Modelos Gráficos Probabilistas

L. Enrique Sucar
INAOE

Sesión 5:
Métodos Básicos y Clasificadores

“ ... tenemos razones para creer que hay en la constutución de las cosas leyes de acuerdo a las cuales suceden los eventos ...”

[Richard Price, 1763]

Métodos Básicos

- Probabilidad conjunta
- Cálculo directo (*fuerza bruta*):
  - Probabilidades marginales / condicionales
  - Eventos más probables
  - Estimación de probabilidades
- Clasificación
  - Clasificador bayesiano simple
  - Extensiones
- Discriminadores lineales y discretización
- Evaluación

Formulación

- Muchos problemas se pueden formular como un conjunto de variables sobre las que tenemos cierta información y queremos obtener otra, por ejemplo:
  - Diagnóstico médico o industrial
  - Percepción (visión, voz, sensores)
  - Clasificación (bancos, empleadores, ...)
  - Modelado de estudiantes, usuarios, etc.

Ejemplo

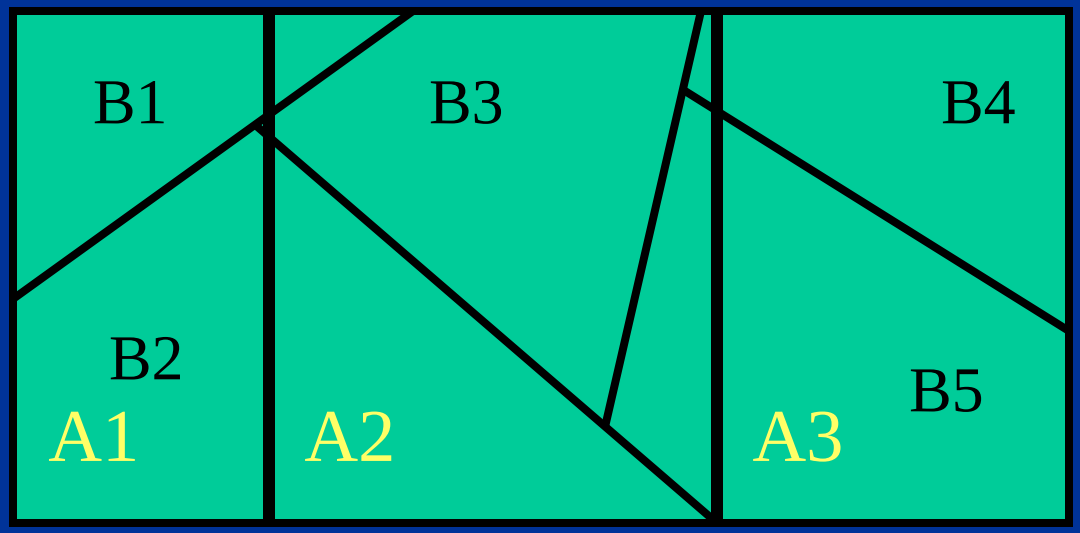
- Determinar si una persona es sujeta de crédito:
  - X1: otorgar crédito (si/no)
  - X2: ingreso anual (entero positivo)
  - X3: créditos anteriores (si/no)
  - X4: edad (entero positivo)
  - X5: ocupación (empleado, empresario, …)

Formulación

- Desde el punto de vista de probabilidad se puede ver como:
  - Un conjunto de variables aleatorias: X1, X2, X3, ...
  - Cada variable es generalmente una partición del espacio
  - Cada variable tiene una distribución de probabilidad (conocida o desconocida)

Variables y Particiones

- A = {A1, A2, A3}
- B = {B1, B2, B3, B4, B5}



Preguntas

- Dada cierta información (como valores de variables y probabilidades), se requiere contestar ciertas preguntas, como:
  - Probabilidad de que una variable tome cierto valor [marginal *a priori*]
  - Probabilidad de que una variable tome cierto valor dada información de otra(s) variable(s) [condicional o *a posteriori*]

Preguntas

- Valor de mayor probabilidad de una o más variables [abducción]
- Valor de mayor probabilidad de una o más variables dada información de otra(s) variable(s) [abducción parcial o explicación]
- Parámetros del modelo dados datos históricos de las variables [estimación o aprendizaje]

Enfoque básico (*fuerza bruta*)

- Dada la probabilidad conjunta de las variables, para todos los posibles valores de cada una (asumimos por ahora que son discretas):

$$P(X1, X2, X3, ..., Xn)$$

- podemos estimar todas las probabilidades requeridas

Inferencia

- Probabilidad marginal (cuál es la probabilidad de las diferentes ocupaciones):

$$p(X) = \Sigma_{Y,\,Z}\, p(X,Y,Z)$$

- Probabilidad condicional (cuál es la probabilidad de otorgar el crédito dado cierto nivel de ingreso) :

$$p(X \mid Y) = p(X,Y) / p(Y)$$

- Donde:

$$p(X,Y) = \Sigma_{Z}\, p(X,Y,Z)$$

Abducción

- Valor más probable (qué tipo de ocupación es el más común):

$$\text{Arg}_X \left[\max p(X) = \max \Sigma_{Y,\,Z}\, p(X,Y,Z)\right]$$

- Valor condicional más probable (debo o no otorgar el crédito):

$$\text{Arg}_X \left[\max p(X \mid y1) = \max p(X,y1) / p(y1)\right]$$

- Valor conjunto más probable (que combinación de ocupación y edad es la más probable):

$$\text{Arg}_{X,Y} \left[\max p(X,Y) = \max \Sigma_Z\, p(X,Y,Z)\right]$$

Ejemplo

- Problema de decidir cuando jugar golf?
- Variables
  - Ambiente
  - Temperatura
  - Viento
  - Humedad
  - Jugar

Ejemplo

- Consideremos inicialmente dos variables: ambiente (S,N,Ll) y temperatura (A,M,B)
- Dada la tabla de P conjunta, encontrar:
  - Probabilidad de ambiente, temperatura
  - Probabilidad de ambiente conocida la temperatura (y viceversa)
  - Combinación de A y T más probable
  - Ambiente más probable dada la temperatura (y viceversa)

Ejemplo

Limitaciones

- El tamaño de la tabla y el número de operaciones crece exponencialmente con el número de variables
- La “tabla” conjunta nos dice poco sobre el fenómeno que estamos analizando
- Puede ser difícil estimar las probabilidades requeridas (por expertos o a partir datos)

Estimación de Parámetros

- Dados un conjunto de valores de las variables (registros), se busca estimar las probabilidades conjuntas requeridas
- Considerando datos completos:
  - Las probabilidades se pueden *estimar* contando el número de casos de cada valor

$$P(Xi, Yj) \sim Ni,j / N$$

  - Esto corresponde al *estimador de máxima verosimilitud* cuando no hay valores faltantes

Ejemplo

- Dados datos sobre lo que “jugadores” han hecho en situaciones pasadas, podemos estimar la probabilidad conjunta
- Consideremos el caso de 2 variables (ambiente y temperatura) y 14 registros de datos

Ejemplos

Ambiente	Temp.	Humedad	Viento	Jugar
soleado	alta	alta	no	N
soleado	alta	alta	si	N
nublado	alta	alta	no	P
lluvia	media	alta	no	P
lluvia	baja	normal	no	P
lluvia	baja	normal	si	N
nublado	baja	normal	si	P
soleado	media	alta	no	N
soleado	baja	normal	no	P
lluvia	media	normal	no	P
soleado	media	normal	si	P
nublado	media	alta	si	P
nublado	alta	normal	no	P
lluvia	media	alta	si	N

Ejemplo

Hoja de cálculo de Microsoft Excel

Limitaciones

- Se requiere una gran cantidad de datos para estimaciones confiables
- Se complica si hay datos faltantes
- Puede ser mejor estimar probabilidades marginales o condicionales (menos datos, más fácil para el experto)
- También puede ser complejo el tener *demasiados* datos (minería de datos)

Alternativas

- El problema de complejidad computacional utilizando el enfoque básico (tanto en espacio para representar el modelo, como en tiempo para el cálculo de probabilidades), nos lleva a buscar alternativas
- Los modelos gráficos probabilistas proveen esta alternativa, mediante representaciones muchos más compactas (y entendibles) y técnicas eficientes para el cálculo de las probabilidades

Clasificadores bayesianos

Clasificación

- El concepto de clasificación tiene dos significados:
  - No supervisada: dado un conjunto de datos, establecer clases o agrupaciones (*clusters*)
  - Supervisada: dadas ciertas clases, encontrar una regla para clasificar una nueva observación dentro de las clases existentes

Clasificación

- El problema de clasificación (supervisada) consiste en obtener el valor más probable de una variable (hipótesis) dados los valores de otras variables (evidencia, atributos)

$$\text{Arg}_H \left[\text{Max } P(H \mid E_1, E_2, \ldots E_N) \right]$$
$$\text{Arg}_H \left[\text{Max } P(H \mid \mathbf{E}) \right]$$

$$\mathbf{E} = \{E_1, E_2, \ldots E_N\}$$

Tipos de Clasificadores

- Métodos estadísticos clásicos
  - Clasificador bayesiano simple (*naive Bayes*)
  - Descriminadores lineales
- Modelos de dependencias
  - Redes bayesianas
- Aprendizaje simbólico
  - Árboles de decisión, reglas, …
- Redes neuronales, SVM, …

Clasificación

- Consideraciones para un clasificador:
  - Exactitud – proporción de clasificaciones correctas
  - Rapidez – tiempo que toma hacer la clasificación
  - Claridad – que tan comprensible es para los humanos
  - Tiempo de aprendizaje – tiempo para obtener o ajustar el clasificador a partir de datos

Regla de Bayes

- La probabilidad posterior se puede obtener en base a la regla de Bayes:

$$P(H \mid \mathbf{E}) = P(H)\, P(\mathbf{E} \mid H) / P(\mathbf{E})$$

$$P(H \mid \mathbf{E}) = P(H)\, P(\mathbf{E} \mid H) / \Sigma_i\, P(\mathbf{E} \mid H_i)\, P(H_i)$$

- Normalmente no se requiere saber el valor de probabilidad, solamente el valor más probable de H

Regla de Bayes

- Para el caso de 2 clases H:{0, 1}, la regla de decisión de Bayes es:

$$H^*(E) = \begin{cases} 1 \text{ si } P(H=1 \mid \mathbf{E}) > 1/2 \\ 0, \text{ de otra forma} \end{cases}$$

- Se puede demostrar que la regla de Bayes es óptima

Valores Equivalentes

- Se puede utilizar cualquier función monotónica para la clasificación:

$$\text{Arg}_H \left[\text{Max } P(H \mid \mathbf{E}) \right]$$

$$\text{Arg}_H \left[\text{Max } P(H)\, P(\mathbf{E} \mid H) / P(\mathbf{E}) \right]$$

$$\text{Arg}_H \left[\text{Max } P(H)\, P(\mathbf{E} \mid H) \right]$$

$$\text{Arg}_H \left[\text{Max } \log \{P(H)\, P(\mathbf{E} \mid H)\} \right]$$

$$\text{Arg}_H \left[\text{Max } (\log P(H) + \log P(\mathbf{E} \mid H)) \right]$$

Clasificador bayesiano simple

- Estimar la probabilidad: $P(\mathbf{E} \mid H)$ es complejo, pero se simplifica si se considera que los atributos son independientes dada la hipotesis:

  $P(E_1, E_2, ...E_N \mid H) = P(E_1 \mid H)\, P(E_2 \mid H) \ldots P(E_N \mid H)$

- Por lo que la probabilidad de la hipótesis dada la evidencia puede estimarse como:

$$P(H \mid E_1, E_2, ...E_N) = \frac{P(H)\, P(E_1 \mid H)\, P(E_2 \mid H) \ldots P(E_N \mid H)}{P(\mathbf{E})}$$

- Esto se conoce como el clasificador bayesiano simple

Clasificador bayesiano simple

- Como veíamos, no es necesario calcular el denominador:

$$P(H \mid E_1, E_2, \ldots E_N) \sim$$

$$P(H)\, P(E_1 \mid H)\, P(E_2 \mid H) \ldots P(E_N \mid H)$$

- P(H) se conoce como la *probabilidad a priori*, $P(E_i \mid H)$ es la probabilidad de los atributos dada la hipótesis (*verosimilitud*), y $P(H \mid E_1, E_2, \ldots E_N)$ es la *probabilidad posterior*

Ejemplo

- Para el caso del golf, cuál es la acción más probable (jugar / no-jugar) dado el ambiente y la temperatura?

o ja d e c á lc u lo d e

M ic ro s o ft E x c e l

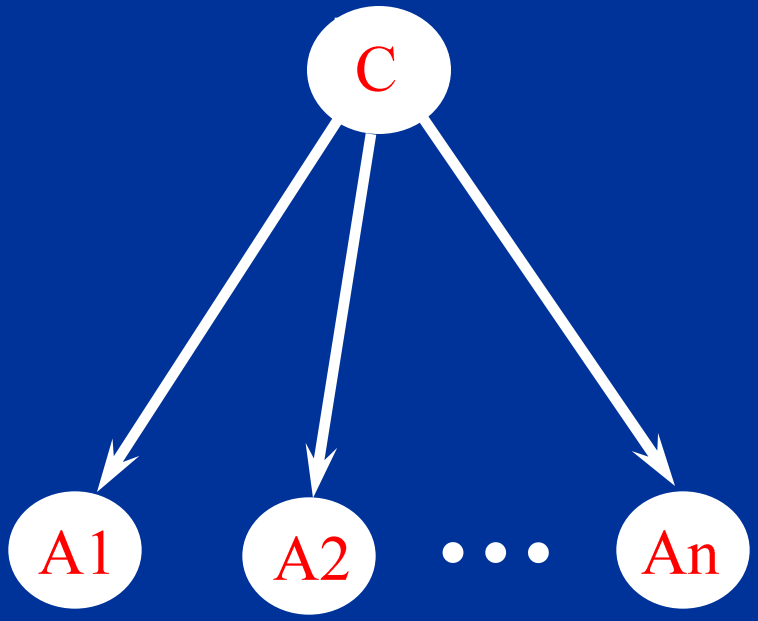
Ventajas

- Bajo tiempo de clasificación
- Bajo tiempo de aprendizaje
- Bajos requerimientos de memoria
- “Sencillez”
- Buenos resultados en muchos dominios

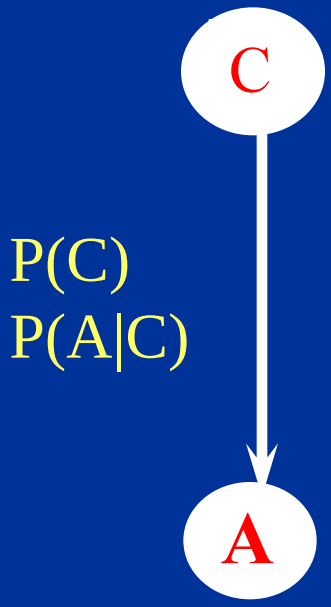
Limitaciones

- En muchas ocasiones la suposición de independencia condicional no es válida
- Para variables continuas, existe el problema de discretización
- Alternativas – dependencias:
  - Estructuras que consideran dependencias
  - Mejora estructural del clasificador
- Alternativas – variables continuas:
  - Descriminador lineal (variables gaussianas)
  - Técnicas de discretización

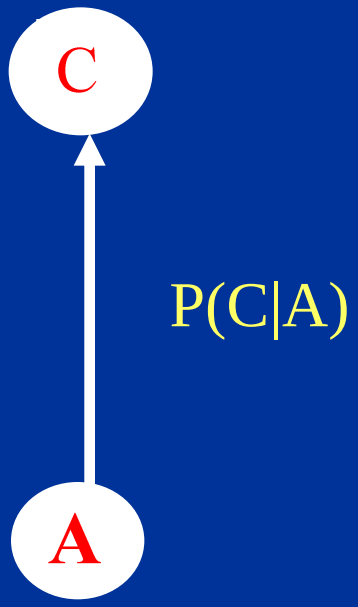
CBS – modelo gráfico



Enfoques para clasificación



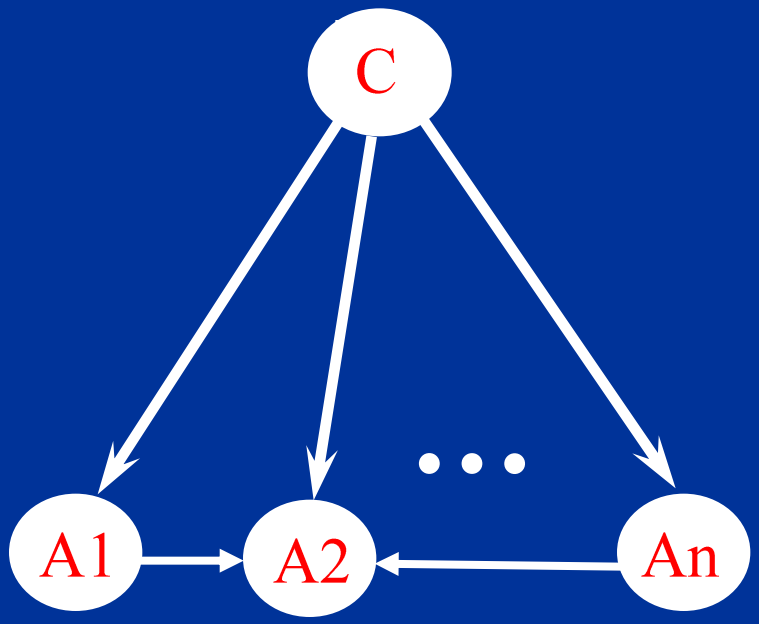
Generativo



Descriminativo

Extensiones

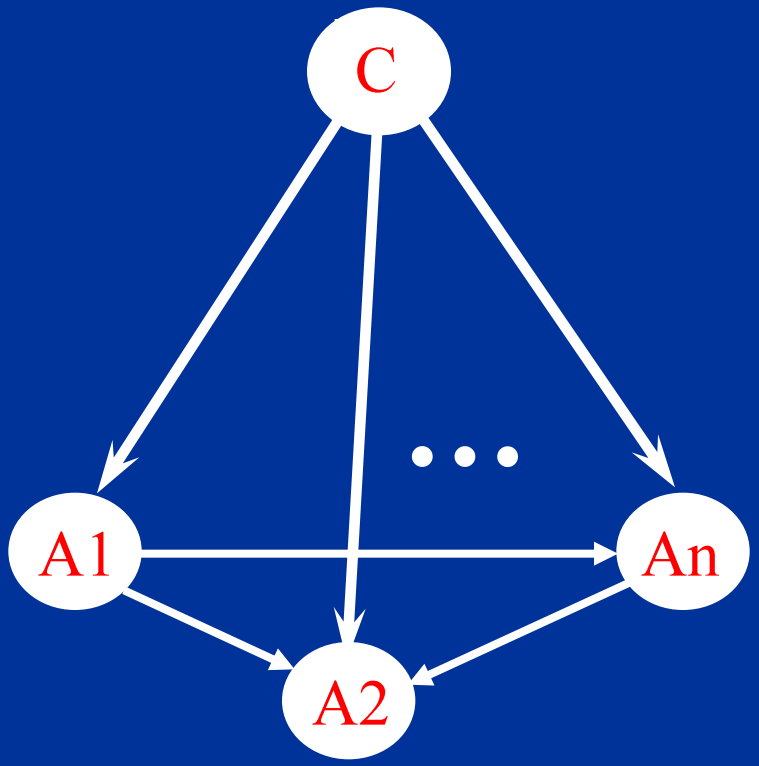
- TAN



Se incorpora algunas dependencias entre atributos mediante la construcción de un "árbol" entre ellos (más adelante veremos como se aprende el árbol)

Extensiones

- BAN



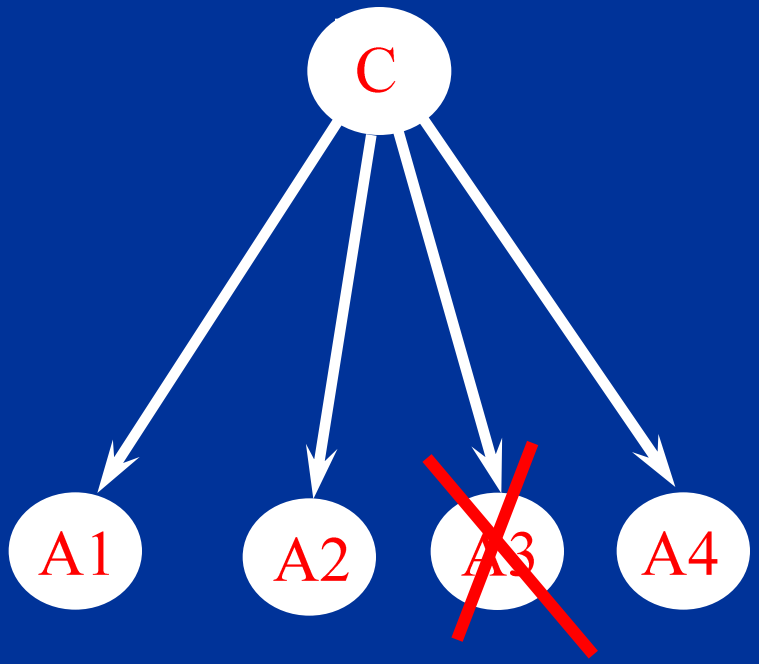
Se incorpora una “red” para modelar las dependencias entre atributos (también más adelante veremos como aprender una red).

Mejora estructural

- Otra alternativa para mejorar el CBS es partir de una estructura “simple” y modificarla mediante:
  - Eliminación de atributos irrelevantes (selección de atributos)
  - Verificación de las relaciones de independencia entre atributos y alterando la estructura:
    - Eliminar nodos
    - Combinar nodos
    - Insertar nodos

Eliminación de atributos

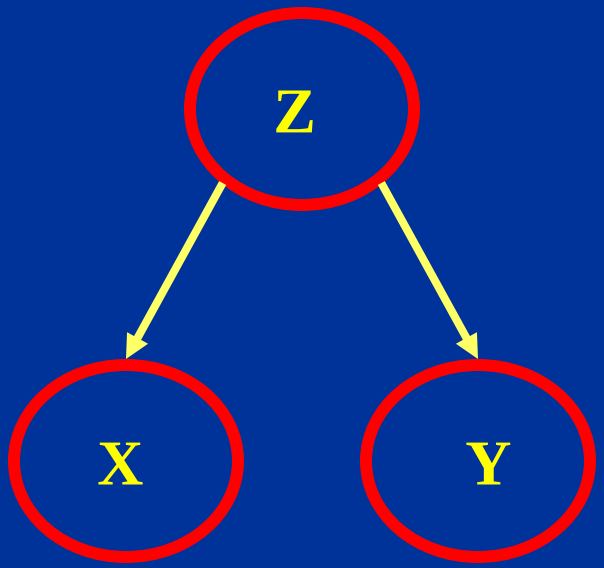
- Medir la “dependencia” entre la clase y atributos (por ejemplo con la información mutua), y eliminar aquellos con “poca” aportación

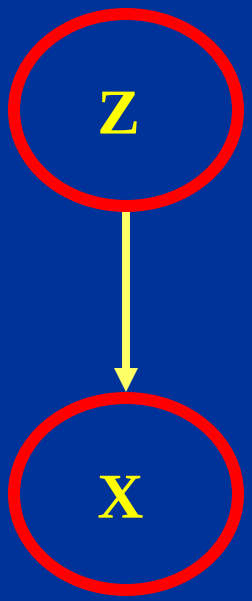


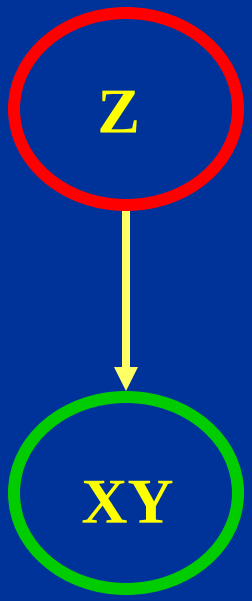
Mejora estructural

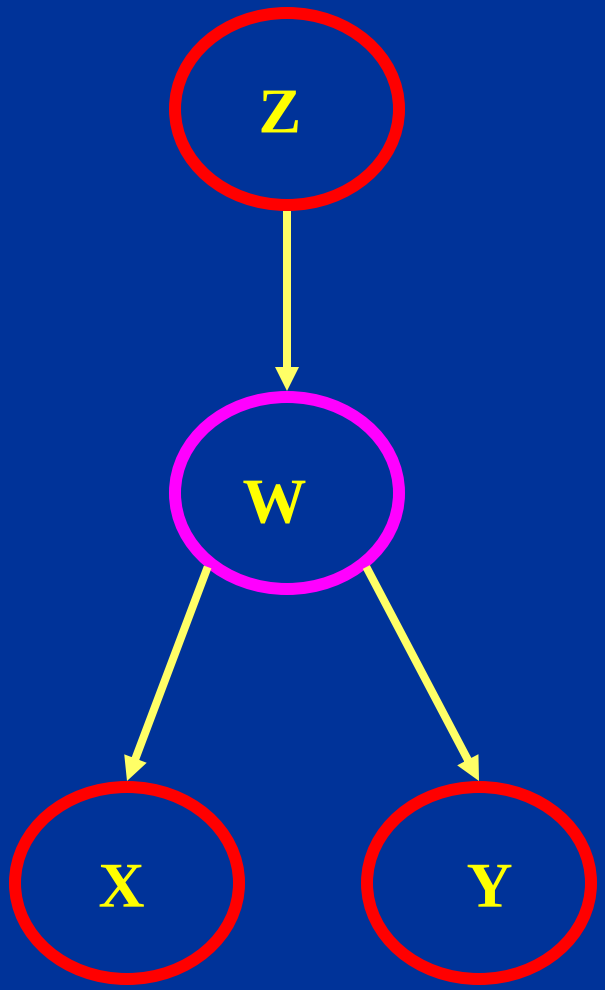
- Medir la dependencia entre pares de atributos dada la clase (por ejemplo mediante la información mutua condicional), alterar la estructura si hay 2 dependientes:
  1. Eliminación: quitar uno de los dos (redundantes)
  2. Unión: juntar los 2 atributos en uno, combinando sus valores
  3. Inserción: insertar un atributo “virtual” entre la clase y los dos atributos que los haga independientes.

Mejora Estructural









Atributos redundantes

- Prueba de dependencia entre cada atributo y la clase
- Información mutua:

$$I(C, Ai) = \Sigma P(C, Ai) \log [P(C, Ai) / P(C) P(Ai)]$$

- Eliminar atributos que no provean información a la clase

Atributos dependientes

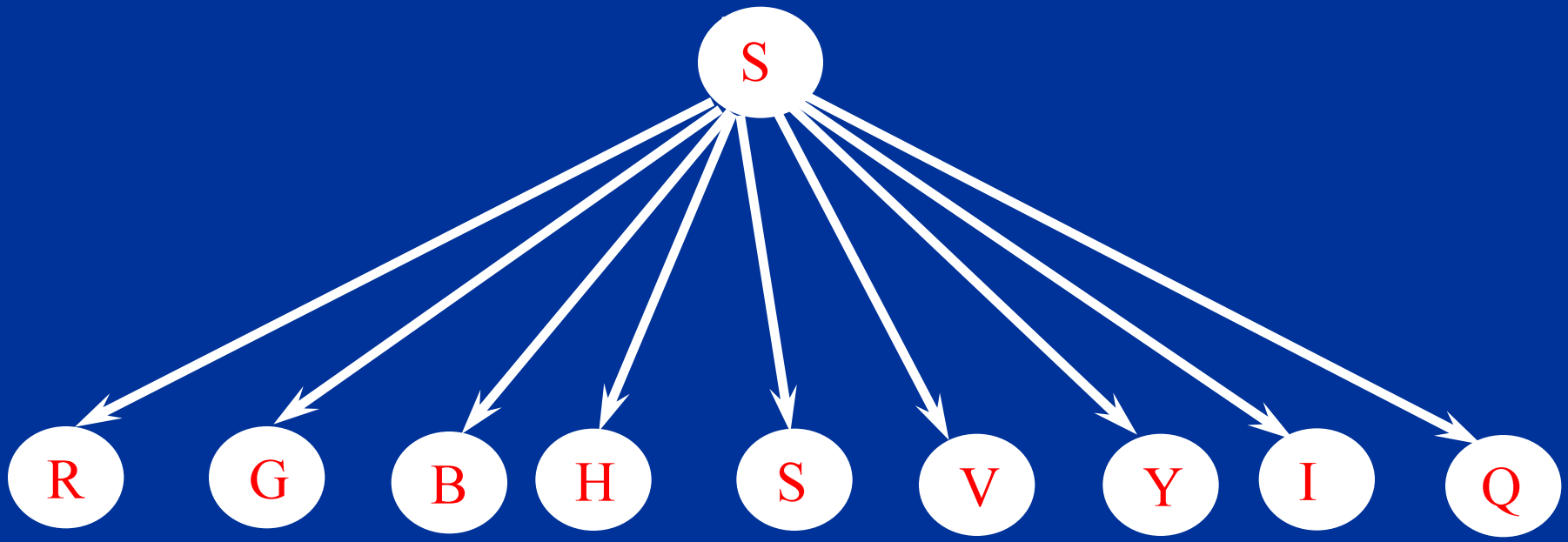
- Prueba de independencia de cada atributo dada la clase
- Información mutua condicional

$$I(A_i, A_j \mid C) = \sum P(A_i,A_j|C) \log [P(A_i,A_j|C) / P(A_i|C)\, P(A_j|C)]$$

- Eliminar, unir o (insertar) atributos

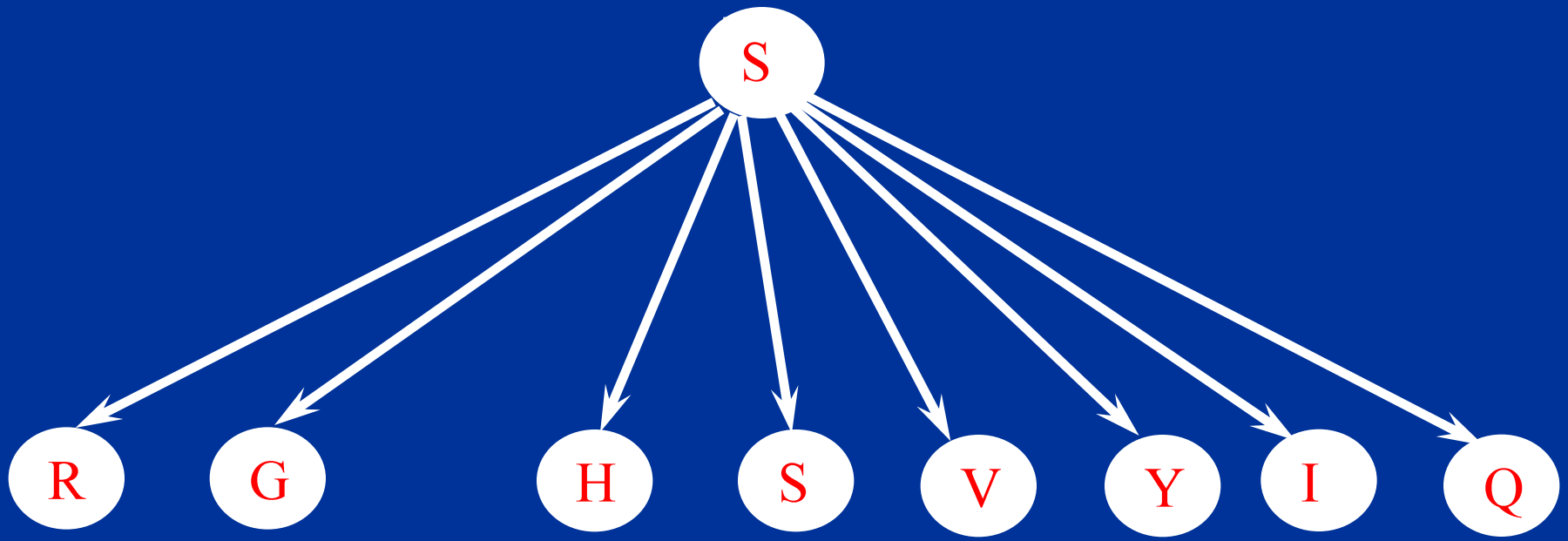
Ejemplo: clasificación de piel

- 9 atributos - 3 modelos de color: RGB, HSV, YIQ

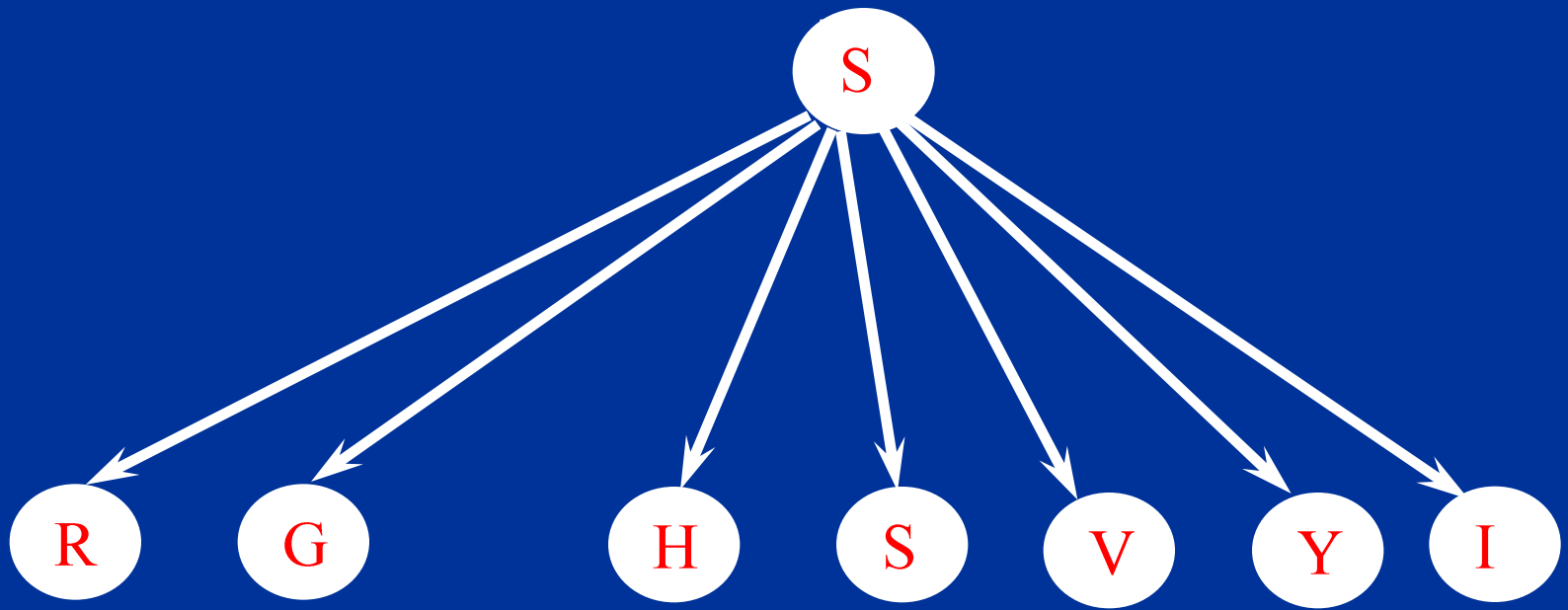


Mejora estructural

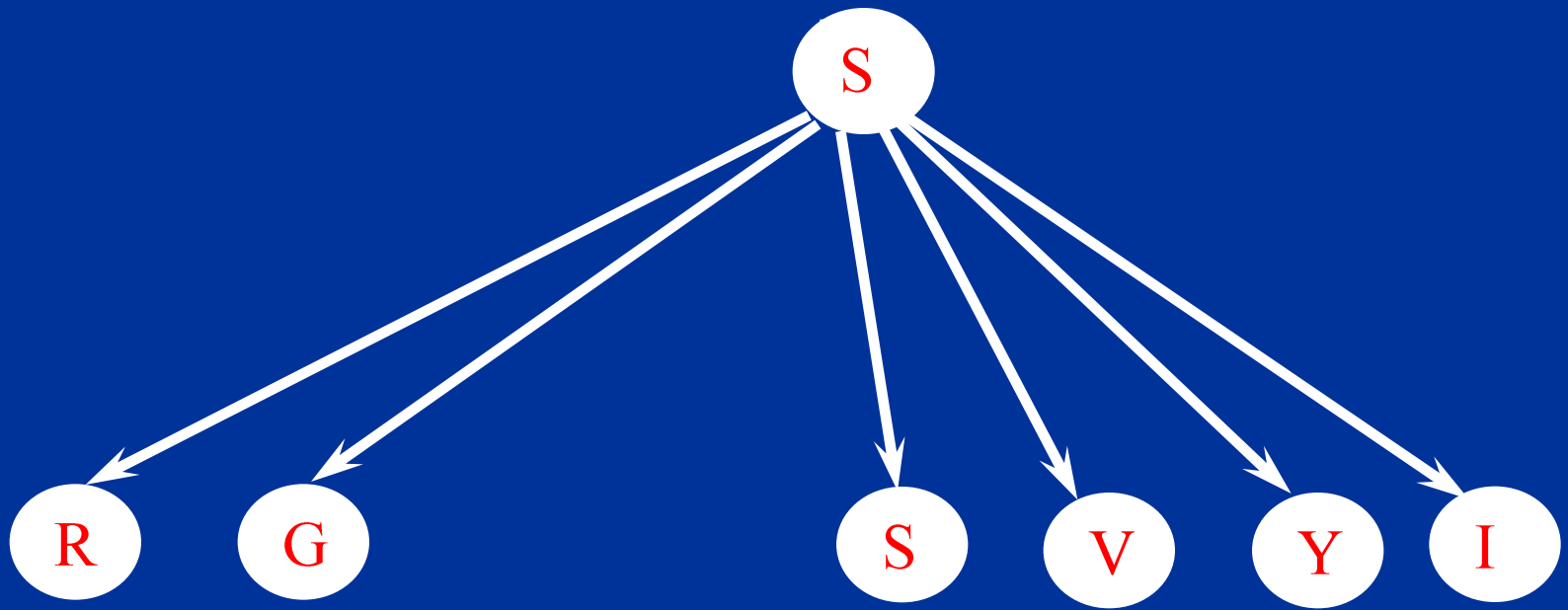
Elimina B



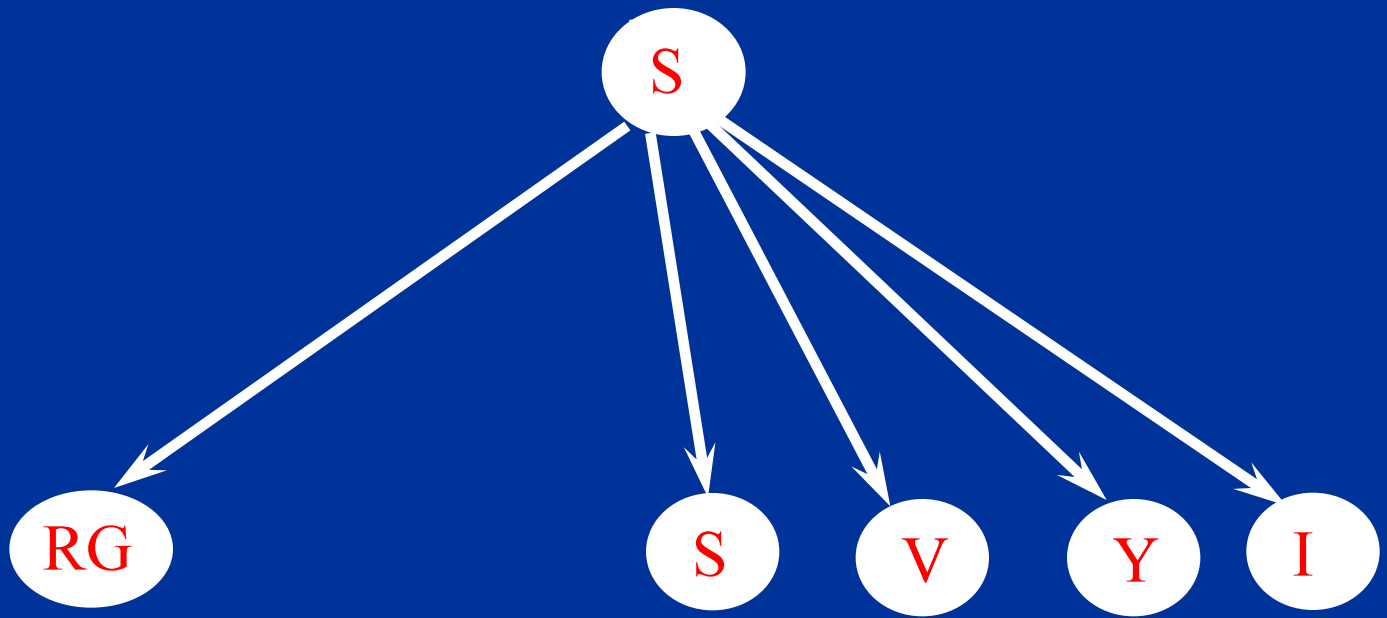
Elimina Q



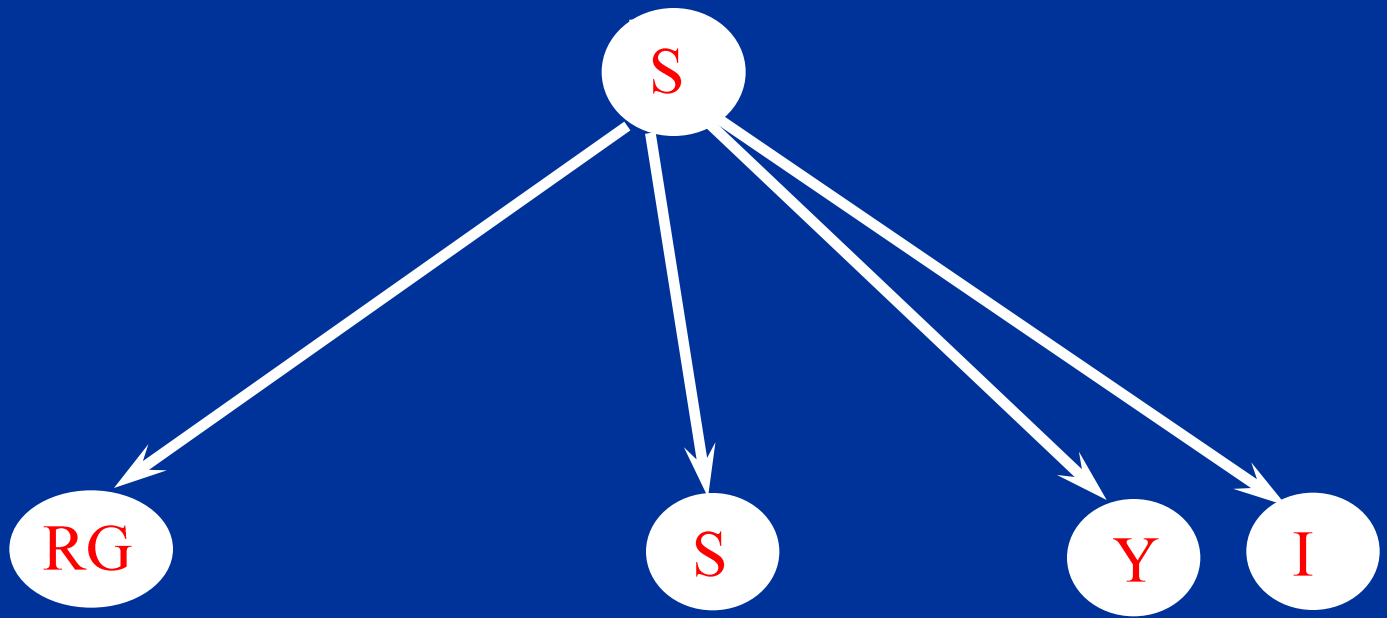
Elimina H



Unir RG

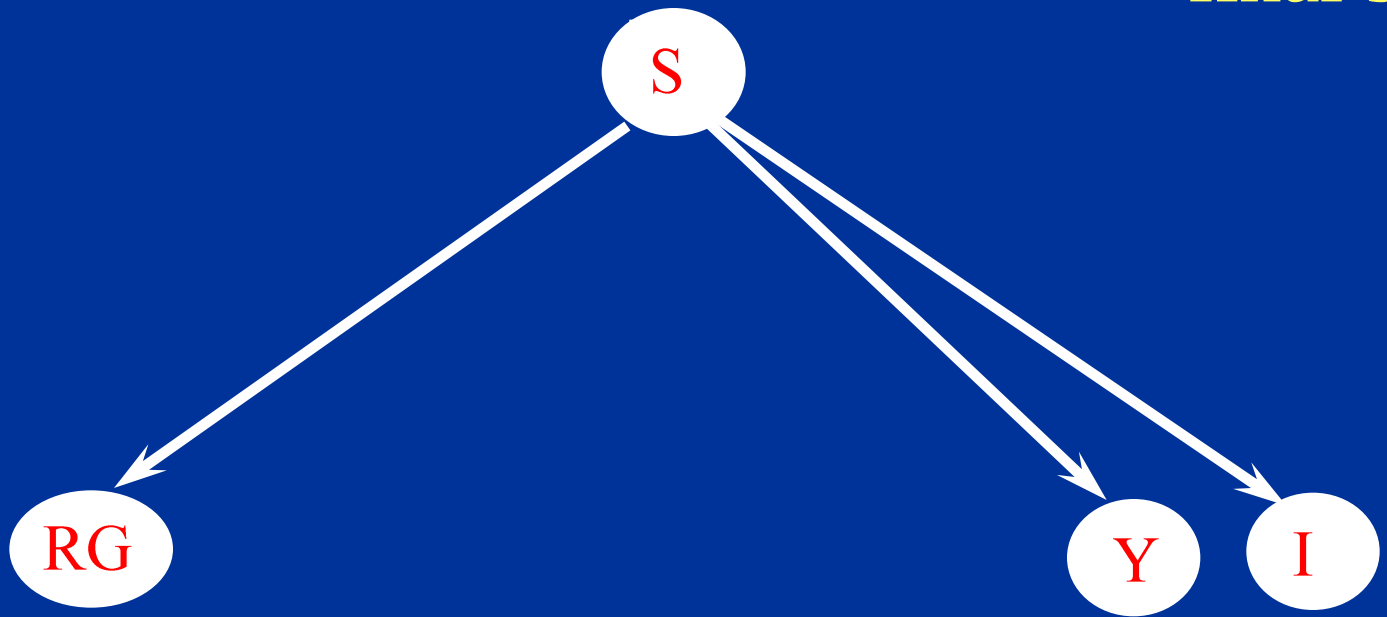


Elimina V



Elimina S

Exactitud: inicial 94%
final 98%



Descriminador lineal

- Se define un hiperplano (descriminante) que es una combinación lineal de los atributos:

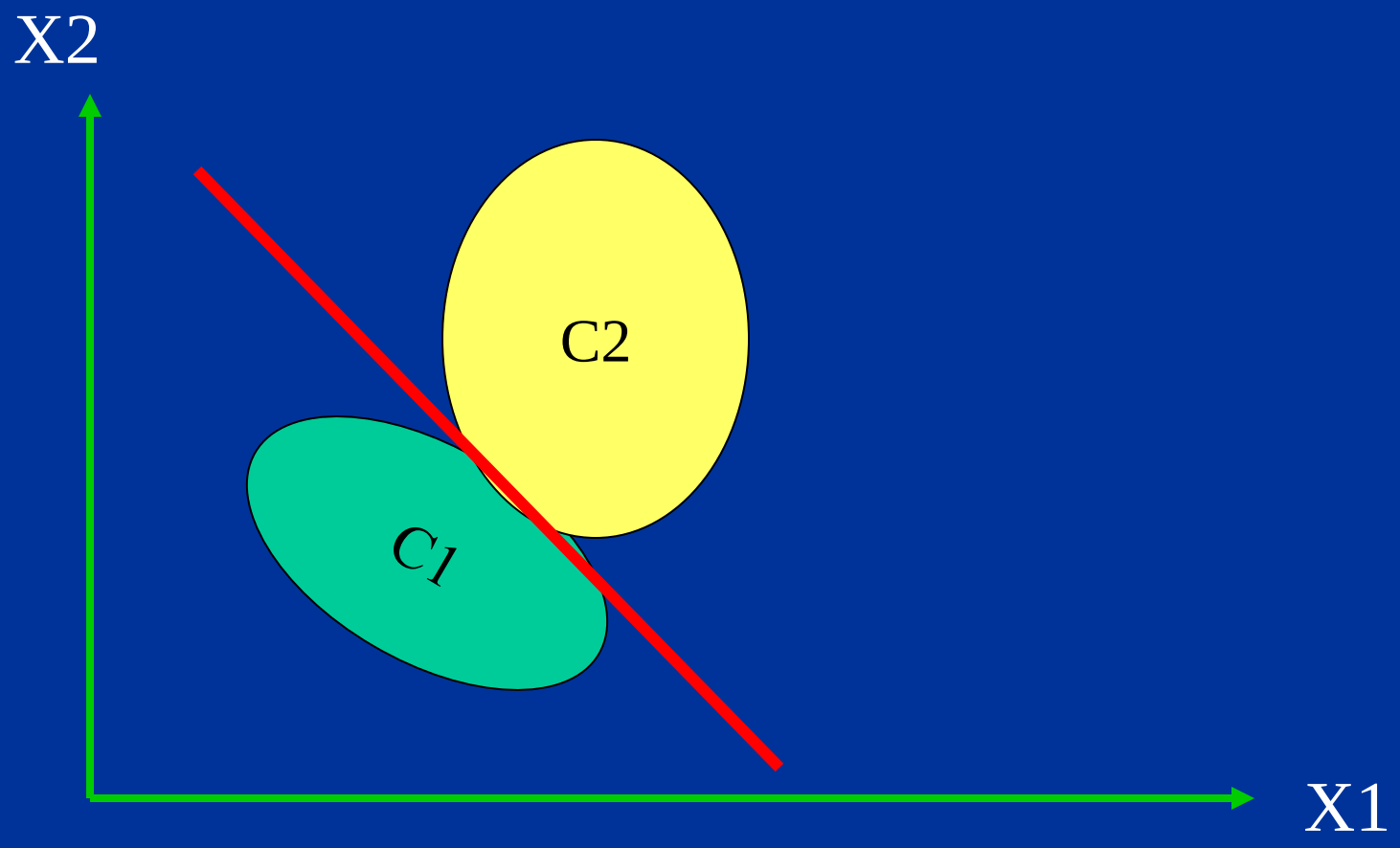
$$g(X) = \Sigma\, a_j\, x_j,$$

x_j – valores de los atributos,

$a_1 \ldots a_n$ - coeficientes

- Asumiendo una distribución normal multivariada, se puede obtener la ecuación del hiperplano en función de los promedios y covarianzas de las clases

Descriminador lineal



Descriminador Lineal

- Para el caso gaussiano, la probabilidad posterior es una función logística (rampa):

$$P(Cn \mid \mathbf{X}) = 1 / [1 + \exp (-\theta^T \mathbf{X})]$$

- Donde el parámetro θ depende de las medias y covarianzas de las distribuciones condicionales de cada clase

Discretización

- Si los atributos no siguen una distribución gaussiana, la alternativa es convertirlos a discretos agrupando los valores en un conjunto de rangos o intervalos
- Dos tipos de técnicas de discretización:
  - No supervisada: no considera la clase
  - Supervisada: en base a la clase

Discretización no supervisada

- Intervalos iguales
- Intervalos con los mismos datos
- En base al histograma

Discretización supervisada

- Considerando los posibles “cortes” entre clases:
  - Probar clasificador (con datos diferentes)
  - Utilizar medidas de información (p. ej., reducir la entropía)
- Problema de complejidad computacional

Costo de mala clasificación

- En realidad, no sólo debemos considerar la clase más probable si no también el costo de una mala clasificación
  - Si el costo es igual para todas las clases, entonces es equivalente a seleccionar la de mayor probabilidad
  - Si el costo es diferente, entonces se debe minimizar el costo esperado

Costo de mala clasificación

- El costo esperado (para dos clases, + y -) está dado por la siguiente ecuación:

$$CE = FN\ p(-)\ C(-|+) + FP\ p(+)\ C(+|-)$$

FN: razón de falsos negativos

FP: razón de falsos positivos

p: probabilidad de negativo o positivo

C(-|+): costo de clasificar un positivo como negativo

C(+|-): costo de clasificar un negativo como positivo

- Considerando esto y también la proporción de cada clase, existen técnicas más adecuadas para comparar clasificadores como la *curva ROC* y las *curvas de costo*

Referencias

- Clasificadores:
  - D. Michie, D.J. Spiegelhalter , C.C. Taylor, “Machine Learning, Neural and Statistical Classification”, Ellis Horwood, 1994
  - L. E. Sucar, D. F. Gillies, D. A. Gillies, "Objective Probabilities in Expert Systems", Artificial Intelligence Journal, Vol. 61 (1993) 187-208.
  - J. Cheng, R. Greiner, “Comparing Bayesian network classifiers”, UAI´99, 101-108.
  - M. Pazzani, “Searching for attribute dependencies in Bayesian classifiers”, Preliminary Papers of Intelligence and Statistics, 424-429.
  - M. Martínez, L.E. Sucar, “Learning an optimal naive Bayesian classifier”, ICPR, 2006

Referencias

- Evaluación:
  - C. Drummond, R. C. Holte, “Explicitly representing expected cost: an alternative to the ROC representation”.

Actividades

- Ejercicios clasificación y métodos básicos
- Ejercicios con Weka (entregar)
- Leer referencias de clasificadores