UNIVERSIDAD
DE LOS ANDES
MERIDA VENEZUELA

Tipos de tareas de Analítica de Datos

¿Qué es la AD?

MODELOS de conocimiento!!!

Modelos de Analítica

	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

Modelos de Analítica

	Optimización	Identificación	Diagnóstico
Preguntas	**Qué puedo mejorar? Cómo mejorarlo?**	**Cómo es el modelo? Qué caracteriza a esos modelos?**	**Por qué sucede? Cuáles son las causas?**
Habilitadores	• Reportes • Modelos de mejora • Simulación	• Simulación • Formulas matemáticas	• Optimización • Simulación • Modelos de Decisión
Resultados	Mejores en la organización	Caracterización	Mejores posibles decisiones y transacciones

La gestión usando AD

El éxito de la analítica sólo puede medirse en términos de lo bien que ayudan a lograr objetivos estratégicos

Por lo tanto, se debe:

- Identificar los objetivos de negocio
- Recoger los datos necesarios para medir sus objetivos
- Analizar los datos
- Sacar conclusiones sobre la base de la información generada

Data-driven decision-making

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!!!

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
  - Pronóstico incluye la duración esperada, la función y la descripción del curso de la enfermedad, como el declive progresivo, la crisis intermitente o una crisis repentina e impredecible.

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 malignano)

Clasificación

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

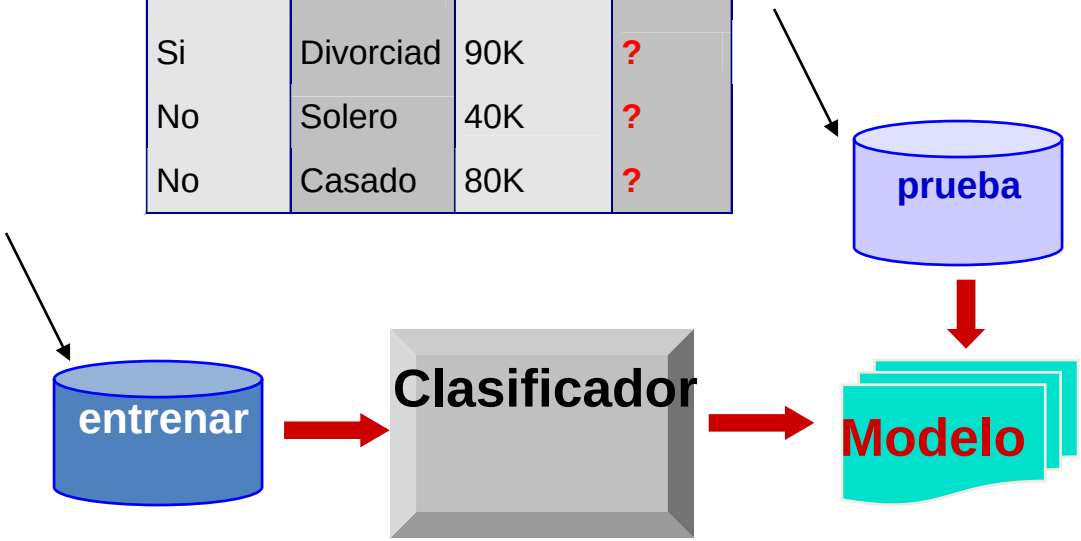
- Para generar dicho modelo o función, es necesario **definir un conjunto de datos de entrenamiento**, compuesto por objetos que ya tienen su clase asignada, también denominados ejemplos etiquetados.
- El modelo o función es creado **analizando las relaciones entre los atributos de los objetos** en el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

Clasificación

ID	Reemb	Edo Civil	pago Impuest	Enga ña
	categoría	categoría	continuo	clase
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 numéricos

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	?



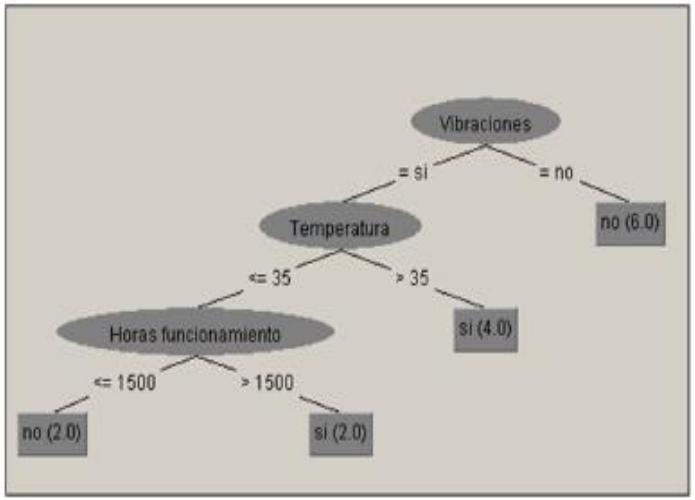
Clasificación

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

Class si: Prior probability = 0.44

Temperatura:  Normal Distribution. Mean = 45.8485 StandardDev = 9.6788 WeightSum = 6 Precision = 3.090909090909091
Vibraciones:  Discrete Estimator. Counts = 7 1  (Total = 8)
Horas funcionamiento:  Normal Distribution. Mean = 1045 StandardDev = 954.7382 WeightSum = 6 Precision = 190.0
Dias desde revisión:  Normal Distribution. Mean = 55.0385 StandardDev = 28.1919 WeightSum = 6 Precision = 6.230769230769231


Class no: Prior probability = 0.56

Temperatura:  Normal Distribution. Mean = 36.3182 StandardDev = 10.1047 WeightSum = 8 Precision = 3.090909090909091
Vibraciones:  Discrete Estimator. Counts = 3 7  (Total = 10)
Horas funcionamiento:  Normal Distribution. Mean = 688.75 StandardDev = 483.8243 WeightSum = 8 Precision = 190.0
Dias desde revisión:  Normal Distribution. Mean = 49.0673 StandardDev = 21.9048 WeightSum = 8 Precision = 6.230769230769231
```

inst#,	actual,	predicted,	error,	probability distribution
1	1:si	1:si		*0.967 0.033
2	2:no	2:no		0.028 *0.972
3	2:no	2:no		0.028 *0.972
4	2:no	2:no		0.028 *0.972
5	2:no	2:no		0.492 *0.508
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
9	2:no	2:no		0.028 *0.972
10	2:no	2:no		0.028 *0.972
11	2:no	2:no		0.028 *0.972
12	2:no	2:no		0.492 *0.508
13	1:si	2:no	+	0.492 *0.508
14	1:si	1:si		*0.967 0.033

Clasificación

Detección de Fraude

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

Enfoque:

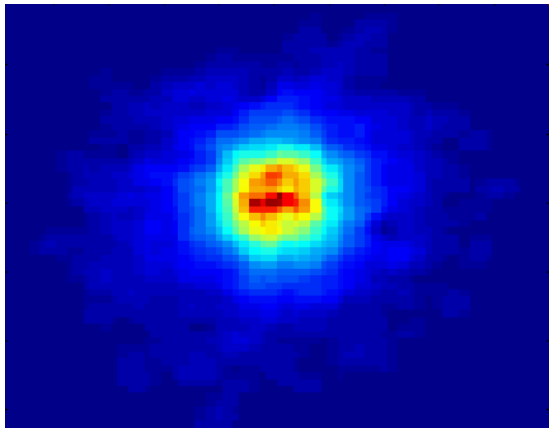
- Utilizar las transacciones con tarjetas de crédito y la información sobre la cuenta del titular como atributos.
  - ¿Cuándo compra un cliente?, ¿qué compra?, ¿con qué frecuencia paga a tiempo?, etc.
- Etiquetar transacciones pasadas: fraudulentas o correctas. Esto forma el atributo de clase.
- Aprender 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

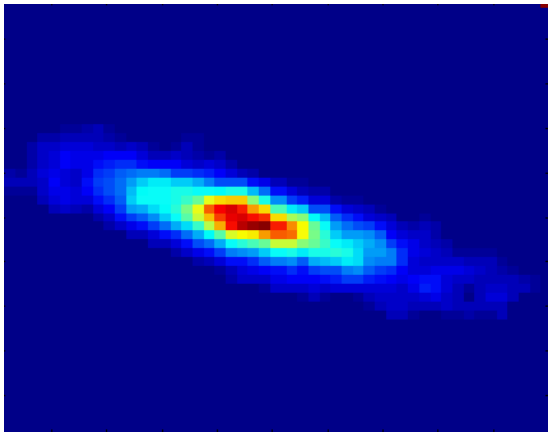
Temprano



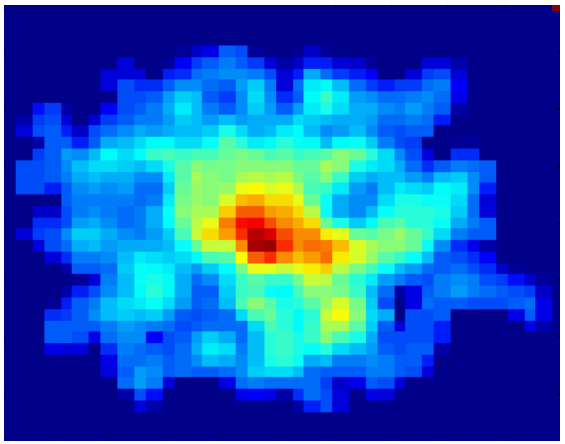
Clases:
- **Estado de Formación**

Atributos:
- **Caract. Imagen,**
- **Caract. Ondas, luz, etc.**

Intermedio

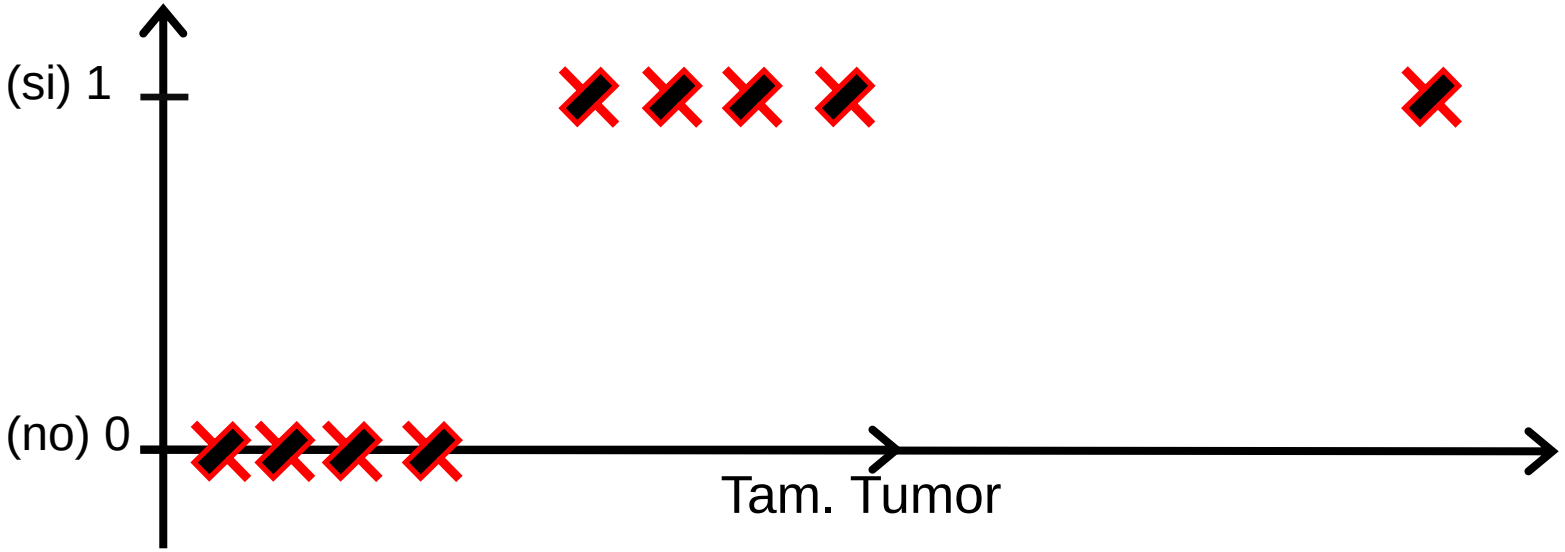


Tarde



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

Clasificación



Umbral clasificador $h_\theta(x)$ = 0.:

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 gaussanianos, Redes bayesianas

- **Basados en Arboles 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, perceptron multicapa, backpropagation, Deep learning

- **Híbridos**

Métricas para evaluar un algoritmo de clasificación

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

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

Modelos de Agrupamiento (Segmentación)

Agrupamiento (Clustering)

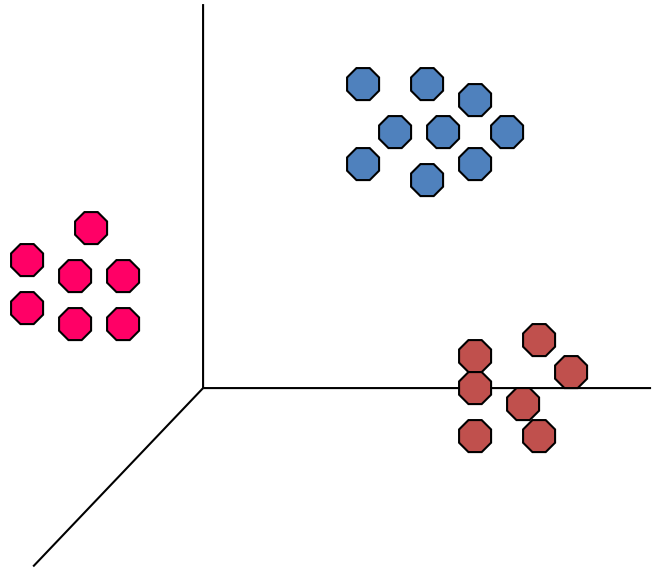
Dado un conjunto de datos, cada una con un conjunto de atributos, y una medida de similitud entre ellos, **encontrar grupos** de tal manera que:

- 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.

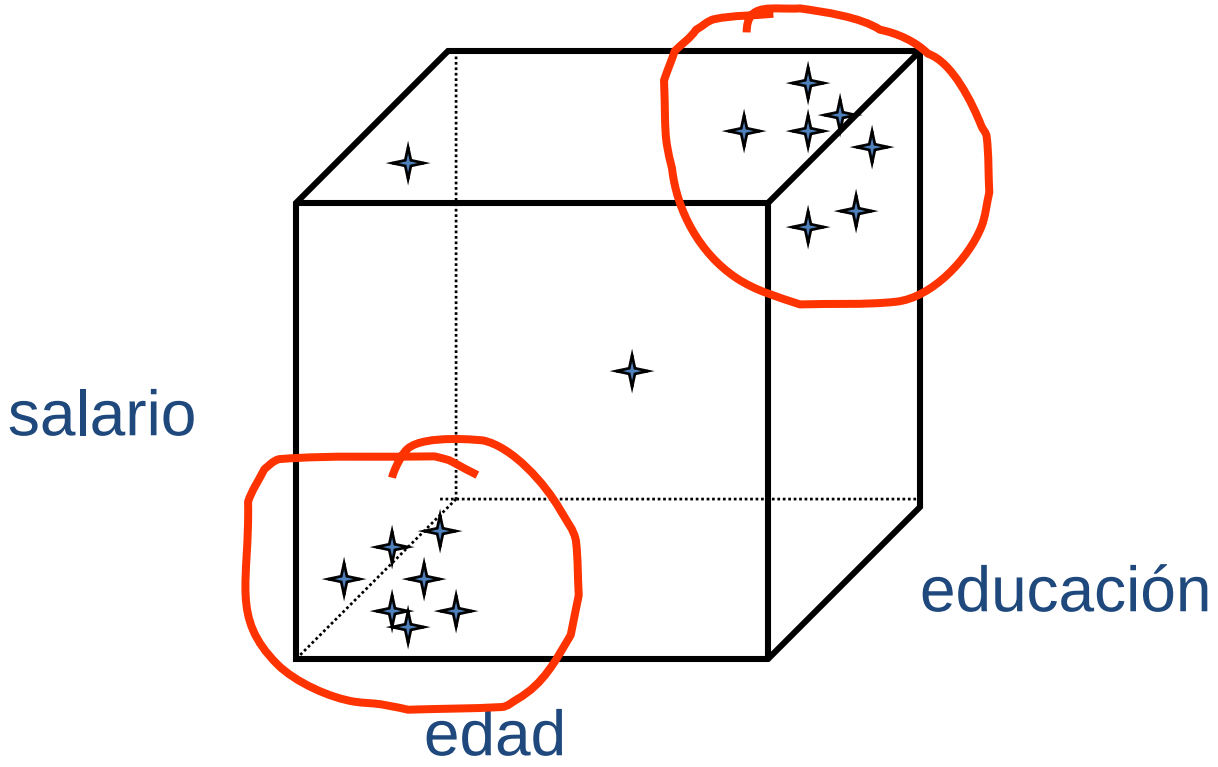
Agrupamiento (Clustering)

Distancias Intracluster son minimizadas

Distancias Intercluster son maximizadas

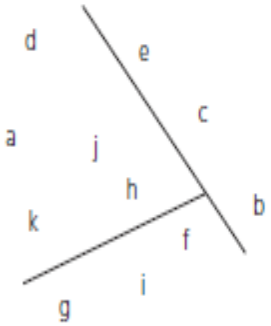


Agrupamiento (Clustering)

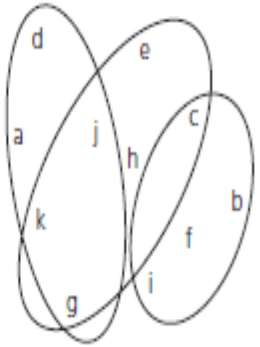


Clusters

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



(a)



(b)

	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)

(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).

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

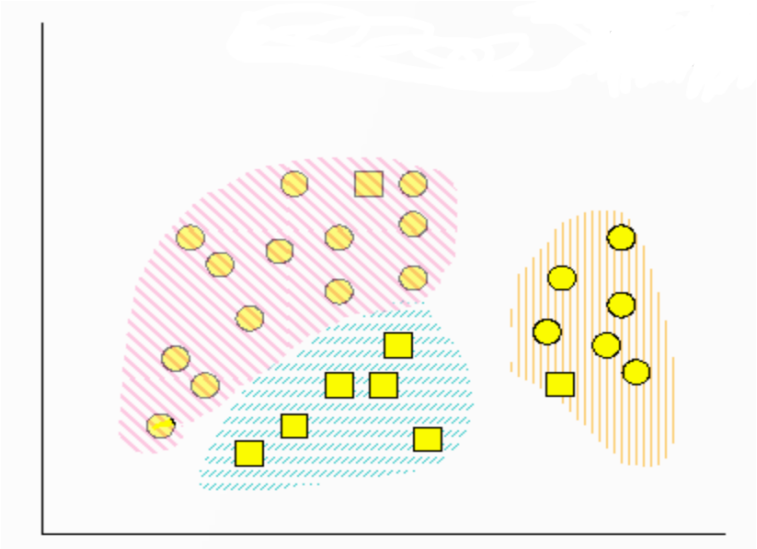
Tipos de clustering

- **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**

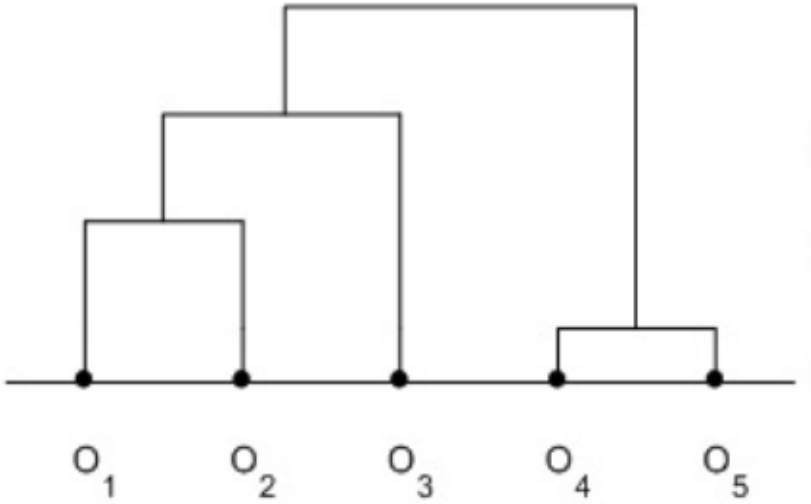
- **Clustering ascendente jerárquico**
  - Crear un dendograma, es decir, crear un conjunto de agrupaciones anidadas hasta construir **un árbol jerárquico**

Clusterización

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



a) Particional



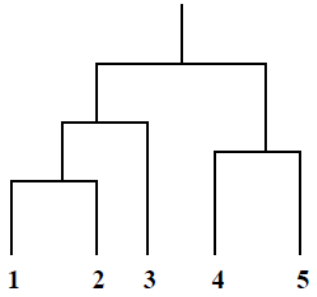
b) Jerárquico

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



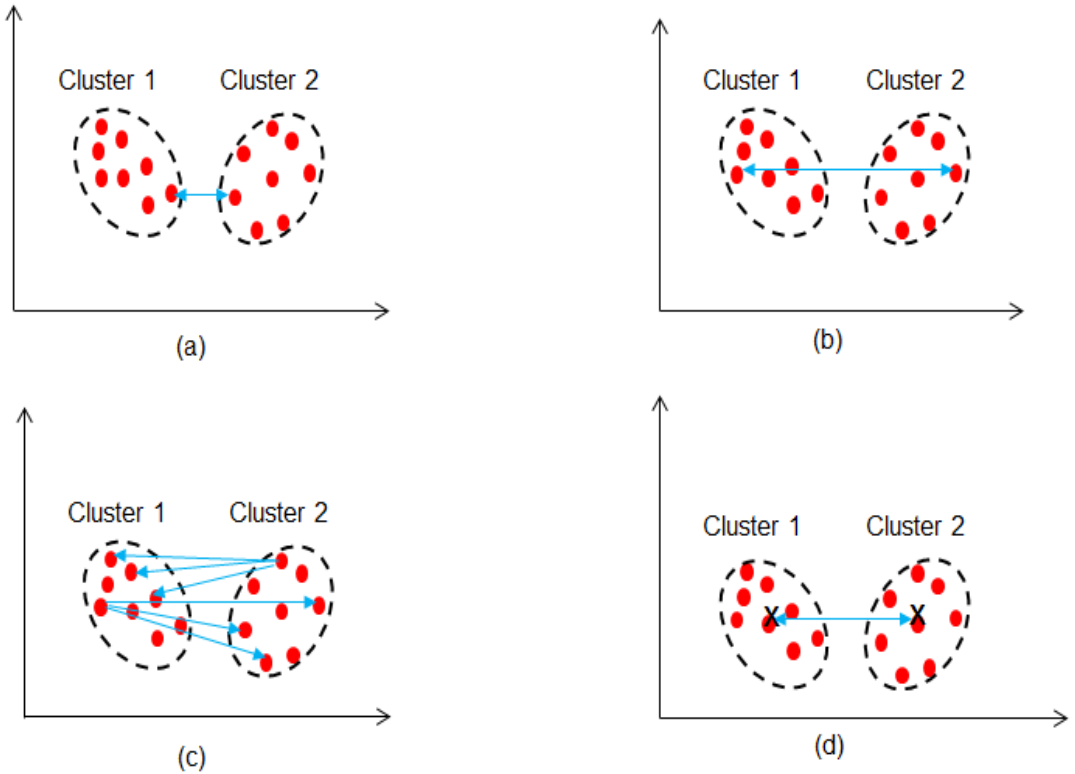
Tipos de clustering: basados en distancia

(a) Distancia mínima

(b) Distancia máxima

(c) Distancia de promedio del grupo

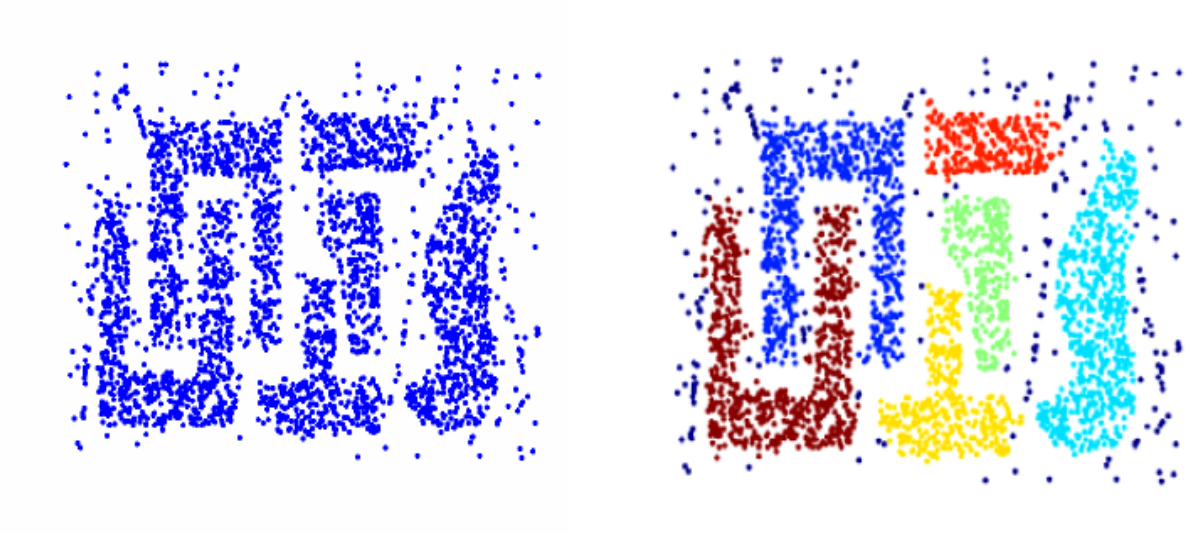
(d) Distancia con respecto al centroide



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 clustering**

Tipos de clustering

Algoritmo de DBSCAN

```
current_cluster_label ← 1
for all core points do
  if the core point has no cluster label then
    current_cluster_label ← 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 restantes**

Tipos de clustering

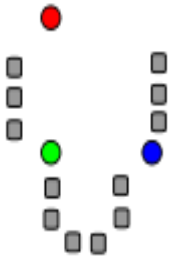
Basados en prototipos

- 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.

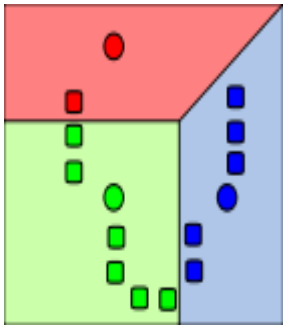
- El prototipo que define al cluster normalmente **denota sus características principales**, por ejemplo, para atributos solo numéricos el prototipo del cluster a menudo se representa como el centroide.

Algoritmo K-medias

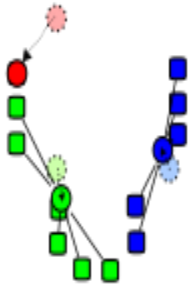
es un método de agrupamiento, que tiene como objetivo la partición de un conjunto (n) en k grupos en el que **cada ejemplo pertenece al grupo más cercano a la media**



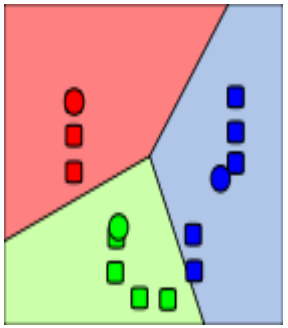
1) *k* centroides iniciales generados aleatoreamente (en este caso *k*=3)



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

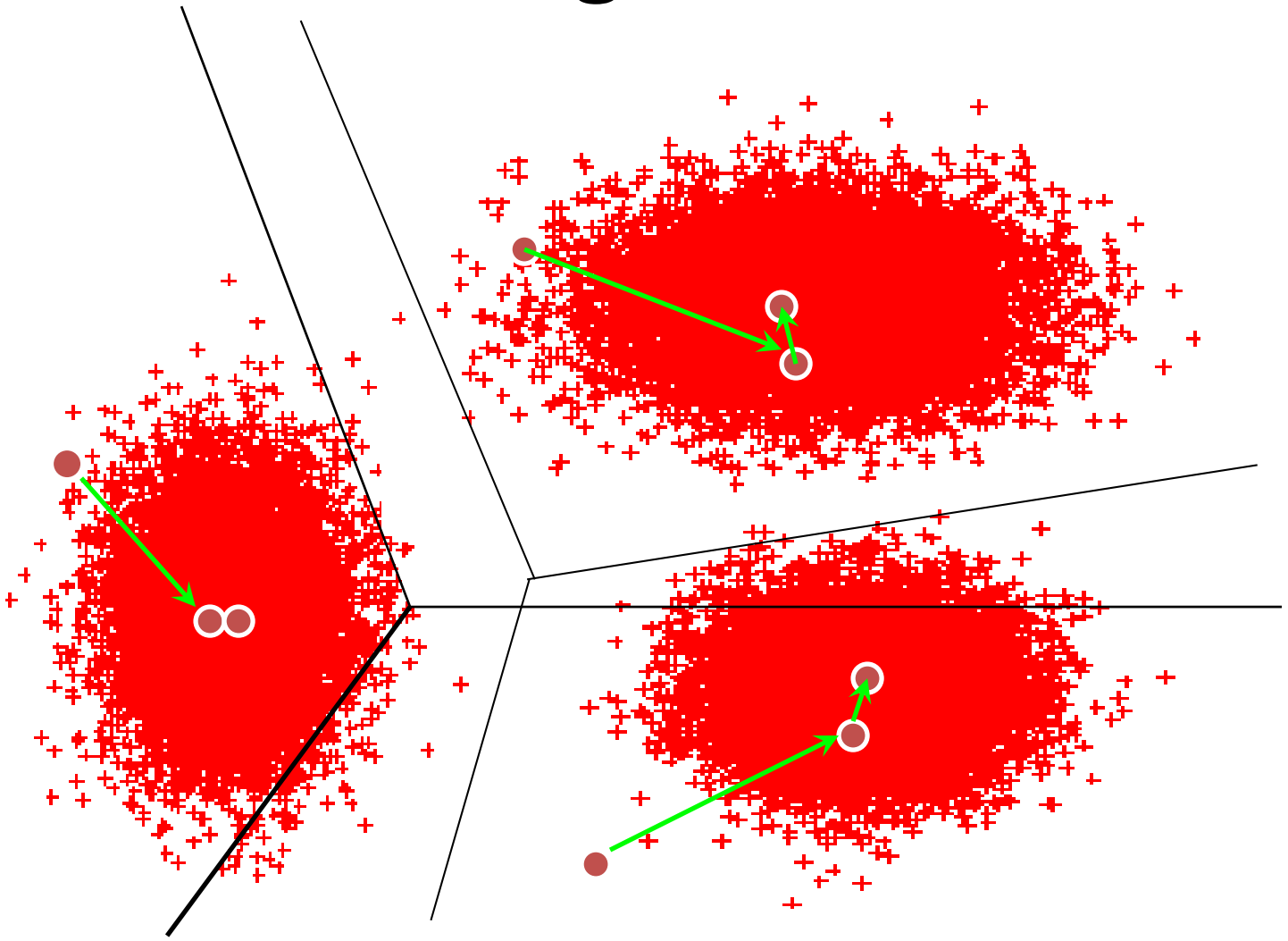
4. Una vez introducidos todos los datos, se **desplazan los prototipos hasta el centro de masas** de su nuevo conjunto, esto se repite hasta que no se desplazan más.

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 $\delta^2 E(Xi, Xj) = ||Xi - Xj||^2$

Clusterización: Algoritmo K-Medias



Corrida en frio K-means

Input:

- K (number of clusters)
- Training set $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

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

Repeat {

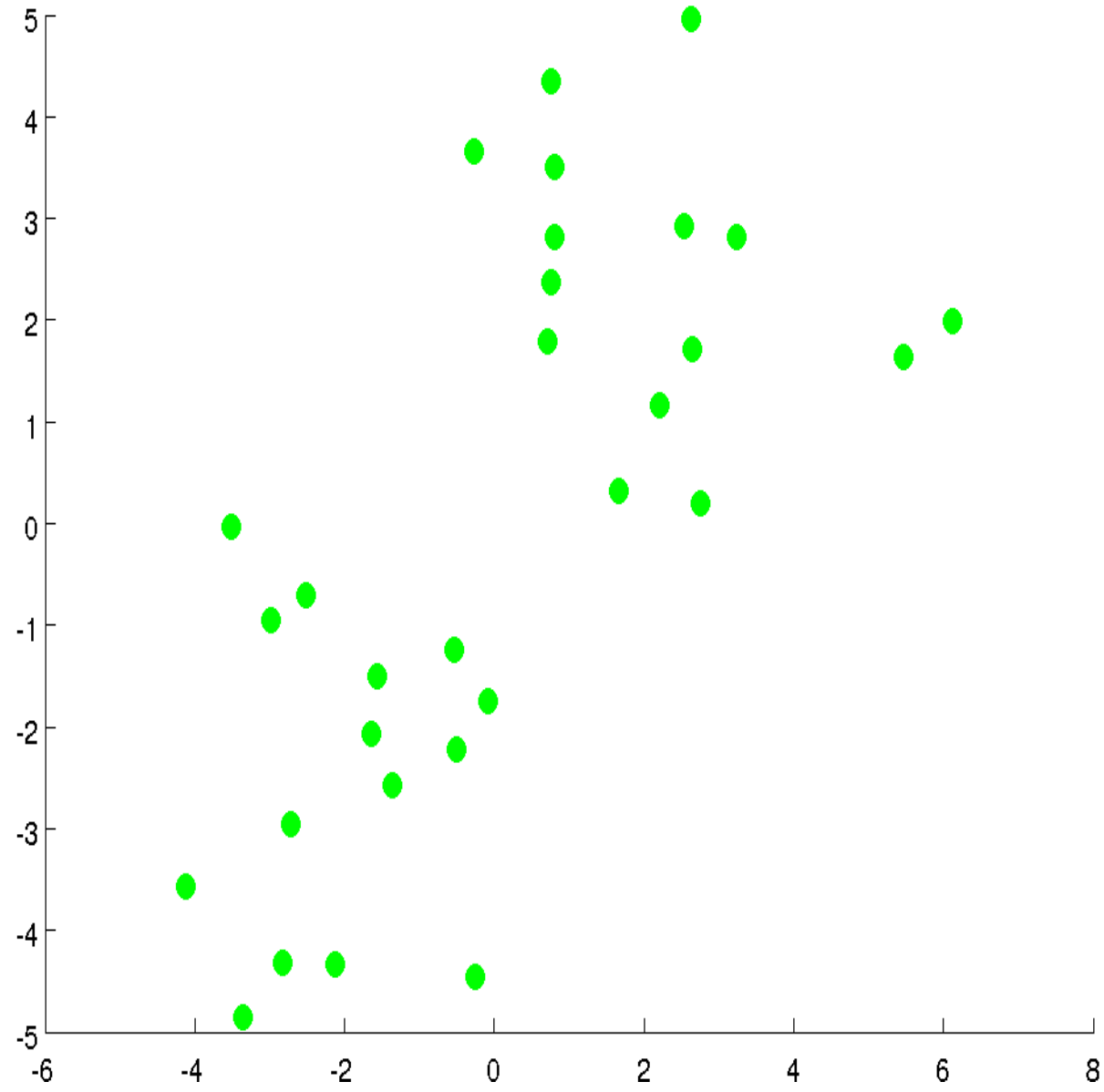
for i =1 to m

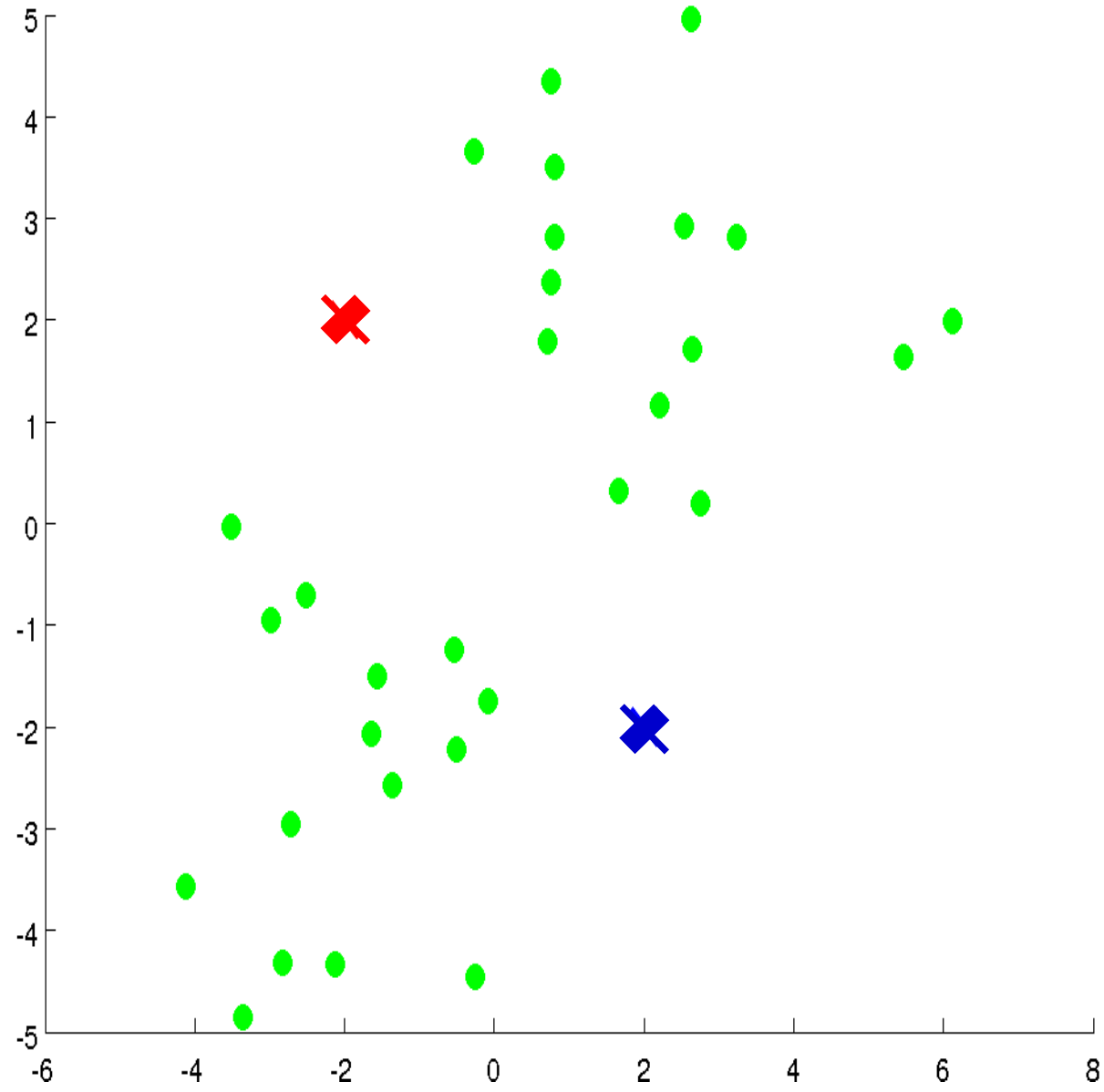
$c^{(i)}$:= index (from 1 to K) of cluster centroid closest to $x^{(i)}$

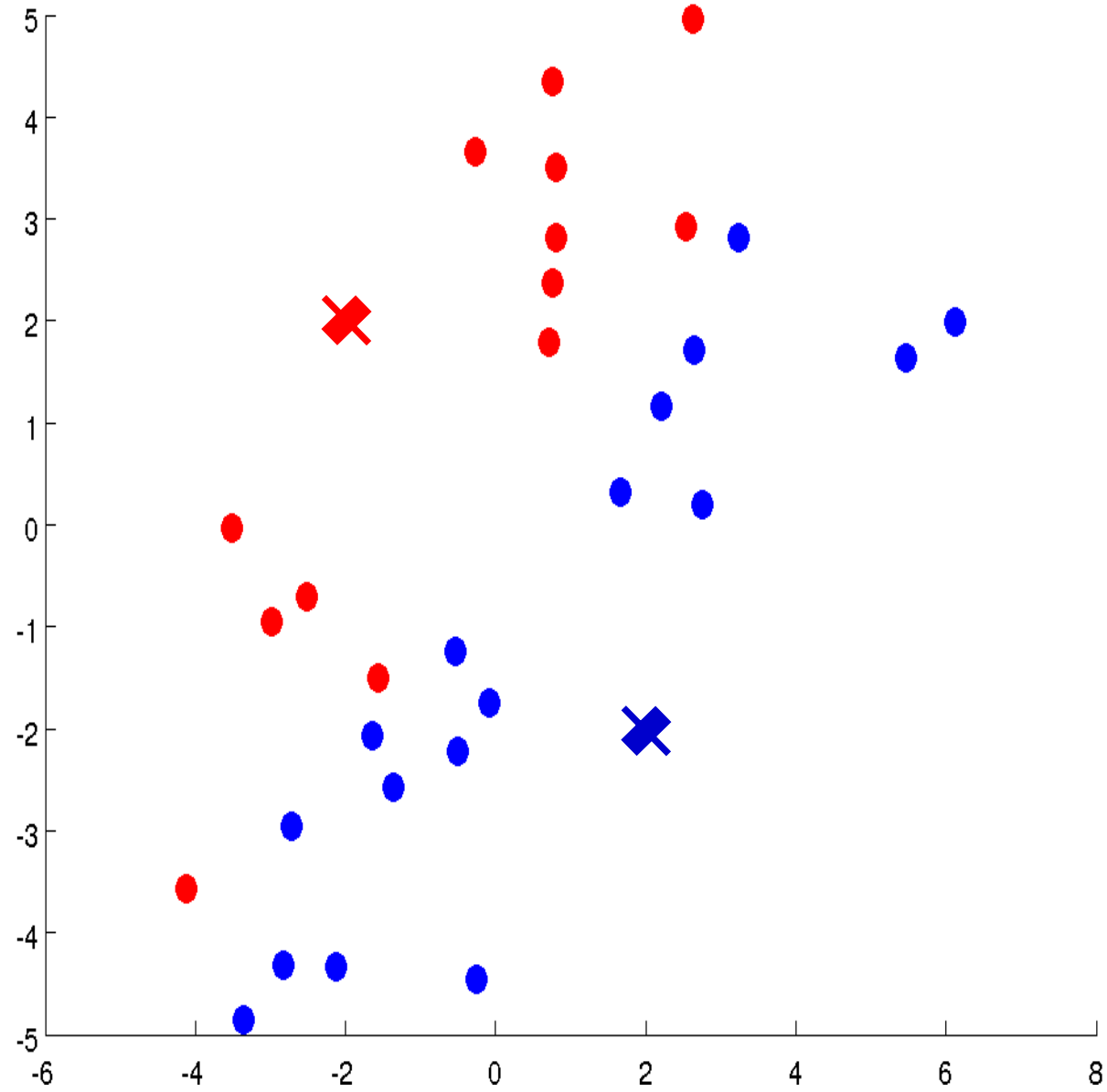
for k = 1 to K

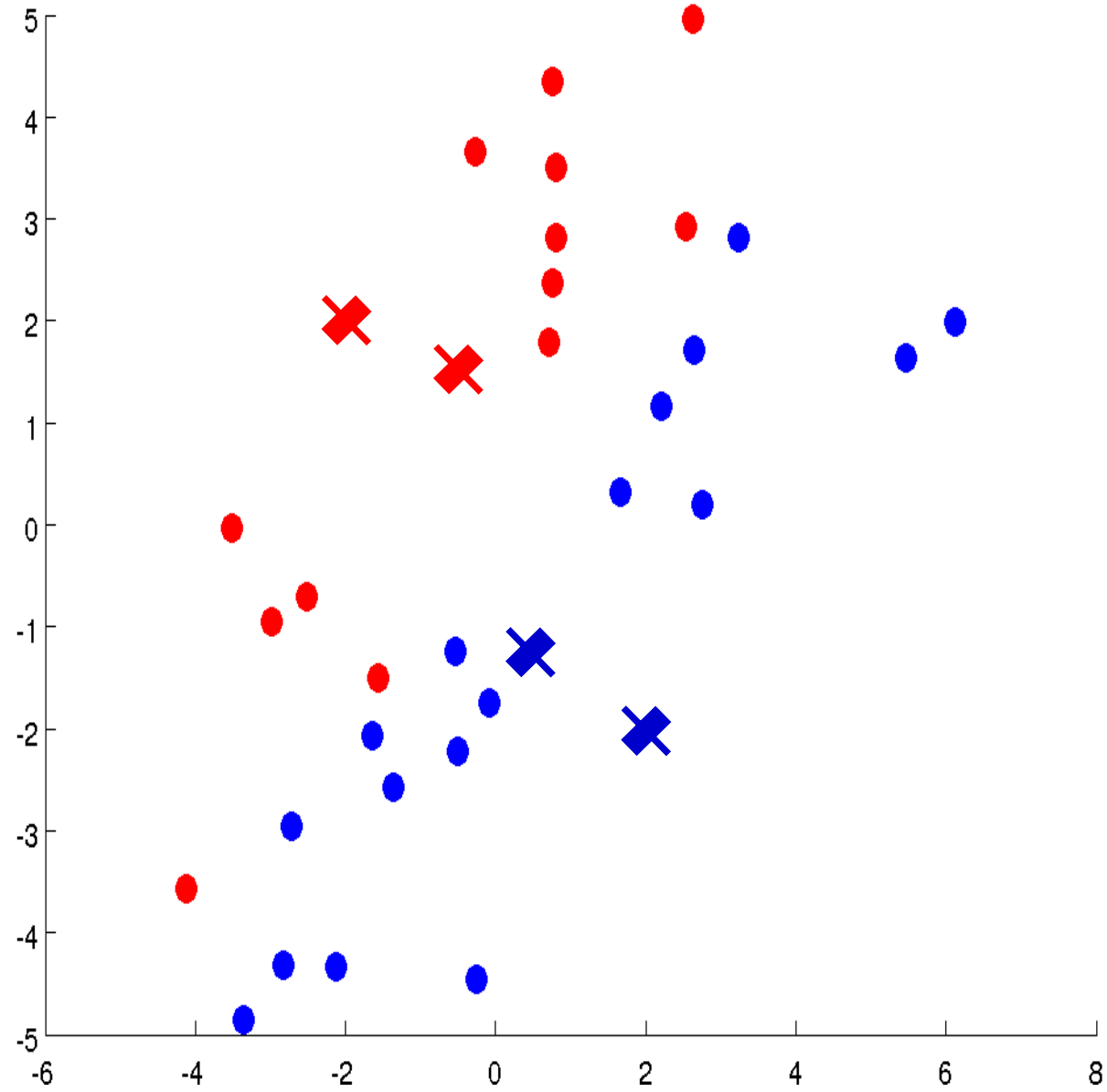
μ_k := average (mean) of points assigned to cluster

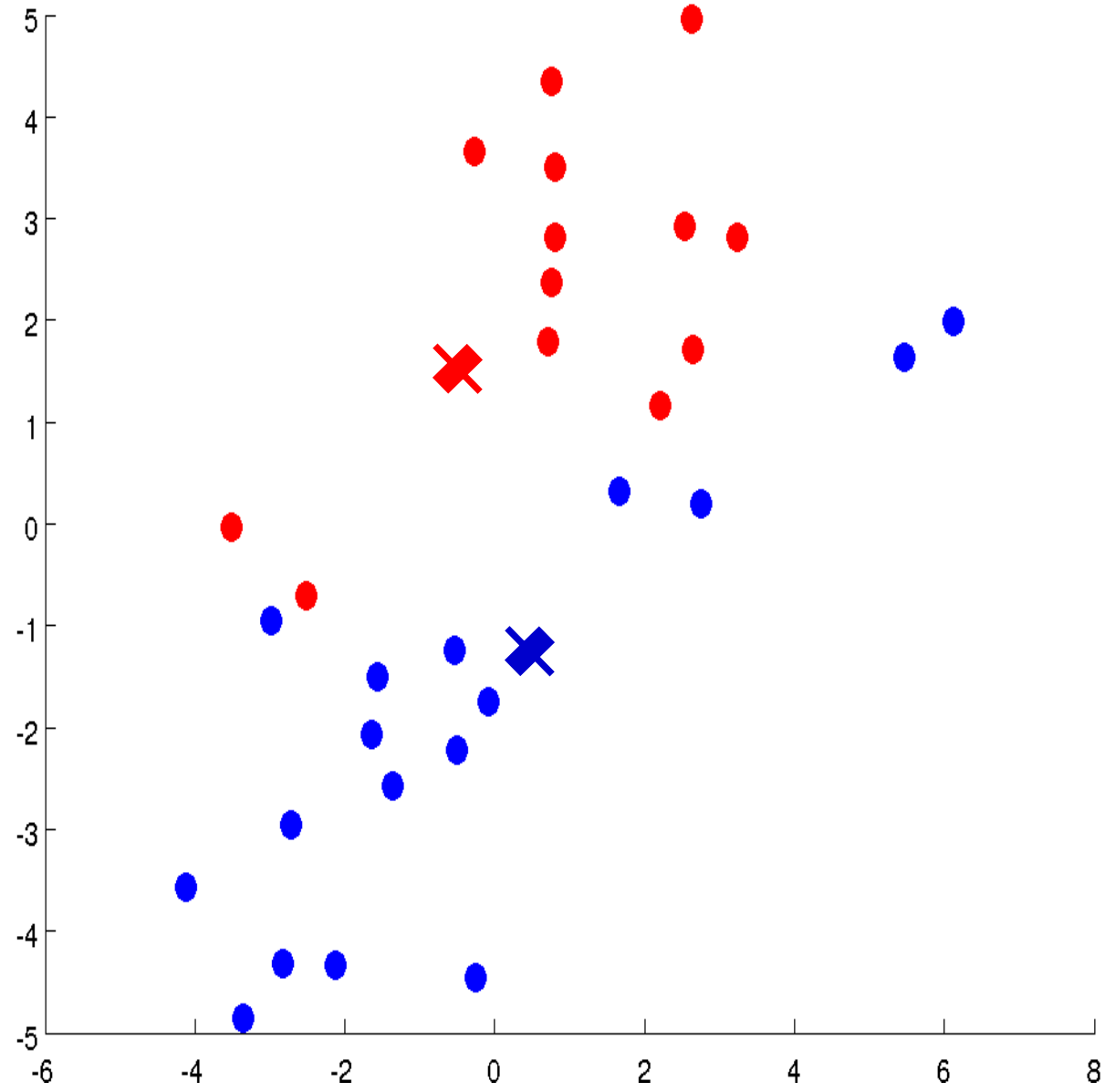
} until convergence criteria is met

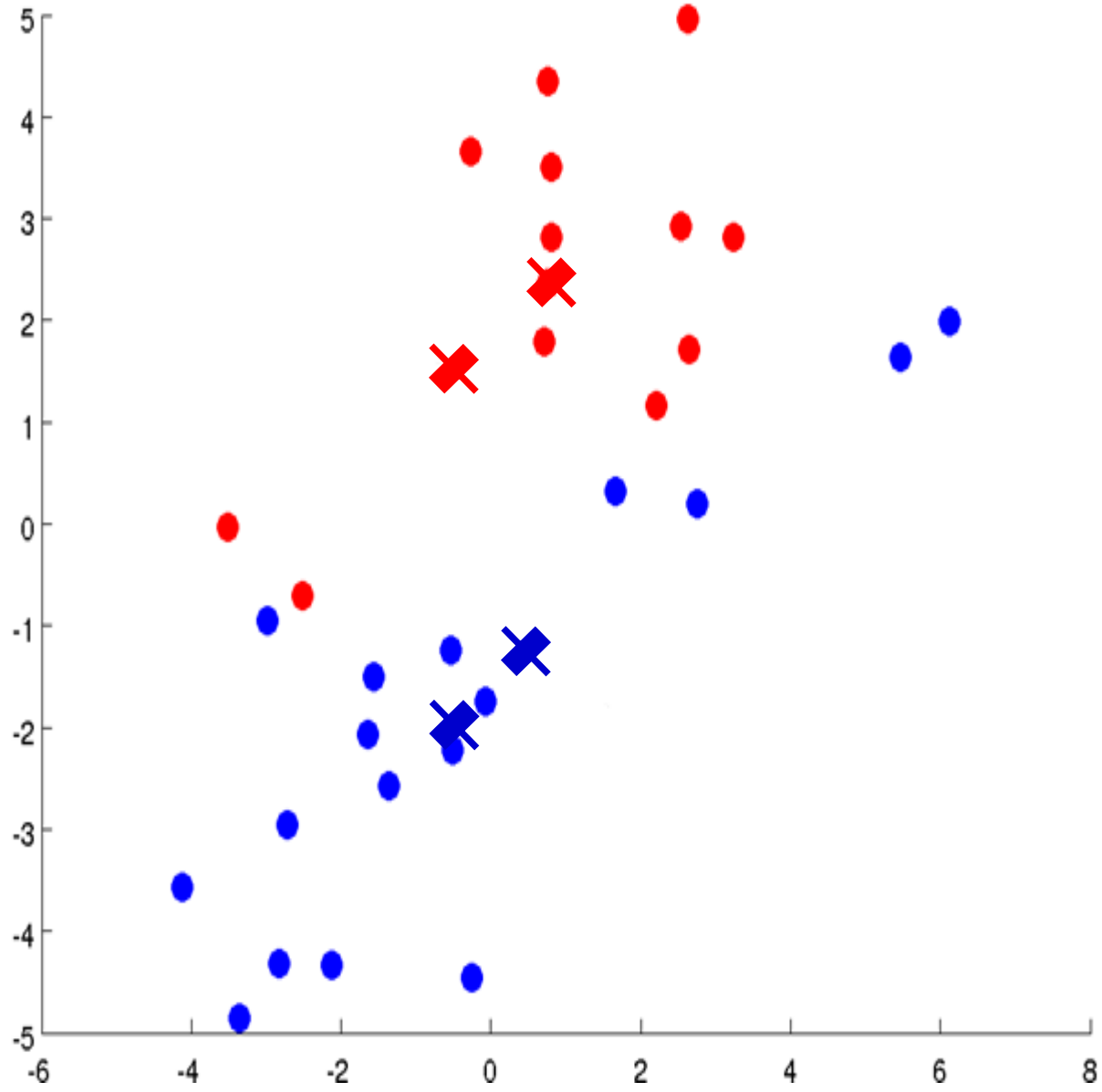

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

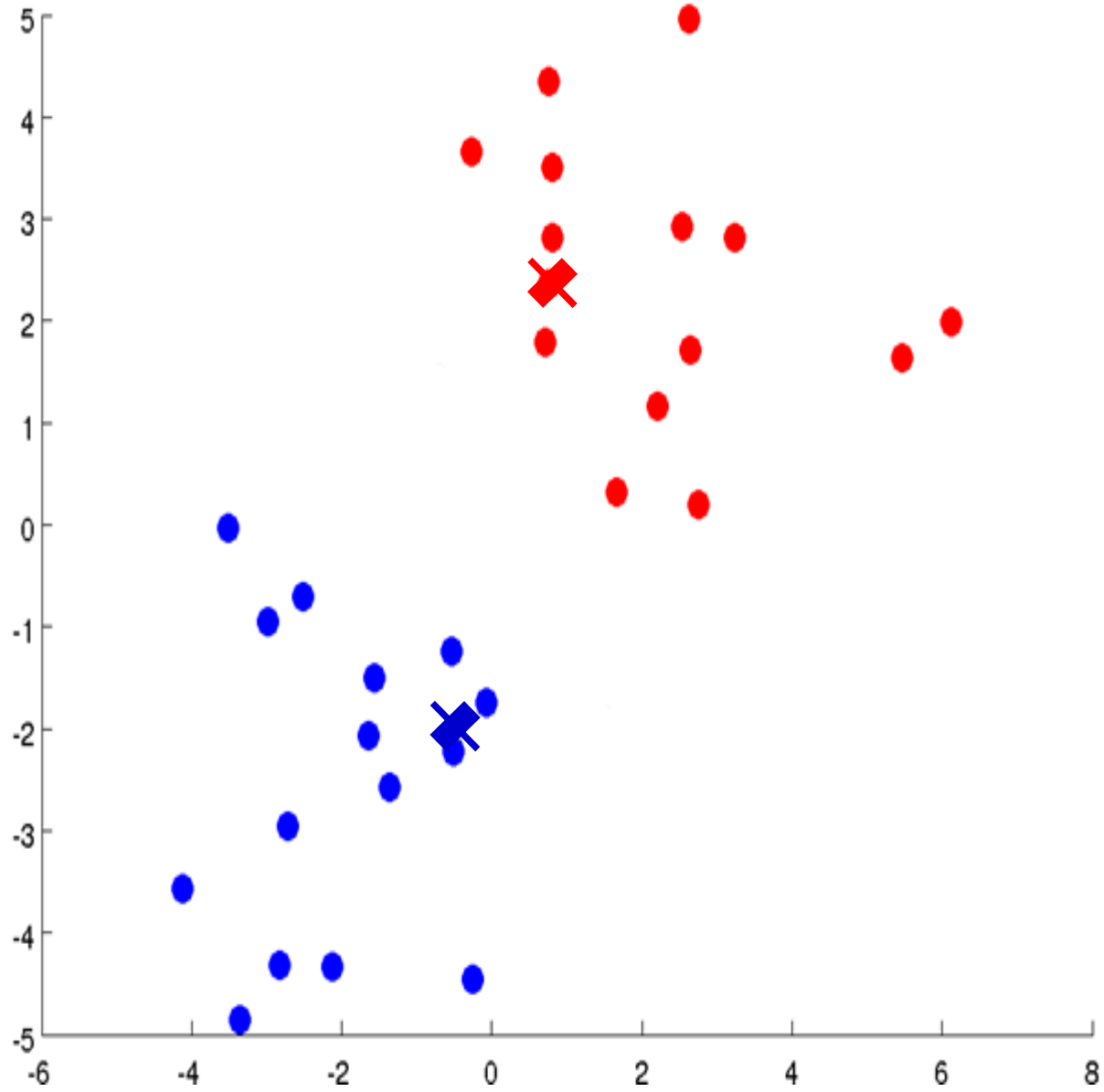

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

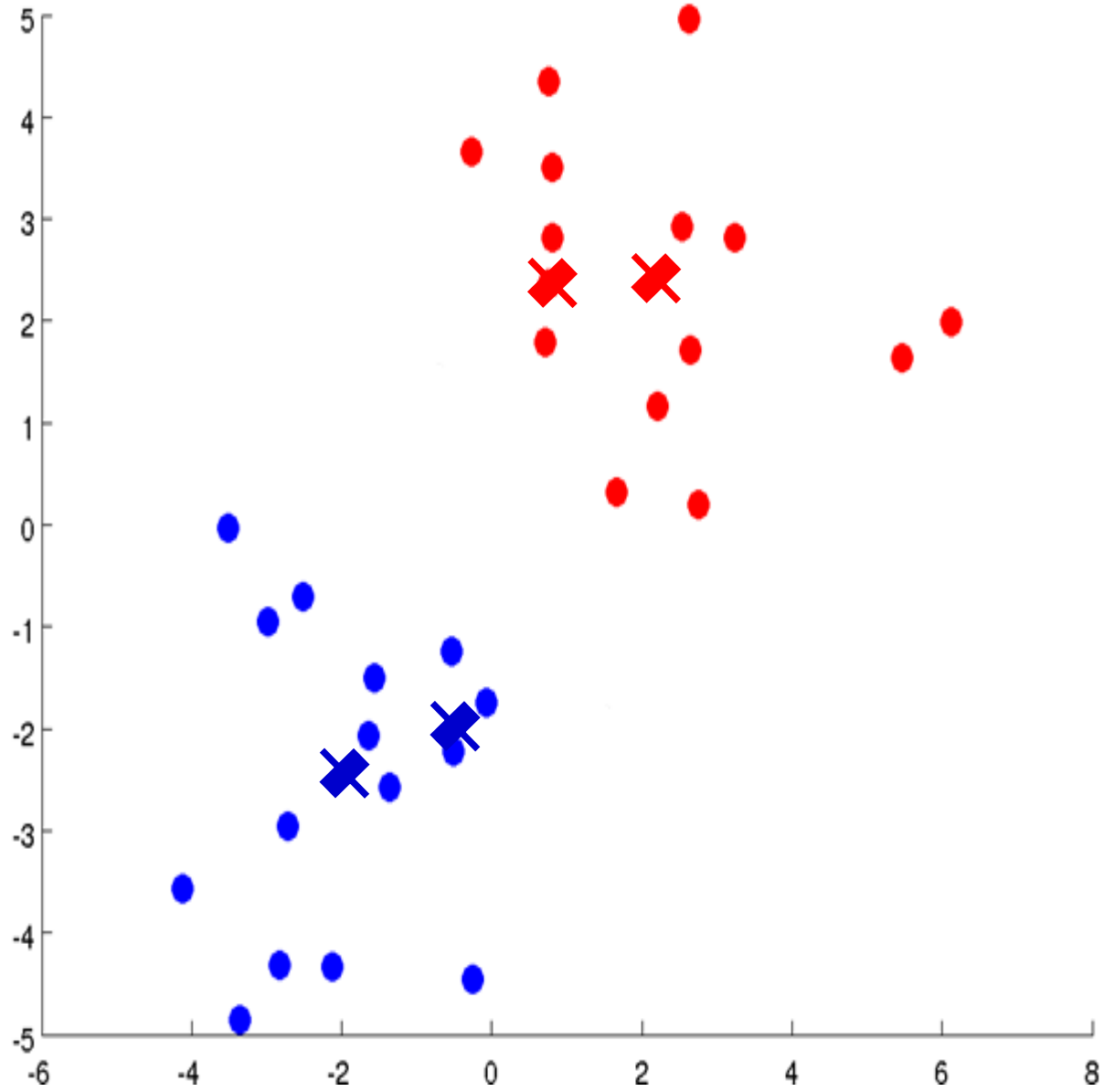

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

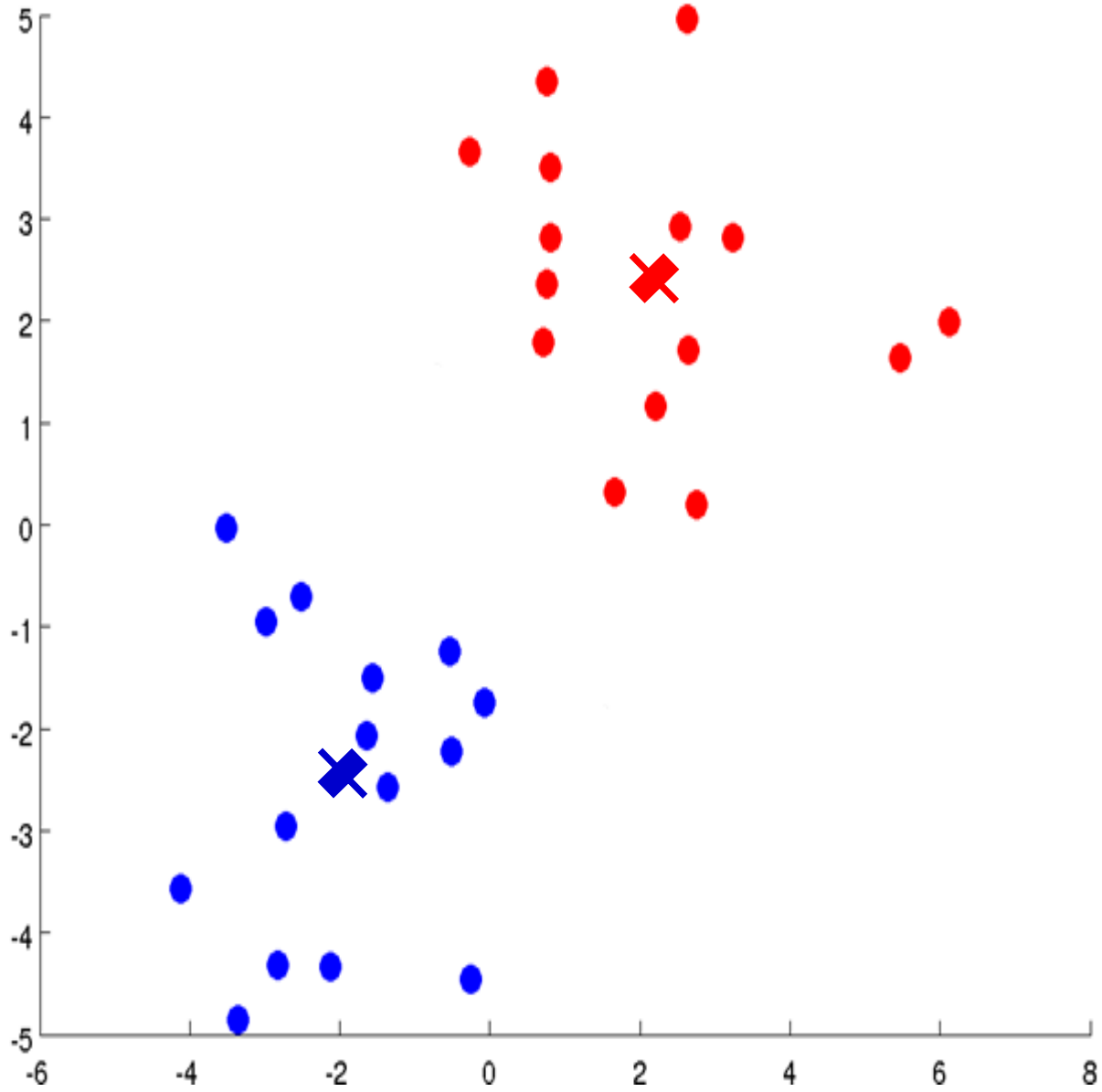

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

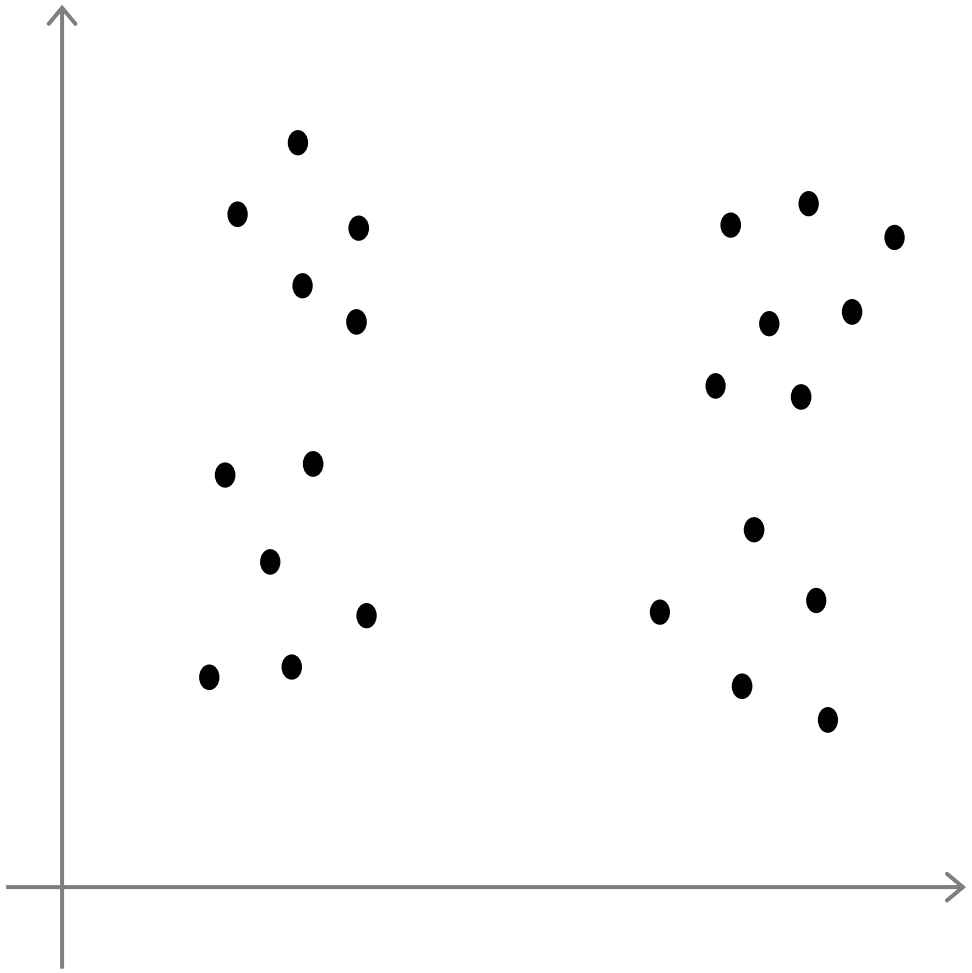

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

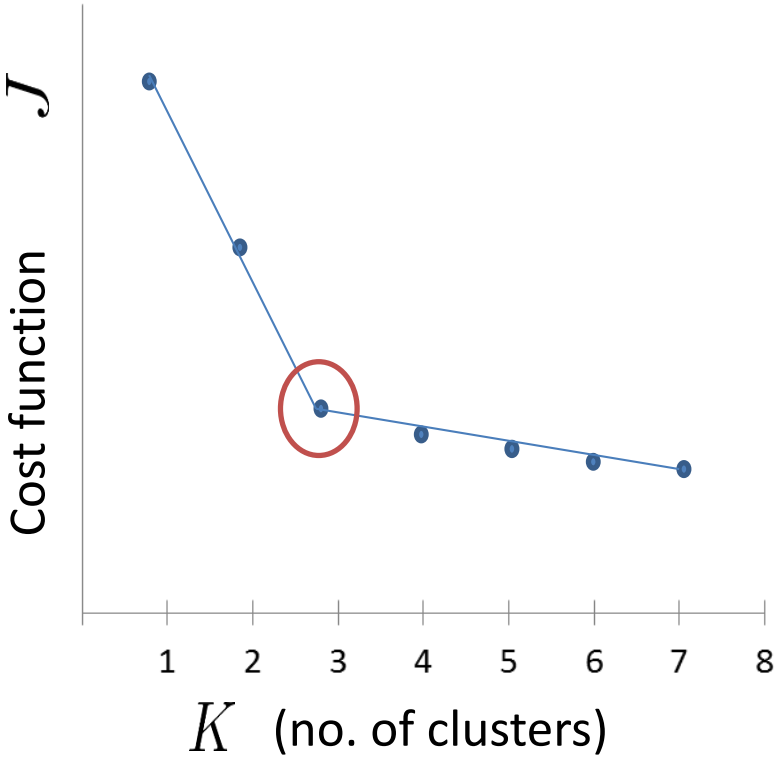

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

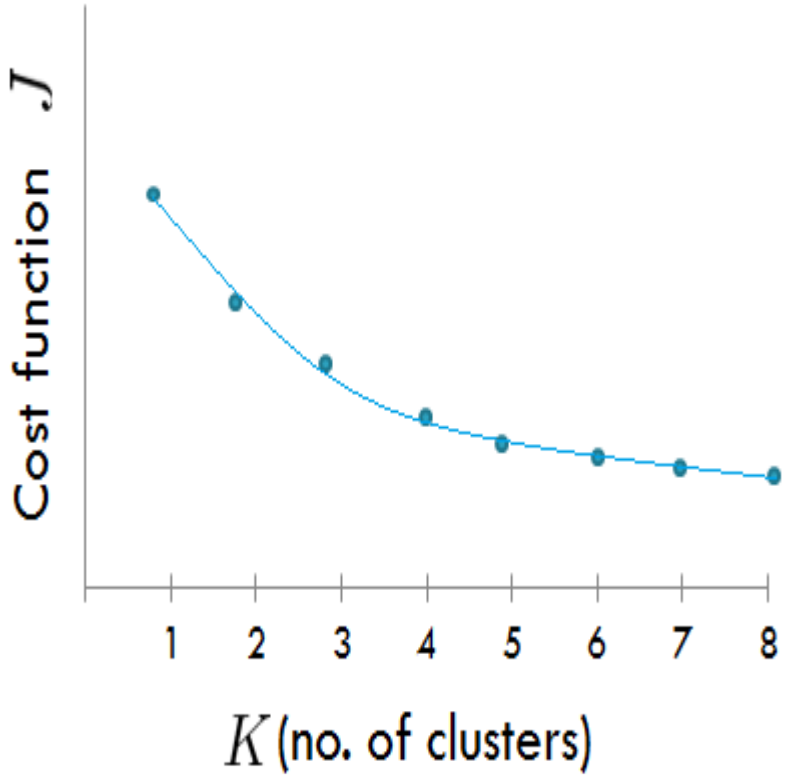
Cual es el correcto valor de K?



Escogiendo el valor de K

Método del codo:





Clustering Incremental

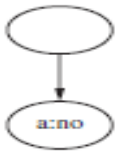
Datos de tiempo

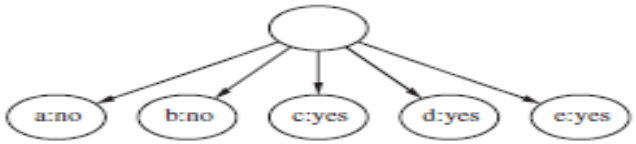
ID code	Outlook	Temperature	Humidity	Windy	Play
a	sunny	hot	high	false	no
b	sunny	hot	high	true	no
c	overcast	hot	high	false	yes
d	rainy	mild	high	false	yes
e	rainy	cool	normal	false	yes
f	rainy	cool	normal	true	no
g	overcast	cool	normal	true	yes
h	sunny	mild	high	false	no
i	sunny	cool	normal	false	yes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

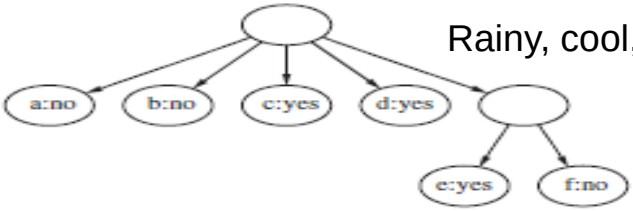
Clustering Incremental

Table 4.6 The weather data with identification codes.

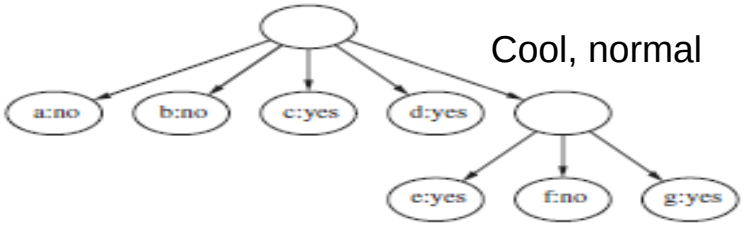
ID code	Outlook	Temperature	Humidity	Windy	Play
a	sunny	hot	high	false	no
b	sunny	hot	high	true	no
c	overcast	hot	high	false	yes
d	rainy	mild	high	false	yes
e	rainy	cool	normal	false	yes
f	rainy	cool	normal	true	no
g	overcast	cool	normal	true	yes
h	sunny	mild	high	false	no
i	sunny	cool	normal	false	yes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



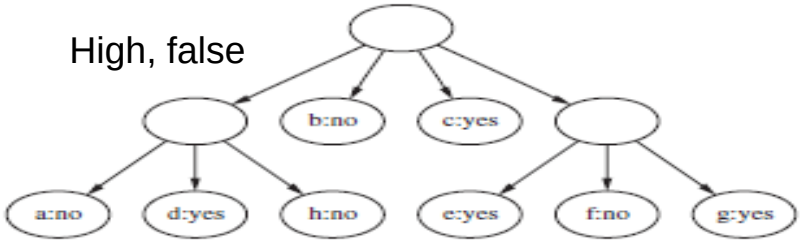




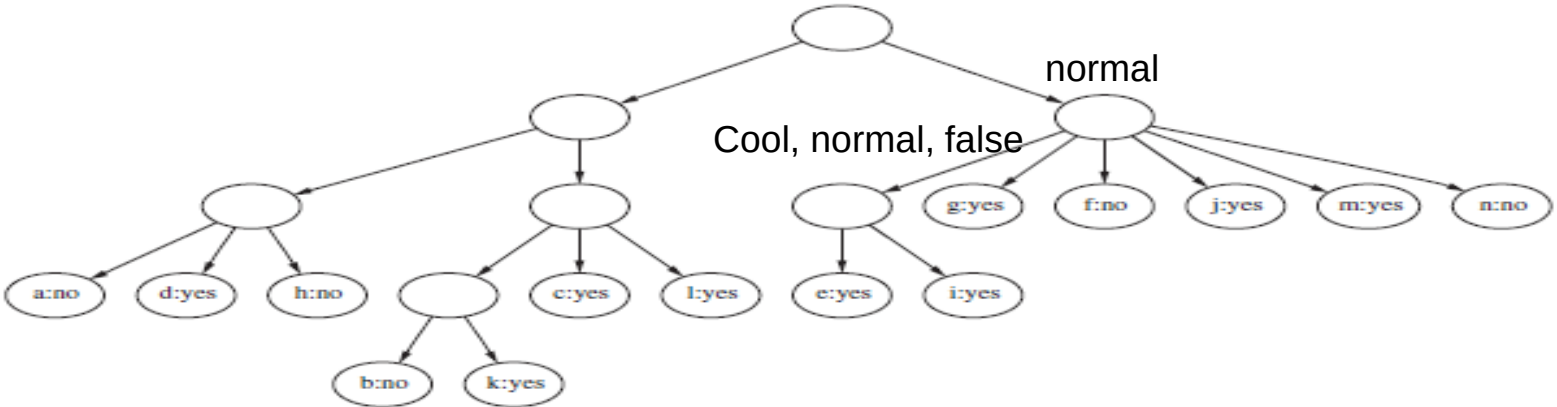
Rainy, cool, normal



Cool, normal



High, false



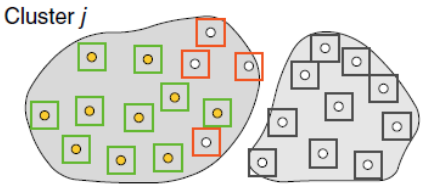
normal

Cool, normal, false

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



		Truth	
		P	N
Hypothesis	P	TP	FP
	N	FN	TN

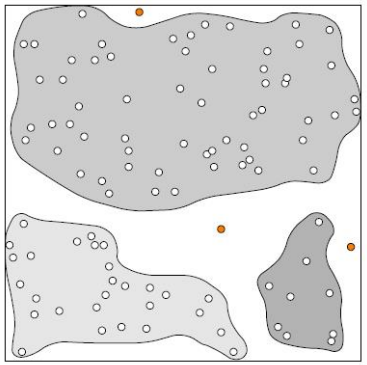
- **Entropía**: Es el grado de coincidencia de los clusters a las clases ya definidas de los datos originales.

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

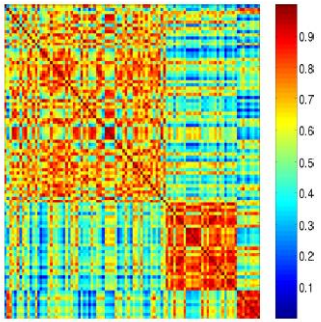
Métricas para evaluar un algoritmo de agrupamiento

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

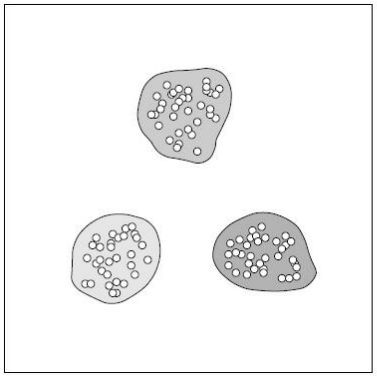
- Coeficiente de correlación:

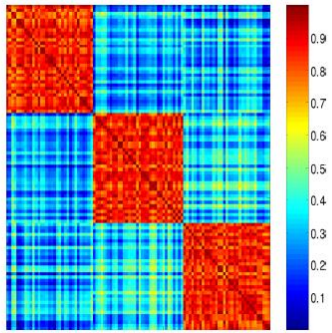


DBSCAN at random data.



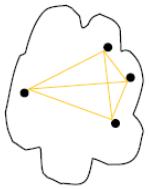
Similarity matrix sorted by cluster label.



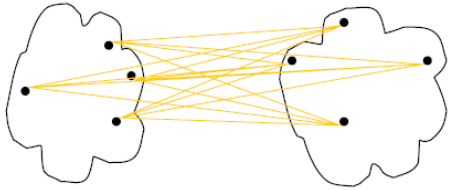


- Cohesión y separación

Separación se mide como la distancia entre los centroides
Cohesión se mide como la distancia promedio entre sus muestras y centroides

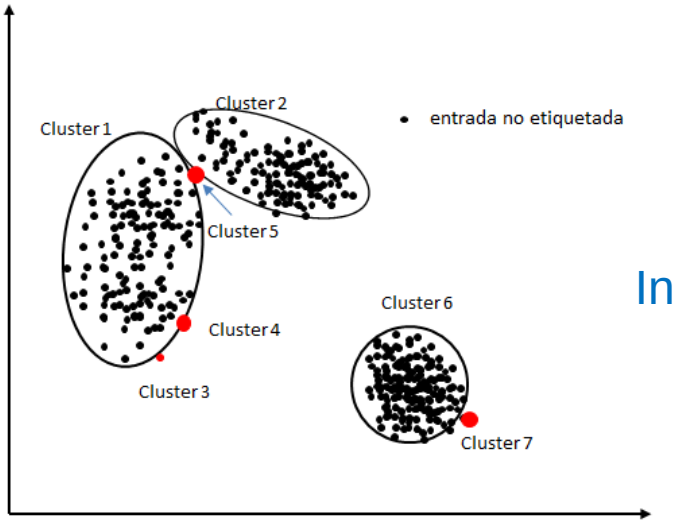


cohesión

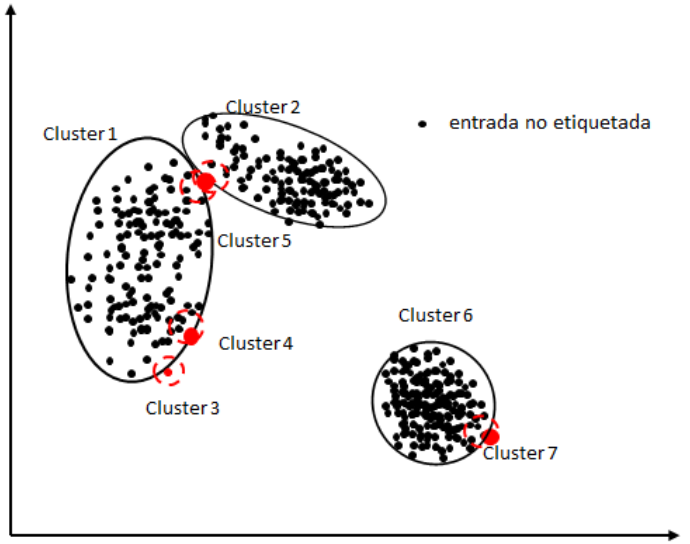


separación

Agrupamiento (Mezcla)



Inicial

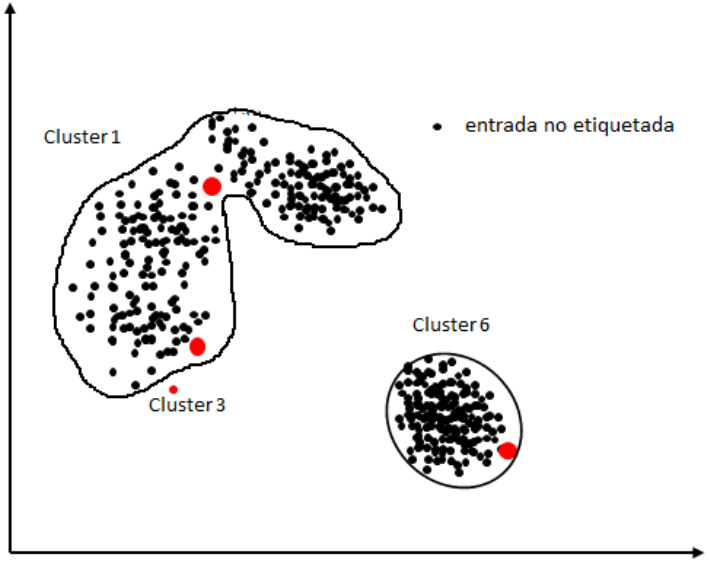


Detección de los grupos candidatos para ser mezclados

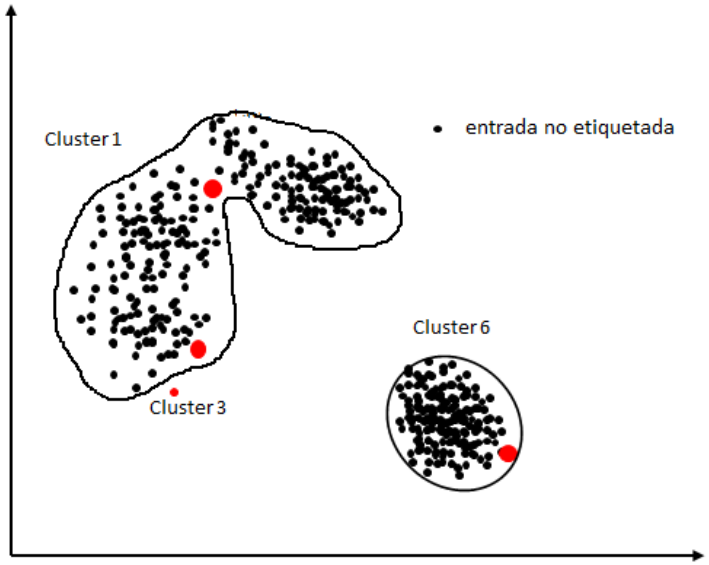
“

1. Alta densidad en su vecindad
2. Intra-distancia promedio mayor a Inter-distancia promedio en la vecindad
3. Distancia promedio a sus centroides mayor a Inter-distancia promedio en la vecindad

Mezcla



Agrupamiento (Dividir)



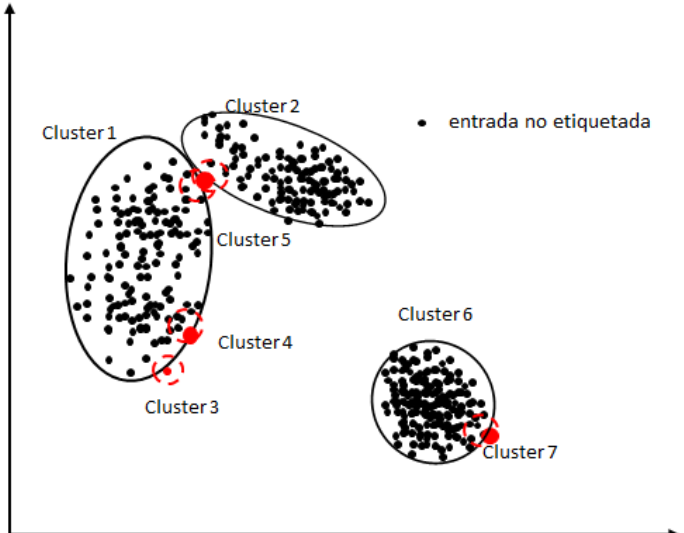
Inicial

Detección grupo candidato a dividir

“

1. Distancia promedio a centroide mayor a Inter-distancia promedio de vecindades
2. Densidad de vecindades mucho mayor a la del cluster

Partir



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.

$$Consecuencia \Leftarrow Antecedente_1\ Antecedente_2\ \ldots\ Antecedente_m.$$

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

- 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***).

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

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

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 = fria entonces humedad = normal**
- **Si viendo = falso y jugar = no entonces pronostico = soleado y humedad = alta**

Reglas de Asociación

Reglas que implican relaciones

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

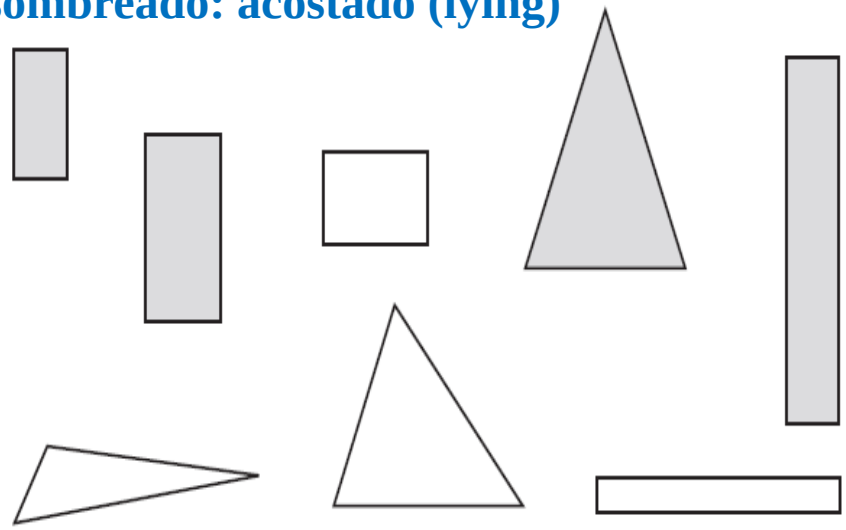


Table 3.2 Training data for the shapes problem.

Tabla con datos de entrenamiento

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

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

	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)
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)
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)
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)
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)
	…	…	…	…

	One-item sets	Two-item sets	Three-item sets	Four-item sets
…		…	…	
38		humidity = normal windy = false (4)	humidity = normal windy = false play = yes (4)	
39		humidity = normal play = yes (6)	humidity = high windy = false play = no (2)	
40		humidity = high windy = true (3)		
…		…		
47		windy = false play = no (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

Árbol de Decisión

Toma como entrada una situación y da como salida una decisión (por ejemplo: si/no)

- Ejemplo: **decidir si esperar o no por una mesa en un restaurant basado en los siguientes criterios**
  1. ¿Hay otro restaurant cerca? (Alternativa)
  2. ¿Hay un bar confortable para esperar? (Bar)
  3. ¿Hoy es Viernes o Sábado? (Día)
  4. ¿Hay hambre? (EdoM)
  5. Numero de personas en el restaurant (Patrón: Vacio, Algo, lleno)
  6. ¿Precio? ($, $$, $$$)
  7. ¿Esta lloviendo? (Edo.D)
  8. ¿Se tiene una reservación?
  9. Tipo de restaurant (Francés, Italiano, Japonés, Hamburguesa)
  10. Tiempo de espera estimado (0-10min, 10-30min, 30-60, >60min)

Árbol de Decisión

Ejemplos

Criterios **¿Qué aprendo?**

Ej	Alt	Bar	Dia	EdM	Patr	Prec	EdD	Tipo	RES	T --->	Espera
X1	S	N	N	S	Alg	$$$	N	Franc	S	0-10	S
X2	S	N	N	S	llen	$	N	Jap	S	10-15	N
X3	N	S	N	N	Alg	$	N	Hamb	N	0	S
...											
X12	S	S	S	S	llen	$	N	Hamb	N	10	S

Tablas de decisión

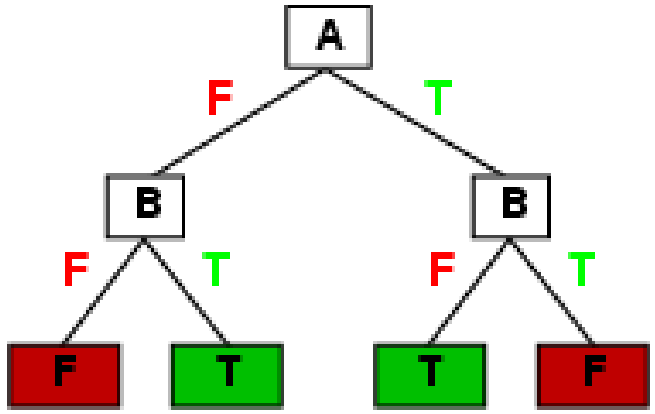
Forma más simple y más rudimentaria para representar la salida de la máquina de aprendizaje.

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Árbol de Decisión

- Puede expresar cualquier función a partir de sus atributos de entrada.
- Un árbol de decisión es consistente para cualquier conjunto de entrenamiento, cuando hay un camino a una hoja para uno o varios ejemplos
- Basado en la idea de tablas de la verdad:

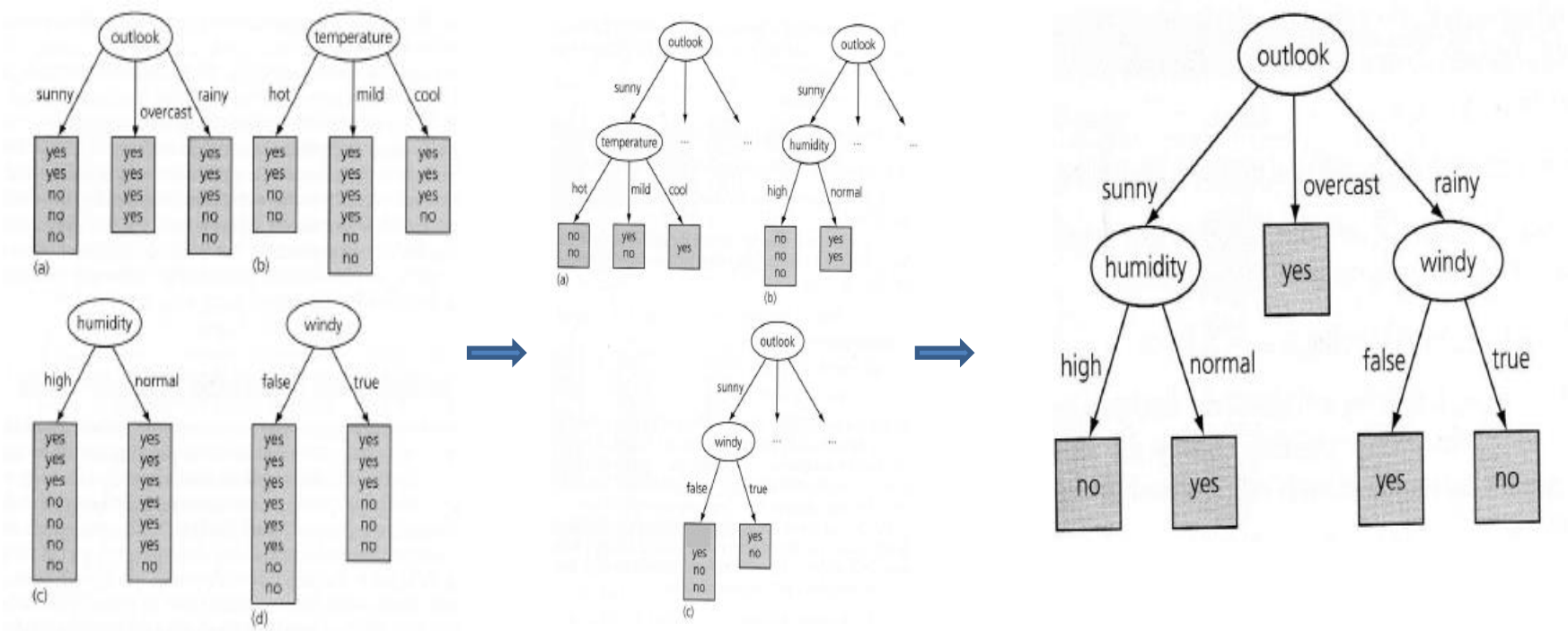
A	B	A xor B
F	F	F
F	T	T
T	F	T
T	T	F



Es una estrategia de aprendizaje inductivo

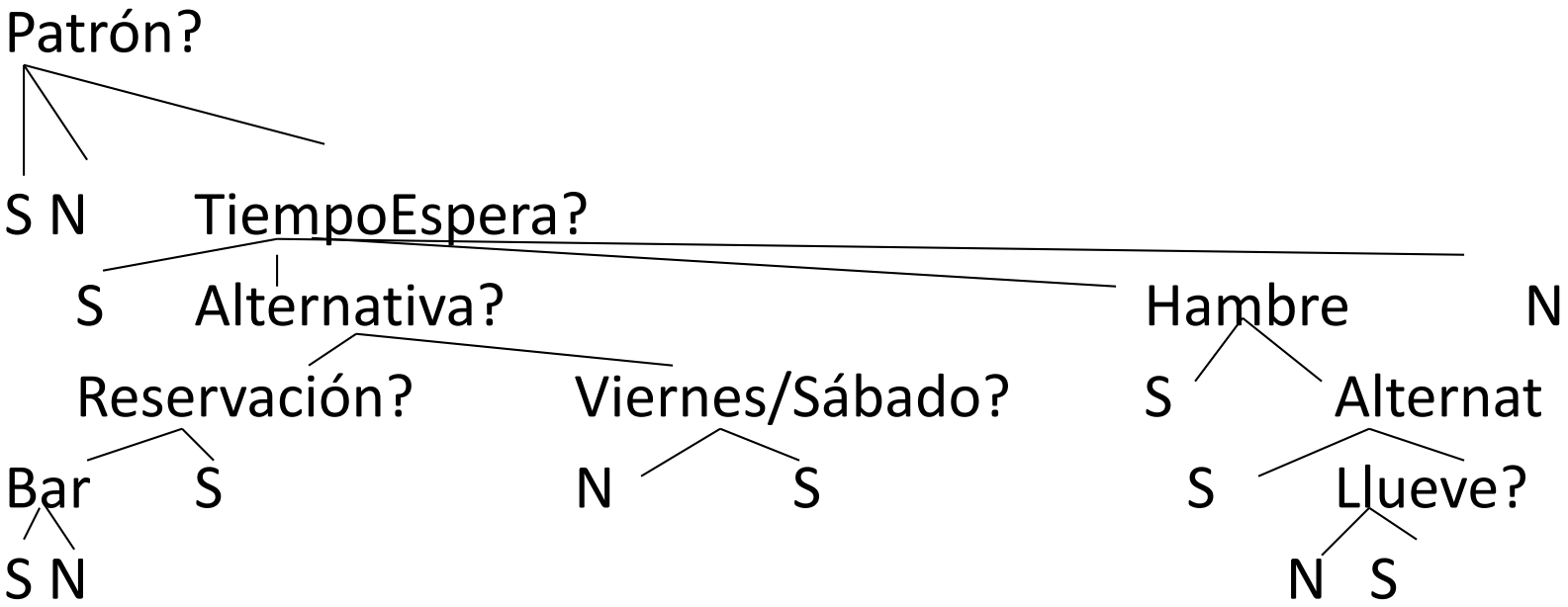
Arboles de decisión

Transformaciones



Árbol de Decisión

Para nuestro ejemplo inicial:



Árbol de Decisión

- Idea: escoger atributo "más significativo" como raíz del (sub)-árbol

¿Cómo?

- Si hay + y - ejemplos escoger atributo que mejor los divida (mayor discriminante)
- Si hay particiones con + y -, buscar un 2do atributo para seguir partiendo

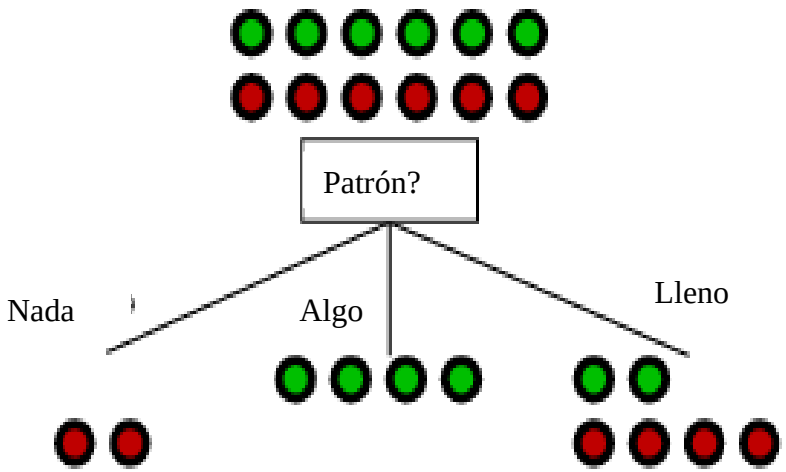
Macroalgoritmo AD(ejemplos, atributos)

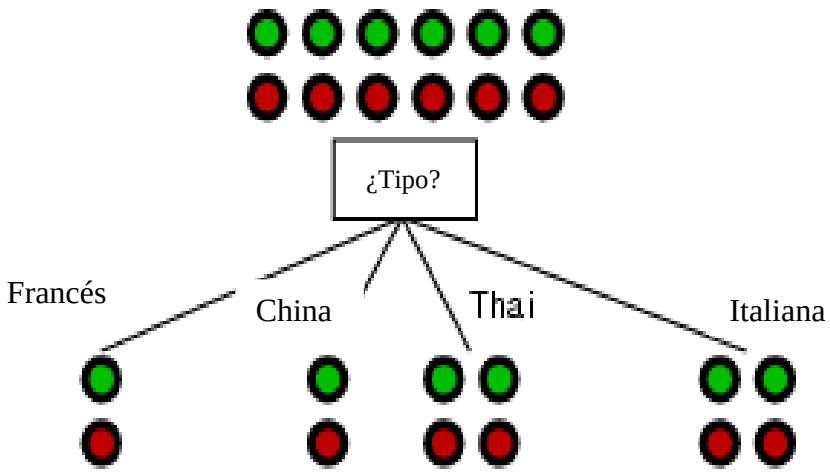
```
Si ejemplos no vacios entonces
    Si ejemplos clasificados entonces
        regresar (clasificación)
    de lo contrario
        mejor: escoger_atributo(atributos, ejemplos)
        arbol: un nuevo árbol de decisión con mejor como raíz
        por cada valor Vi de mejor
            Subejemplos:ejemplos con mejor=Vi
            Subarbol: AD(Subejemplos, atributos)
            Arbol: actualizar(nueva rama con etiqueta Vi y Subarbol)
Regresa(árbol)
```

Escoger un atributo

aprender reglas (clases)

¿***Patrón*** es una mejor escogencia que ***Tipo***?





Basado en conceptos vinculados a *contenidos de información,* p.ej.:

$$Info(p,n) = -p\log_2(p) - n\log_2(n)$$

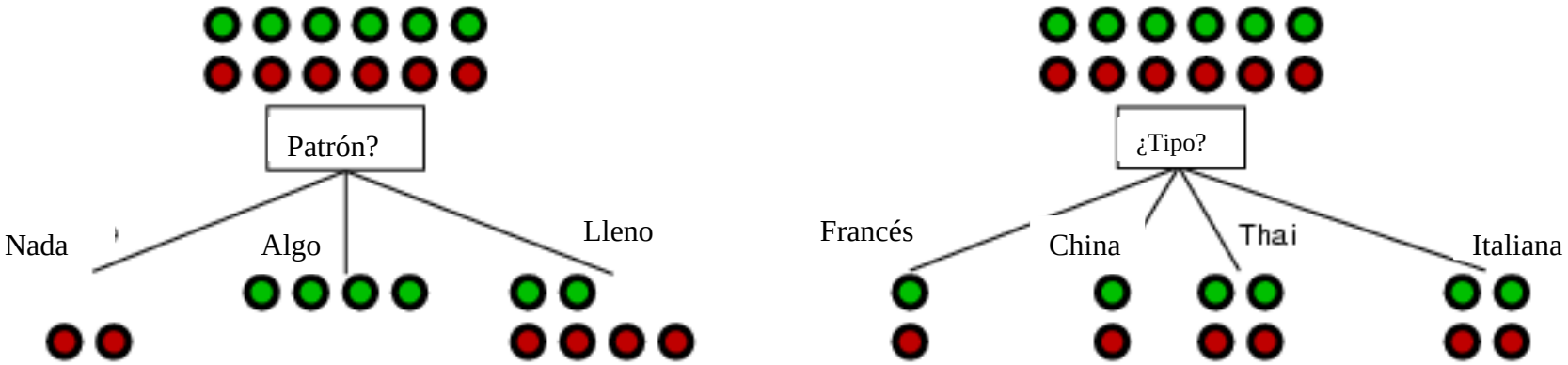
Es una medida de la entropía (grado de desorden) de los ejemplos

n: numero de ejemplos - p: numero de ejemplos +

Escoger un atributo

aprender reglas (clases)

¿***Patrón*** es una mejor escogencia que ***Tipo***?



Escoger atributo *A* con mas grande IG (ganancia en información)

$$IG(A) = I(\frac{p}{p+n}, \frac{n}{p+n}) - resto(A)$$

Donde:

I es entropía de los ejemplos:

$$I(\frac{p}{p+n}, \frac{n}{p+n}) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

y

$$\text{resto}(A) = \sum_{i=1}^{v} \left|\frac{p_i - n_i}{p+n}\right| I(\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i})$$

v: posibles valores de A

p_i y n_i? ver siguiente lamina

Escoger un atributo

aprender reglas (clases)

¿Quién es p_i? p_i puede ser $p_i = \frac{\left|E_i^+\right|}{\left|E_i^+\right| + \left|E_i^-\right|}$

Donde E_i^+ es el porcentaje de ejemplos clasificados como + por el valor v_i del atributo A

Una Formula general para escoger a los atributos:

Como hay que elegir el atributo con mayor información (menor entropía), otra posibilidad es calcular una **función de merito (FM)**

$$FM(A) = \sum_{i=1}^{v} r_i \inf o(p_i, n_i)$$

p_i = % ejemplos clasificados como + en la rama i

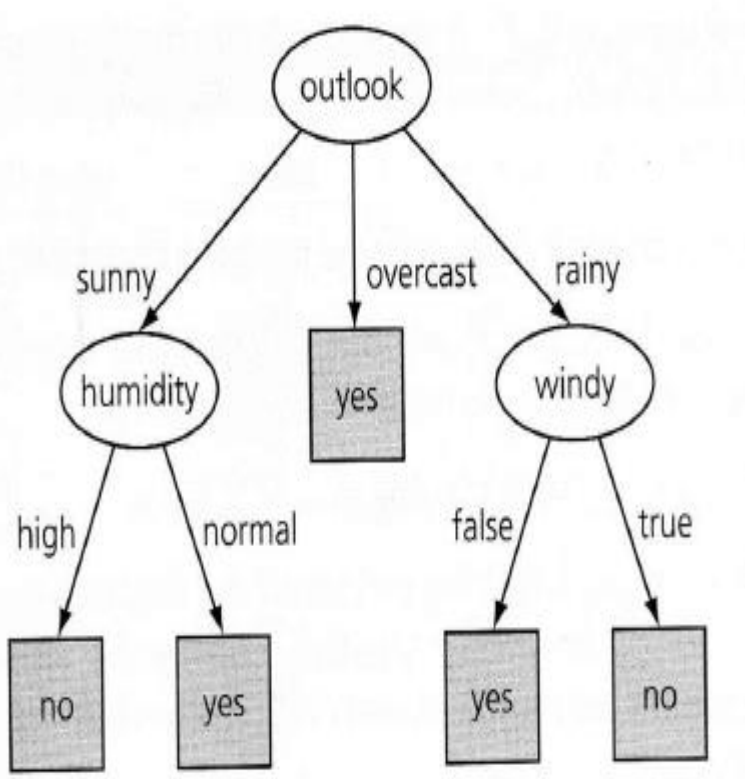
$$r_i = \left| \frac{p_i - n_i}{p + n} \right|$$

Construcción de árboles de decisión

Se completa el árbol completando cada rama hasta cumplir un ciertos compromisos:

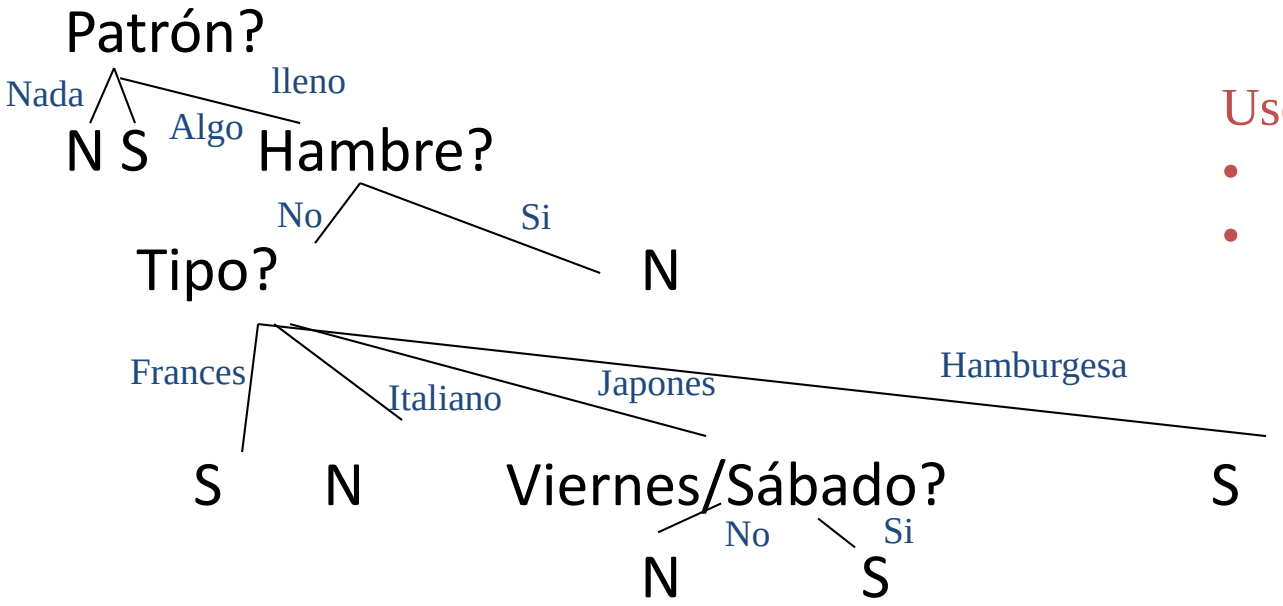
- Número mínimo de hojas.
- Cobertura: Mínimo número (o porcentaje) de casos posibles cubiertos correctamente de la BD.

- Precisión: Error de clasificación menor de un umbral puesto. Por ejemplo: precisión del 80%. Significa, que pararemos en esa hoja cuando el número de clases clasificadas correctamente sea mayor o igual al 80%.



Arbol de Decisión y Lógica de Predicado

∀r espera(r) => Patrón(r, algo) O (Patrón(r, full) Y NoHambre(r) Y tipo(r, francés)) O (Patrón(r, full) Y NoHambre(r) Y tipo(r, hamburguesa)) O (Patrón(r, full) Y NoHambre(r) Y tipo(r, Japones) Y viernes/Sabado(r))

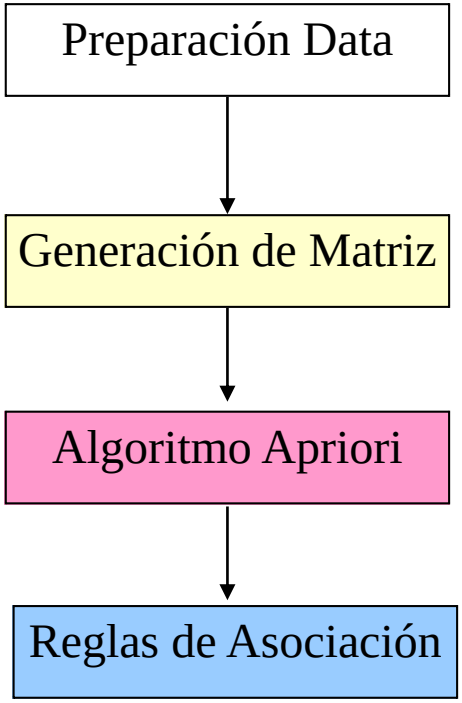


Uso de operadores:

- Para unir ramas O
- Para seguir una rama Y

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_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

Registro_Paginas

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+…+0)/# páginas

Reglas de Asociación

X → Y

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

Métricas

Soporte:

Soporte (X → Y) = Probabilidad (X U Y)

Confianza:

Confianza (X → 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"*

1

Candidatos antecedentes
(Sop > Soporte)

2

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

S1 S2 ……S5 ……Sn

3

1→2 Solo si conf(1→2) > confianza confianza=Prob (X / Y)

4

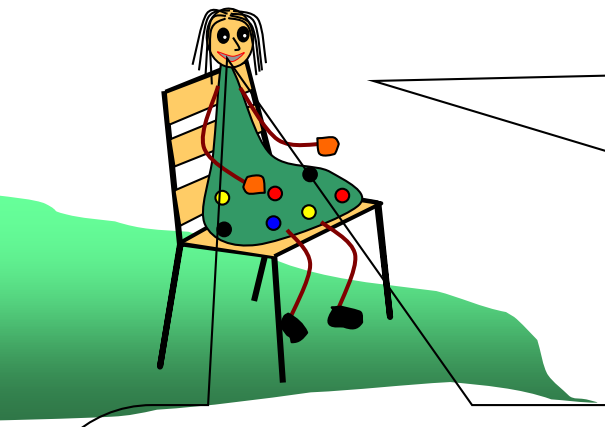
Confianza: hallar esa prob. condicional desde los nodos soportes que pasen un umbral

Reglas de Asociación

Nº	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 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 Prediciones

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 Prediciones

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

Error cuadrático medio:

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

Raíz cuadrada del MSE

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

Error absoluto medio

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

donde $e_{t+h,t}=y_{t+h}-y_{t+h,t}$ 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.
- Compara la predicción con valores reales (error del pronóstico)

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

ATRIBUTE	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

Prediccion 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	Recall
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

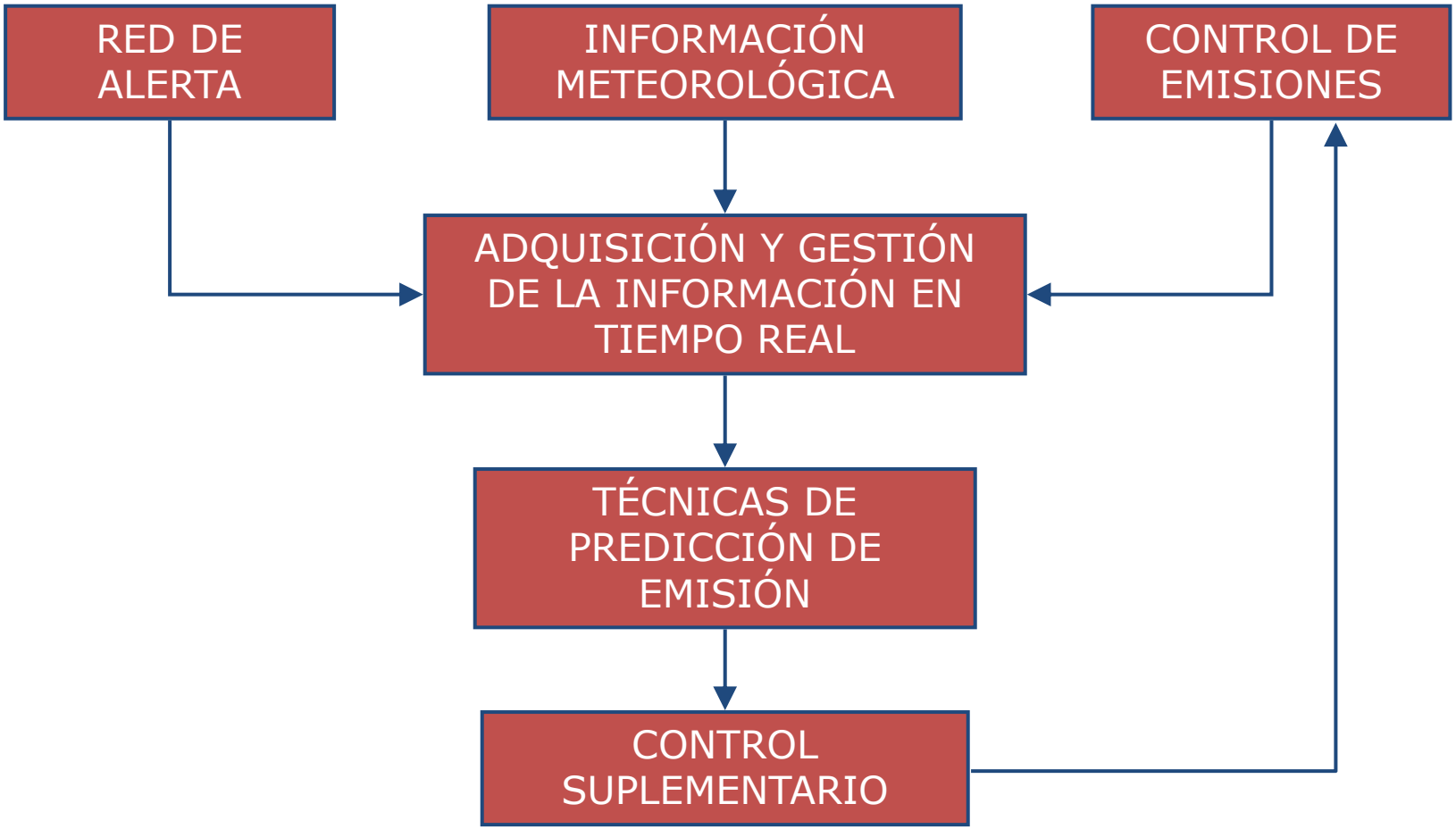
- **Objetivo:** predecir con media hora de anticipación la evolución de los niveles de dióxido de azufre en un entorno y sugerir una línea de actuación

- Utilización de modelos estadísticos a partir de la información en tiempo real de emisiones, calidad de aire y meteorología

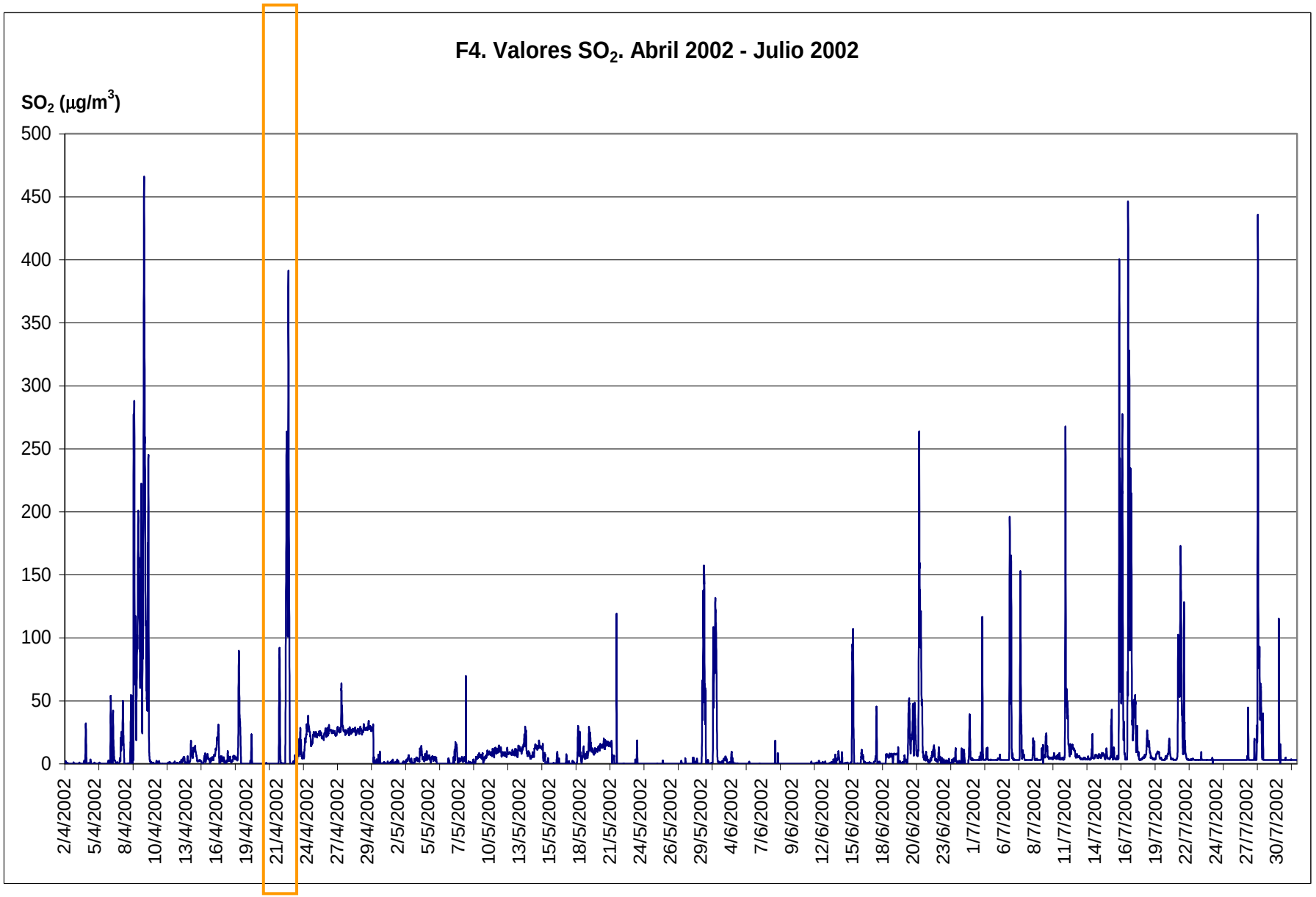
- Se recogen datos de calidad de aire, en particular de SO_2, de las estaciones de la Red de Vigilancia de Calidad de Aire (*frecuencia pentaminutal*).

-

Sistemas de control suplementario de la contaminación atmosférica



Datos


F4. Valores SO2. Abril 2002 - Julio 2002
SO2 (µg/m3)
500
450
400
350
300
250
200
150
100
50
0
2/4/2002
5/4/2002
8/4/2002
10/4/2002
13/4/2002
16/4/2002
19/4/2002
21/4/2002
24/4/2002
27/4/2002
29/4/2002
2/5/2002
5/5/2002
7/5/2002
10/5/2002
13/5/2002
15/5/2002
18/5/2002
21/5/2002
24/5/2002
26/5/2002
29/5/2002
4/6/2002
7/6/2002
9/6/2002
12/6/2002
15/6/2002
18/6/2002
20/6/2002
23/6/2002
1/7/2002
6/7/2002
8/7/2002
11/7/2002
14/7/2002
16/7/2002
19/7/2002
22/7/2002
24/7/2002
27/7/2002
30/7/2002

Principal objetivo:

prevenir episodios de alteración de la calidad del aire

- La instalación necesita disponer de información, al menos, con ***½ hora*** de antelación.

Solución:

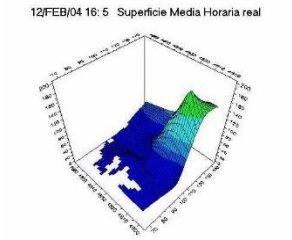
Herramientas de predicción de valores de SO_2, en media horaria, basados en modelos estadísticos

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



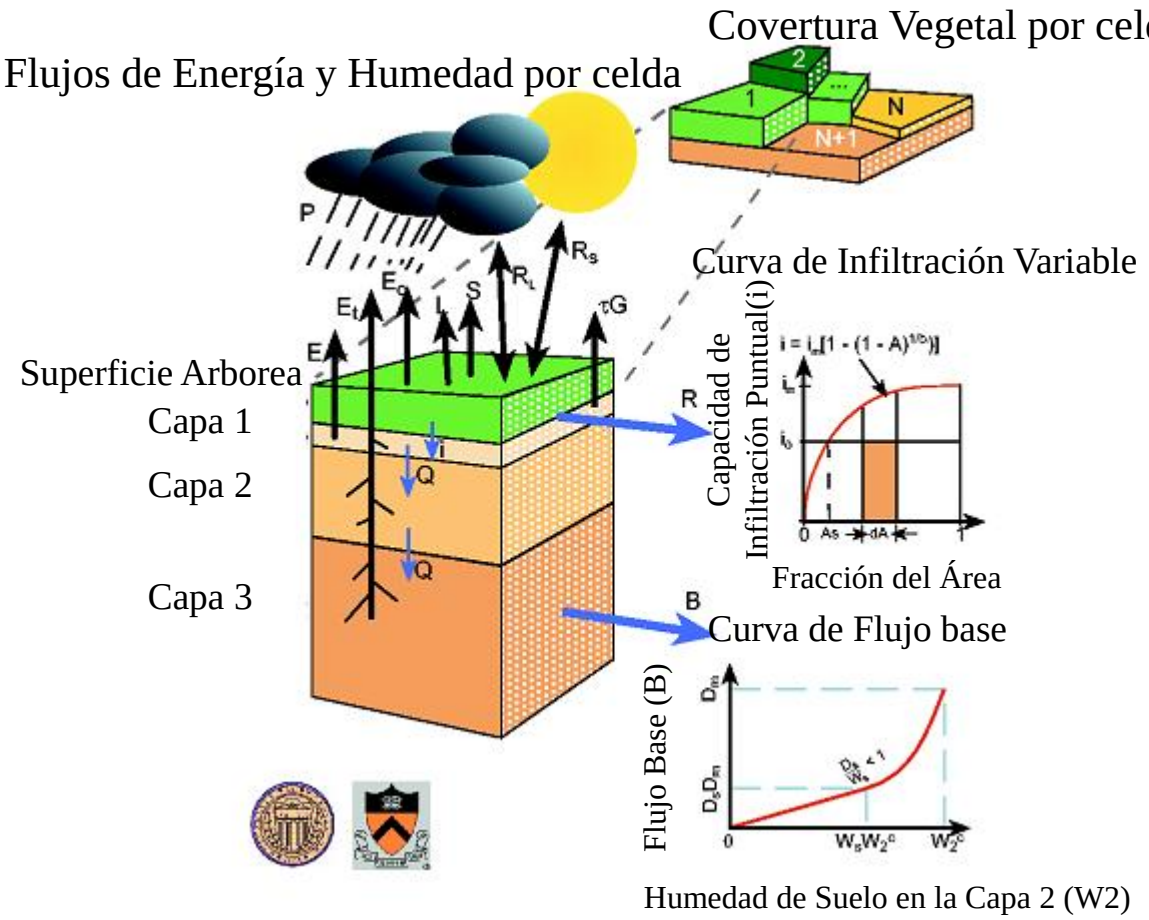
II. Predicción funcional: curva completa de niveles de SO_2 para un cierto intervalo de tiempo

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

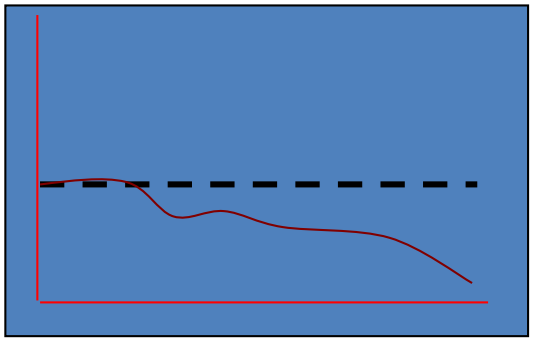
- Modelo Hidrológico Superficial de Capacidades de Infiltración Variable
- Condiciones cuasi actuales
- Índices de Sequía (SMI, SRI, y percentiles)

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)

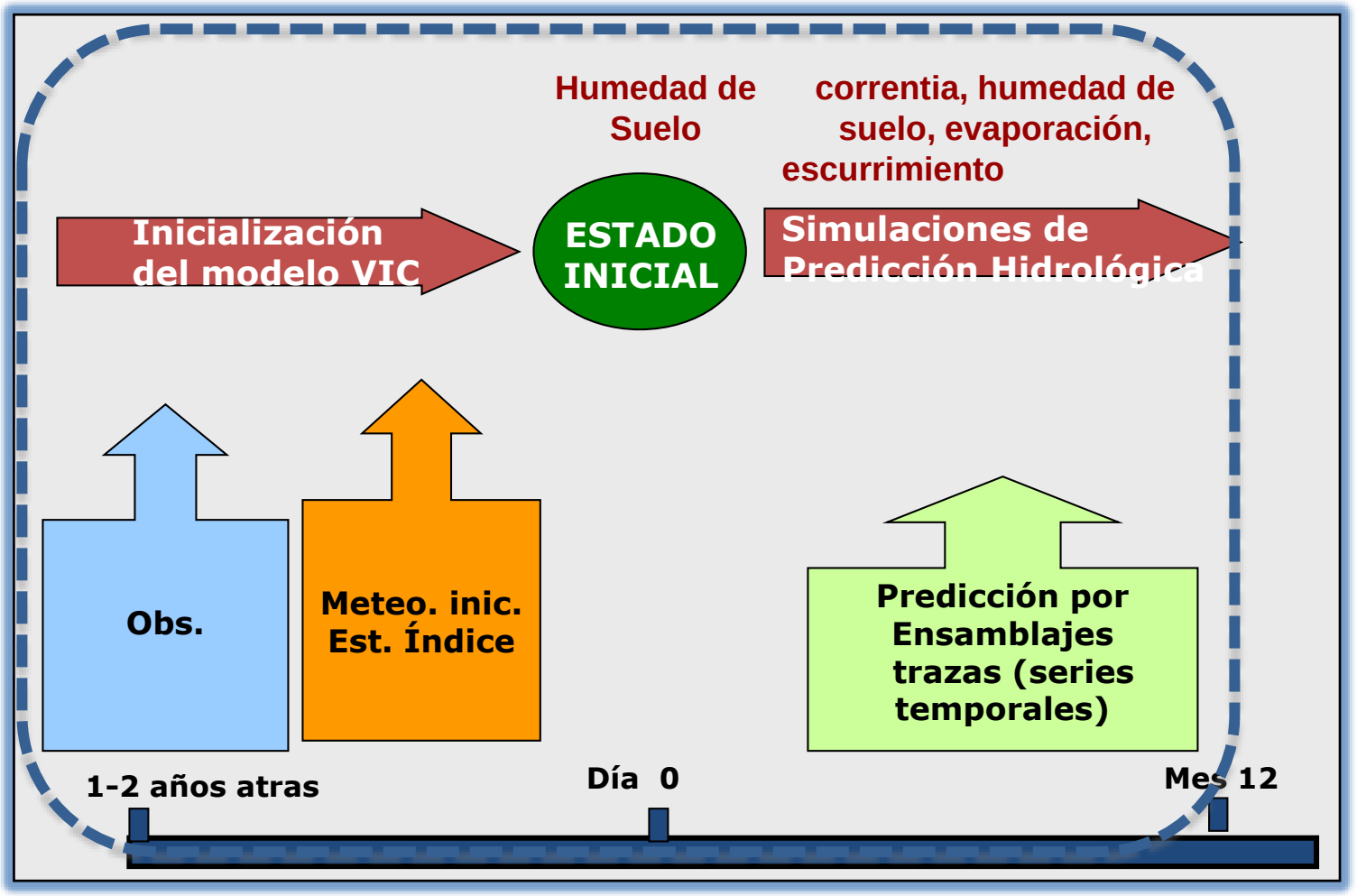
INDICE ESTANDARIZADO DE PRECIPITACIÓN (SPI)

INDICE DE ABASTECIMIENTO DE AGUA SUPERFICIAL (SWSI)

INDICES BASADOS EN MODELAJE HIDROLÓGICO

- ✓ *Precipitación*
- ✓ *Precipitación*
- ✓ *Precipitación*
- ✓ *Temperatura*
- ✓ *Esquema complejo de balance*
- ✓ *de Energía y agua*

Sistema de Predicción Hidrológica Extendido de la UW



PRONÓSTICO DEL TIEMPO

PRONÓSTICO DEL TIEMPO

Pronóstico para 10 días

Madrid, España

Ultima 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.

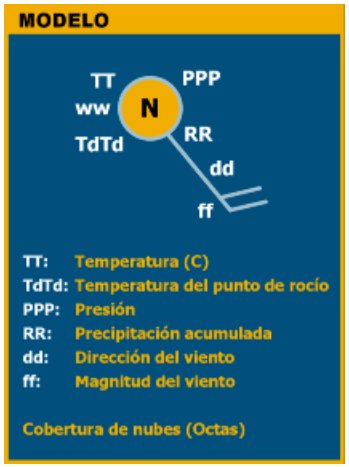
Predicción Numérica

- 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

OPTIMIZACIÓN

OPTIMIZACIÓN

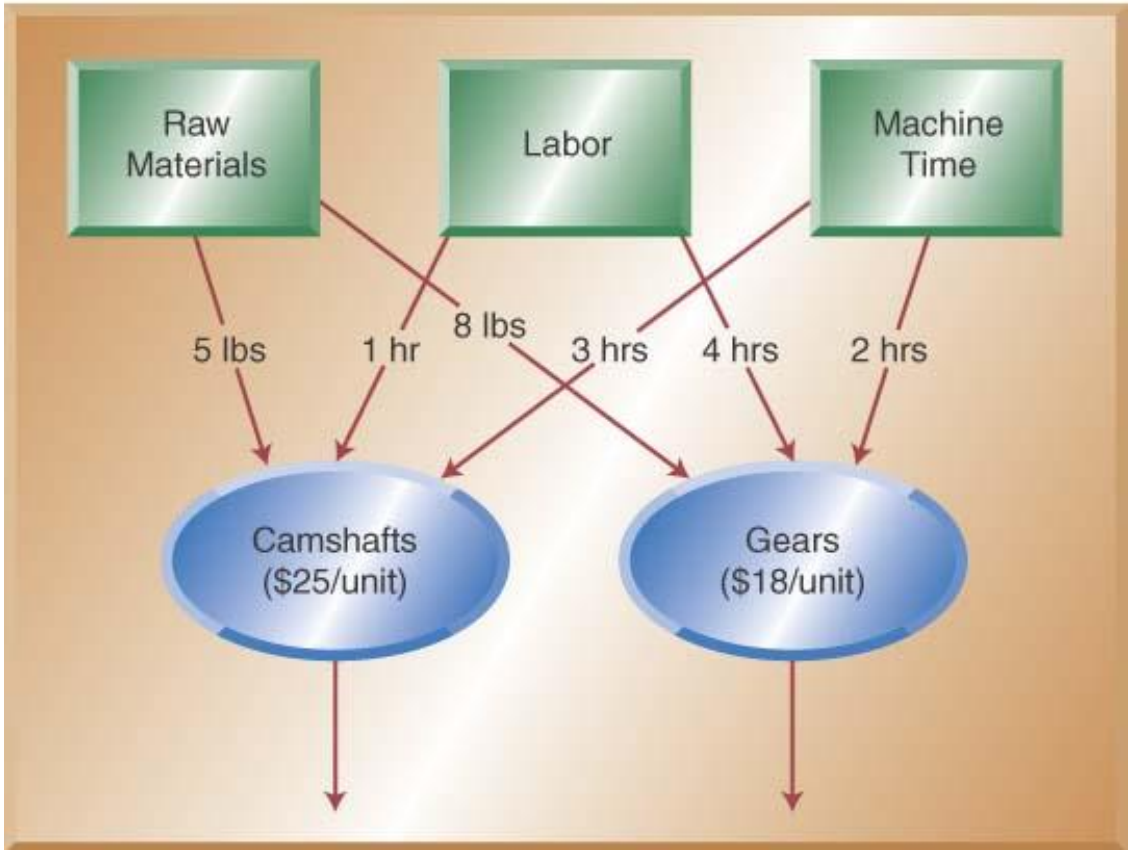
- Comprender el problema;
- Formular el problema en palabras
- Formular algebraicamente el problema
- Definir las variables de decisión
- Escribir la función objetivo en términos de las variables de decisión
- Escribir las restricciones en términos de las variables de decisión

Ejemplo: decisión de mezcla de producto

- DJJ Enterprises hace piezas automotrices, árboles de levas y engranajes
- Beneficio unitario: árboles de levas $ 25 / unidad, engranajes $ 18 / unidad
- Recursos necesarios: acero, mano de obra, tiempo de la máquina.
- En total, 5000 libras de acero disponibles, 1500 horas de mano de obra y 1000 horas de máquina.
- Los árboles de levas necesitan 5 libras de acero, 1 hora de mano de obra, 3 horas de tiempo de máquina.
- Los engranajes necesitan 8 libras de acero, 4 horas de trabajo, 2 horas de tiempo de máquina.

¿Cuántos árboles de levas y engranajes debe hacer para maximizar las ganancias?

Comprender el problema



Formular pb:

- **Variables de decisión:** número de árboles de levas a hacer, número de engranajes a hacer
- **Función Objetivo:** Maximizar ganancias Restricciones, pero sin exceder las cantidades disponibles de acero, mano de obra y tiempo de la máquina.

Formulación algebraica

Variables de decisión

- C = número de árboles de levas a hacer
- G = número de engranajes a hacer

Función objetivo

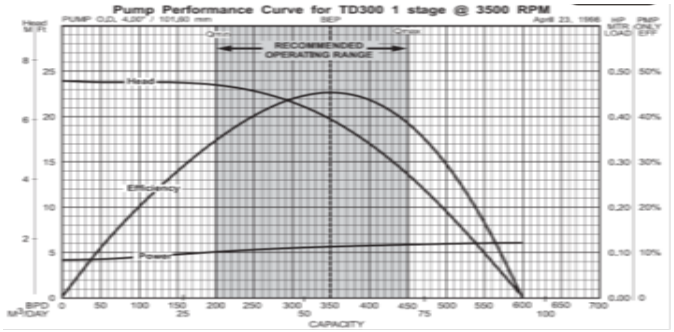
- Maximizar 25C + 18G (ganancia en $)

Restricciones

- 5C + 8G <= 5000 (acero en libras)
- 1C + 4G <= 1500 (trabajo en horas)
- 3C + 2G <= 1000 (tiempo de la máquina en horas)
- C> = 0, G> = 0 (no negatividad)

Proceso de optimización

de los pozos de petróleo implican una función objetiva que maximiza la producción y minimiza la energía de la elevación.



Generación del modelo de producción de pozos

$$\frac{dq}{dt} = \frac{1}{M}(P_{bh} - P_{wh} - (\rho_1 g h_1 + \rho_2 g h_2) + \Delta P_p - \Delta P_f)$$

Pbh = Bottom Pressure
Pwh = Head Pressure
β1, β2 = Volumetric Module
V1, V2 = Volume Multiphase
qr = Input Flow
q = Flow of well

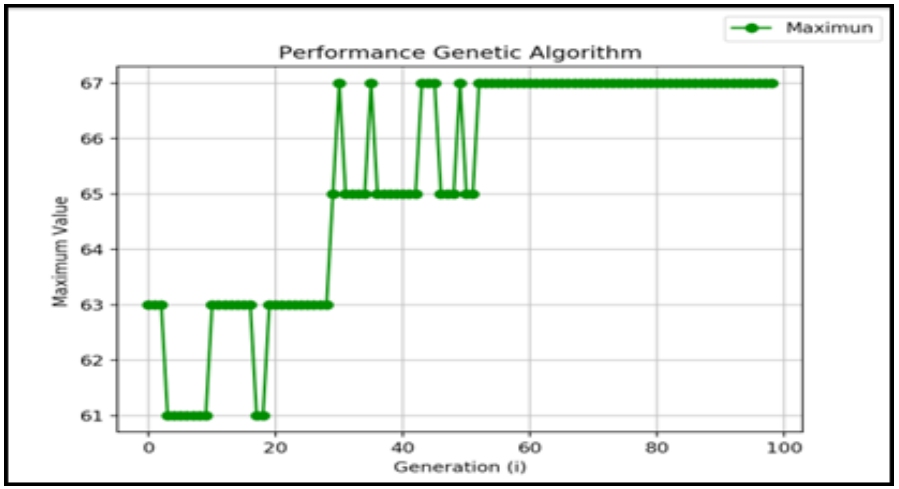
qc = Flow of Valve
ρ1, ρ2 = Density of flow
g = gravity
h1,h2 = Vertical Distances
ΔPp = Differential of Pressures
ΔPf = Total Pressure Loss in the Well

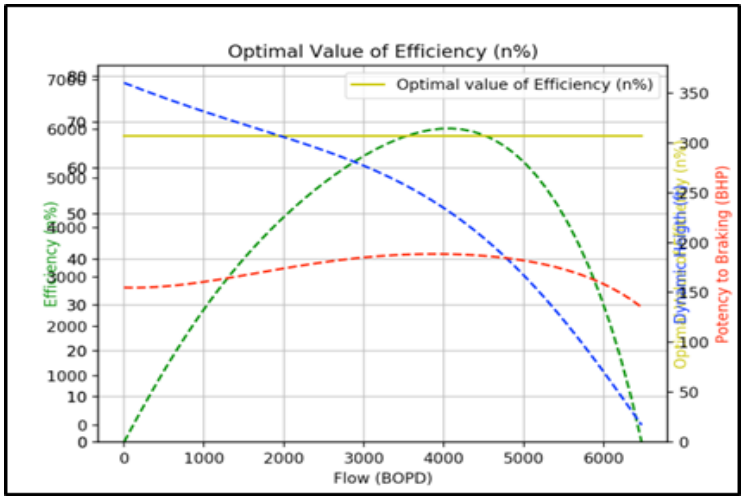
Procesos de optimización

maximiza la producción y minimiza la energía de la elevación.

Función objetivo

$$\eta(\%) = \frac{\rho g Q H}{Potencypowerforthepump * 1{,}000W \ / \ KW}$$



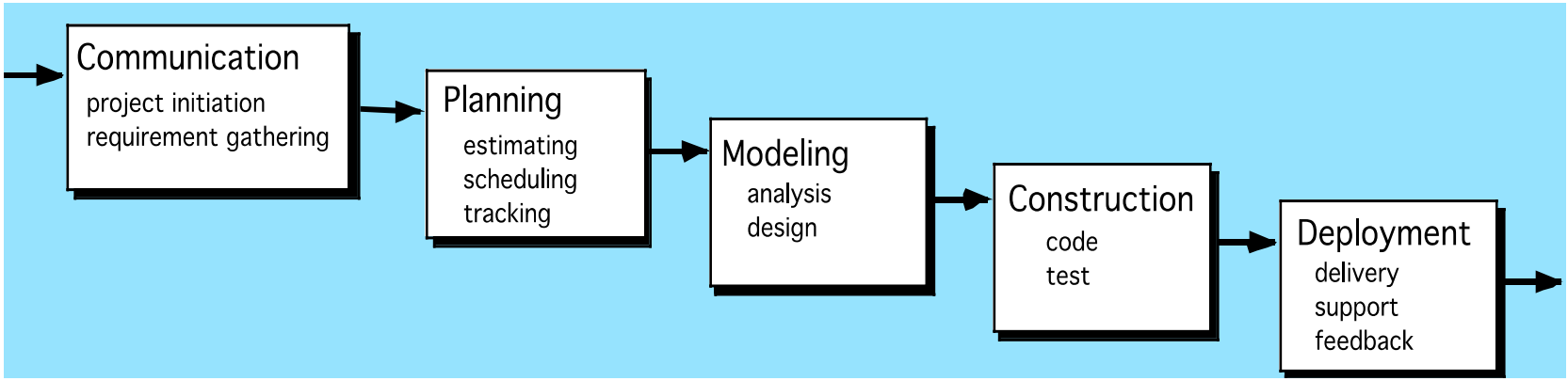


PRESCRIPTIVO

Modelos prescriptivos

- Los modelos de procesos prescriptivos abogan por un enfoque ordenado
  Eso lleva a algunas preguntas ...
- Si los modelos de procesos prescriptivos luchan por la estructura y el orden, ¿son inapropiados para un mundo que prospera con el cambio?
- Sin embargo, si rechazamos los modelos de procesos tradicionales (y el orden que implican) y los reemplazamos por algo menos estructurado, ¿hacemos imposible lograr coordinación y coherencia en el trabajo?

El modelo de cascada



Es el paradigma más antiguo

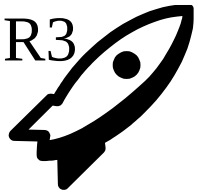
Cuando los requisitos están bien definidos y razonablemente estables, esto conduce a un proceso lineal.

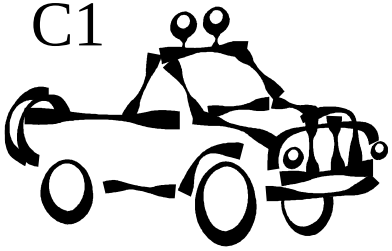
Problemas: 1. raramente lineal, iteración necesaria. 2. difícil de establecer todos los requisitos explícitamente. Bloqueo de estado. 3. código no se lanza hasta muy tarde.

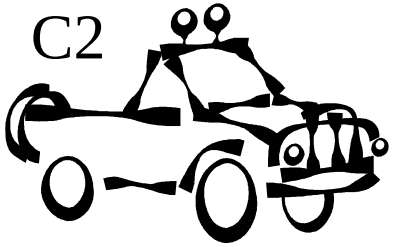
El ciclo clásico sugiere un enfoque sistemático y secuencial

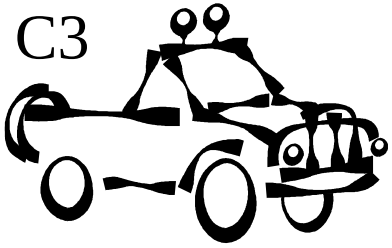
Ejemplo traslado cohete

St. Louis









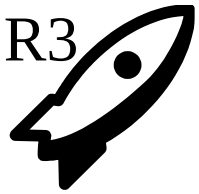
San Francisco

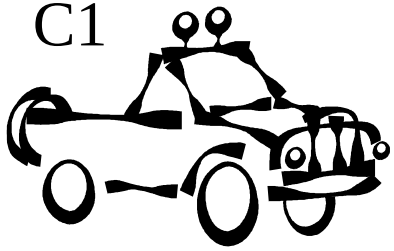
Seattle

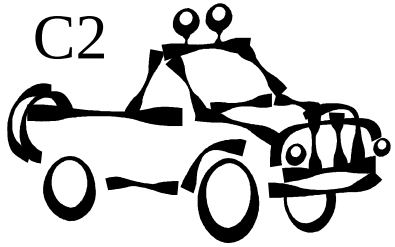
- La solución se puede generalizar en 3 pasos

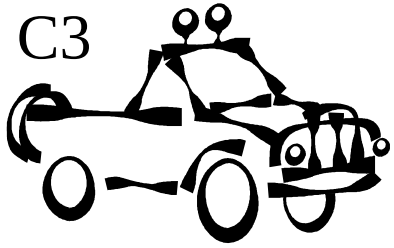
Ejemplo traslado cohete

St. Louis









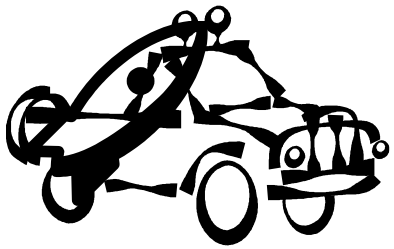
San Francisco

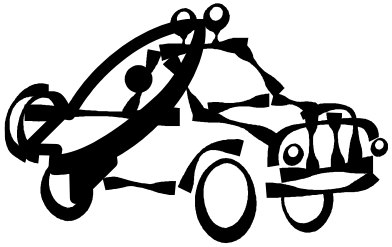
Seattle

- Paso 1: Cargar todos los cohetes

Ejemplo traslado cohete

St. Louis





San Francisco

Seattle

- Paso 2: Mueva todos los cohetes

Ejemplo traslado cohete

St. Louis

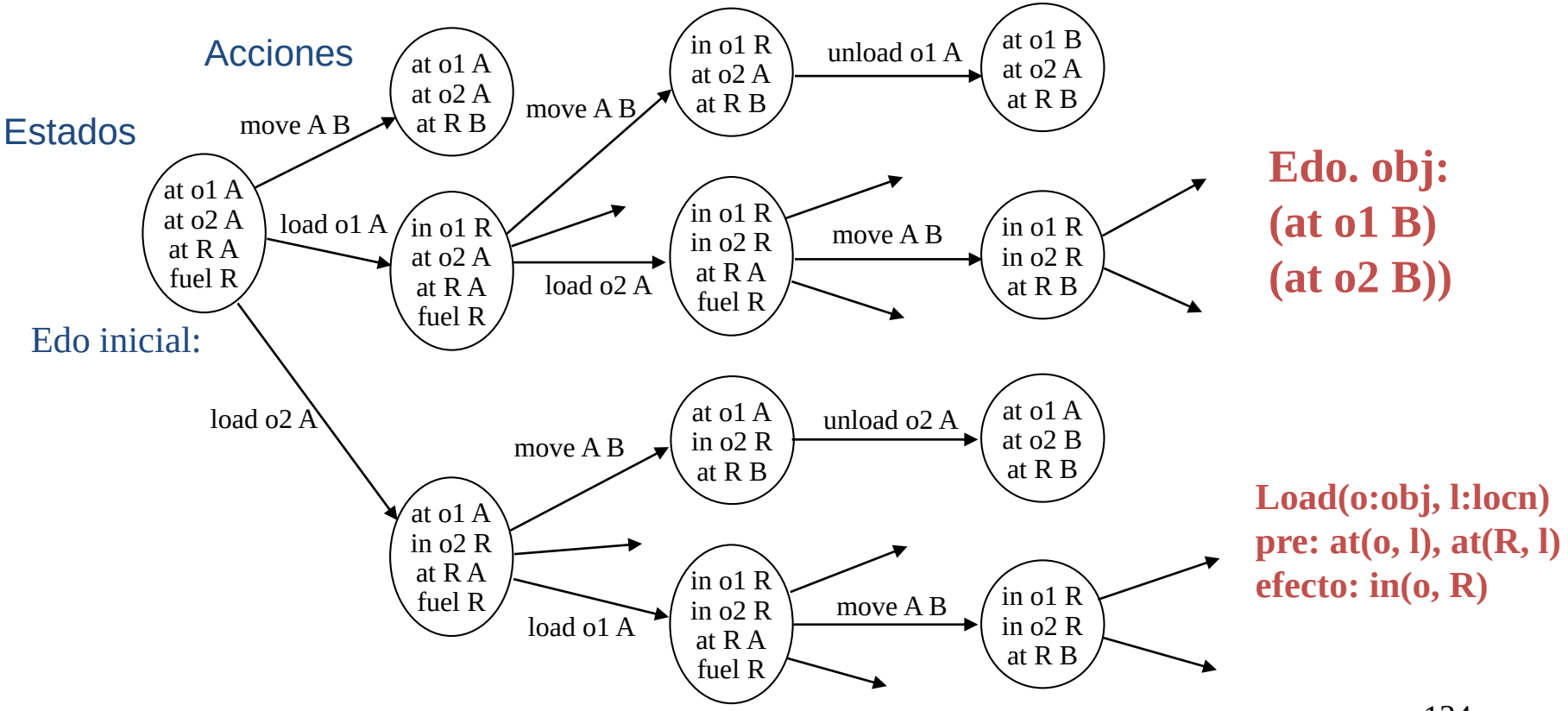
San Francisco

Seattle

- Paso 3: lanzar todos los cohetes

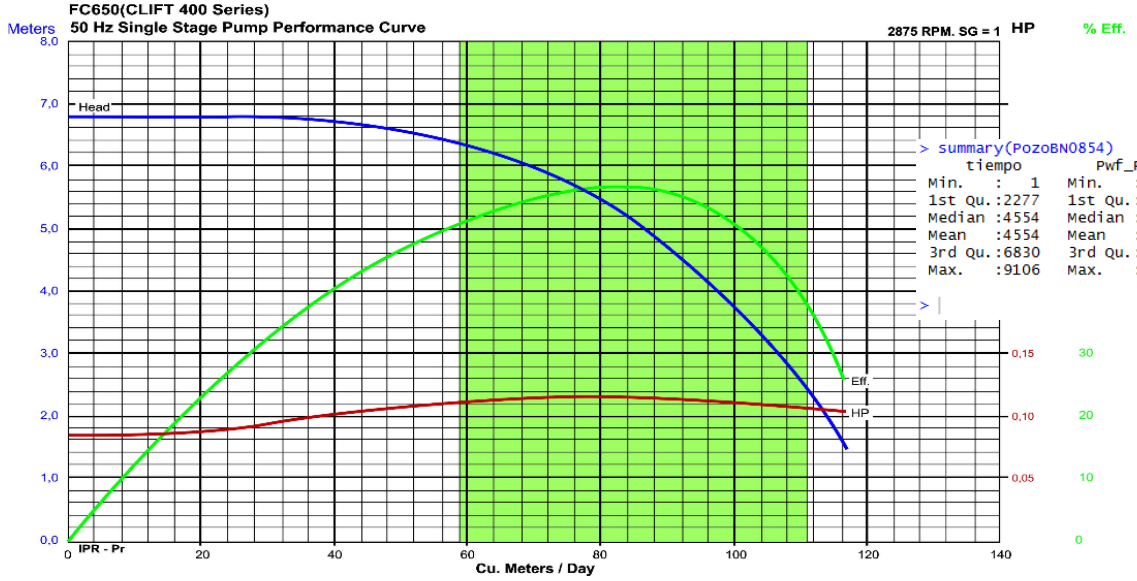
Plan

- Encontrar todas las metas alcanzables desde el estado inicial
- Exponencial en tiempo y memoria



IDENTIFICACIÓN

desarrollo de un modelo matemático que describe la eficiencia de un ESP, utilizando sus variables operativas, como presión de flujo (Pwf) del pozo, presión en la cabeza de producción (Php) , temperatura en el Cabezal de producción (Thp) y flujo producido (Qprod), capturados en el campo por sensores de temperatura y presión. El comportamiento del sistema ESP se puede describir mediante una serie de curvas características, como carga vs. flujo, eficiencia vs. flujo, potencia vs. flujo



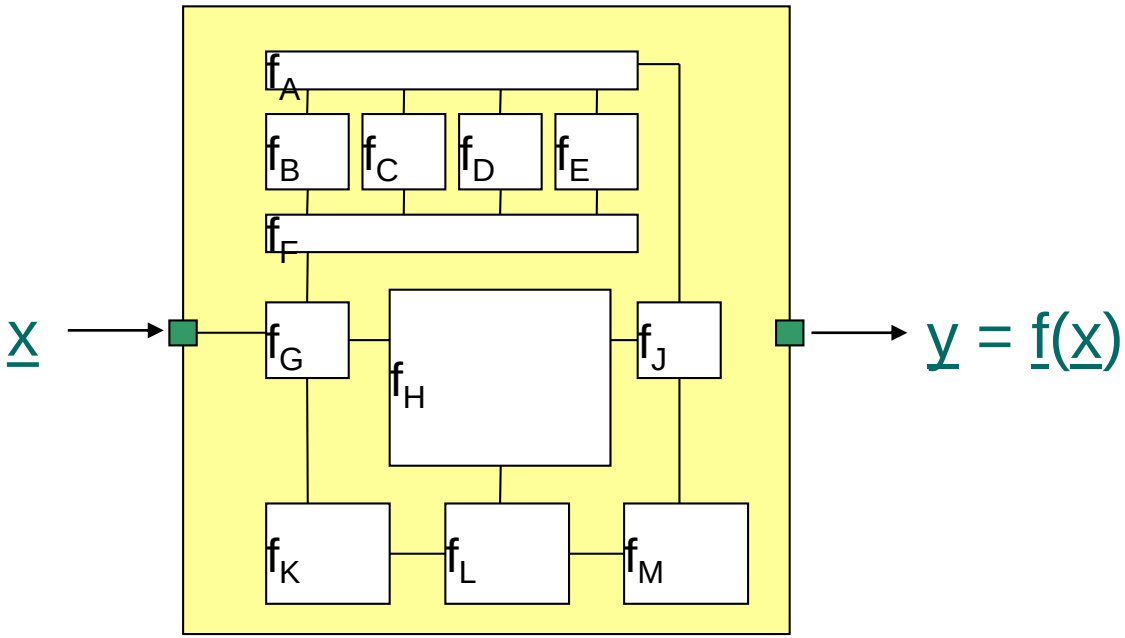

```
> summary(PozoBN0854)
     tiempo         Pwf_Psi         Thp_Psi         Qprod_Bpd        Frec_Hz         Plp_Psi
Min.   :   1   Min.   :639.8   Min.   :117.9   Min.   :494.4   Min.   : 7.20   Min.   : 690.0
1st Qu.:2277   1st Qu.:639.8   1st Qu.:149.2   1st Qu.:866.3   1st Qu.:53.90   1st Qu.: 800.0
Median :4554   Median :639.8   Median :149.2   Median :866.3   Median :53.90   Median : 820.0
Mean   :4554   Mean   :640.1   Mean   :149.1   Mean   :865.7   Mean   :53.93   Mean   : 832.5
3rd Qu.:6830   3rd Qu.:639.8   3rd Qu.:149.2   3rd Qu.:866.3   3rd Qu.:54.00   3rd Qu.: 860.0
Max.   :9106   Max.   :784.4   Max.   :159.3   Max.   :872.7   Max.   :54.00   Max.   :1990.0
                                                               NA's   :13
> |
```

RGP Mathematical models	Error	Complexity of the structure	Variables
exp(sqrt(sqrt(5.1154277799651 * sqrt(exp(5.1154277799651 + 6.8138162791729) + (5.1154277799651 * sqrt(exp(5.1154277799651 + 6.8138162791729)) + (Frec_Hz + exp(6.59457715693861) + exp(6.59457715693861))))))) + Frec_Hz	0.15%	High	Frec
Frec_Hz + (Thp_Psi + Pwf_Psi + sqrt(Pwf_Psi))	0.36%	Low	Frec, Thp, Pwf

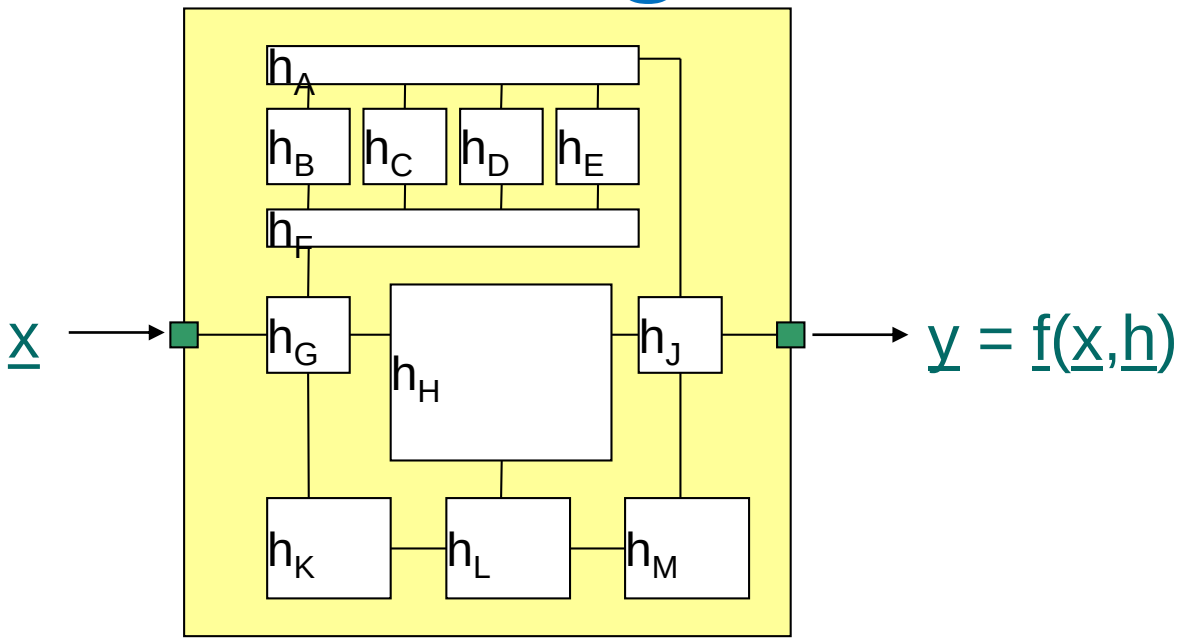
DIAGNÓSTICO

Modelos diagnóstico

funcionalidad nominal



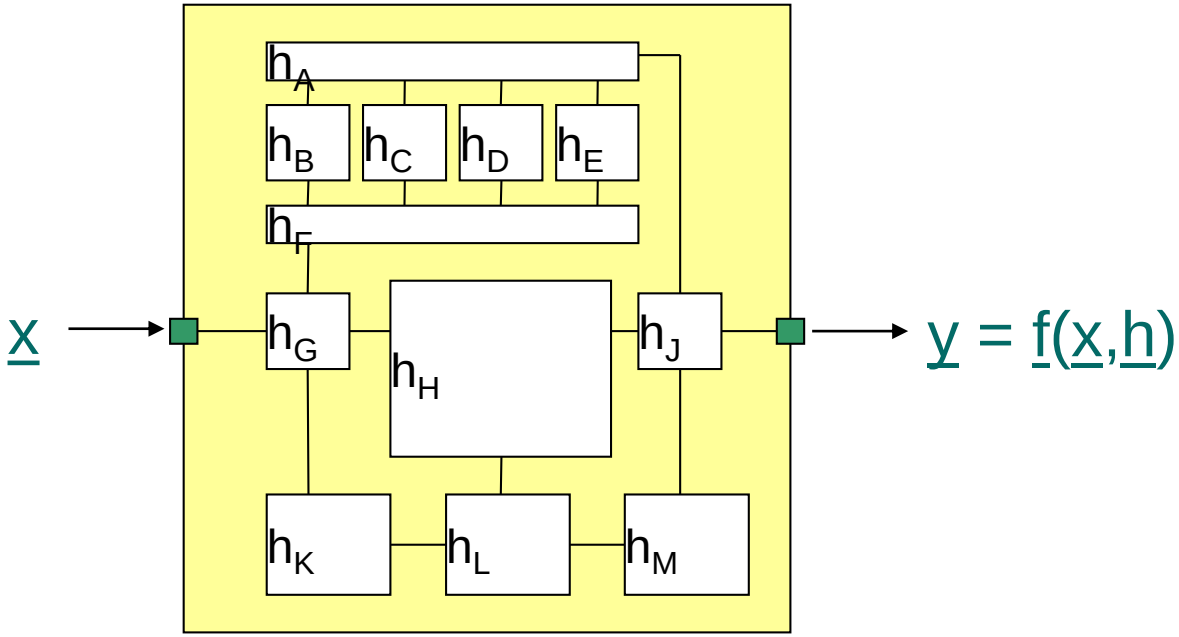
Modelos diagnóstico



$h_i = 1$ significa fi es saludable,
$h_i = 0$ significa fi falla

Nos gustaría encontrar:
$\underline{h} = \underline{f}^{-1}(\underline{x},\underline{y})$
Pero, en general, $\underline{f}^{-1}(\underline{x},\underline{y})$ no puede determinarse.
En la práctica, calculamos soluciones consistentes para h con un algoritmo de búsqueda eficiente.

Modelos diagnóstico



Process:

1. map f to propositional logic
2. observe $\underline{x}$ and $\underline{y}$
3. find all $\underline{h}$ for which $\underline{y} = \underline{f}(\underline{x},\underline{h})$ is consistent
   (i.e., the diagnosis or "numeric solution" for by $\underline{h} = \underline{f}^{-1}(\underline{x},\underline{y})$)

Problema: Control automático de una lavadora

- La naturaleza de las decisiones que realizan los seres humanos en este problema es fácil de entender y modelar.
- **Tarea:** Se desea automatizar la selección del ciclo y el tiempo de lavado basado en la cantidad de ropa y lo sucia que esta la ropa, lo cual es proporcionado por dos transductores.

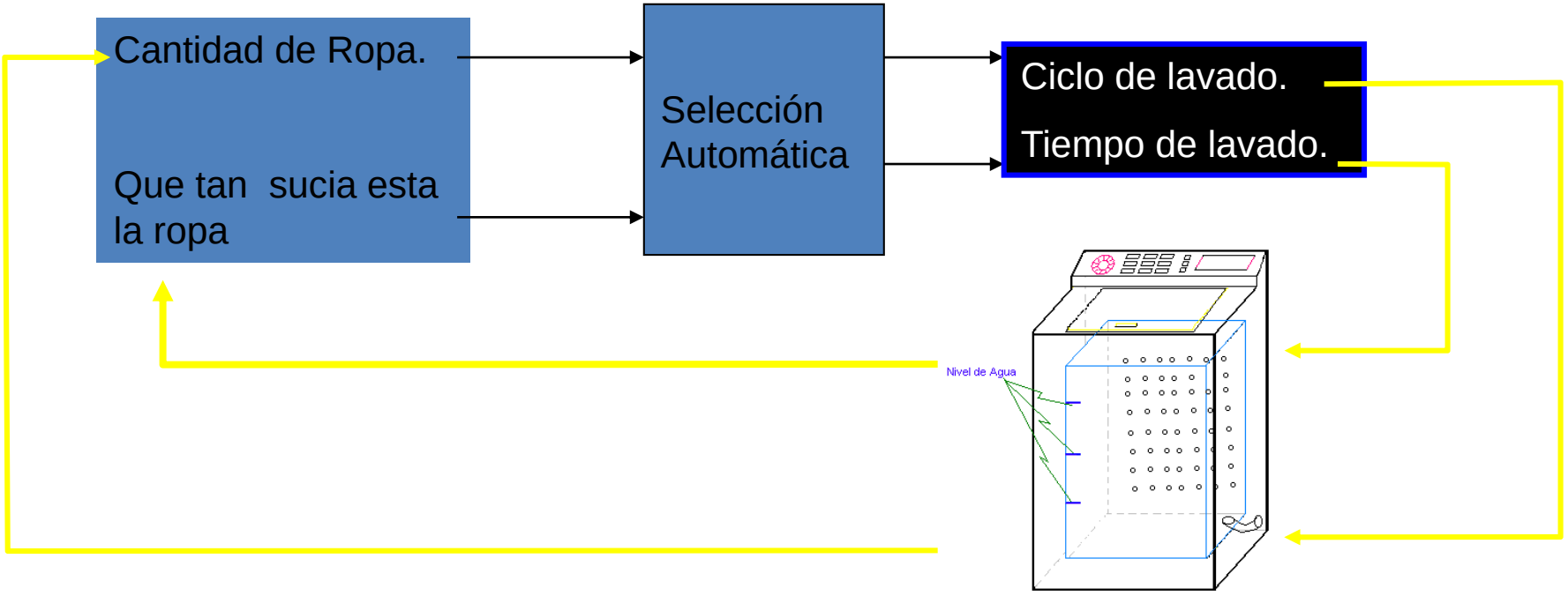


Tabla de Reglas Difusas para el ciclo de lavado

CANT. DE ROPA / MUGROSIDAD	Poca	Medio	Mucha
Suave	Delicado	Ligero	Normal
Normal Suave	Ligero	Normal	Normal
Normal Rudo	Ligero	Normal	Fuerte
Rudo	Ligero	Normal	Fuerte

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.