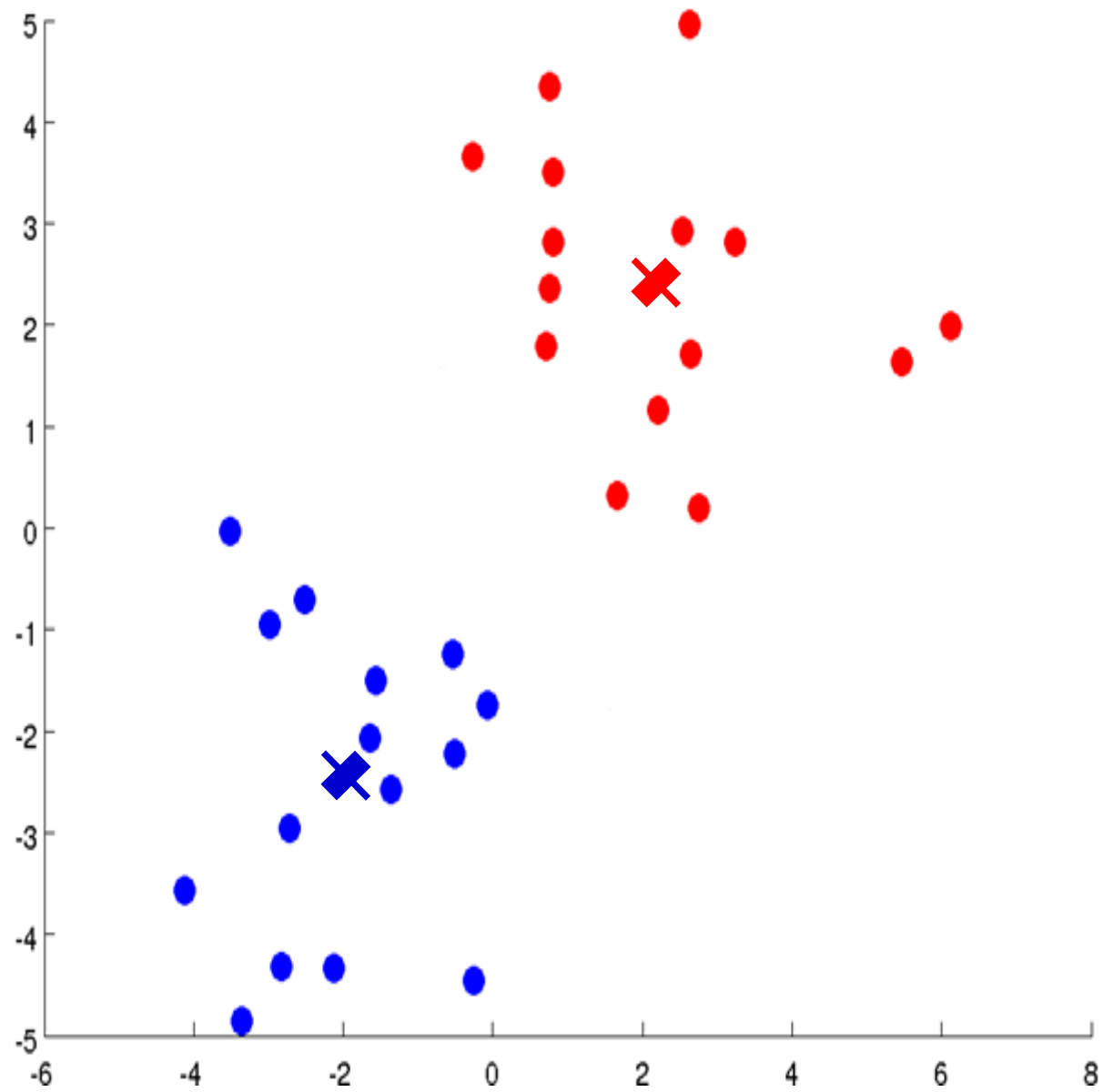




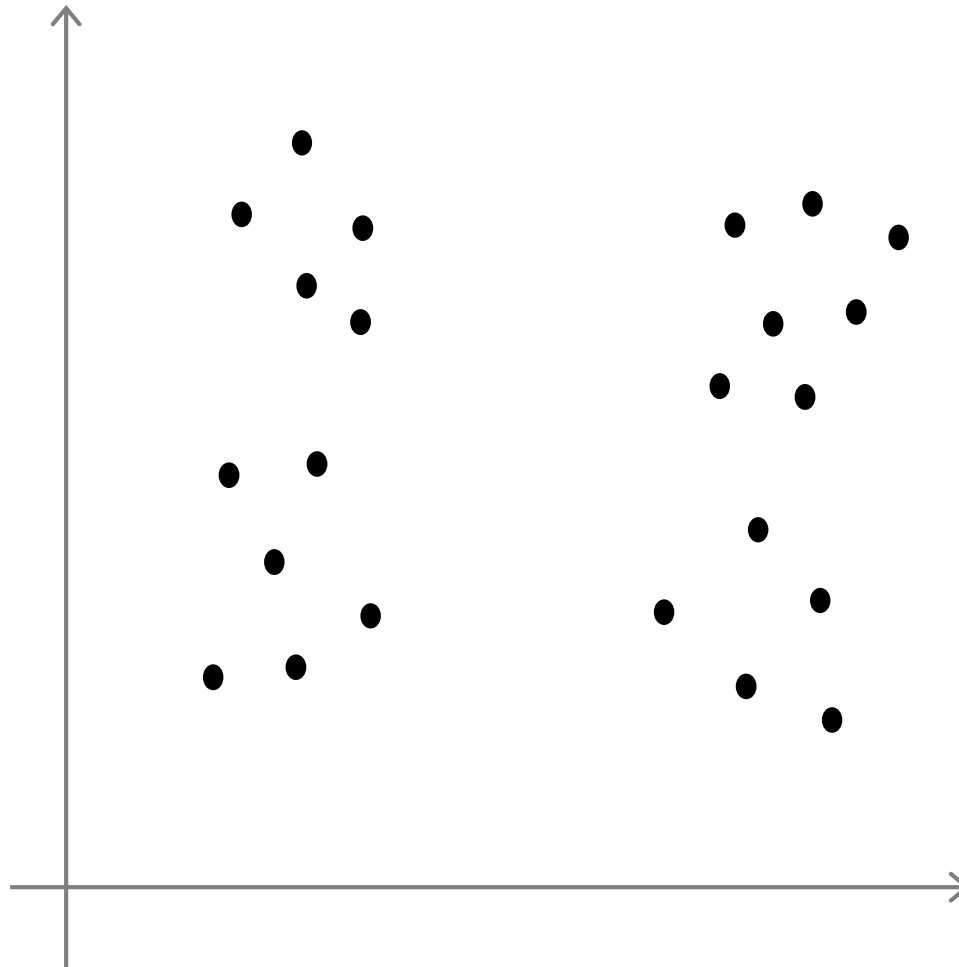






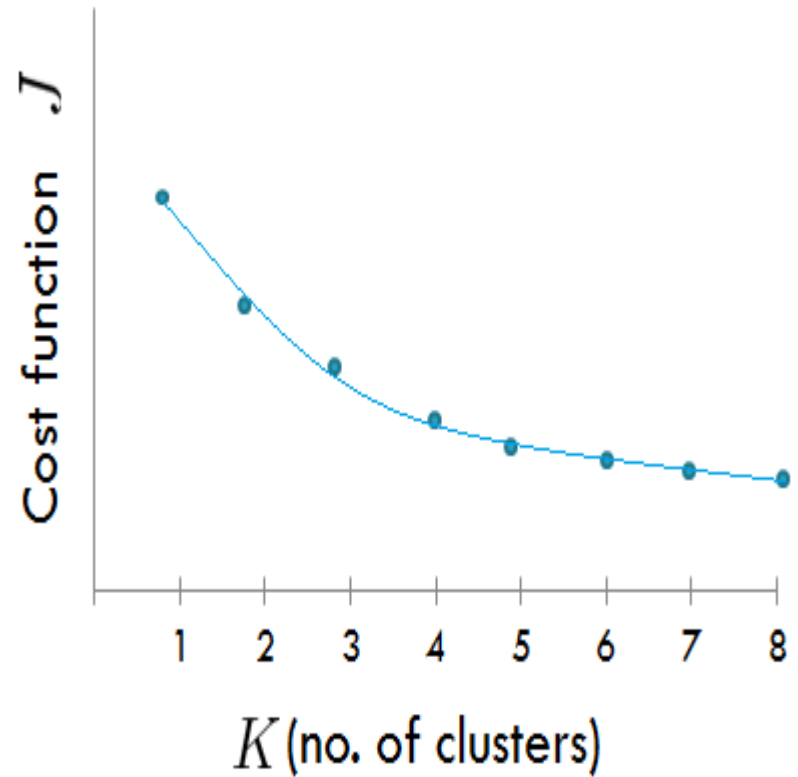
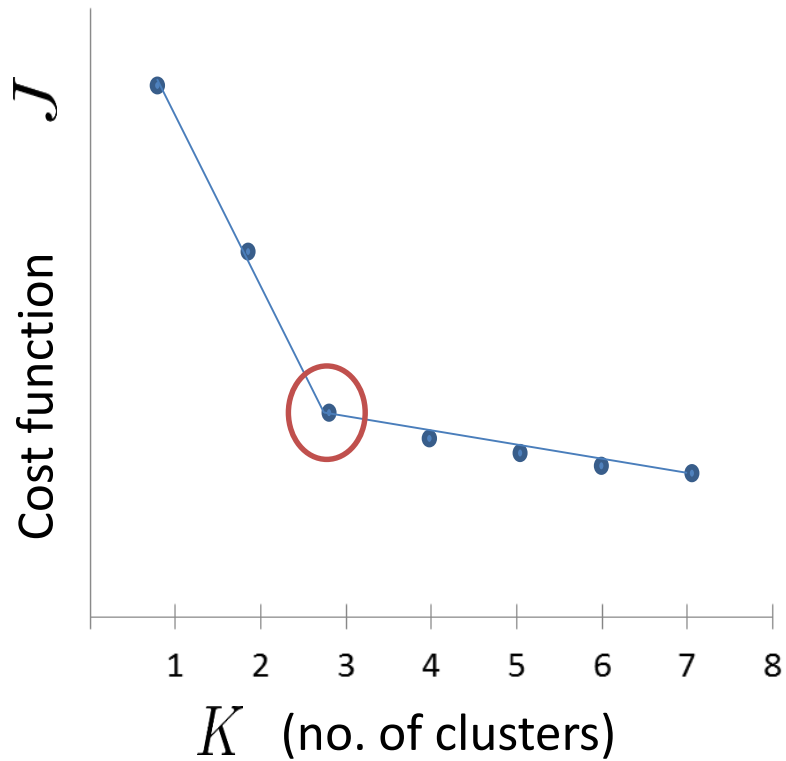


Cual es el correcto valor de K?



Escogiendo el valor de K

Método del codo:



Clustering Incremental

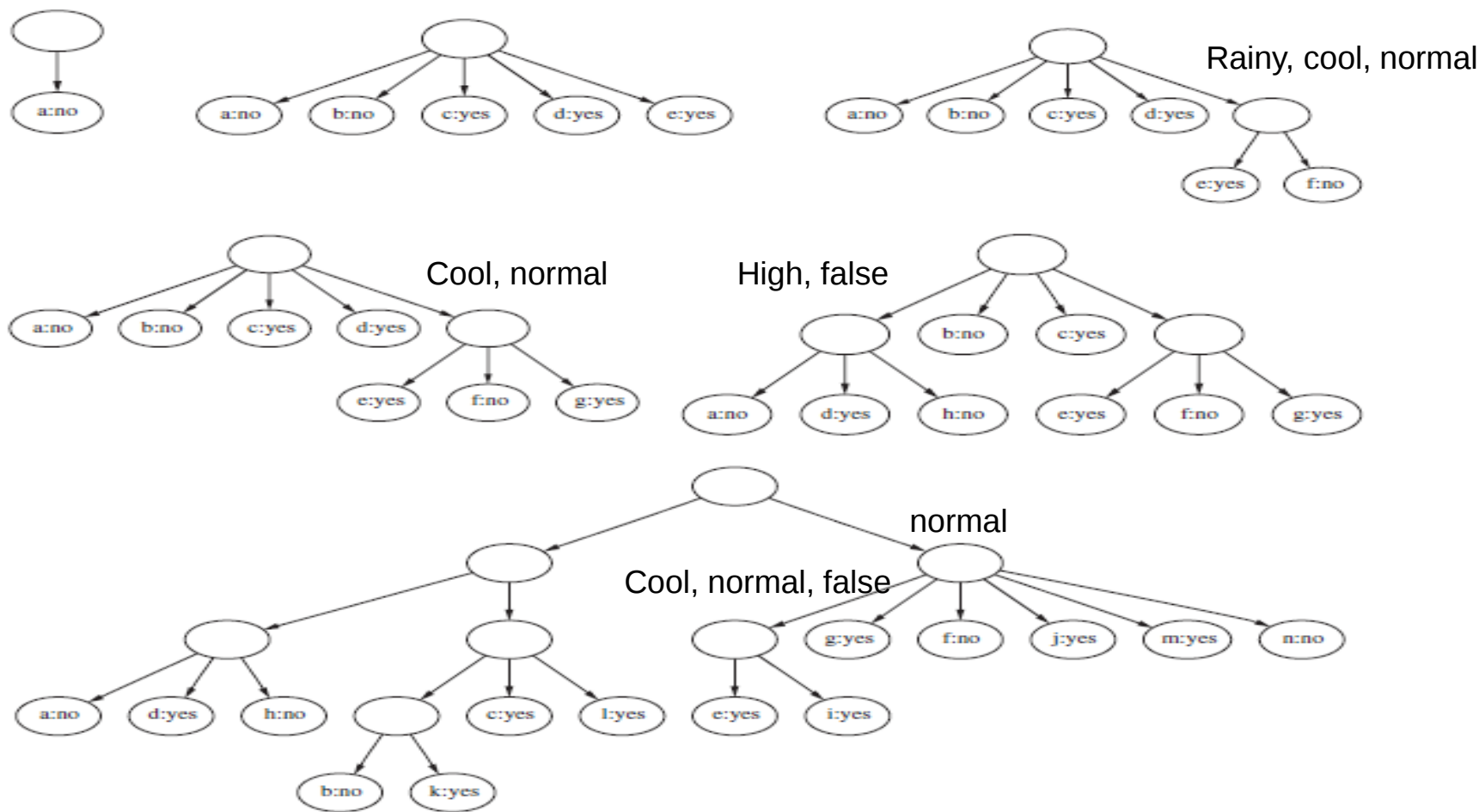
Datos de tiempo

ID code	Outlook	Temperature	Humidity	Windy	Play
a	sunny	hot	high	false	no
b	sunny	hot	high	true	no
c	overcast	hot	high	false	yes
d	rainy	mild	high	false	yes
e	rainy	cool	normal	false	yes
f	rainy	cool	normal	true	no
g	overcast	cool	normal	true	yes
h	sunny	mild	high	false	no
i	sunny	cool	normal	false	yes
j	rainy	mild	normal	false	yes
k	sunny	mild	normal	true	yes
l	overcast	mild	high	true	yes
m	overcast	hot	normal	false	yes
n	rainy	mild	high	true	no

Clustering Incremental

Table 4.6 The weather data with identification codes.

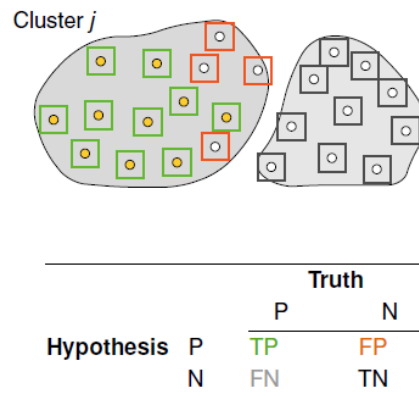
ID code	Outlook	Temperature	Humidity	Windy	Play
a	sunny	hot	high	false	no
b	sunny	hot	high	true	no
c	overcast	hot	high	false	yes
d	rainy	mild	high	false	yes
e	rainy	cool	normal	false	yes
f	rainy	cool	normal	true	no
g	overcast	cool	normal	true	yes
h	sunny	mild	high	false	no
i	sunny	cool	normal	false	yes
j	rainy	mild	normal	false	yes
k	sunny	mild	normal	true	yes
l	overcast	mild	high	true	yes
m	overcast	hot	normal	false	yes
n	rainy	mild	high	true	no



Métricas para evaluar un algoritmo de agrupamiento

Índice Externo: usado para medir el grado en que las etiquetas de un cluster coinciden con etiquetas de clases externas

- **F-measure:** esta métrica está basada en la precisión y recall,



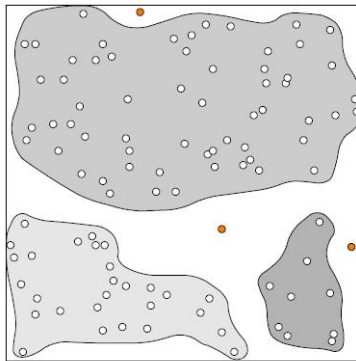
- **Entropía:** Es el grado de coincidencia de los clusters a las clases ya definidas de los datos originales.

$$e_i = - \sum_{j=1}^c p_{ij} \log_2 p_{ij} \quad (2.8)$$

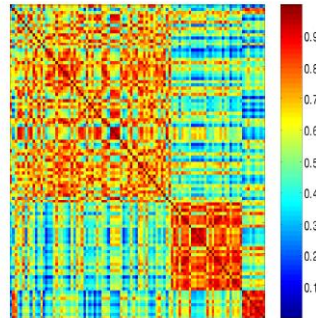
Métricas para evaluar un algoritmo de agrupamiento

Índice Interno: usado para medir la “bondad” de un estructura agrupada sin tener información de referencia externa

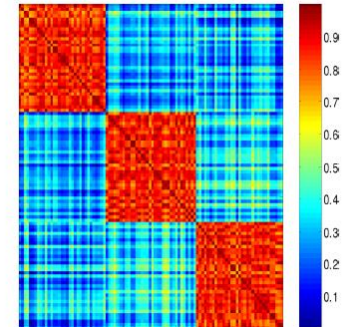
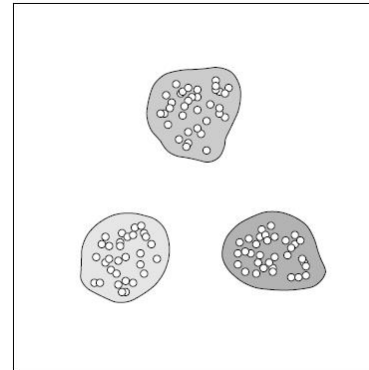
- Coeficiente de correlación:



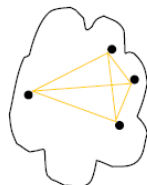
DBSCAN at random data.



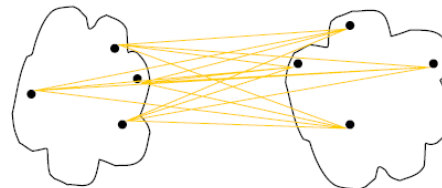
Similarity matrix sorted by cluster label.



- Cohesión y separación

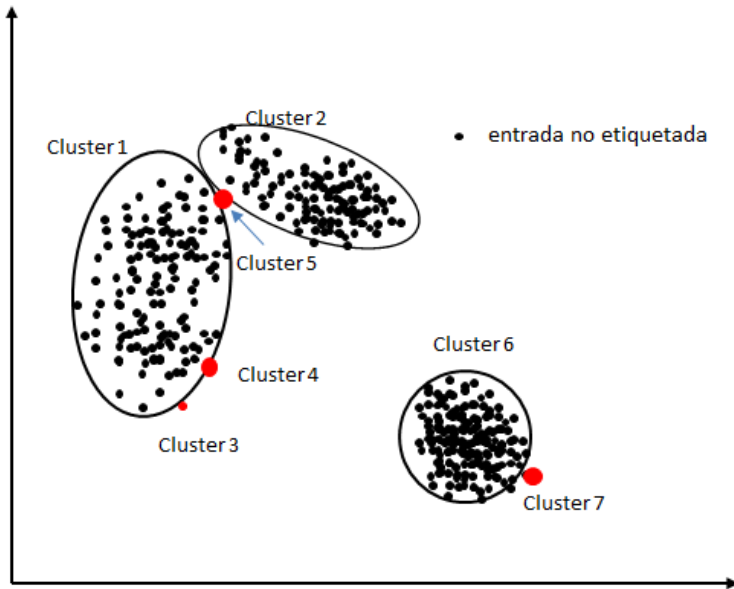


cohesión

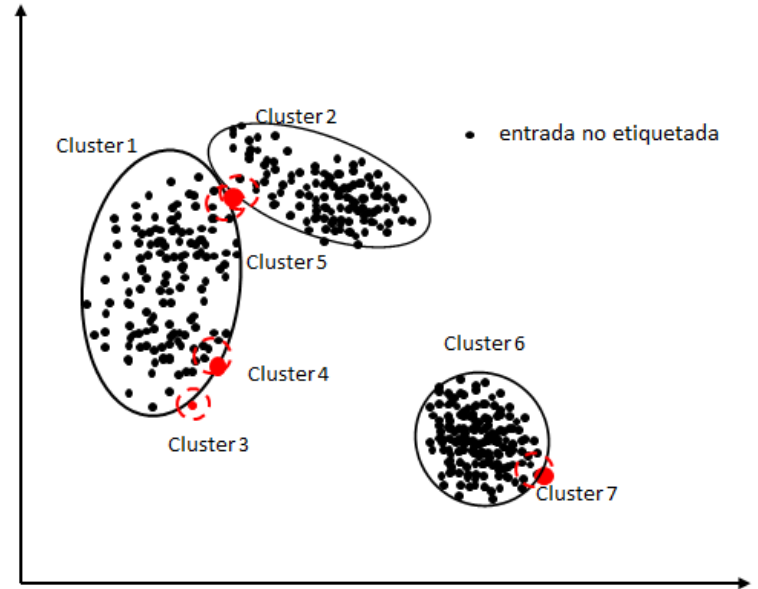


separación

Agrupamiento

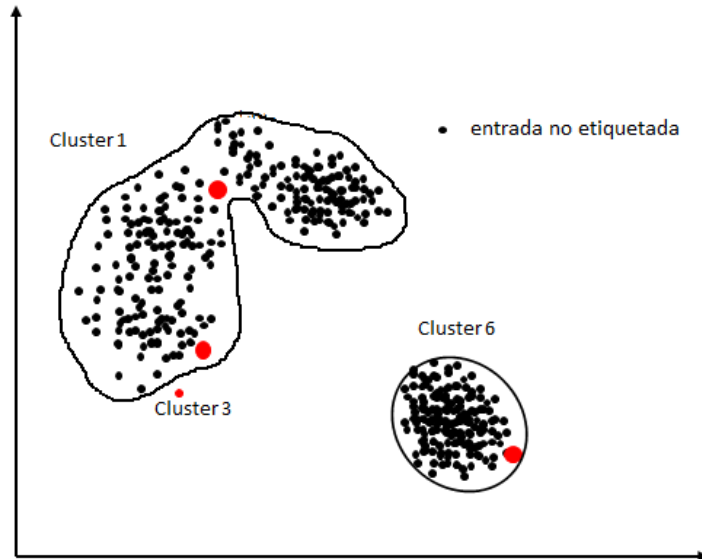


Inicial



Detección de los grupos
candidatos para ser
mezclados

Mezcla



ASOCIACION

ASOCIACION

Es el descubrimiento de relaciones entre las características (atributos) que conforman la base de datos,

Dichas asociaciones es el conocimiento

REGLAS DE ASOCIACION

Técnica no supervisada que permite predecir patrones de comportamientos futuros **basado en las ocurrencias simultaneas** de valores de variables.

Una asociación entre dos atributos ocurre cuando la **frecuencia con la que se dan dos o más valores determinados de cada uno conjuntamente es relativamente alta.**

Las reglas de asociación intentan descubrir asociaciones o conexiones entre objetos.

Consecuencia $\Leftarrow$ Antecedente₁ Antecedente₂ ... Antecedente_m.

Ejemplo, en un supermercado se analiza si los pañales y las compotas se compran conjuntamente.

REGLAS DE ASOCIACION: ejemplo

Gestión Estantes del supermercado.

- **Objetivo:** Identificar los elementos que compran juntos muchos clientes.
- **Enfoque:** encontrar dependencias entre elementos.
- **Un ejemplo de regla:**
 - Si un cliente compra pañales y leche, entonces es muy probable que compre compotas.

Reglas de Asociación

Reglas que implican relaciones

Sombreado: parado (standing)
No sombreado: acostado (lying)

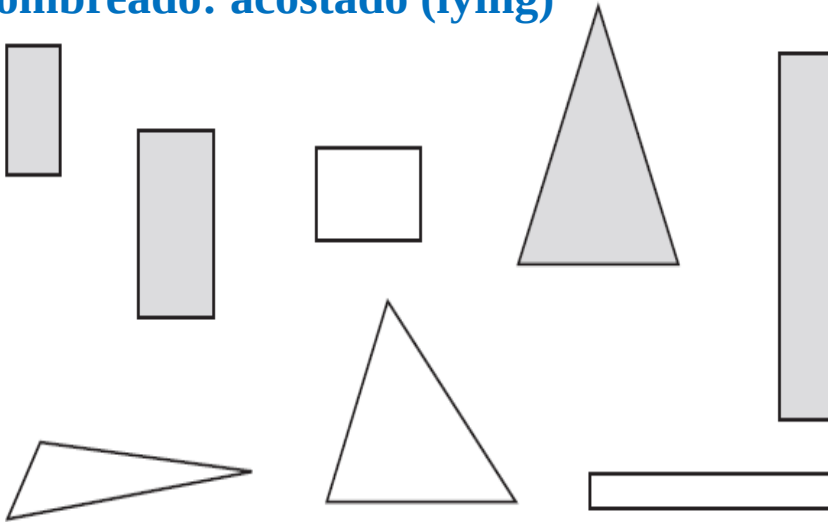


Tabla con datos de entrenamiento

Width	Height	Sides	Class
2	4	4	standing
3	6	4	standing
4	3	4	lying
7	8	3	standing
7	6	3	lying
2	9	4	standing
9	1	4	lying
10	2	3	lying

Reglas



if width ≥ 3.5 and height < 7.0 then
lying

if height ≥ 3.5 then standing

Reglas de Asociación

- Pueden predecir cualquier atributo, o combinaciones de atributos.

- La **cobertura** de una regla de asociación es el número de instancias para las cuales ella predice correctamente (**sopORTE**).

Pronostico	Temperatura	Humedad	Viento	Jugar
soleado	caliente	alta	falso	no
soleado	caliente	alta	verdadero	no
nublado	caliente	alta	falso	si
lluvioso	templado	alta	falso	si
lluvioso	fresco	normal	falso	si
lluvioso	fresco	normal	falso	si
nublado	fresco	normal	verdadero	si
soleado	templado	alta	falso	no
soleado	fresco	normal	falso	si
lluvioso	templado	normal	falso	si
soleado	templado	normal	verdadero	si
nublado	templado	alta	verdadero	si
nublado	caliente	normal	falso	si
lluvioso	templado	alta	verdadero	no

- La **precisión (confianza)** es el número de instancias que predice correctamente, expresado como una proporción de todas las instancias a las que se aplica.

Se utilizan para descubrir **hechos que ocurren en común** dentro de un determinado conjunto de datos

Por ejemplo, en la tabla anterior las reglas:

- Si temperatura = fría entonces humedad = normal
- Si viento = falso y jugar = no entonces pronostico = soleado y humedad = alta

Reglas de Asociación

Reglas con diferentes antecedentes y valores de cobertura (>1)

	One-item sets	Two-item sets	Three-item sets	Four-item sets		One-item sets	Two-item sets	Three-item sets	Four-item sets
1	outlook = sunny (5)	outlook = sunny temperature = mild (2)	outlook = sunny temperature = hot humidity = high (2)	outlook = sunny temperature = hot humidity = high play = no (2)	...	...	humidity = normal windy = false (4)	humidity = normal windy = false play = yes (4)	
2	outlook = overcast (4)	outlook = sunny temperature = hot (2)	outlook = sunny temperature = hot play = no (2)	outlook = sunny humidity = high windy = false play = no (2)	38	39	humidity = normal play = yes (6)	humidity = high windy = false play = no (2)	
3	outlook = rainy (5)	outlook = sunny humidity = normal (2)	outlook = sunny humidity = normal play = yes (2)	outlook = overcast temperature = hot windy = false play = yes (2)	40	47	humidity = high windy = true (3)		
4	temperature = cool (4)	outlook = sunny humidity = high (3)	outlook = sunny humidity = high windy = false (2)	outlook = rainy temperature = mild windy = false play = yes (2)			...	windy = false play = no (2)	
5	temperature = mild (6)	outlook = sunny windy = true (2)	outlook = sunny humidity = high play = no (3)	outlook = rainy humidity = normal windy = false play = yes (2)					
	...	...	...	...					

Reglas de Asociación

- Las reglas se obtienen a partir de valores de las variables

humidity = normal, windy = false, play = yes

- Esto nos lleva a las 7 reglas potenciales:

If humidity = normal and windy = false → play = yes 4/4

If humidity = normal and play = yes → windy = false 4/6

If windy = false and play = yes → humidity = normal 4/7

If humidity = normal → windy = false and play = yes 4/6

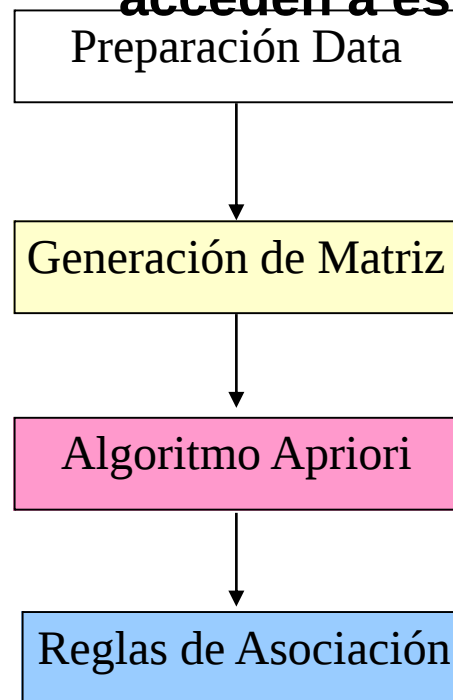
If windy = false → humidity = normal and play = yes 4/8

If play = yes → humidity = normal and windy = false 4/9

If → humidity=normal and windy=false and play=yes 4/12

Método para determinar Reglas de Asociación (Algoritmo Apriori)

Encontrar las asociaciones que se producen entre los diferentes sitios de la página Web cuando los usuarios acceden a ésta.



Reglas de Asociación

Preparación de Data

Registro_Log

id	Id_se...	id_user	ip	Solicitud	fecha	bytes
8	2	11	200.110.86.82	/loginError.jsp	2006-02-26 00:03:00	3641
13	2	11	200.110.86.82	/private/mycourses/	2006-02-26 00:04:00	4785
16	2	11	200.110.86.82	/private/mycourses/website/...	2006-02-26 00:05:00	5717
19	2	11	200.110.86.82	/private/download/1048/3676...	2006-02-26 00:09:00	50688
24	2	11	200.110.86.82	/private/mycourses/website/...	2006-02-26 00:10:00	0
25	2	11	200.110.86.82	/private/mycourses/website/...	2006-02-26 00:10:00	4100
44	2	11	200.110.86.82	/private/mycourses/website/...	2006-02-26 00:19:00	5717
53	2	11	200.110.86.82	/js/tiny_mce/plugins/previe...	2006-02-26 00:21:00	0
110	4	11	200.110.86.82	/js/util.js	2006-02-26 01:03:00	0
176	4	11	200.110.86.82	/private/mycourses/website/...	2006-02-26 01:08:00	8778

Registro_Paginas

Registro_Sesion

id_sesion	id_user	ip	hora_inicio	hora_fin	num_pag..
3	31	201.2...	2006-02-26 00:54:00	2006-02-26 01:24:00	10
14	30	201.2...	2006-02-26 11:20:00	2006-02-26 11:27:00	9
30	23	200.6...	2006-02-26 16:41:00	2006-02-26 16:43:00	2
38	17	200.2...	2006-02-26 18:46:00	2006-02-26 18:46:00	0
1	6	200.1...	2006-02-26 00:01:00	2006-02-26 00:02:00	1
2	11	200.1...	2006-02-26 00:01:00	2006-02-26 00:29:00	42
7	11	200.1...	2006-02-26 01:36:00	2006-02-26 01:44:00	14
10	3	200.1...	2006-02-26 10:17:00	2006-02-26 10:23:00	3
11	32	201.2...	2006-02-26 10:33:00	2006-02-26 10:33:00	2
13	1	200.1...	2006-02-26 11:14:00	2006-02-26 11:15:00	4

id_pagina	url
2	/index.jsp
7	/private/mybriefc...
16	/private/mycourse...
20	/private/mycourse...
22	/private/mycourse...
26	/private/mycourse...
30	/private/mycourse...
32	/private/myprofil...
35	/public/findUsers...
36	/public/portalDoc...

Reglas de Asociación

Generación Matriz

Sesión / Página	1	2	3	4	5		# sesiones
1	0	1	0	1	0		2
2	1	0	1	1	0		3
3	1	1	0	1	0		3
4	0	1	1	1	0		3
5	1	0	0	0	0		1
6	0	1	0	0	1		2
:	:	:	:	:	:		0
:	:	:	:	:	:		0
# páginas	3	3	2	4	1		


$$S1 = (0 + 1 + 1 + 0 + 1 + 0 + \dots + 0) / \# \text{ páginas}$$

Reglas de Asociación

$$X \rightarrow Y$$

[/public/about.jsp]---->/public/team.jsp

Métricas

Soporte:

Soporte ($X \rightarrow Y$) = Probabilidad ($X \cup Y$)

Confianza:

Confianza ($X \rightarrow Y$) = Probabilidad (X / Y)

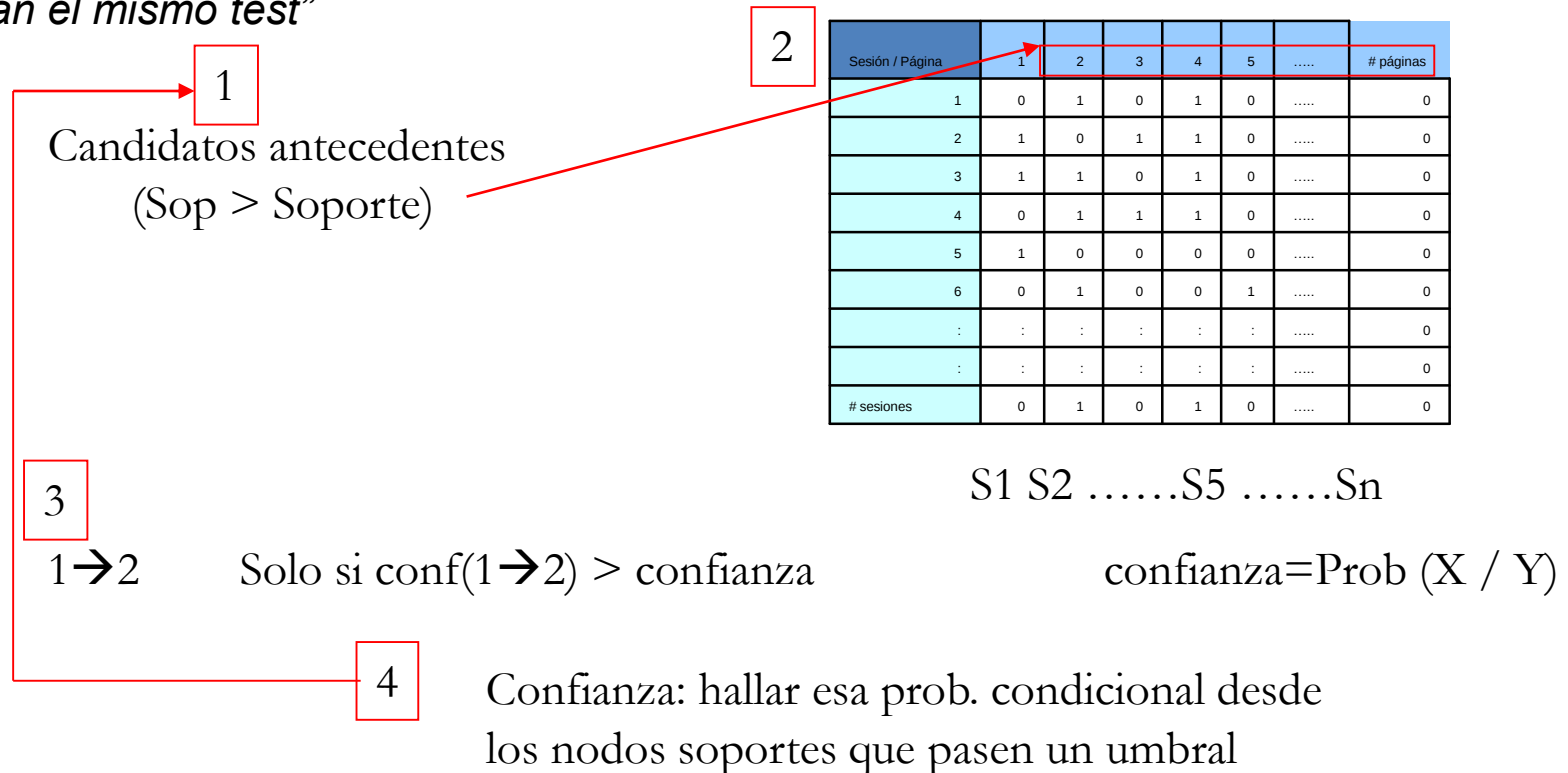
prob. condicional

Reglas de Asociación

Algoritmo Apriori (matriz , soporte, confianza)

Usa *conocimiento a priori* de las propiedades de los ítems (páginas) frecuentes que ya se han encontrado.

Premisa: “Si un conjunto no pasa un test, todos sus súper conjuntos tampoco pasarán el mismo test”



Reglas de Asociación



MineroWeb

Ingreso | **Procesamiento** | **Salidas**

[Estadísticas de Uso](#) | [K-Medias](#) | [Patrones Secuenciales](#) | [Reglas de Asociación](#)

Reglas Generadas: 14

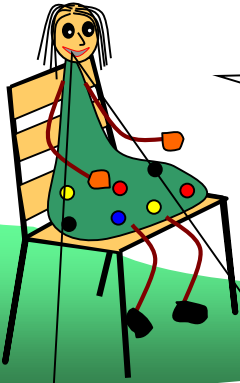
Para mejor entendimiento de la regla, siga el esquema:
*El (Confianza)% de usuarios que visitaron
(antecedente), visitarán (consecuente)*

Soporte: % Confianza: %

N°	Regla	Soporte	Confianza
1	[/public/about.jsp]---->/public/team.jsp	15,21%	57,14%
2	[/public/findUsers.jsp]---->/public/portalDocument.jsp	13,04%	83,33%
3	[/public/findUsers.jsp]---->/public/team.jsp	13,04%	83,33%
4	[/public/portalDocument.jsp]---->/index.jsp	15,21%	57,14%
5	[/public/portalDocument.jsp]---->/public/team.jsp	15,21%	71,42%
6	[/index.jsp]---->/public/team.jsp	17,39%	62,5%
7	[/loginError.jsp]---->/private/mycourses/index.jsp	10,86%	80%

Predicción

Pregunta



Se dispone de dos libros con muchas hojas, por ejemplo, dos guías de teléfono.

Se ponen frente a frente los dos libros y luego, con mucha paciencia, se intercalan las hojas de ellos.

Una vez que se intercalaron todas las hojas, se intenta separarlos tirándolas desde sus respectivos lomos.

¿Se pueden separar con facilidad?



plantee una predicción al respecto.

Aplicando fuerzas en los lomos respectivos de los libros,

¿será fácil o difícil separar los libros una vez que tienen mezcladas las hojas?

Modelos de Predicción

Piensa en una variable que quieras predecir. ***Que necesitas?***

- **Objeto a predecir:** Una serie temporal, un suceso, ...etc.
- **Formato de la Predicción:** Puntual, Intervalo, Densidad, ...etc.
- **Horizonte de la predicción:** Corto, Medio o Largo Plazo
- **Conjunto de Información:** Univariante o Multivariante
- **Metodos y Complejidad:** Modelos, ...etc.

Predicción

- Predice un valor de una variable dada, sobre la base de los valores de otras variables, suponiendo un modelo lineal o no lineal de dependencia.
- **Ejemplos:**
 - Predecir las ventas de nuevos productos basados en gastos de publicidad.
 - Predecir la velocidad del viento como una función de la temperatura, humedad, presión de aire, etc.
 - Predecir series de tiempo de los índices bursátiles.

Predicción

- Predice un valor de una variable dada, sobre la base de los valores de otras variables, suponiendo un modelo lineal o no lineal de dependencia.
- Ejemplos:
 - Predecir las ventas de nuevos productos basados en gastos de publicidad.
 - Predecir la velocidad del viento como una función de la temperatura, humedad, presión de aire, etc.
 - Predecir comportamiento en el tiempo de los índices bursátiles (series de tiempo).

Modelos de Predicción

Las predicciones ayudan a la toma de decisiones en una gran variedad de areas.

- **Planificación y Control de Operaciones:** Las empresas usan predicciones para decidir que producir, cuando y donde.
- **Mercadeo:** Decisiones de precios, de gastos en publicidad, ...dependen fuertemente de las previsiones que se tengan sobre como van a responder las ventas a los diferentes esquemas de marketing.
- **Economía:** Predicciones de las variables macro-económicas claves como el PNB, Paro, Consumo, Inversión, Tipos de Interés, etc... son usadas por el gobierno para fijar su política monetaria y fiscal.
- **Financiera:** actores de los mercados financieros tienen un gran interés en la predicción de los rendimientos de activos financieros (acciones, tipos de interés, tipos de cambio, etc...).
- **Demografía:** La predicción de la población es crucial para planificar el gasto publico en sanidad, infraestructuras, educación, etc.

Modelos de Predicción

Hay muchas formas de hacer predicciones; pero todas ellas tienen en común los siguientes ingredientes:

- 1. que hay ciertas regularidades que captar*
- 2. que tales regularidades son informativas sobre el futuro*
- 3. están encapsuladas en el método seleccionado para predecir*
- 4. normalmente se excluyen las no-regularidades*

Los principales métodos son:

- **Adivinación**
- **Extrapolación**
- **Encuestas**
- **Modelos de Series Temporales**

Evaluación de Predicciones

Al menos hay tres fuentes de error

- **Incertidumbre en la Especificación:** Todos los modelos están equivocados!!!! (algunos mas que otros)
- **Incertidumbre en la Innovación:** Innovaciones futuras son desconocidas cuando se hace la predicción.
- **Incertidumbre en los Parámetros:** Los coeficientes que usamos para producir las predicciones son *estimaciones*, y por lo tanto están sujetas a variabilidad muestral.

Diferente medidas de errores de predicción

- **Error de Especificación**
- **Error de Aproximación**
- **Error de Estimación**

Evaluación de Predicciones

Las medidas mas comunes de la precisión de la predicción son:

Error cuadrático medio:

$$\text{MSE} = \frac{1}{T} \sum_{t=1}^T e_{t+h,t}^2$$

Raíz cuadrada del MSE

$$\text{RMSE} = \sqrt{\frac{1}{T} \sum_{t=1}^T e_{t+h,t}^2}$$

Error absoluto medio

$$\text{MAE} = \frac{1}{T} \sum_{t=1}^T |e_{t+h,t}|$$

donde $e_{t+h,t} = y_{t+h} - \hat{y}_{t+h,t}$ son los errores de predicción.

Modelos de Predicción

El modelo de regresión es un modelo explícitamente multi-variable, en donde la variable a explicar **se explica y se predice** en base a su **propia historia pasada y la historia pasada de otras variables relacionadas**.

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

$$\varepsilon_t \text{ es } WN(0, \sigma^2)$$

Predicción

Evaluación de la capacidad de predecir

- Dividir la muestra en dos partes; una para estimación del modelo y una para evaluar la capacidad de predecir.
- Estimar el modelo
- Calcular la predicción para los periodos no usadas.
- Comparar la predicción con valores reales (error del pronóstico)

Modelo predictivo del éxito de una tutoría en línea para un año y estudiante dado

ATTRIBUTE	DESCRIPTION
ECV_ID	ID Estudiante
COE_ID	Identificador de componente educativo
ETR_CODIGO	Aprobado, Reprob, otro
GCR_CODIGO	Créditos
DISCAPACIDAD	
ESTADO_CIVIL	
GENERO	
CENTRO MATRICULA	

Precision 90% y Recall 93%.

MATRIZ DE CONFUSIÓN

a	b	c	<-- classified as
9915	19	634	approved
130	8758	130	reprobate
918	0	2566	other

Predicción de deserción estudiantil

- No es posible crear un modelo predictivo universal,
- Se deben crear múltiples modelos aplicables a diferentes contextos: Carreras específicas, ubicaciones geográficas, etc.

N.	Condiciones	Instancias	Precisión	<u>Recall</u>
1	2;14,12,2,1,52 4;5,2 6;1,2	2529	0.6965	0.8947
2	4;1,2,4 6;2	1637	0.7108	0.8439
3	4;5,2,4 6;1,2	4490	0.6971	0.8422
4	4;5,1 6;1	1869	0.6708	0.8278

Redes bayesianas

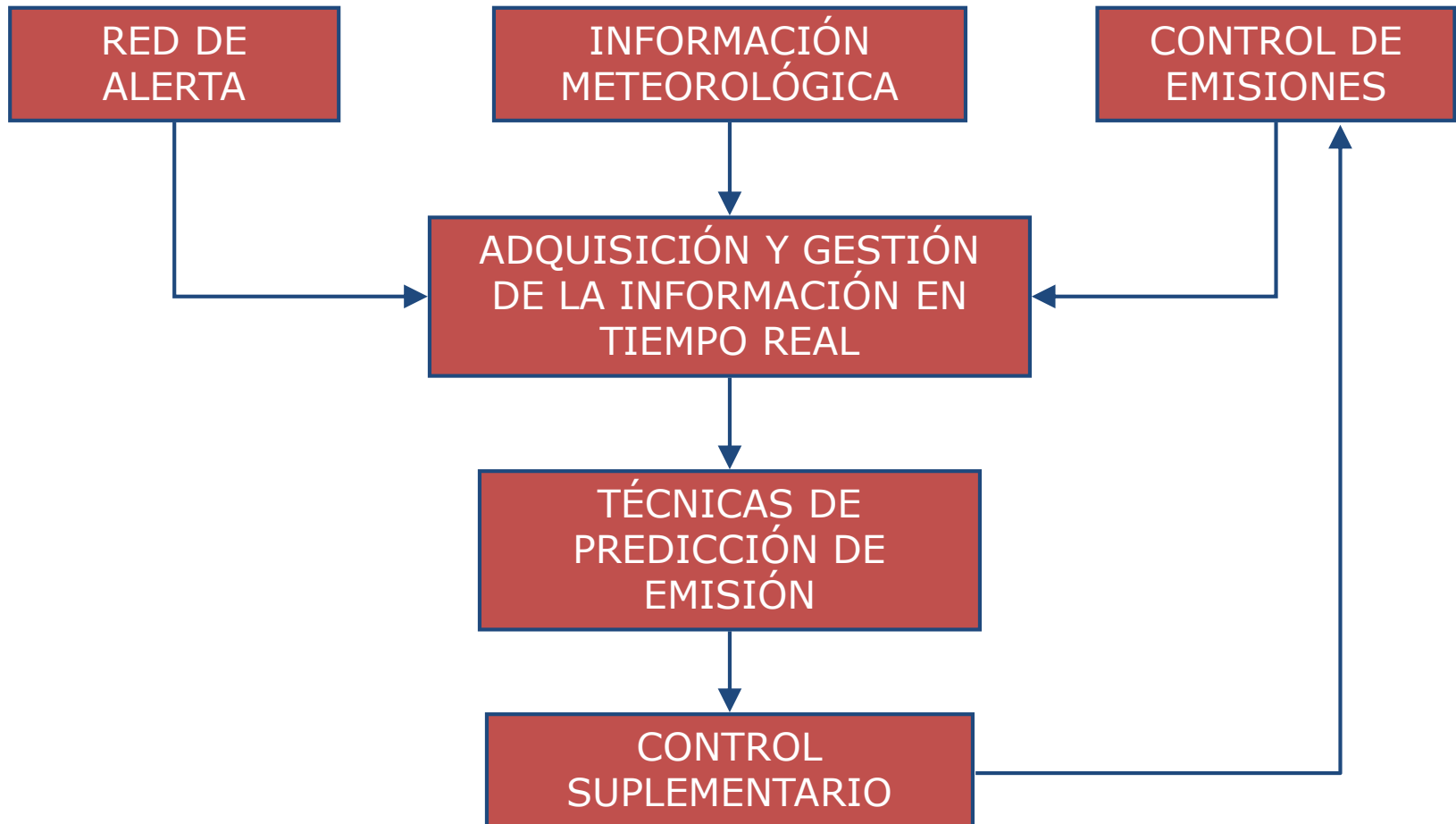
Modelos aplicables específicamente para los datos que cumplen las condiciones (filtros) codificadas
Primer número es el atributo filtro y el segundo valores que toma dicho atributo
Por ejemplo: << 6;2 >> indica atributo “Genero” (6) y valor sexo “Masculino” (2).

**Sistemas de control
suplementario de la
contaminación atmosférica:
predicción con modelos
estadísticos**

Sistema de predicción estadística de inmisión

- Objetivo: predecir con media hora de anticipación la evolución de los niveles de dióxido de azufre en un entorno y sugerir una línea de actuación
- Utilización de modelos estadísticos a partir de la información en tiempo real de emisiones, calidad de aire y meteorología
- Se recogen datos de calidad de aire, en particular de SO_2 , de las estaciones de la Red de Vigilancia de Calidad de Aire (*frecuencia pentaminutal*).

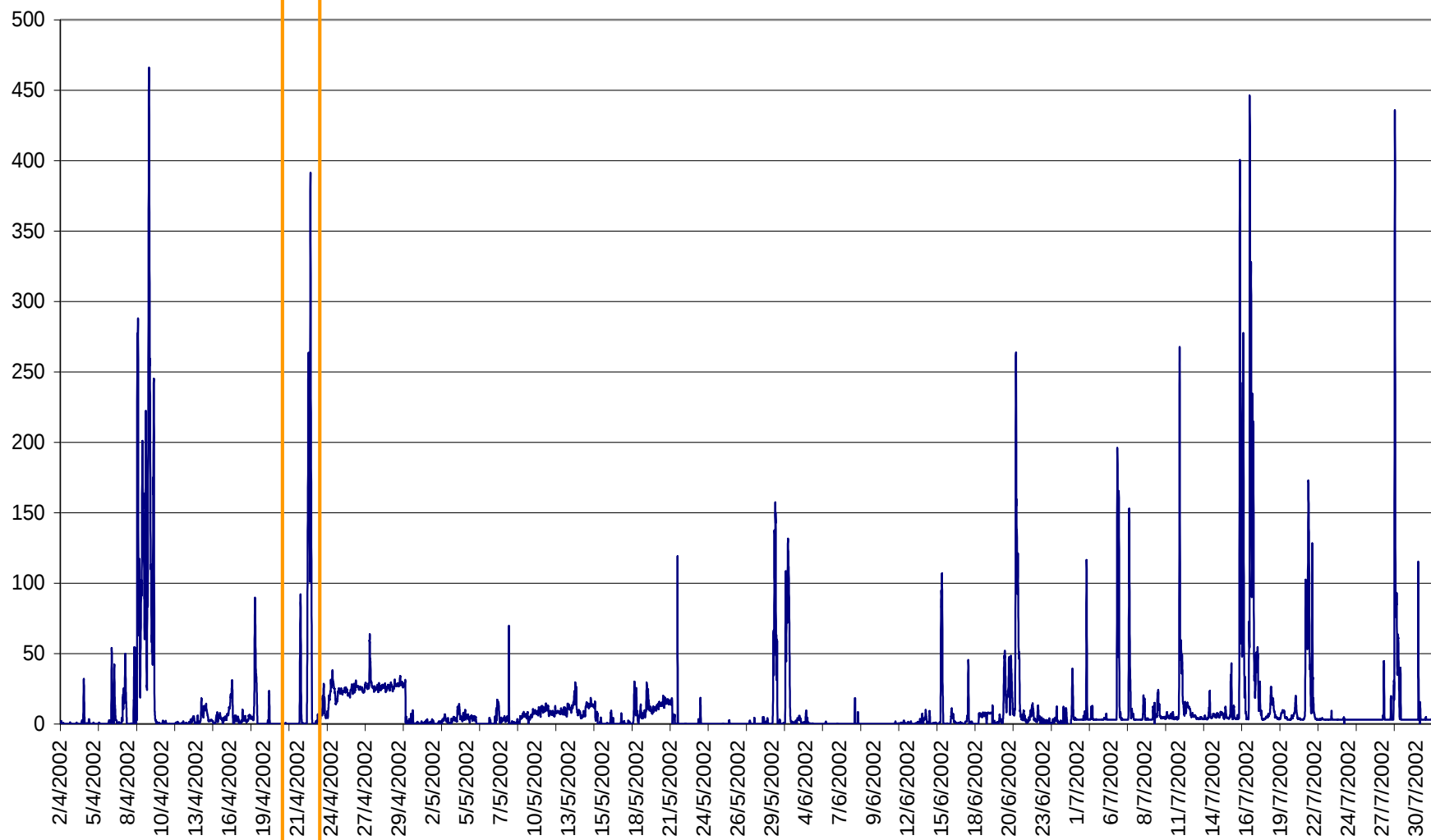
Sistemas de control suplementario de la contaminación atmosférica



Datos

F4. Valores SO₂. Abril 2002 - Julio 2002

SO₂ (µg/m³)



Principal objetivo:

prevenir episodios de alteración de la calidad del aire

- La instalación necesita disponer de información, al menos, con ***1/2 hora*** de antelación.

Solución:

Herramientas de predicción de valores de SO₂, en media horaria, basados en modelos estadísticos

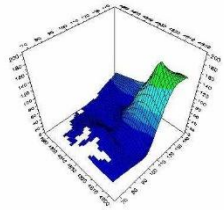
Las ***diferentes metodologías de predicción*** utilizadas aportan soluciones desde varios puntos de vista

I. Predicción puntual: predicción del nivel de SO_2

- Semi-paramétrico
- Redes Neuronales

I. Predicción espacial: construir una superficie de predicción de niveles de SO_2 para el entorno

12/FEB/04 16: 5 Superficie Media Horaria real



II. Predicción funcional: curva completa de niveles de SO_2 para un cierto intervalo de tiempo

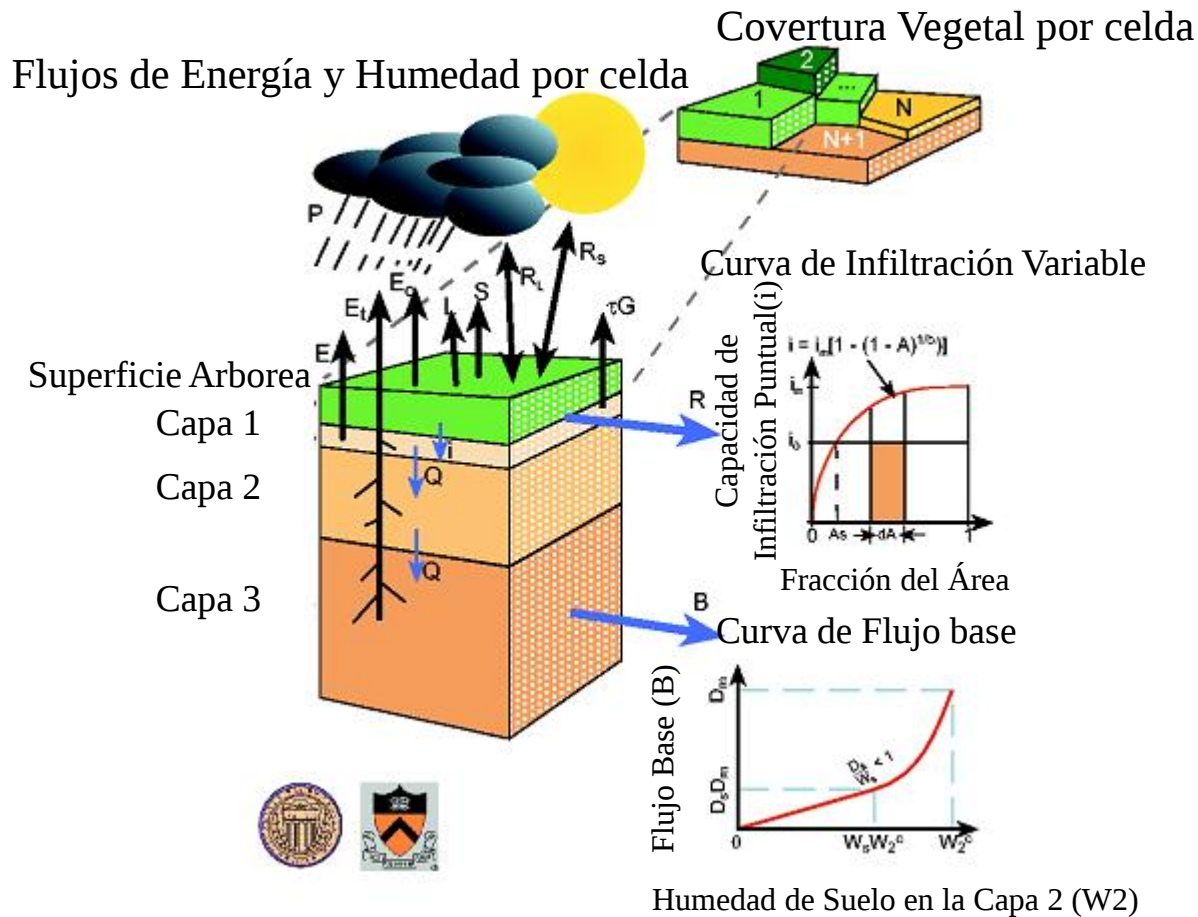
Monitoreo y Predicción de Sequías : Aplicaciones del
Sistema de Predicción Hidrológica Estacional *West-
wide* de la Universidad de Washington

Sistema de Predicción Hidrológica

West-wide de la UW

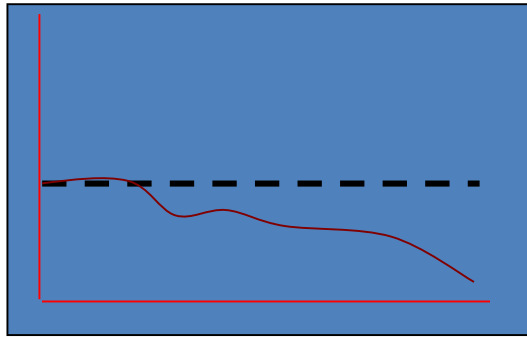
- Modelo Hidrológico Superficial de Capacidades de Infiltración Variable
- Condiciones cuasi actuales
- Índices de Sequía (SMI, SRI, y percentiles)

Modelo Superficial Terrestre de Capacidades de Infiltración Variable(VIC)



Monitoreo de Sequía

Falta de Precipitación



- Humedad de Suelo Precedente
- Condición Hidrológica

INDICE PALMER DE SEQUÍA (PDI)

INDICE ESTANDARIZADO DE PRECIPITACIÓN (SPI)

INDICE DE ABASTECIMIENTO DE AGUA SUPERFICIAL (SWSI)

INDICES BASADOS EN MODELAJE HIDROLÓGICO

✓ Precipitación

✓ Precipitación

✓ Precipitación

✓ Temperatura

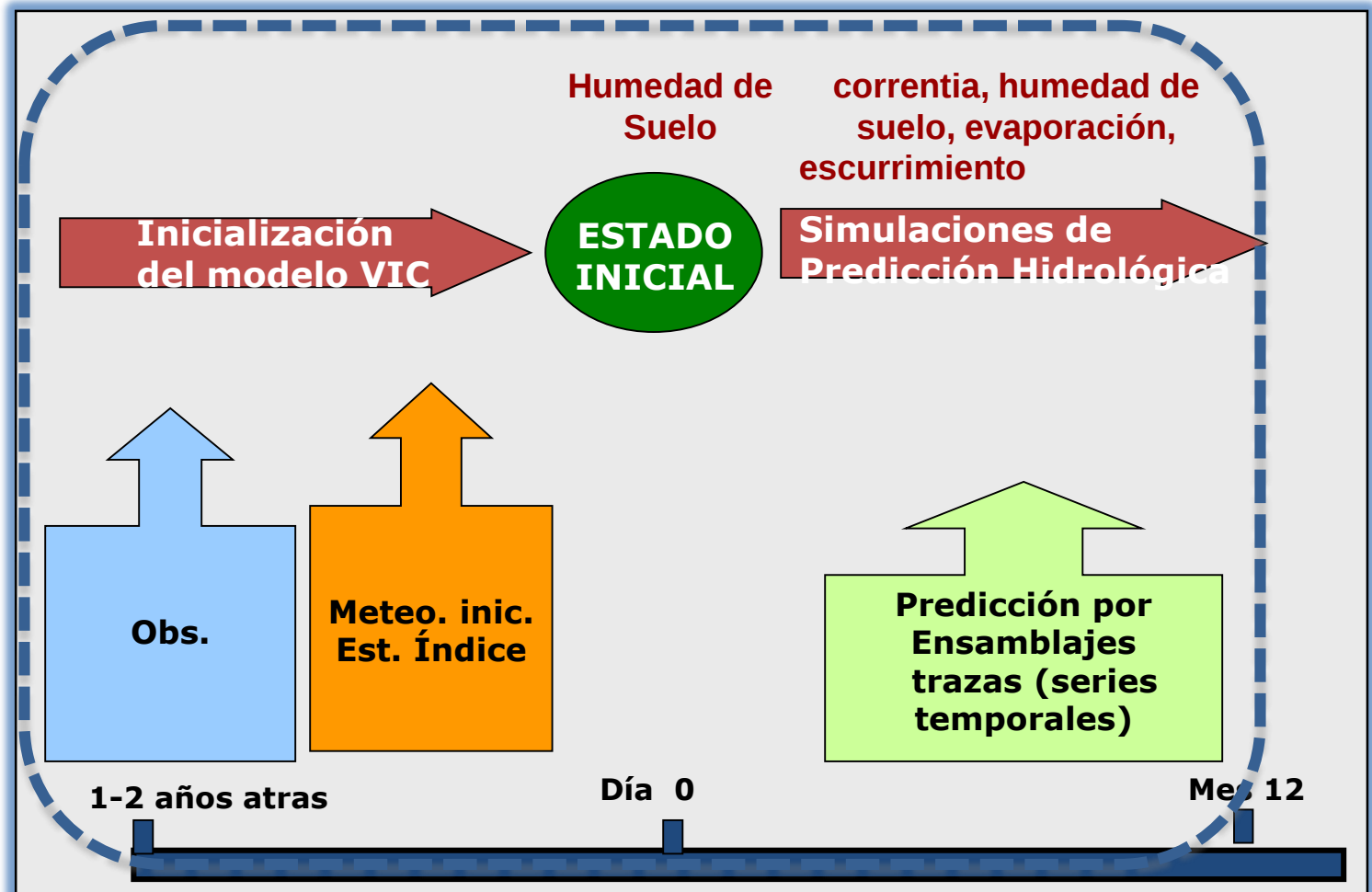
✓ Procesos de estacionales fríos

✓ Esquema complejo de balance

✓ de Energía y agua

OS

Sistema de Predicción Hidrológica Extendido de la UW



PRONÓSTICO DEL TIEMPO

PRONÓSTICO DEL TIEMPO



Pronóstico para 10 días ?				
Madrid, España				
Última actualización martes 16 de octubre de 2007, a las 7:31 Hora de Verano de Europa Central (martes, 5:31 GMT)				
			Máx (C)	Mín (C)
Hoy 16 oct		Parcialmente nuboso	20°C	9°C
		Índice UV: 4 Bajo		
mié 17 oct		Parcialmente nuboso	21°C	9°C
		Índice UV: 5 Moderado		
jue 18 oct		Soleado	22°C	8°C
		Índice UV: 4 Bajo		
vie 19 oct		Soleado	22°C	8°C
		Índice UV: 4 Bajo		
sáb 20 oct		Soleado	22°C	7°C
		Índice UV: 4 Bajo		
dom 21 oct		Soleado	22°C	7°C
		Índice UV: 4 Bajo		
lun 22 oct		Soleado	21°C	7°C
		Índice UV: 4 Bajo		
mar 23 oct		Soleado	19°C	6°C
		Índice UV: 4 Bajo		
mié 24 oct		Soleado	18°C	6°C
		Índice UV: 4 Bajo		
jue 25 oct		Parcialmente nuboso	17°C	7°C
		Índice UV: 4 Bajo		

METODOS DE PRONOSTICO:

El Método de la persistencia (Hoy es igual a mañana)

- Este método asume que las condiciones atmosféricas no cambiarán en el tiempo.
- Este método trabaja bien cuando los patrones atmosféricos cambian poco

El Método de la tendencia (Usando matemáticas)

- Este método involucra el cálculo de la velocidad de centros de altas y bajas presiones, frentes y áreas de nubes y precipitación
- Este método es bueno para predecir dentro de un lapso de tiempo corto

El Método climatológico

- Este método involucra el uso de promedios estadísticos de las variables atmosféricas, acumulados de muchos años.
- El método climatológico trabajará bien mientras que los patrones climatológicos sean similares para la fecha escogida,

METODOS DE PRONOSTICO:

El Método análogo

- Supone examinar el escenario del pronóstico actual y recordar un día en el pasado en el cual el escenario meteorológico fue muy similar (un análogo).
- El pronosticador podría predecir que el tiempo en este pronóstico será muy similar al ocurrido en el pasado.

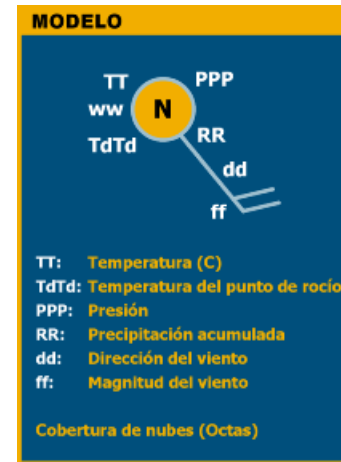
Predicción Numérica

- Usa complejos programas de cómputo, conocidos como modelos numéricos, que procesan datos en supercomputadoras y proporcionan predicciones de las variables meteorológicas: temperatura, presión atmosférica, viento, humedad y precipitación.
- Un modelo numérico es un conjunto de ecuaciones matemáticas cuya solución requiere de métodos numéricos.
- Las ecuaciones básicas son aquellas que rigen el movimiento del aire (horizontal y vertical), conservación de la masa y la energía, los efectos termodinámicas, los procesos de desarrollo de las nubes, etc.
- Los métodos numéricos más comunes usados para resolver el sistema de ecuaciones diferenciales en derivadas parciales (modelo numérico del tiempo) son: métodos espectrales y elementos finitos.

METODOS DE PRONOSTICO:

Predicción Numérica

- El Modelo ETA-SENAMHI
- El Modelo RAMS
- El Modelo Climático CCM3.



ECUACIONES QUE GOBIERNAN LOS MODELOS NUMÉRICOS:

El Movimiento horizontal

La Ecuación hidrostática

La Ecuación Termodinámica

La Ecuación de Continuidad

La Ecuación del Estado

La Ecuación de Vapor de H₂O

Un sistema moderno diario de pronóstico del tiempo consiste en cinco componentes:

- Recopilación de datos
- Preparación de datos
- Predicción numérica del tiempo
- Postprocesamiento de modelos
- Presentación del pronóstico al usuario final

Aplicación de la Predicción

Las predicciones fundamentales en muchas áreas!!!

- **Planificación estratégica**
- **Mundo Financiera**
- **Demografía**
- **Economía**

PRECIOS: TIPO DE CAMBIO $t-1$ + PRECIOS $t-1$ + PRECIOS $t-2$ + ERRORES

PRECIOS: TIPO DE CAMBIO $t-1$ + TIPO DE CAMBIO $t-2$ + PRECIOS $t-1$ + PRECIOS $t-2$ + E

TIPO DE CAMBIO: TIPO DE CAMBIO $t-1$ + TIPO DE CAMBIO $t-2$ + PRECIOS $t-1$ + PRECIOS $t-2$ + E

Resumen

- Para realizar una predicción es recomendable buscar y analizar la información disponible.
- Para realizar una predicción es necesario argumentarla, independientemente de que dichos argumentos sean correctos o erróneos.
- Es necesario poner a prueba la validez de la predicción.
- Hay que idear un procedimiento experimental, o teórico, para validar la predicción.
- Hay que estar dispuesto a modificar la predicción cuando la evidencia experimental arroja un resultado diferente al propuesto.