



DECSAI

Departamento de Ciencias de la Computación e I.A.

Universidad de Granada



Clasificación y predicción

© Fernando Berzal, berzal@acm.org

Clasificación y predicción



- **Introducción**
 - Uso y construcción de modelos de clasificación
 - Evaluación de la precisión de un modelo de clasificación
 - El problema del sobreaprendizaje
- **Modelos de clasificación**
 - Árboles de decisión
 - Inducción de reglas
- **Evaluación**
 - Métricas
 - Métodos de evaluación
- **Técnicas de regresión**
- **Apéndice: Otros modelos de clasificación**

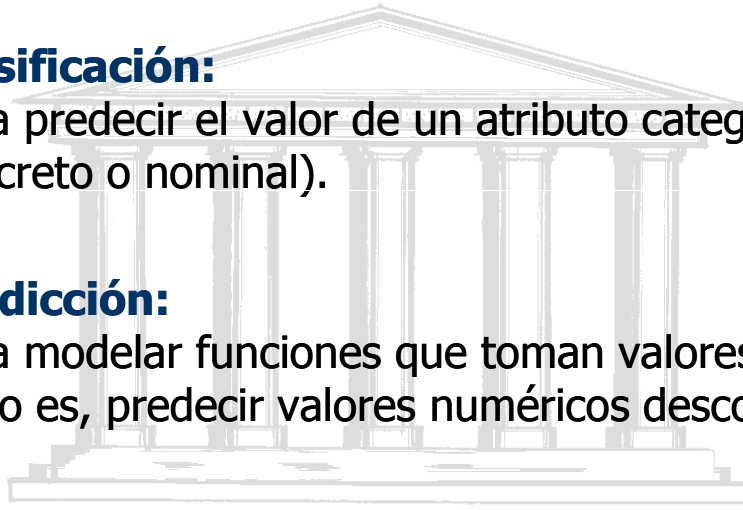


Introducción



Clasificación vs. Predicción

- **Clasificación:**
Para predecir el valor de un atributo categórico (discreto o nominal).
- **Predicción:**
Para modelar funciones que toman valores continuos (esto es, predecir valores numéricos desconocidos).

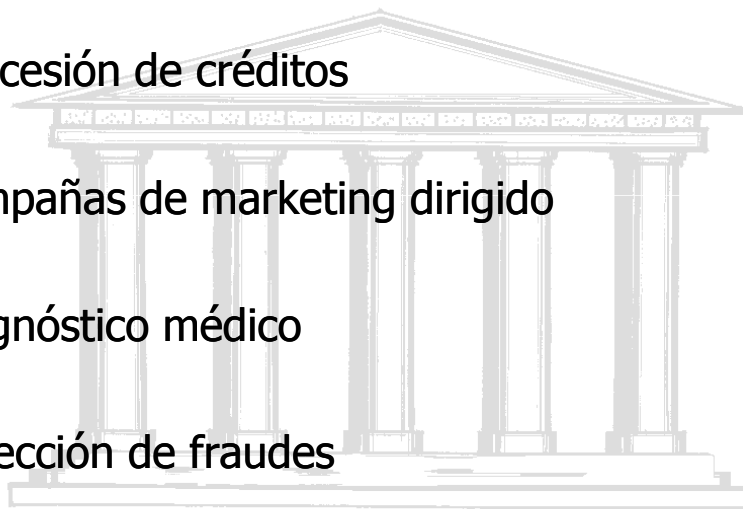


Introducción



Aplicaciones

- Concesión de créditos
- Campañas de marketing dirigido
- Diagnóstico médico
- Detección de fraudes
- ...



Introducción



Construcción del modelo

- El conjunto de datos utilizado para construir el modelo de clasificación se denomina **conjunto de entrenamiento**.
- Cada caso/tupla/muestra corresponde a una clase predeterminada: los casos de entrenamiento vienen etiquetados por su atributo de clase.

Uso del modelo

- El modelo construido a partir del conjunto de entrenamiento se utiliza para clasificar nuevos datos.



Introducción



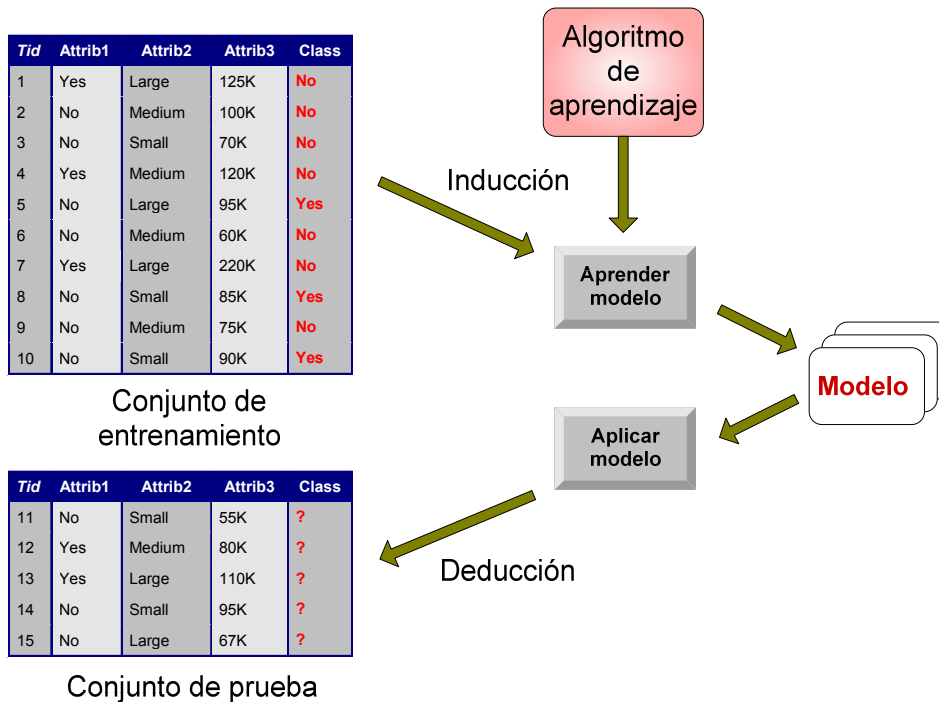
Aprendizaje

Supervisado vs. No Supervisado

- **Aprendizaje supervisado (clasificación)**: Los casos del conjunto de entrenamiento aparecen etiquetados con la clase a la que corresponden.
- **Aprendizaje no supervisado (clustering)**: No se conocen las clases de los casos del conjunto de entrenamiento (ni siquiera su existencia).



Introducción



Introducción



Estimación de la precisión del modelo

Antes de construir el modelo de clasificación, se divide el conjunto de datos disponible en

- un **conjunto de entrenamiento** (para construir el modelo) y
- un **conjunto de prueba** (para evaluar el modelo).





Estimación de la precisión del modelo

- Una vez construido el modelo a partir del conjunto de entrenamiento, se usa dicho modelo para clasificar los datos del conjunto de prueba:
- Comparando los casos etiquetados del conjunto de prueba con el resultado de aplicar el modelo, se obtiene un **porcentaje de clasificación**.
- Si la precisión del clasificador es aceptable, podremos utilizar el modelo para clasificar nuevos casos (de los que desconocemos realmente su clase).



El problema del sobreaprendizaje

- Cuanto mayor sea su complejidad, los modelos de clasificación tienden a ajustarse más al conjunto de entrenamiento utilizado en su construcción (**sobreaprendizaje**), lo que los hace menos útiles para clasificar nuevos datos.
- En consecuencia, el conjunto de prueba debe ser siempre independiente del conjunto de entrenamiento.
- El error de clasificación en el conjunto de entrenamiento **NO** es un buen estimador de la precisión del clasificador.

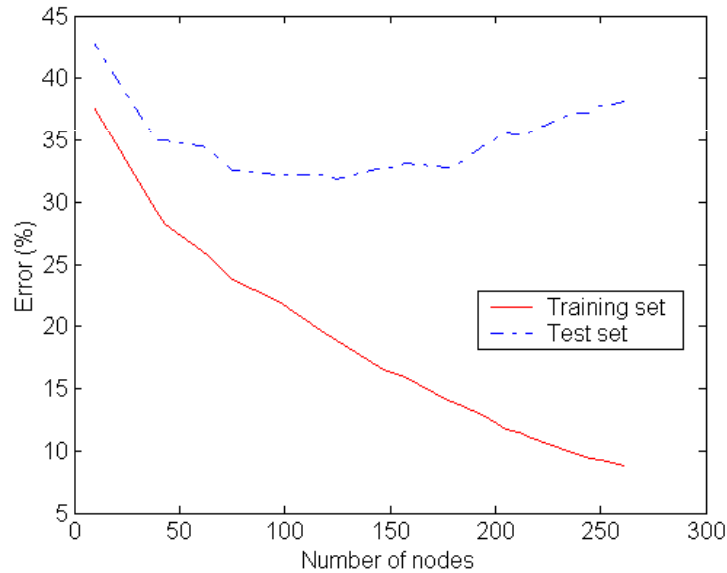


Introducción



Sobreaprendizaje

debido a la complejidad del clasificador

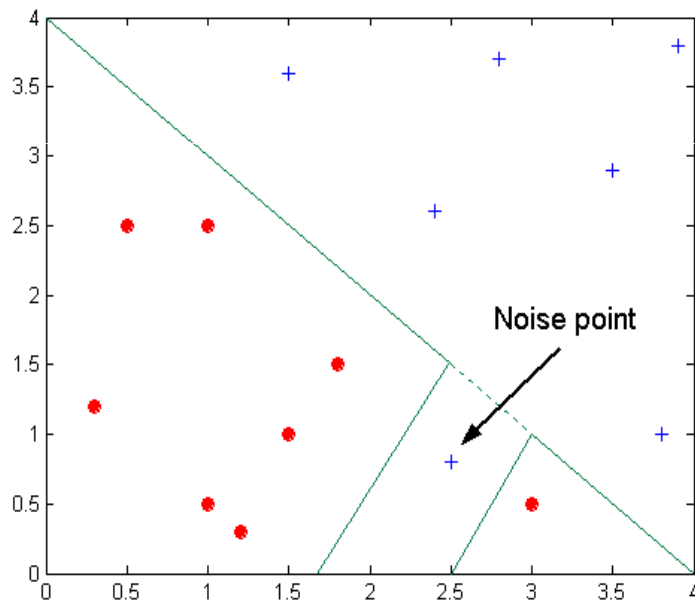


Introducción



Sobreaprendizaje

debido a la presencia de ruido en los datos:

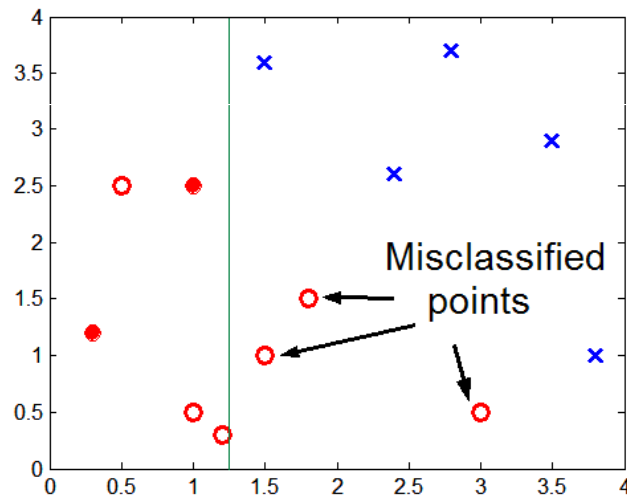


Introducción



Sobreaprendizaje

debido a la escasez de muestras:

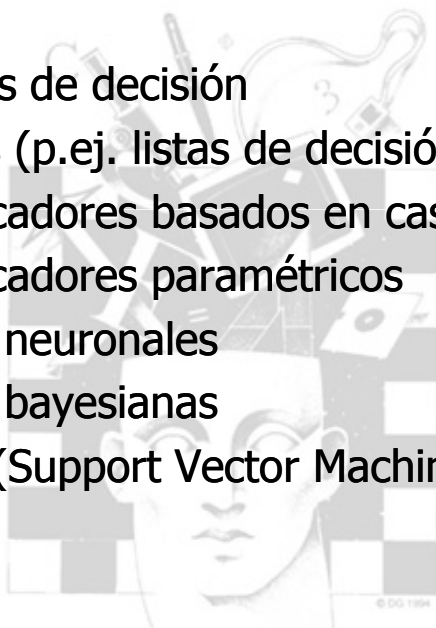


Modelos de clasificación

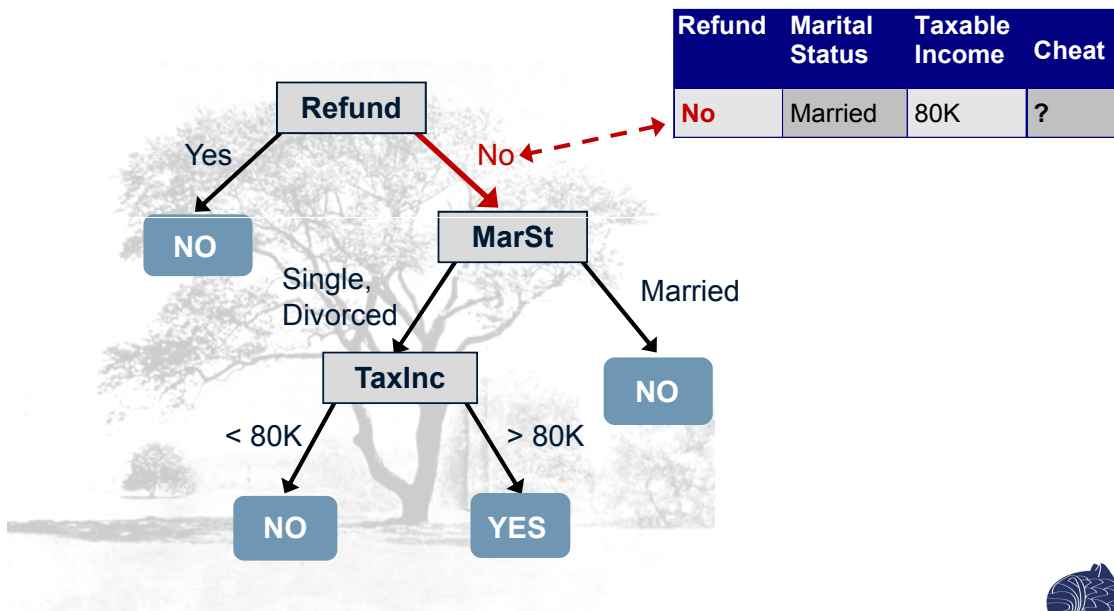


Se pueden construir distintos tipos de clasificadores:

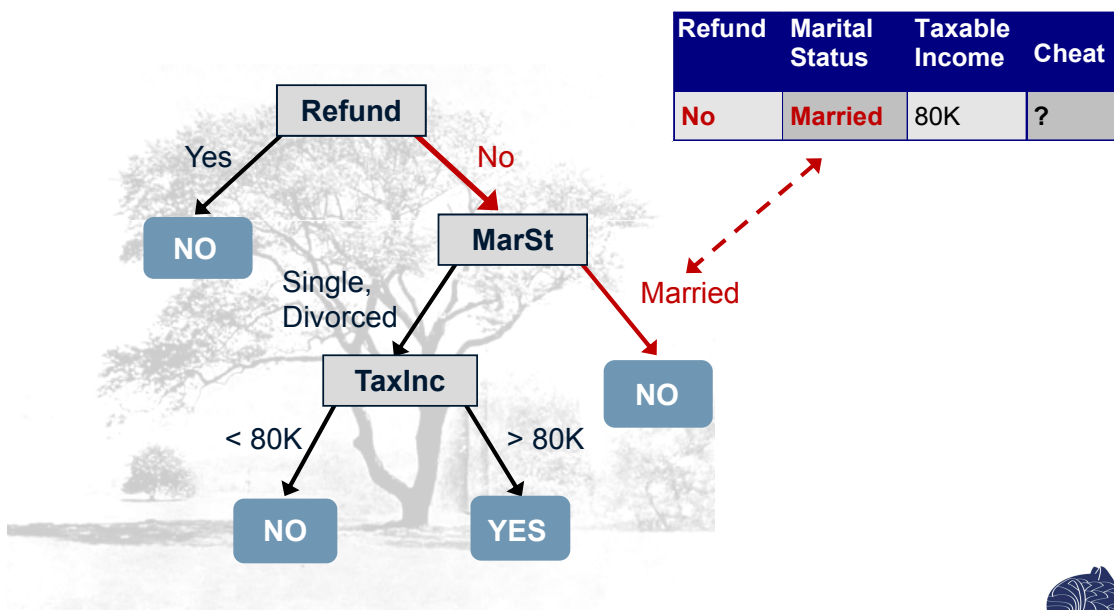
- Árboles de decisión
- Reglas (p.ej. listas de decisión)
- Clasificadores basados en casos
- Clasificadores paramétricos
- Redes neuronales
- Redes bayesianas
- SVMs (Support Vector Machines)
- ...



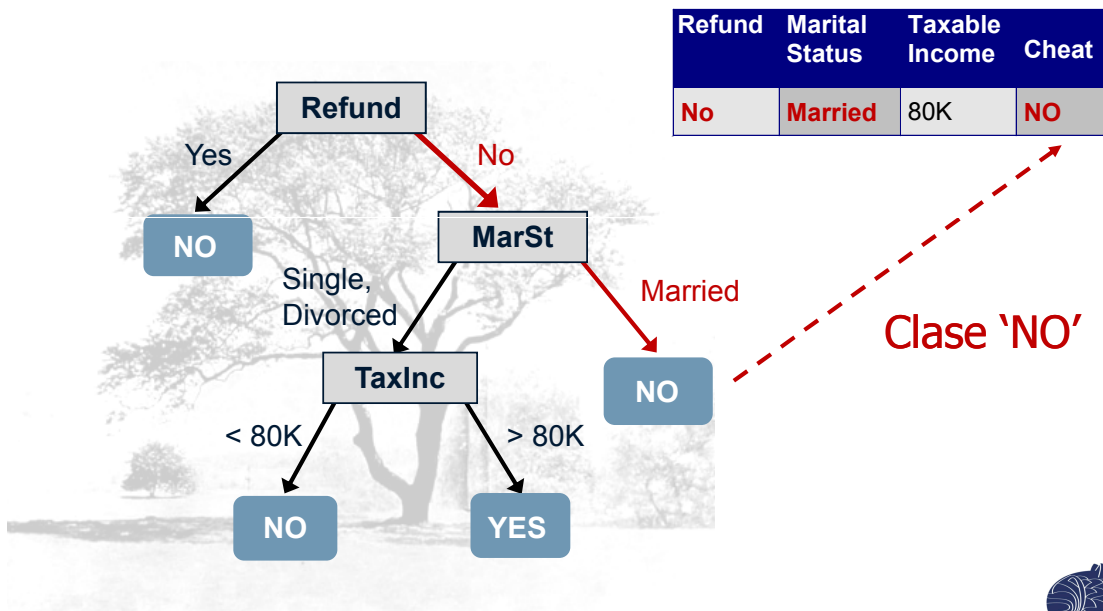
Árboles de decisión



Árboles de decisión



Árboles de decisión

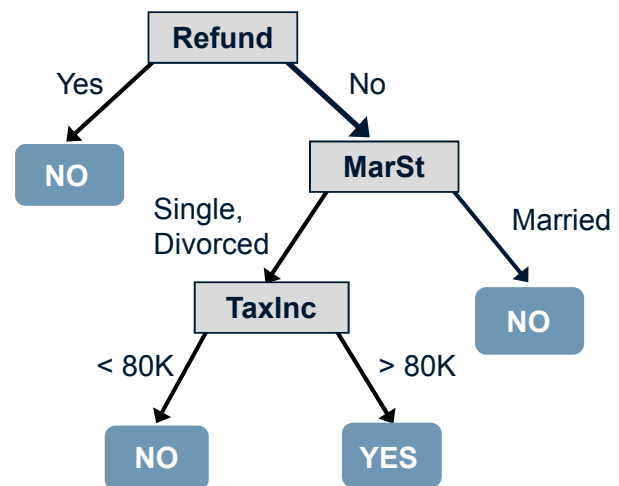


Árboles de decisión



categórico
categórico
continuo
clase

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|-----|--------|----------------|----------------|-------|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |



Conjunto de entrenamiento



Modelo de clasificación:
Árbol de decisión



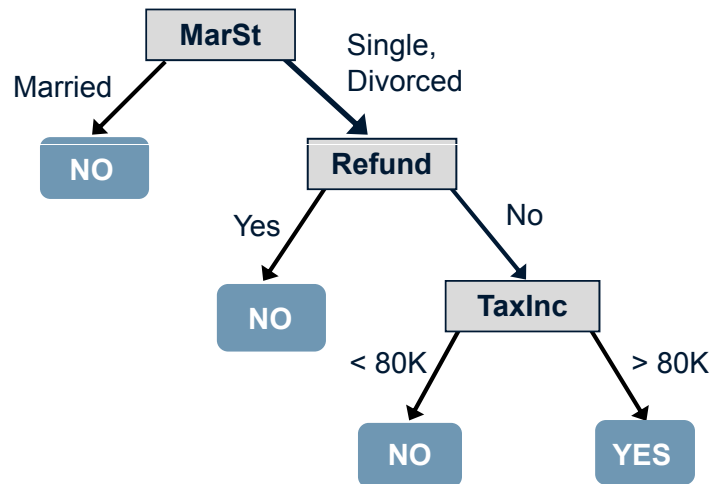
Árboles de decisión



categórico
categórico
continuo
clase

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|-----|--------|----------------|----------------|-------|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

Podemos construir distintos árboles:
¿cuál es mejor?



Conjunto de
entrenamiento



Modelo de clasificación:
Árbol de decisión



18

Árboles de decisión



Construcción de árboles de decisión

- Estrategia greedy (problema NP)
- Algoritmo "divide y vencerás":
 - Comenzamos con todos los ejemplos de entrenamiento en la raíz del árbol de decisión.
 - Los ejemplos se van dividiendo en función del atributo que se seleccione para ramificar el árbol en cada nodo.
 - Los atributos que se usan para ramificar se eligen en función de una heurística (**regla de división**).



19

Árboles de decisión



Construcción de árboles de decisión

- ¿Cuándo se detiene la construcción del árbol de decisión? **Criterios de parada:**
 - Cuando todos los ejemplos que quedan pertenecen a la misma clase (se añade una hoja al árbol con la etiqueta de la clase).
 - Cuando no quedan atributos por los que ramificar (se añade una hoja etiquetada con la clase más frecuente en el nodo).
 - Cuando no nos quedan datos para clasificar.

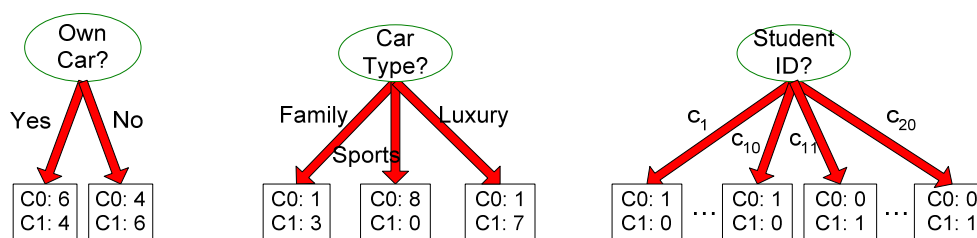


Árboles de decisión



Construcción de árboles de decisión

- ¿Qué heurísticas se pueden utilizar para decidir cómo ramificar el árbol?



¿Cuál es mejor?

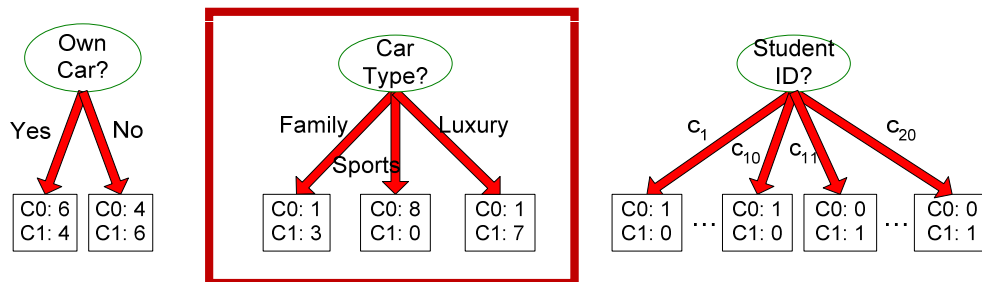


Árboles de decisión



Construcción de árboles de decisión

- ¿Qué heurísticas se pueden utilizar para decidir cómo ramificar el árbol?



La que nos proporciona nodos más homogéneos.

Necesitamos medir la impureza de un nodo.



22

Árboles de decisión



Construcción de árboles de decisión

- **Reglas de división**
(heurísticas para la selección de atributos):

- Ganancia de información (ID3, C4.5)
- Índice de Gini (CART, SLIQ, SPRINT)

Existen otras muchas reglas de división:
 χ^2 , MDL (Minimum Description Length)...



23

Árboles de decisión



Entropía

Teoría de la Información: $Info(D) = -\sum_{i=1}^m p_i \log_2(p_i)$

| | |
|----|----------|
| C1 | 0 |
| C2 | 6 |

$$\text{Entropía} = 0 \\ = -0 \log_2 0 - 1 \log_2 1 = 0$$

| | |
|----|----------|
| C1 | 1 |
| C2 | 5 |

$$\text{Entropía} = 0.65 \\ = - (1/6) \log_2 (1/6) - (5/6) \log_2 (5/6)$$

| | |
|----|----------|
| C1 | 2 |
| C2 | 4 |

$$\text{Entropía} = 0.92 \\ = - (2/6) \log_2 (2/6) - (4/6) \log_2 (4/6)$$

| | |
|----|----------|
| C1 | 3 |
| C2 | 3 |

$$\text{Entropía} = 1 \\ = - (1/2) \log_2 (1/2) - (1/2) \log_2 (1/2)$$



24

Árboles de decisión



Ganancia de información (ID3)

p_i Estimación de la probabilidad de que un ejemplo de D pertenezca a la clase C_i

Entropía

(información necesaria para clasificar un ejemplo en D)

$$Info(D) = -\sum_{i=1}^m p_i \log_2(p_i)$$



25

Árboles de decisión



Ganancia de información (ID3)

Información necesaria para clasificar D después de usar el atributo A para dividir D en v particiones:

$$Info_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \times I(D_j)$$

Ganancia obtenida al ramificar utilizando el atributo A:

$$Gain(A) = Info(D) - Info_A(D)$$



Árboles de decisión



Criterio de proporción de ganancia (Gain Ratio, C4.5)

ID3 tiende a ramificar el árbol utilizando los atributos que tengan más valores diferentes, por lo que se "normaliza" la ganancia de información usando la entropía de la partición (que será mayor cuantas más particiones pequeñas haya):

$$SplitInfo_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \times \log_2 \left(\frac{|D_j|}{|D|} \right)$$

$$GainRatio(A) = Gain(A) / SplitInfo(A)$$



Árboles de decisión



Índice de Gini (CART, SLIQ, SPRINT)

Medida estadística de impureza:

$$gini(D) = 1 - \sum_{j=1}^n p_j^2$$

| | |
|------------|---|
| C1 | 0 |
| C2 | 6 |
| Gini=0.000 | |

| | |
|------------|---|
| C1 | 1 |
| C2 | 5 |
| Gini=0.278 | |

| | |
|------------|---|
| C1 | 2 |
| C2 | 4 |
| Gini=0.444 | |

| | |
|------------|---|
| C1 | 3 |
| C2 | 3 |
| Gini=0.500 | |

- Para construir el árbol, elegimos el atributo que proporciona la mayor reducción de impureza

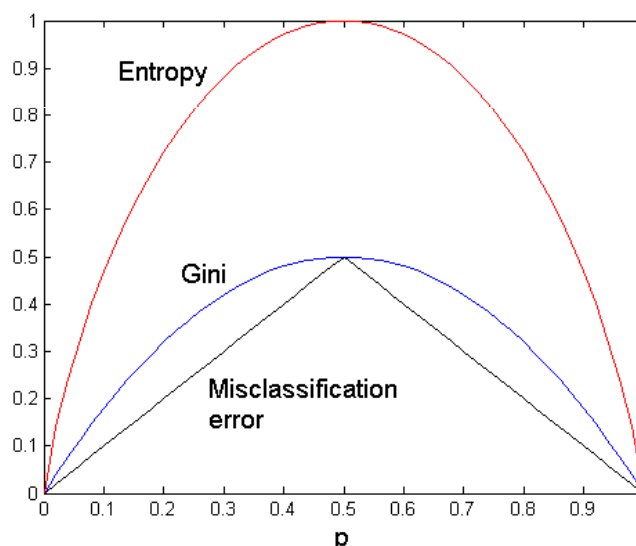


Árboles de decisión



Comparación de reglas de división

Para problemas con dos clases:



Árboles de decisión



Comparación de reglas de división

- **Ganancia de información**
Sesgado hacia atributos con muchos valores diferentes.
- **Criterio de proporción de ganancia**
Tiende a preferir particiones poco balanceadas (con una partición mucho más grande que las otras)
- **Índice de Gini**
Funciona peor cuando hay muchas clases y tiende a favorecer particiones de tamaño y pureza similares.

Ninguna regla de división es significativamente mejor que los demás.



Árboles de decisión



Otros aspectos de interés

- **¿Árboles binarios o n-arios?**
(CART binario; C4.5 n-ario para atributos categóricos, binario para atributos continuos).
- **Manejo de atributos continuos**
(selección del conjunto de tests candidatos para ramificar el árbol, p.ej. discretización previa).
- **Manejo de valores nulos**
(cómo se tratan los valores nulos/desconocidos).



Árboles de decisión



Ejemplo

| 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 |



Árboles de decisión



Ejemplo

Para el cálculo de las entropías...

| n | $\log_2(n)$ |
|----|-------------|
| 1 | 0,000 |
| 2 | 1,000 |
| 3 | 1,585 |
| 4 | 2,000 |
| 5 | 2,322 |
| 6 | 2,585 |
| 7 | 2,807 |
| 8 | 3,000 |
| 9 | 3,170 |
| 10 | 3,322 |
| 11 | 3,459 |
| 12 | 3,585 |
| 13 | 3,700 |
| 14 | 3,807 |
| 15 | 3,907 |
| 16 | 4,000 |



Árboles de decisión



Ejemplo

Cálculo de las entropías $E(+,-)$

$$E(+,-) = -P(+)\log_2 P(+)-P(-)\log_2 P(-)$$

| $E(+,-)$ | 0- | 1- | 2- | 3- | 4- | 5- |
|----------|-------|-------|-------|-------|-------|-------|
| 0+ | | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| 1+ | 0,000 | 1,000 | 0,918 | 0,811 | 0,722 | 0,650 |
| 2+ | 0,000 | 0,918 | 1,000 | 0,971 | 0,918 | 0,863 |
| 3+ | 0,000 | 0,811 | 0,971 | 1,000 | 0,985 | 0,954 |
| 4+ | 0,000 | 0,722 | 0,918 | 0,985 | 1,000 | 0,991 |
| 5+ | 0,000 | 0,650 | 0,863 | 0,954 | 0,991 | 1,000 |
| 6+ | 0,000 | 0,592 | 0,811 | 0,918 | 0,971 | 0,994 |
| 7+ | 0,000 | 0,544 | 0,764 | 0,881 | 0,946 | 0,980 |
| 8+ | 0,000 | 0,503 | 0,722 | 0,845 | 0,918 | 0,961 |
| 9+ | 0,000 | 0,469 | 0,684 | 0,811 | 0,890 | 0,940 |



Árboles de decisión



Ejemplo

Raíz del árbol (9+,5-)

$$\text{Info}(D) = E(9+,5-) = 0.940 \text{ bits}$$

Ramificando por el atributo "Outlook"

Info_{Outlook}(D)

$$\begin{aligned} &= (5/14) \text{Info}(D_{\text{sunny}}) + (4/14) \text{Info}(D_{\text{overcast}}) + (5/14) \text{Info}(D_{\text{rainy}}) \\ &= (5/14) E(2+,3-) + (4/14) E(4+,0-) + (5/14) E(3+,2-) \\ &= (5/14) \cdot 0.971 + (4/14) \cdot 0 + (5/14) \cdot 0.971 = 0.693 \text{ bits} \end{aligned}$$

$$\text{Gain}(\text{Outlook}) = \text{Info}(D) - \text{Info}_{\text{Outlook}}(D) = \mathbf{0.247 \text{ bits}}$$



Árboles de decisión



Ejemplo

Raíz del árbol (9+,5-)

$$\text{Info}(D) = E(9+,5-) = 0.940 \text{ bits}$$

Ramificando por el atributo **"Temperature"**

Info_{Temperature}(D)

$$\begin{aligned} &= (4/14) \text{Info}(D_{\text{cool}}) + (6/14) \text{Info}(D_{\text{mild}}) + (4/14) \text{Info}(D_{\text{hot}}) \\ &= (4/14) E(3+,1-) + (6/14) E(4+,2-) + (4/14) E(2+,2-) \\ &= (4/14) \cdot 0.811 + (6/14) \cdot 0.918 + (4/14) \cdot 1 = 0.911 \text{ bits} \end{aligned}$$

$$\text{Gain}(\text{Temperature}) = \text{Info}(D) - \text{Info}_{\text{Temperature}}(D) = \mathbf{0.029 \text{ bits}}$$



Árboles de decisión



Ejemplo

Raíz del árbol (9+,5-)

$$\text{Info}(D) = E(9+,5-) = 0.940 \text{ bits}$$

Ramificando por el atributo **"Humidity"**

Info_{Humidity}(D)

$$\begin{aligned} &= (7/14) \text{Info}(D_{\text{high}}) + (7/14) \text{Info}(D_{\text{normal}}) \\ &= (7/14) E(3+,4-) + (7/14) E(6+,1-) \\ &= (7/14) \cdot 0.985 + (7/14) \cdot 0.592 = 0.789 \text{ bits} \end{aligned}$$

$$\text{Gain}(\text{Humidity}) = \text{Info}(D) - \text{Info}_{\text{Humidity}}(D) = \mathbf{0.151 \text{ bits}}$$



Árboles de decisión



Ejemplo

Raíz del árbol (9+,5-)

$$\text{Info}(D) = E(9+,5-) = 0.940 \text{ bits}$$

Ramificando por el atributo **"Windy"**

Info_{Windy}(D)

$$\begin{aligned} &= (8/14) \text{Info}(D_{\text{false}}) + (6/14) \text{Info}(D_{\text{true}}) \\ &= (8/14) E(6+,2-) + (6/14) E(3+,3-) \\ &= (8/14) \cdot 0.811 + (6/14) \cdot 1 = 0.892 \text{ bits} \end{aligned}$$

$$\text{Gain}(\text{Windy}) = \text{Info}(D) - \text{Info}_{\text{Windy}}(D) = \mathbf{0.048 \text{ bits}}$$



Árboles de decisión



Ejemplo

Raíz del árbol (9+,5-)

$$\text{Gain}(\text{Outlook}) = \text{Info}(D) - \text{Info}_{\text{Outlook}}(D) = \mathbf{0.247 \text{ bits}}$$

$$\text{Gain}(\text{Temperature}) = \text{Info}(D) - \text{Info}_{\text{Temperature}}(D) = \mathbf{0.029 \text{ bits}}$$

$$\text{Gain}(\text{Humidity}) = \text{Info}(D) - \text{Info}_{\text{Humidity}}(D) = \mathbf{0.151 \text{ bits}}$$

$$\text{Gain}(\text{Windy}) = \text{Info}(D) - \text{Info}_{\text{Windy}}(D) = \mathbf{0.048 \text{ bits}}$$

Por tanto, ramificamos usando el atributo "Outlook"...

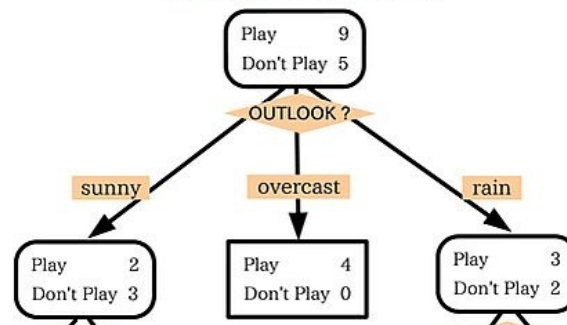


Árboles de decisión



Ejemplo

Nuestro árbol de decisión está así ahora mismo...



... pero aún tenemos que seguir construyéndolo.



Árboles de decisión



Ejemplo

Nodo "Outlook = sunny" (2+,3-)

$$\text{Info}(D_s) = E(2+,3-) = 0.971$$

Temperature: { (0+,2-), (1+,1-), (1+,0-) }

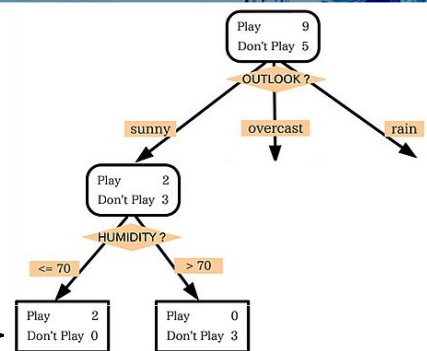
$$\text{Gain}(\text{Temperature}) = \text{Info}(D_s) - \text{Info}_{\text{Temperature}}(D_s) = \mathbf{0.571 \text{ bits}}$$

Humidity: { (0+,3-), (2+,0-) }

$$\text{Gain}(\text{Humidity}) = \text{Info}(D_s) - \text{Info}_{\text{Humidity}}(D_s) = \mathbf{0.971 \text{ bits}}$$

Windy: { (1+,2-), (1+,1-) }

$$\text{Gain}(\text{Windy}) = \text{Info}(D_s) - \text{Info}_{\text{Windy}}(D_s) = \mathbf{0.019 \text{ bits}}$$



Árboles de decisión

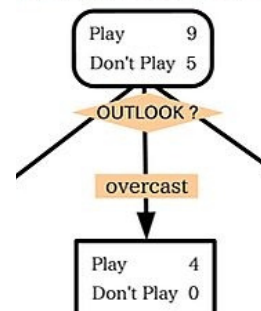


Ejemplo

Nodo "Outlook = overcast" (4+,0-)

$$\text{Info}(D_o) = E(4+,0-) = 0.000$$

Creamos un nodo hoja directamente, ya que todos los ejemplos son de la misma clase.



Árboles de decisión



Ejemplo

Nodo "Outlook = rainy" (3+,2-)

$$\text{Info}(D_r) = E(3+,2-) = 0.971$$

Temperature: { (0+,0-), (2+,1-), (1+,1-) }

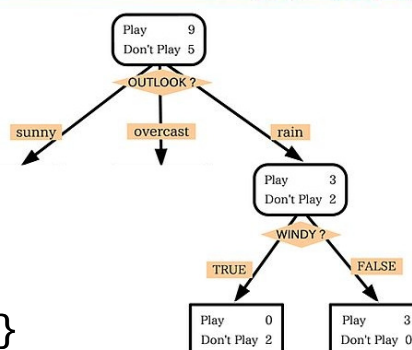
$$\text{Gain}(\text{Temperature}) = \text{Info}(D_r) - \text{Info}_{\text{Temperature}}(D_r) < 0$$

Humidity: { (2+,1-), (1+,1-) }

$$\text{Gain}(\text{Humidity}) = \text{Info}(D_r) - \text{Info}_{\text{Humidity}}(D_r) < 0$$

Windy: { (0+,2-), (3+,0-) }

$$\text{Gain}(\text{Windy}) = \text{Info}(D_r) - \text{Info}_{\text{Windy}}(D_r) = \mathbf{0.971 \text{ bits}}$$

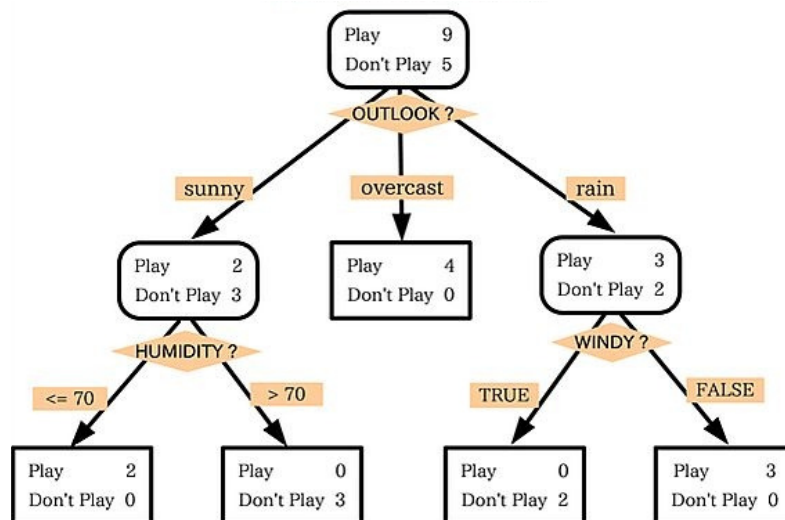


Árboles de decisión



Ejemplo

Resultado final...



Árboles de decisión



El problema del sobreaprendizaje

Los árboles de decisión tienden a ajustarse demasiado al conjunto de entrenamiento utilizado para construir el árbol:

- Demasiadas ramas del árbol reflejan anomalías del conjunto de entrenamiento (ruido y outliers).
- El árbol resultante es más complejo de lo que debería ser.
- Como consecuencia, disminuye la precisión del clasificador de cara a situaciones nuevas.



Árboles de decisión



El problema del sobreaprendizaje

Una solución al problema del sobreaprendizaje:

Técnicas de poda

Una vez construido el árbol, se van eliminando ramas: utilizando un conjunto de datos distinto al conjunto de entrenamiento [CART: Poda por coste-complejidad] o no [C4.5: Poda pesimista].



Árboles de decisión



El problema del sobreaprendizaje

Técnicas de poda

Para podar un árbol de decisión, se sustituye...

- un subárbol por un nodo hoja (correspondiente a la clase más frecuente en el subárbol), o bien,
- un subárbol por otro subárbol contenido en el primero.

Por tanto, se introducirán errores de clasificación adicionales en el conjunto de entrenamiento (aunque, si la poda se realiza correctamente, la precisión del clasificador aumentará).

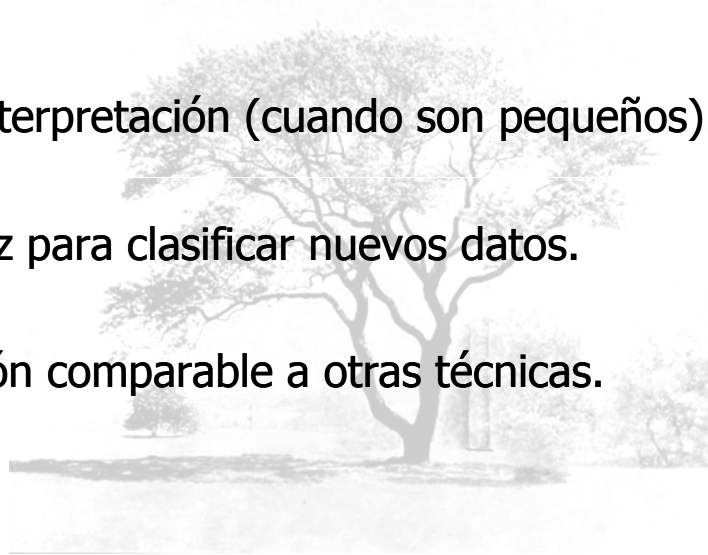


Árboles de decisión



Ventajas de los árboles de decisión

- Fácil interpretación (cuando son pequeños).
- Rapidez para clasificar nuevos datos.
- Precisión comparable a otras técnicas.

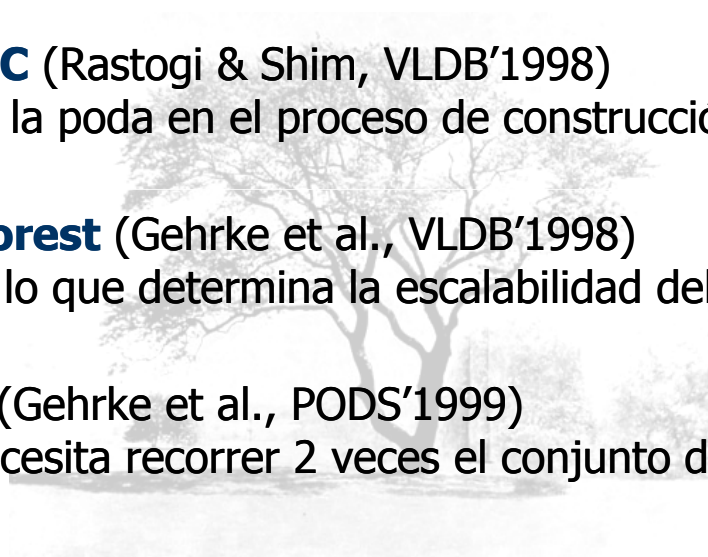


Árboles de decisión



Algoritmos eficientes y escalables

- **PUBLIC** (Rastogi & Shim, VLDB'1998)
integra la poda en el proceso de construcción del árbol
- **RainForest** (Gehrke et al., VLDB'1998)
separa lo que determina la escalabilidad del algoritmo
- **BOAT** (Gehrke et al., PODS'1999)
sólo necesita recorrer 2 veces el conjunto de datos



Árboles de decisión



DEMO



TDIDT

Top-Down Induction of Decision Trees



Reglas



Existen muchas formas de construir modelos de clasificación basados en reglas:

- A partir de un árbol de decisión.
- Diseñando algoritmos específicos de inducción de reglas:
 - Metodología STAR de Michalski
 - Listas de decisión (p.ej. RIPPER).
- A partir de reglas de asociación.



Reglas



A partir de un árbol de decisión

¿Por qué?

Las reglas son más fáciles de interpretar que un árbol de decisión complejo.

¿Cómo?

Se crea una regla para cada hoja del árbol.

Las reglas resultantes son

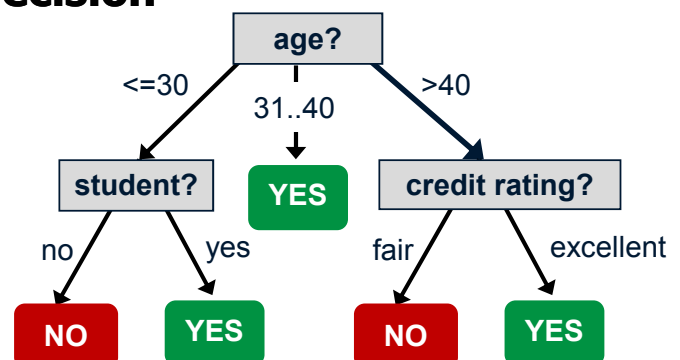
- mutuamente excluyentes y
- exhaustivas.



Reglas



A partir de un árbol de decisión



IF (age<=30) AND (student=no)

THEN buys_computer = NO

IF (age<=30) AND (student=yes)

THEN buys_computer = YES

IF (30<age<=40)

THEN buys_computer = YES

IF (age>40) AND (credit_rating=excellent)

THEN buys_computer = YES

IF (age>40) AND (credit_rating=fair)

THEN buys_computer = NO





A partir de un árbol de decisión

Las reglas que se derivan de un árbol se pueden simplificar (generalizar), aunque entonces:

- Dejan de ser mutuamente excluyentes: varias reglas pueden ser válidas para un mismo ejemplo (hay que establecer un orden entre las reglas [lista de decisión] o realizar una votación).
- Dejan de ser exhaustivas: puede que ninguna regla sea aplicable a un ejemplo concreto (hace falta incluir una clase por defecto).



Inducción de reglas

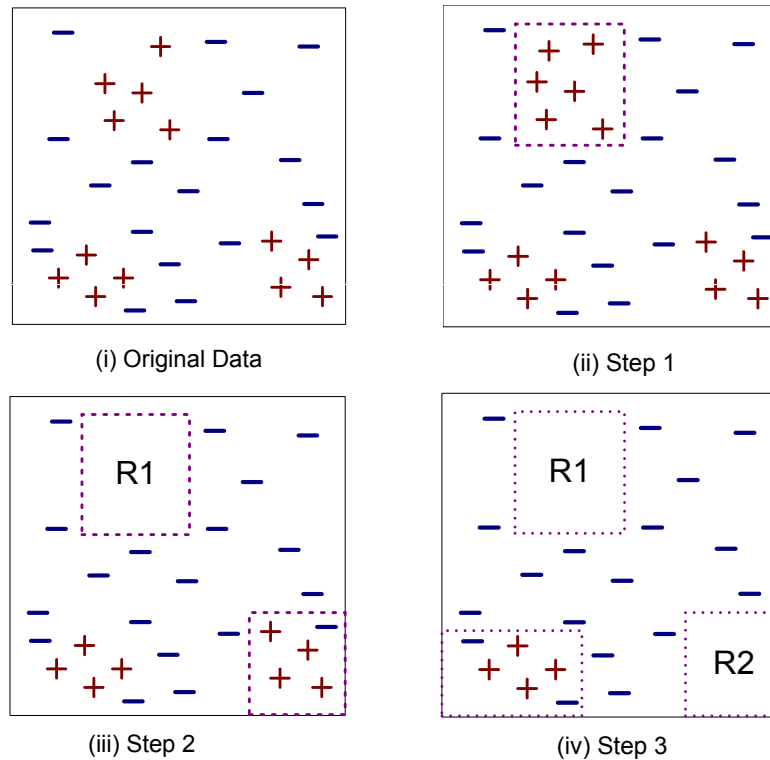
(directamente a partir del conjunto de entrenamiento)

p.ej. **LISTAS DE DECISIÓN**

¿Cómo?

- Las reglas se aprenden de una en una.
- Cada vez que se escoge una regla, se eliminan del conjunto de entrenamiento todos los casos cubiertos por la regla seleccionada.
- El proceso se repite iterativamente hasta que se cumpla alguna condición de parada.





Inducción de reglas

(directamente a partir del conjunto de entrenamiento)

p.ej. **LISTAS DE DECISIÓN**

¿Cómo se aprende una regla?

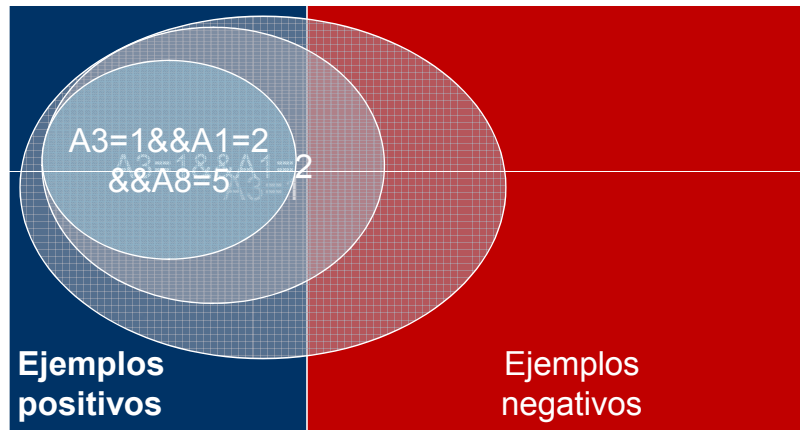
- Se empieza con la regla más general posible.
- Se le van añadiendo antecedentes a la regla para maximizar la "calidad" de la regla (cobertura y precisión).



Reglas



Inducción de reglas



Reglas



Inducción de reglas

(directamente a partir del conjunto de entrenamiento)

p.ej. **LISTAS DE DECISIÓN**

Algoritmos de inducción de reglas

- FOIL (Quinlan, Machine Learning, 1990)
- CN2 (Clark & Boswell, EWSL'1991)
- RIPPER (Cohen, ICML'1995)
- PNrul (Joshi, Agarwal & Kumar, SIGMOD'2001)





DEMO

CN2

- Metodología STAR: Unordered CN2
- Listas de decisión: Ordered CN2



RIPPER

Repeated Incremental Pruning to Produce Error Reduction
(basado en **IREP**, Iterative Reduced Error Pruning)



Evaluación



La evaluación de un algoritmo de construcción de modelos de clasificación se puede realizar atendiendo a distintos aspectos:

- **Precisión**
(porcentaje de casos clasificados correctamente).
- **Eficiencia**
(tiempo necesario para construir/usar el clasificador).
- **Robustez**
(frente a ruido y valores nulos)
- **Escalabilidad**
(utilidad en grandes bases de datos)
- **Interpretabilidad**
(el clasificador, ¿es sólo una caja negra?)
- **Complejidad**
(del modelo de clasificación) → Navaja de Occam.





Métricas

Cómo evaluar la "calidad" de un modelo de clasificación.

Métodos

Cómo estimar, de forma fiable, la calidad de un modelo.

Comparación

Cómo comparar el rendimiento relativo de dos modelos de clasificación alternativos



Matriz de confusión (confusion matrix)

| | | Predicción | |
|------------|-------|--------------------|--------------------|
| | | C_p | C_N |
| Clase real | C_p | TP: True positive | FN: False negative |
| | C_N | FP: False positive | TN: True negative |

Precisión del clasificador

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$



Evaluación: Métricas



Limitaciones de la precisión ("accuracy") :

Supongamos un problema con 2 clases no equilibradas:

- 9990 ejemplos de la clase 1
- 10 ejemplos de la clase 2

Si el modelo de clasificación siempre dice que los ejemplos son de la clase 1, su precisión es

$$9990/10000 = \mathbf{99.9\%}$$

Totalmente engañosa, ya que nunca detectaremos ningún ejemplo de la clase 2.



Evaluación: Métricas



Alternativa: Matriz de costes

| $C(i j)$ | | Predicción | |
|------------|-------|------------|----------|
| | | C_p | C_N |
| Clase real | C_p | $C(P P)$ | $C(N P)$ |
| | C_N | $C(P N)$ | $C(N N)$ |

El coste de clasificación será proporcional a la precisión del clasificador sólo si

$$\forall i, j: i \neq j \quad \begin{aligned} C(i|j) &= C(j|i) \\ C(i|i) &= C(j|j) \end{aligned}$$



Evaluación: Métricas



Medidas "cost-sensitive"

| | | Predicción | |
|------------|-------|--------------------|--------------------|
| | | C_p | C_N |
| Clase real | C_p | TP: True positive | FN: False negative |
| | C_N | FP: False positive | TN: True negative |

$$\text{precision} = \text{TP}/(\text{TP}+\text{FP})$$

True positive recognition rate

$$\text{recall} = \text{sensitivity} = \text{TP}/P = \text{TP}/(\text{TP}+\text{FN})$$

True negative recognition rate

$$\text{specificity} = \text{TN}/N = \text{TN}/(\text{TN}+\text{FP})$$



Evaluación: Métricas



Medidas "cost-sensitive"

| | | Predicción | |
|------------|-------|--------------------|--------------------|
| | | C_p | C_N |
| Clase real | C_p | TP: True positive | FN: False negative |
| | C_N | FP: False positive | TN: True negative |

F-measure

Media armónica de precisión y recall:

$$F = 2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$$

$$F = 2TP / (2TP + FP + FN)$$



Evaluación: Métricas



Medidas "cost-sensitive"

| | | Predicción | |
|------|-------|------------|-------|
| | | C_P | C_N |
| Real | C_P | TP | FN |
| | C_N | FP | TN |

Accuracy

| | | Predicción | |
|------|-------|------------|-------|
| | | C_P | C_N |
| Real | C_P | TP | FN |
| | C_N | FP | TN |

Recall

| | | Predicción | |
|------|-------|------------|-------|
| | | C_P | C_N |
| Real | C_P | TP | FN |
| | C_N | FP | TN |

Precision

| | | Predicción | |
|------|-------|------------|-------|
| | | C_P | C_N |
| Real | C_P | TP | FN |
| | C_N | FP | TN |

F-measure



Evaluación: Métodos



Para evaluar la precisión de un modelo de clasificación nunca debemos utilizar el conjunto de entrenamiento (lo que nos daría el "**error de resustitución**" del clasificador), sino un conjunto de prueba independiente:

Por ejemplo, podríamos reservar 2/3 de los ejemplos disponibles para construir el clasificador y el 1/3 restante lo utilizaríamos de **conjunto de prueba** para estimar la precisión del clasificador.



Evaluación: Métodos



Validación cruzada

[k-CV: k-fold Cross-Validation]

- Se divide aleatoriamente el conjunto de datos en k subconjuntos de intersección vacía (más o menos del mismo tamaño). Típicamente, $k=10$.
- En la iteración i , se usa el subconjunto i como conjunto de prueba y los $k-1$ restantes como conjunto de entrenamiento.
- Como medida de evaluación del método de clasificación se toma la media aritmética de las k iteraciones realizadas.



Evaluación: Métodos



Validación cruzada

Variantes de la validación cruzada

- **“Leave one out”:**
Se realiza una validación cruzada con k particiones del conjunto de datos, donde k coincide con el número de ejemplos disponibles.
- **Validación cruzada estratificada:**
Las particiones se realizan intentando mantener en todas ellas la misma proporción de clases que aparece en el conjunto de datos completo.



Evaluación: Métodos



Bootstrapping

Muestreo uniforme con reemplazo de los ejemplos disponibles (esto es, una vez que se escoge un ejemplo, se vuelve a dejar en el conjunto de entrenamiento y puede que se vuelva a escoger).

NOTA: Método utilizado en "ensembles".



Evaluación: Métodos



Bootstrapping

0.632 bootstrap

- Dado un conjunto de d datos, se toman d muestras. Los datos que no se escojan formarán parte del conjunto de prueba.
- En torno al 63.2% de las muestras estarán en el "bootstrap" (el conjunto de entrenamiento) y el 36.8% caerá en el conjunto de prueba, ya que $(1-1/d)^d \approx e^{-1} = 0.368$
- Si repetimos el proceso k veces, tendremos:

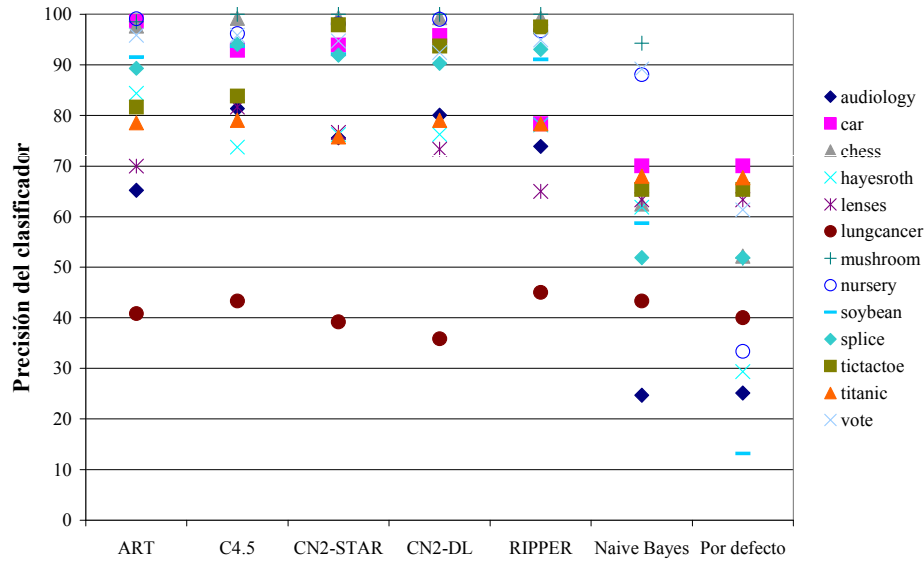
$$acc(M) = \sum_{i=1}^k (0.632 \times acc(M_i)_{test_set} + 0.368 \times acc(M_i)_{train_set})$$



Evaluación: Comparación



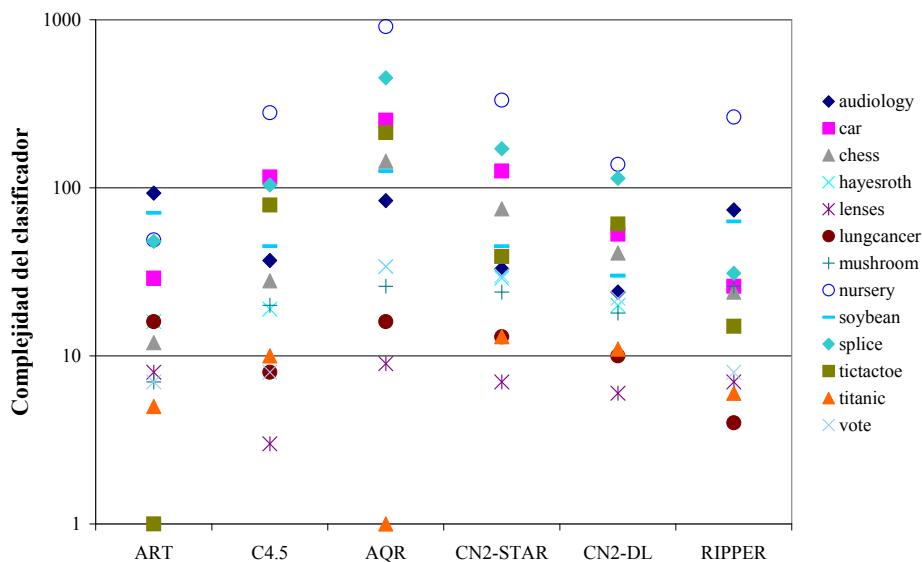
Precisión [accuracy]



Evaluación: Comparación



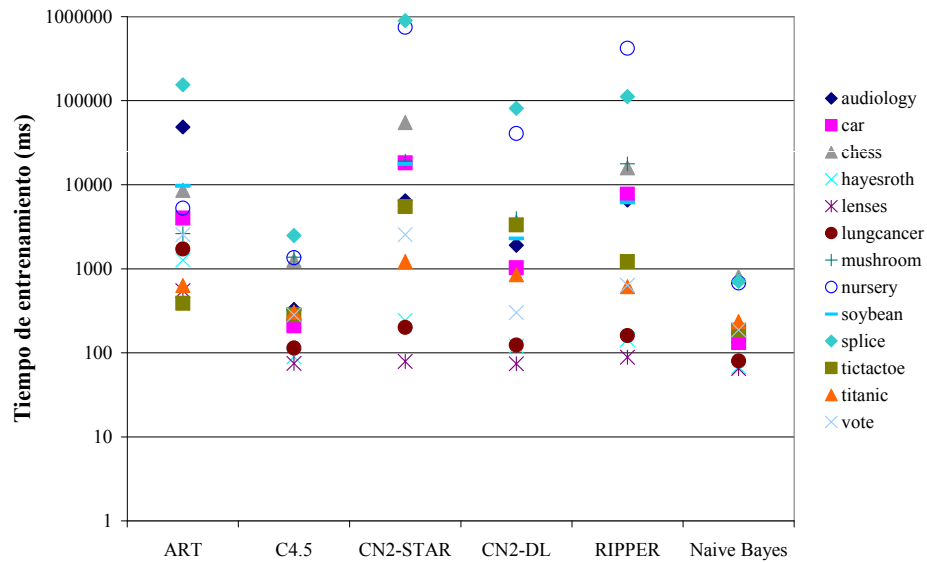
Complejidad del clasificador



Evaluación: Comparación



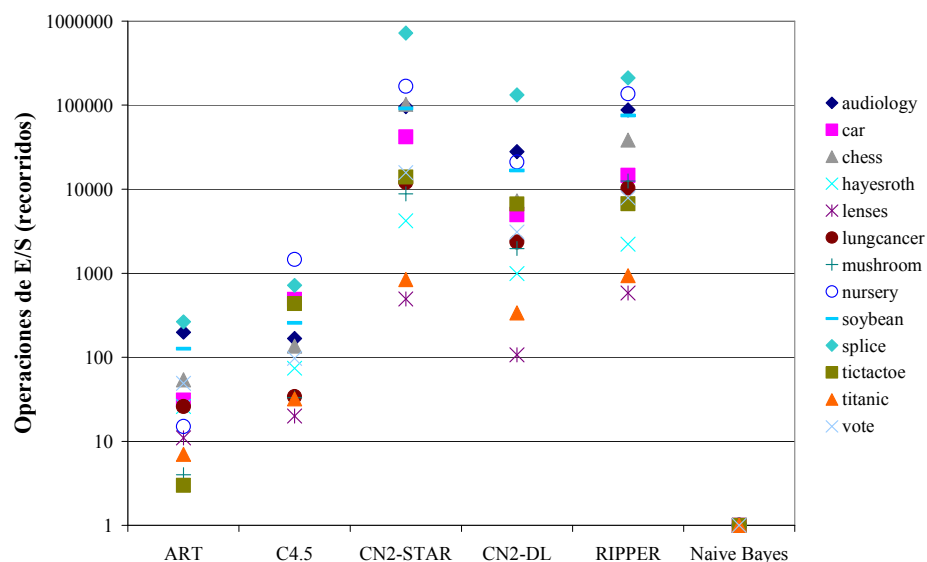
Tiempo de entrenamiento



Evaluación: Comparación



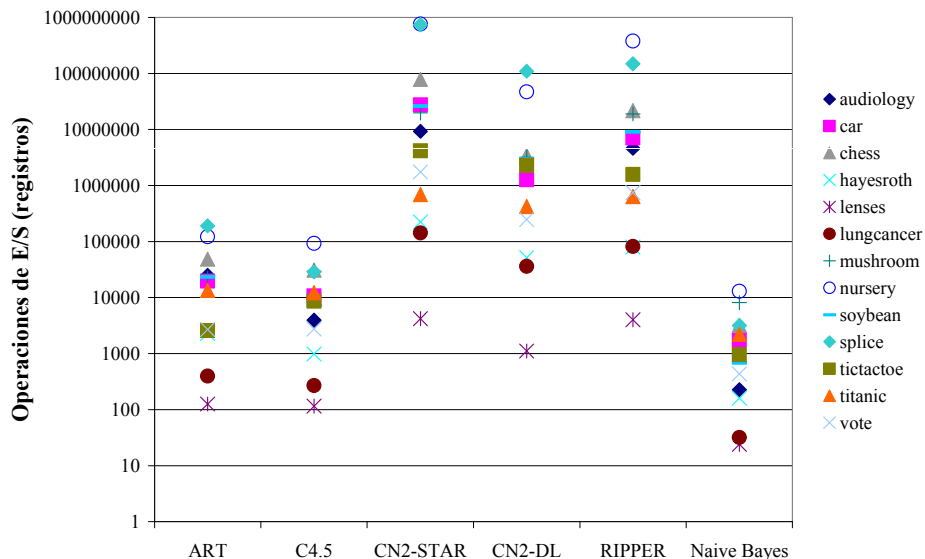
Operaciones de E/S: Recorridos



Evaluación: Comparación



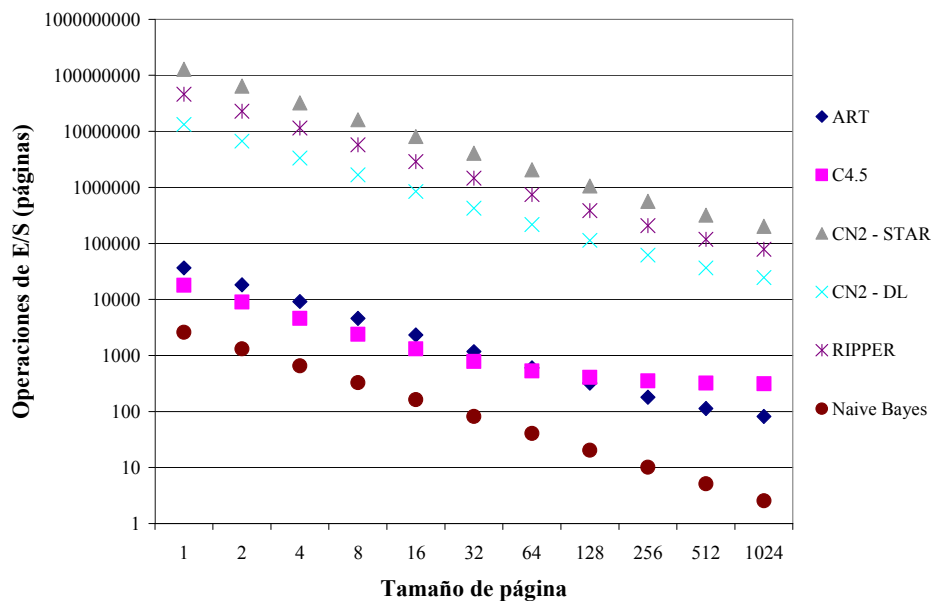
Operaciones de E/S: Registros



Evaluación: Comparación



Operaciones de E/S: Páginas de disco

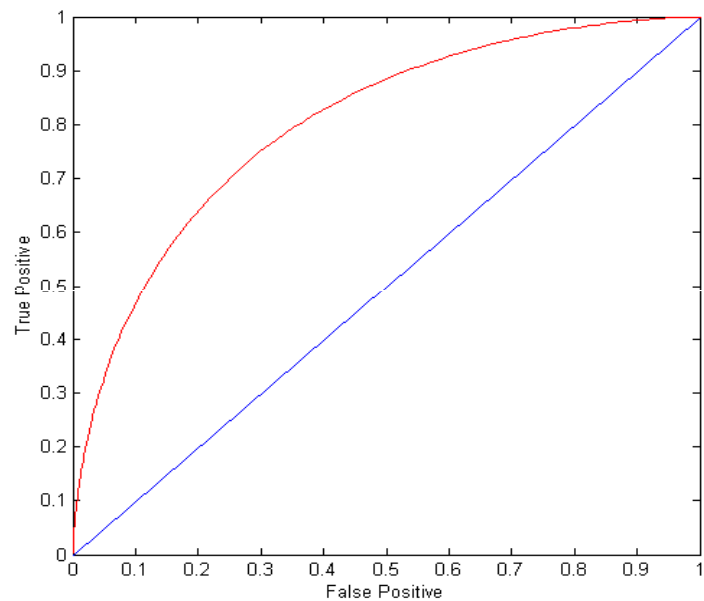


Evaluación: Comparación



Curvas ROC

Receiver Operating Characteristics



$TPR = TP/(TP+FN)$ Eje vertical: "true positive rate"

$FPR = FP/(FP+TN)$ Eje horizontal: "false positive rate"



Evaluación: Comparación



Curvas ROC

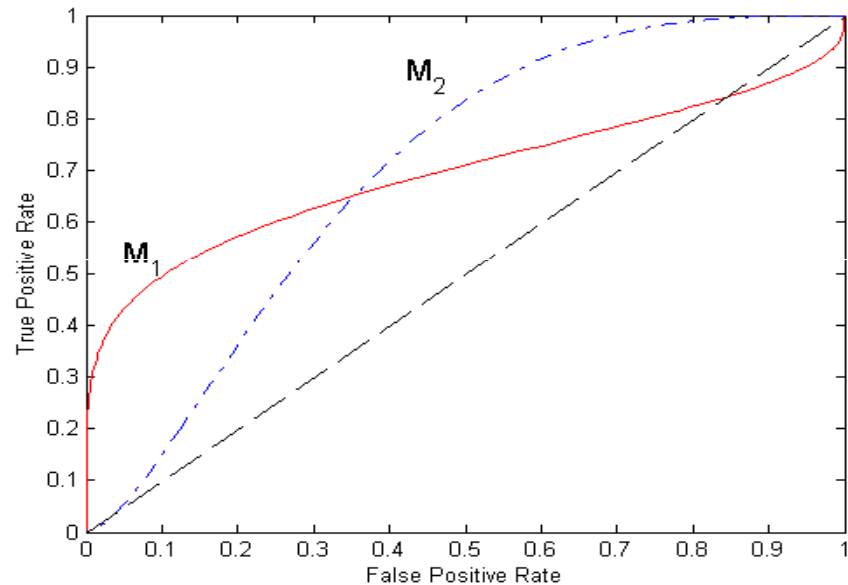
- Desarrolladas en los años 50 para analizar señales con ruido: caracterizar el compromiso entre aciertos y falsas alarmas.
- Permiten comparar visualmente distintos modelos de clasificación.
- El área que queda bajo la curva es una medida de la precisión [accuracy] del clasificador:
 - ❖ Cuanto más cerca estemos de la diagonal (área cercana a 0.5), menos preciso será el modelo.
 - ❖ Un modelo "perfecto" tendrá área 1.



Evaluación: Comparación



Curvas ROC



Ningún modelo es consistentemente mejor que el otro:
 M_1 es mejor para FPR bajos, M_2 para FPR altos.



82

Evaluación: Comparación



Curvas ROC

¿Cómo se construye la curva ROC?

- Se usa un clasificador que prediga la probabilidad de que un ejemplo E pertenezca a la clase positiva $P(+|E)$
- Se ordenan los ejemplos en orden decreciente del valor estimado $P(+|E)$
- Se aplica un umbral para cada valor distinto de $P(+|E)$, para el que se cuenta el número de TP, FP, TN y FN.

$$TPR = TP/(TP+FN)$$

$$FPR = FP/(FP+TN)$$

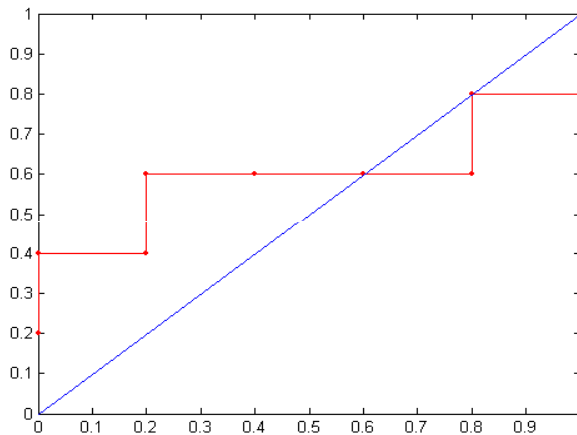


83

Evaluación: Comparación



Curvas ROC



| Ejemplo | P(+ E) | Clase |
|---------|--------|-------|
| 1 | 0.95 | + |
| 2 | 0.93 | + |
| 3 | 0.87 | - |
| 4 | 0.85 | - |
| 5 | 0.85 | - |
| 6 | 0.85 | + |
| 7 | 0.76 | - |
| 8 | 0.53 | + |
| 9 | 0.43 | - |
| 10 | 0.25 | + |

| Clase | + | - | + | - | - | - | + | - | + | + | |
|-------|------|------|------|------|------|------|------|------|------|------|------|
| | 0.25 | 0.43 | 0.53 | 0.76 | 0.85 | 0.85 | 0.85 | 0.87 | 0.93 | 0.95 | 1.00 |
| TP | 5 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | 2 | 1 | 0 |
| FP | 5 | 5 | 4 | 4 | 3 | 2 | 1 | 1 | 0 | 0 | 0 |
| TN | 0 | 0 | 1 | 1 | 2 | 3 | 4 | 4 | 5 | 5 | 5 |
| FN | 0 | 1 | 1 | 2 | 2 | 2 | 2 | 3 | 3 | 4 | 5 |
| TPR | 1 | 0.8 | 0.8 | 0.6 | 0.6 | 0.6 | 0.6 | 0.4 | 0.4 | 0.2 | 0 |
| FPR | 1 | 1 | 0.8 | 0.8 | 0.6 | 0.4 | 0.2 | 0.2 | 0 | 0 | 0 |



Técnicas de regresión



La predicción (numérica) es...

- Similar a la clasificación:
 - Se construye un modelo a partir de un conjunto de entrenamiento.
 - Se utiliza el modelo para predecir el valor de una variable (continua u ordenada).
- Diferente a la clasificación:
 - El modelo define una función continua.

Método más empleado: **Regresión**



Técnicas de regresión



Las técnicas de regresión modelan la relación entre una o más variables independiente (predictores) y una variable dependiente (variable de respuesta).

Métodos de regresión

- Regresión lineal
- Regresión no lineal
- Árboles de regresión (p.ej. CART)
- ...



Técnicas de regresión



Regresión lineal simple

Una única variable independiente:

$$y = w_0 + w_1 x$$

donde w_0 (desplazamiento) y w_1 (pendiente) son los coeficientes de regresión.

■ Método de los mínimos cuadrados

(estima la línea recta que mejor se ajusta a los datos):

$$w_0 = \bar{y} - w_1 \bar{x} \quad w_1 = \frac{\sum_{i=1}^{|D|} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|} (x_i - \bar{x})^2}$$

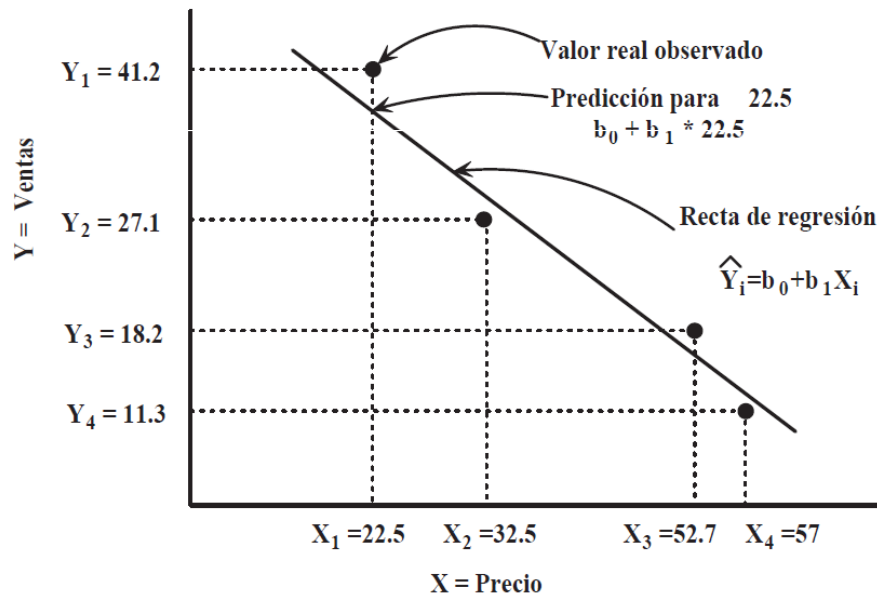


Técnicas de regresión



Regresión lineal simple

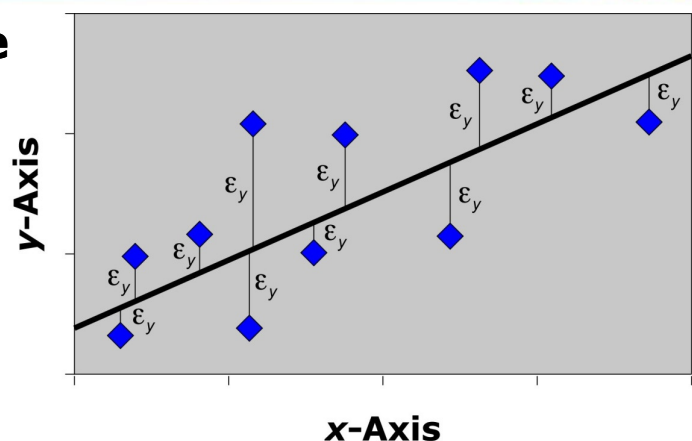
$$\hat{Y}_i = b_0 + b_1 X_i$$



Técnicas de regresión



Regresión lineal simple

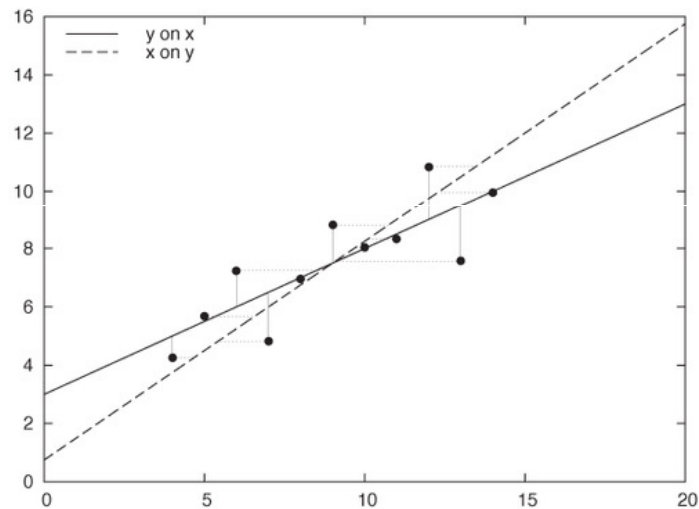


El método de los mínimos cuadrados minimiza la suma de los cuadrados de los residuos ϵ_i (las diferencias entre las predicciones y los valores observados).





Regresión lineal simple



¡OJO! Al utilizar regresión lineal, la recta $y=f(x)$ que se obtiene es distinta a la que obtenemos si $x=f(y)$.



Regresión lineal múltiple

Varias variables independientes:

$$y = w_0 + w_1 x_1 + w_2 x_2 + \dots$$

- Resoluble por métodos numéricos de optimización.
- Muchas funciones no lineales pueden transformarse en una expresión lineal.

p.ej. Un modelo de regresión polinomial

$$y = w_0 + w_1 x + w_2 x^2 + w_3 x^3$$

puede transformarse en un modelo lineal

definiendo las variables $x_2 = x^2$, $x_3 = x^3$:

$$y = w_0 + w_1 x + w_2 x_2 + w_3 x_3$$



Técnicas de regresión



Regresión lineal

Condiciones necesarias para aplicar regresión lineal:

- Obviamente, la muestra ha de ser aleatoria.
- El tipo de dependencia descrita ha de ser lineal.
- Fijado un valor de la(s) variable(s) independiente(s), la variable dependiente se distribuye según una distribución normal.
- Los errores han de tener la misma varianza (nube de puntos homogénea).

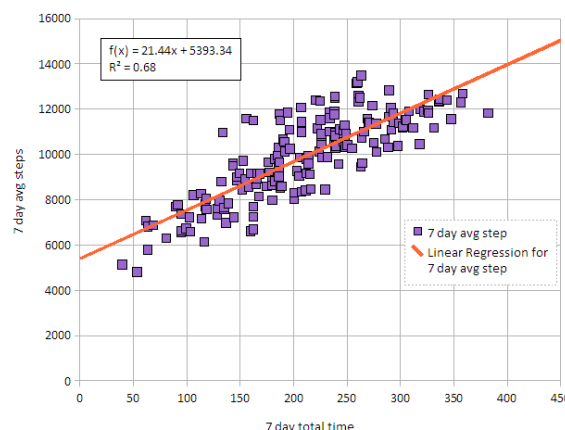


Técnicas de regresión



Regresión lineal simple

1. Mediante un diagrama de dispersión, comprobamos visualmente si existe una relación lineal entre las variables X (predictor) e Y (respuesta):





Regresión lineal simple

2. Cuantificamos la relación construyendo la recta que resume la dependencia y damos una medida de cómo se ajusta la recta a los datos (correlación):

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \in [-1, 1]$$



Coefficiente de correlación

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \in [-1, 1]$$

r=+1 Dependencia lineal total en sentido positivo (cuanto mayor es X, mayor es Y).

r=-1 Dependencia lineal total en sentido negativo (cuanto mayor es X, menor es Y).



Técnicas de regresión



Coefficiente de correlación

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \in [-1, 1]$$

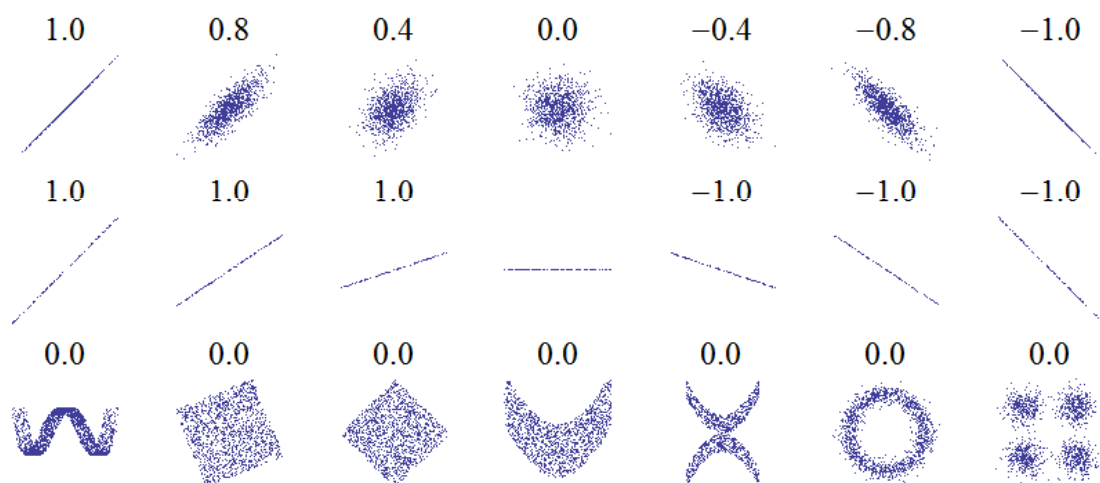
- $r > 0$** Existe una dependencia positiva.
Cuanto más se acerque a 1, mayor es ésta.
- $r < 0$** Existe una dependencia negativa.
Cuanto más se acerque a -1, mayor será.
- $r = 0$** No podemos afirmar nada.



Técnicas de regresión



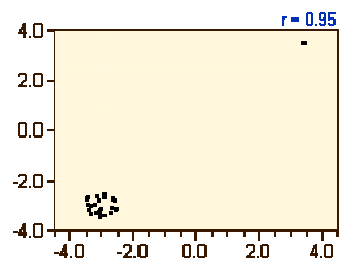
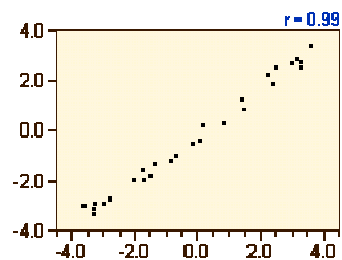
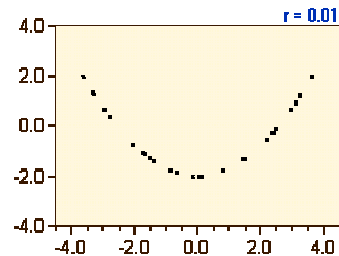
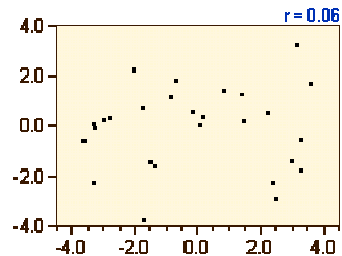
Coefficiente de correlación



Técnicas de regresión



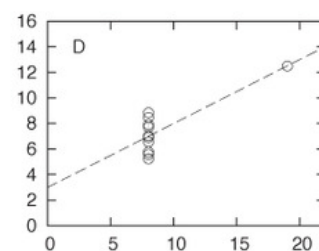
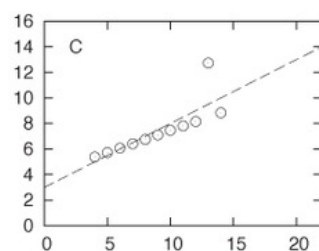
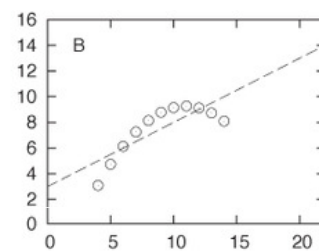
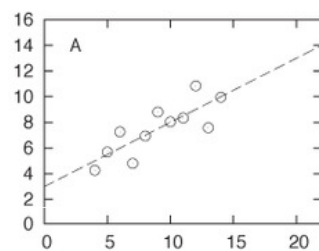
Coefficiente de correlación



Técnicas de regresión



Coefficiente de correlación



El cuarteto de Anscombe

(4 conjuntos de datos con el mismo coeficiente de correlación)



Técnicas de regresión



Coefficiente de correlación

Ventaja de r

- No depende de las unidades usadas en la medición.

Limitaciones de r

- Sólo mide dependencia lineal entre las variables.

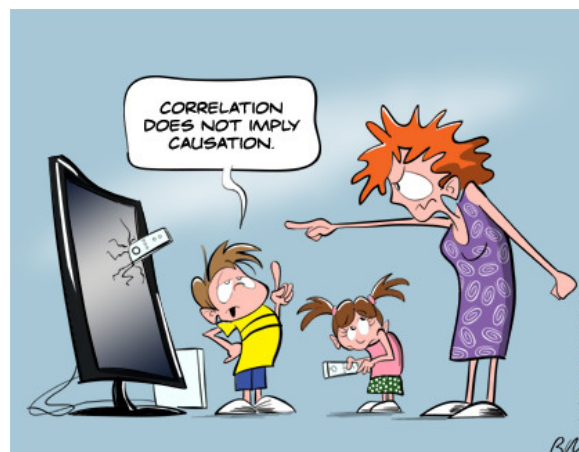
¡OJO! La correlación no implica causalidad...



Técnicas de regresión



Coefficiente de correlación



"Correlation is not causation but it sure is a hint."
-- Edward Tufte



Técnicas de regresión



Más técnicas de predicción...



Forecasting

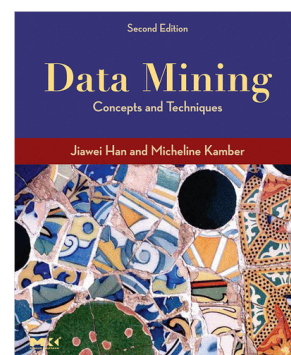
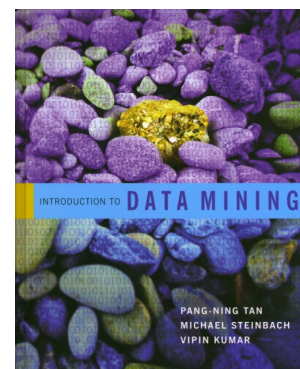
<http://en.wikipedia.org/wiki/Forecasting>



Bibliografía



- Pang-Ning Tan,
Michael Steinbach
& Vipin Kumar:
Introduction to Data Mining
Addison-Wesley, 2006.
ISBN 0321321367 [capítulos 4&5]
- Jiawei Han
& Micheline Kamber:
**Data Mining:
Concepts and Techniques**
Morgan Kaufmann, 2006.
ISBN 1558609016 [capítulo 6]



Bibliografía



- F. Berzal, J.C. Cubero, D. Sánchez, and J.M. Serrano: **ART: A hybrid classification method**. Machine Learning, 2004
- L. Breiman, J. Friedman, R. Olshen, and C. Stone. **Classification and Regression Trees**. Wadsworth International Group, 1984.
- W. Cohen. **Fast effective rule induction**. ICML'95
- R. O. Duda, P. E. Hart, and D. G. Stork. **Pattern Classification**, 2ed. John Wiley and Sons, 2001
- U. M. Fayyad. **Branching on attribute values in decision tree generation**. AAAI'94
- Y. Freund and R. E. Schapire. **A decision-theoretic generalization of on-line learning and an application to boosting**. J. Computer and System Sciences, 1997.
- J. Gehrke, V. Gant, R. Ramakrishnan, and W.-Y. Loh, **BOAT -- Optimistic Decision Tree Construction**. SIGMOD'99.
- J. Gehrke, R. Ramakrishnan, and V. Ganti. **Rainforest: A framework for fast decision tree construction of large datasets**. VLDB'98.



Bibliografía



- T.-S. Lim, W.-Y. Loh, and Y.-S. Shih. **A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms**. Machine Learning, 2000.
- S. K. Murthy, **Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey**, Data Mining and Knowledge Discovery 2(4): 345-389, 1998
- J. R. Quinlan. **Induction of decision trees**. *Machine Learning*, 1:81-106, 1986.
- J. R. Quinlan and R. M. Cameron-Jones. **FOIL: A midterm report**. ECML'93.
- J. R. Quinlan. **C4.5: Programs for Machine Learning**. Morgan Kaufmann, 1993.
- J. R. Quinlan. **Bagging, boosting, and c4.5**. AAAI'96.
- R. Rastogi and K. Shim. **Public: A decision tree classifier that integrates building and pruning**. VLDB'98
- H. Yu, J. Yang, and J. Han. **Classifying large data sets using SVM with hierarchical clusters**. KDD'03.



Modelos basados en reglas de asociación**¿Por qué?**




Buscando entre las mejores reglas de asociación, se superan algunas limitaciones de los árboles de decisión (p.ej. sólo consideran los atributos de uno en uno).

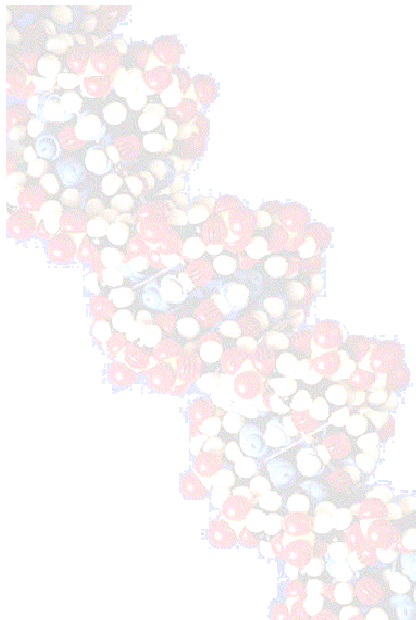
**Modelos basados en reglas de asociación**

- Modelos de clasificación parcial
Bayardo, KDD'1997
- Modelos "asociativos" de clasificación
CBA (Liu, Hsu & Ma, KDD'1998)
RCBT (Cong et al., SIGMOD'2005)
- Patrones emergentes
CAEP (Dong et al., ICDS'1999)
- Árboles de reglas
Wang et al., KDD'2000
- Reglas con excepciones
Liu et al., AAI'2000



Modelos basados en reglas de asociación

- **CMAR**
Classification based on Multiple Association Rules
 Li, Han & Pei, ICDM'2001
- **CPAR**
Classification based on Predictive Association Rules
 Yin & Han, SDM'2003
- **ART**
Association Rule Trees
 Berzal et al., Machine Learning, 2004

**Modelos basados en reglas de asociación****ART [Association Rule Trees]**

```

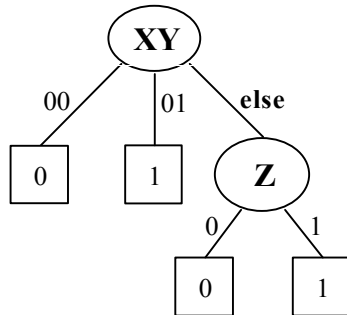
P30 = A : TYPE = N (473|62)
P30 = C : TYPE = N (441|24)
P30 = T : TYPE = N (447|57)
else
P28 = A and P32 = T : TYPE = EI (235|33)
P28 = G and P32 = T : TYPE = EI (130|20)
P28 = C and P32 = A : TYPE = IE (160|31)
P28 = C and P32 = C : TYPE = IE (167|35)
P28 = C and P32 = G : TYPE = IE (179|36)
else
P28 = A : TYPE = N (106|14)
P28 = G : TYPE = N (94|4)
else
P29 = C and P31 = G : TYPE = EI (40|5)
P29 = A and P31 = A : TYPE = IE (86|4)
P29 = A and P31 = C : TYPE = IE (61|4)
P29 = A and P31 = T : TYPE = IE (39|1)
else
P25 = A and P35 = G : TYPE = EI (54|5)
P25 = G and P35 = G : TYPE = EI (63|7)
else
P23 = G and P35 = G : TYPE = EI (40|8)
P23 = T and P35 = C : TYPE = IE (37|7)
else
P21 = G and P34 = A : TYPE = EI (41|5)
else
P28 = T and P29 = A : TYPE = IE (66|8)
else
P31 = G and P33 = A : TYPE = EI (62|9)
else
P28 = T : TYPE = N (49|6)
else
P24 = C and P29 = A : TYPE = IE (39|8)
else
TYPE = IE (66|39)

```

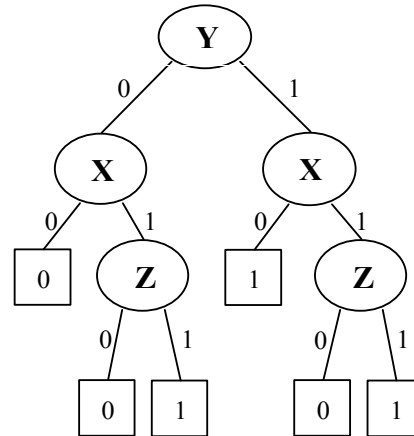


Modelos basados en reglas de asociación

ART



TDIDT



DEMO



ART

Association Rule Trees



Clasificadores bayesianos**Naïve Bayes**

Aplicando el Teorema de Bayes, se maximiza:

$$P(C_i | \mathbf{X}) = P(\mathbf{X} | C_i) P(C_i)$$

$$P(\mathbf{X} | C_i) = \prod_{k=1}^n P(x_k | C_i) = P(x_1 | C_i) \times P(x_2 | C_i) \times \dots \times P(x_n | C_i)$$

Ventaja

- Basta con recorrer los datos una sola vez.

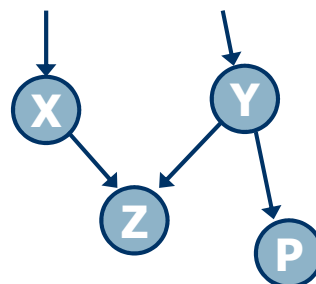
Desventajas

- Interpretabilidad del modelo.
- Supone que las variables son independientes.

**Clasificadores bayesianos****Redes Bayesianas**

Representan mediante un grafo dirigido acíclico dependencias entre variables, especificando sus distribuciones de probabilidad conjuntas.

Nodos: Variables
Enlaces: Dependencias

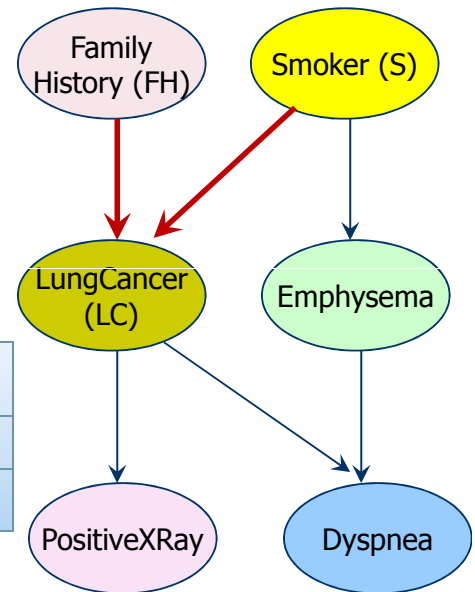


Clasificadores bayesianos**Redes Bayesianas**

CPT [Conditional Probability Table]
para la variable LungCancer:

| P(LC ...) | (FH,S) | (FH, ~S) | (~FH,S) | (~FH, ~S) |
|-----------|--------|----------|---------|-----------|
| LC | 0.8 | 0.5 | 0.7 | 0.1 |
| ~LC | 0.2 | 0.5 | 0.3 | 0.9 |

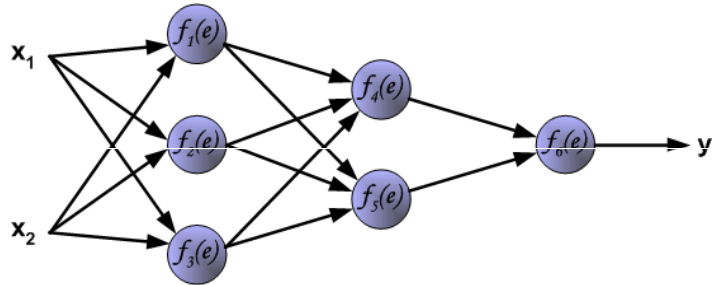
Muestra la probabilidad condicional de que alguien desarrolle cáncer de pulmón para combinación de las variables que lo "causan".

**Clasificadores bayesianos****Redes Bayesianas**

Entrenamiento de las redes bayesianas:

- Dada la estructura de la red, calcular CPTs (sencillo, como en Naïve Bayes).
- Dada la estructura de la red, con algunas variables "ocultas" (desconocidas), buscar una configuración adecuada de la red que encaje con nuestro conjunto de entrenamiento (usando técnicas de optimización como el gradiente descendente).
- Dadas las variables observables, determinar la topología óptima de la red (muy ineficiente).

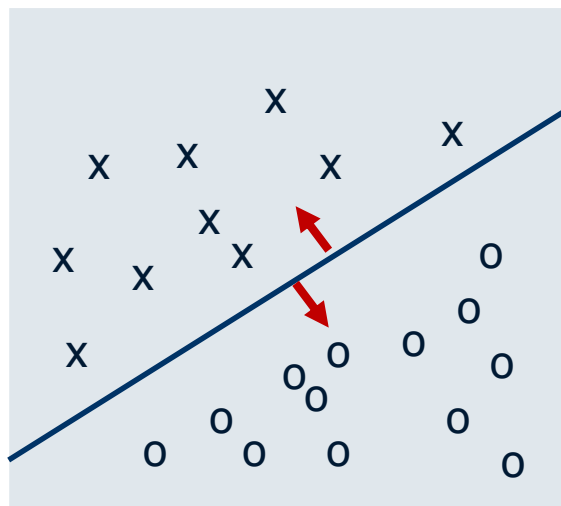


Redes neuronales**p.ej. Backpropagation**

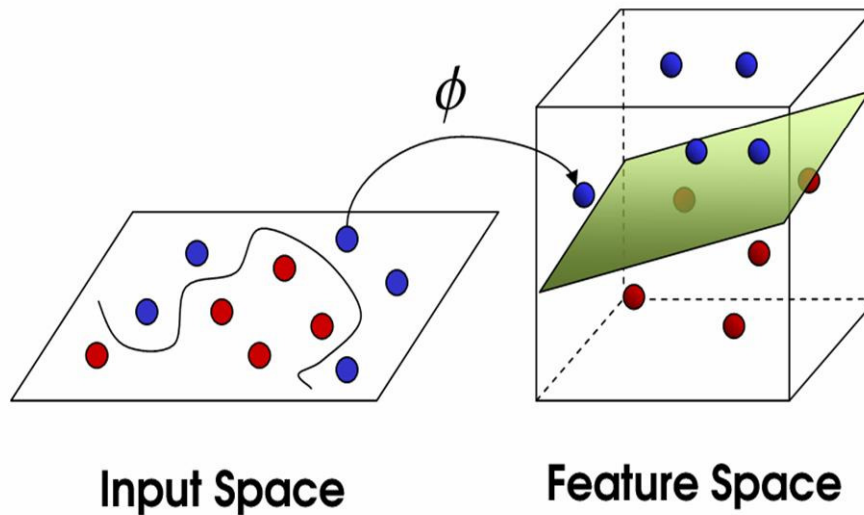
- Como “aproximadores universales”, pueden aplicarse para predecir el valor de un atributo (tanto nominal como numérico).
- Ejemplo de modelo predictivo pero no descriptivo (podemos verlo como una caja negra).



116

SVMs [Support Vector Machines]

117

SVMs [Support Vector Machines]**SVMs [Support Vector Machines]****Ventajas**

- Precisión generalmente alta.
- Robustez frente a ruido.

Desventajas

- Costosos de entrenar (eficiencia y escalabilidad).
- Difíciles de interpretar (basados en transformaciones matemáticas para conseguir que las clases sean linealmente separables)



Clasificadores basados en casos [lazy learners]

Almacenan el conjunto de entrenamiento (o parte de él) y lo utilizan directamente para clasificar nuevos datos.

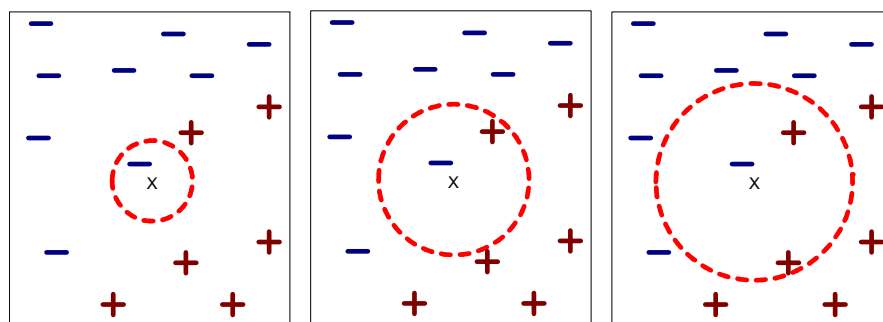
Ejemplos

- k-NN (k Nearest Neighbors)
- Razonamiento basado en casos (CBR)



Clasificadores basados en casos

k-NN



(a) 1-nearest neighbor

(b) 2-nearest neighbor

(c) 3-nearest neighbor

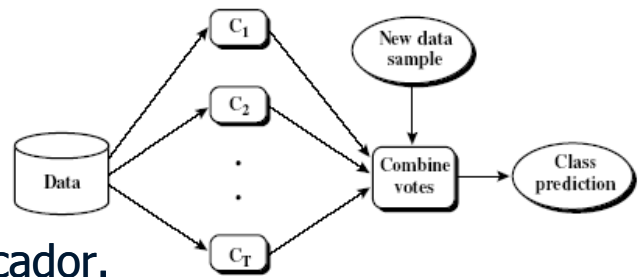
Se escoge la clase más común entre los k vecinos más cercanos:

- k demasiado pequeño
→ Sensible a ruido.
- k demasiado grande
→ El vecindario puede incluir puntos de otras clases.



“Ensembles”

Combinan varios modelos con el objetivo de mejorar la precisión final del clasificador.



- **Bagging:** Votación por mayoría.
Varios clasificadores diferentes votan para decidir la clase de un caso de prueba (usa bootstrapping).
- **Boosting:** Votación ponderada.
Los clasificadores tienen distintos pesos en la votación (en función de su precisión), vg: AdaBoost.

