Modelos Gráficos Probabilistas

L. Enrique Sucar
INAOE

Sesión 6: Clasificadores Bayesianos

" La teoría de probabilidades en el fondo no es nada más que sentido común reducido a cálculos ..."

[Pierre Simon Laplace, 1819]

Clasificadores

- Introducción a Clasificación
- Clasificador bayesiano simple
- Extensiones - TAN, BAN
- Mejora estructural
- Clasificadores multidimensionales
- Discriminadores lineales y discretización
- Evaluación

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:

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

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

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

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

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

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?

Hoja de cálculo de Microsoft Excel

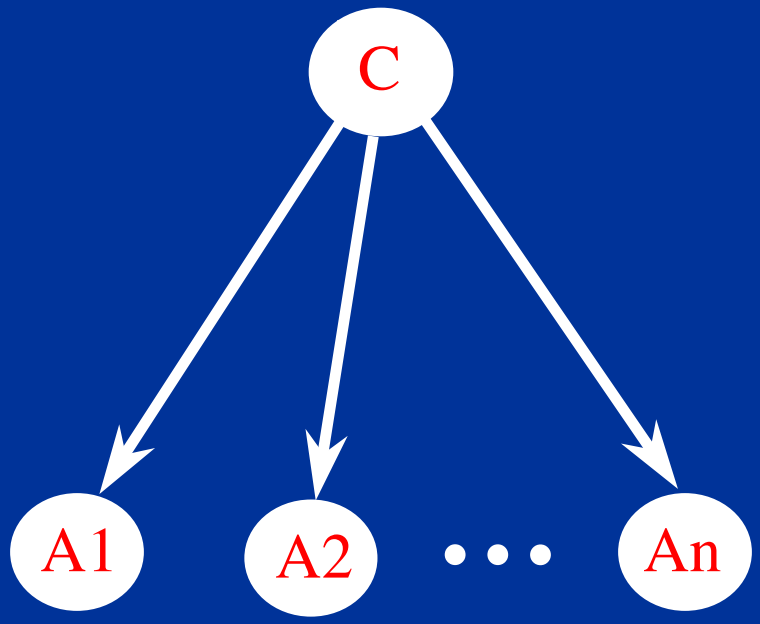
Ventajas

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

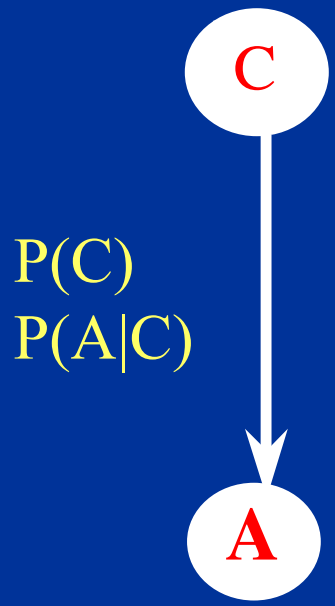
Limitaciones

- 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

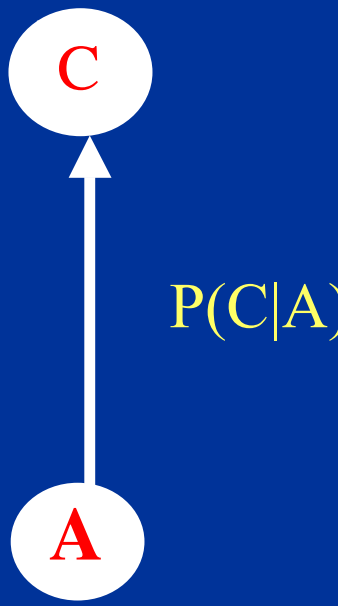
CBS – modelo gráfico



Enfoques para clasificación



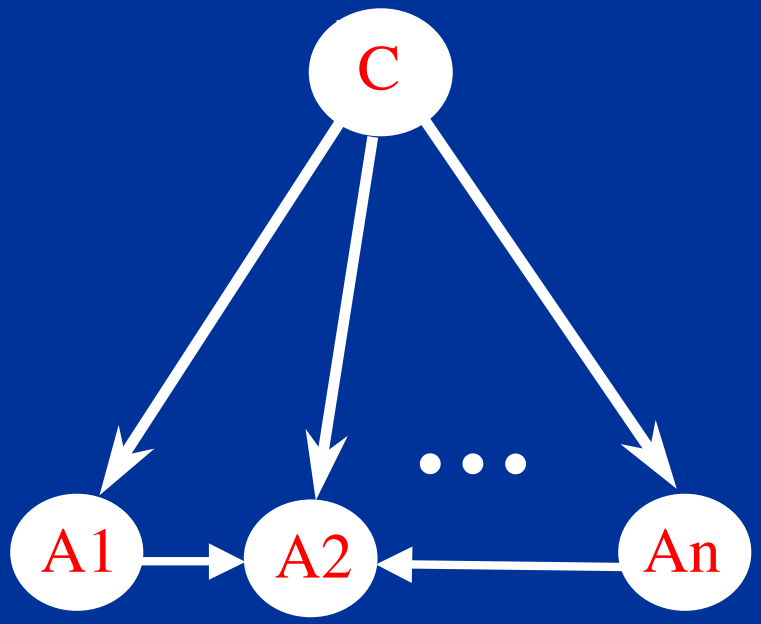
Generativo



Descriminativo

Extensiones

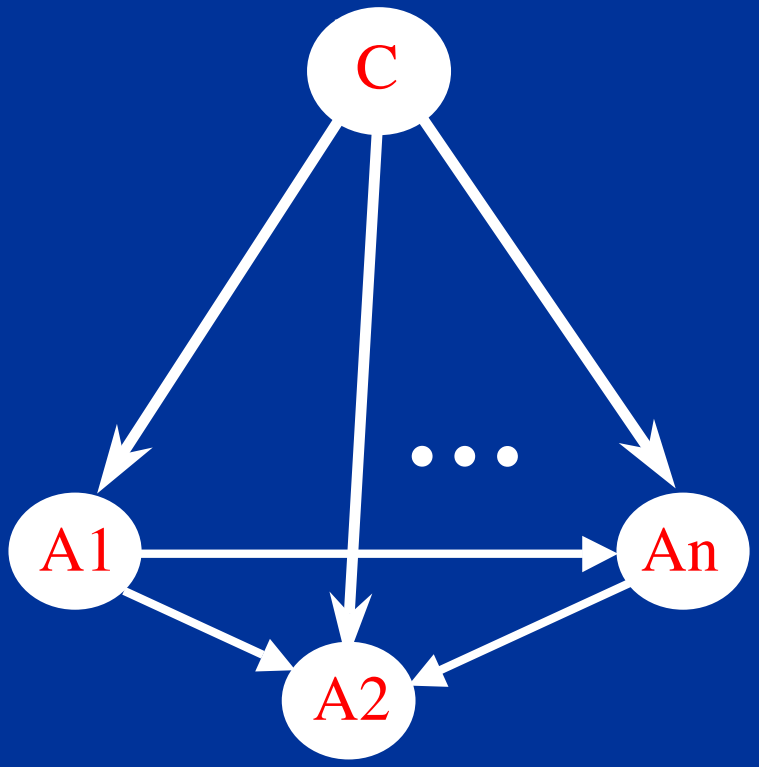
- TAN



Se incorpora algunas dependencias entre atributos mediante la construcción de un "árbol" entre ellos

Extensiones

- BAN



Se incorpora una “red” para modelar las dependencias entre atributos

Extensiones

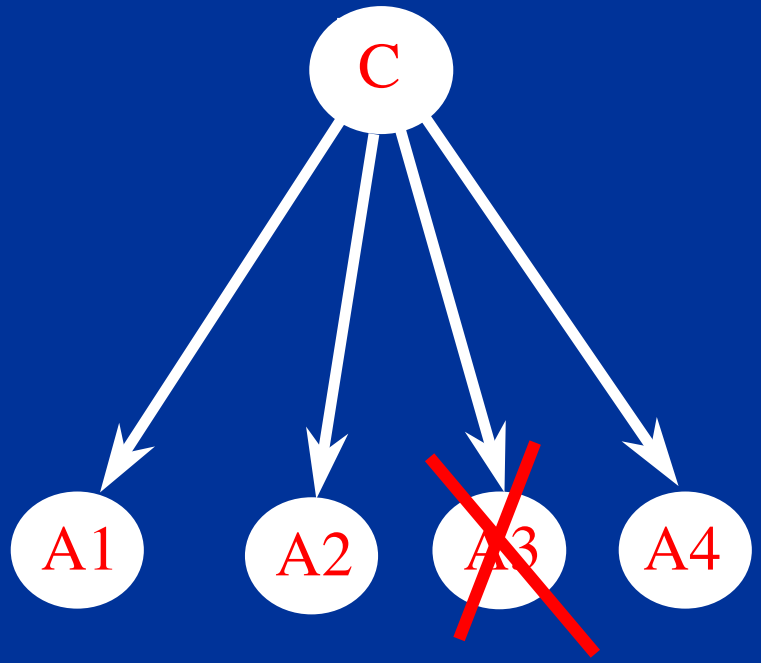
- Los clasificadores TAN y BAN son en casos particulares de redes bayesianas, por lo que para realizar el aprendizaje y clasificación en base a estos modelos se utilizan técnicas de redes bayesianas

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

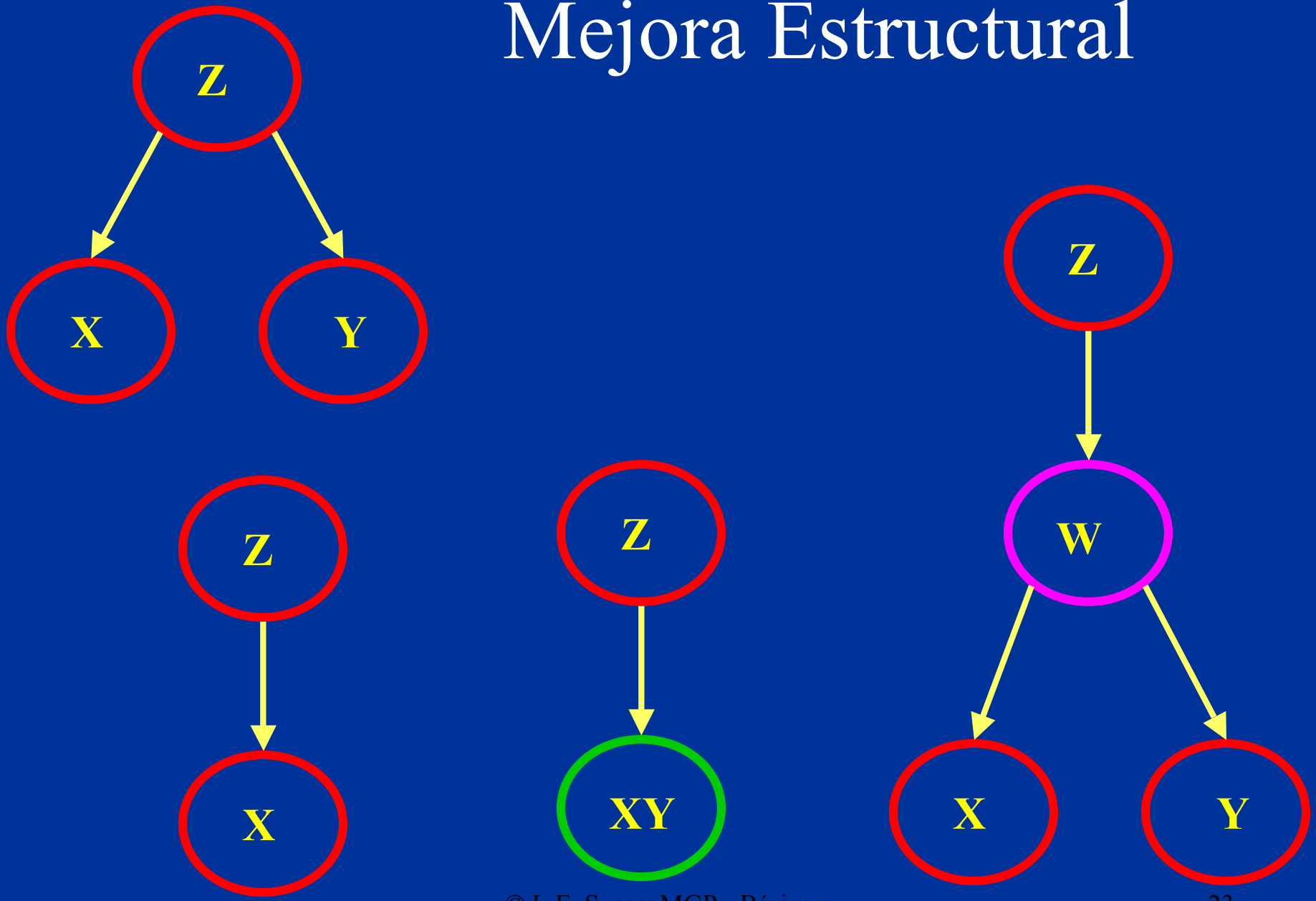
- Medir la "dependencia" entre la clase y atributos (por ejemplo con la información mutua), y eliminar aquellos con "poca" aportación



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

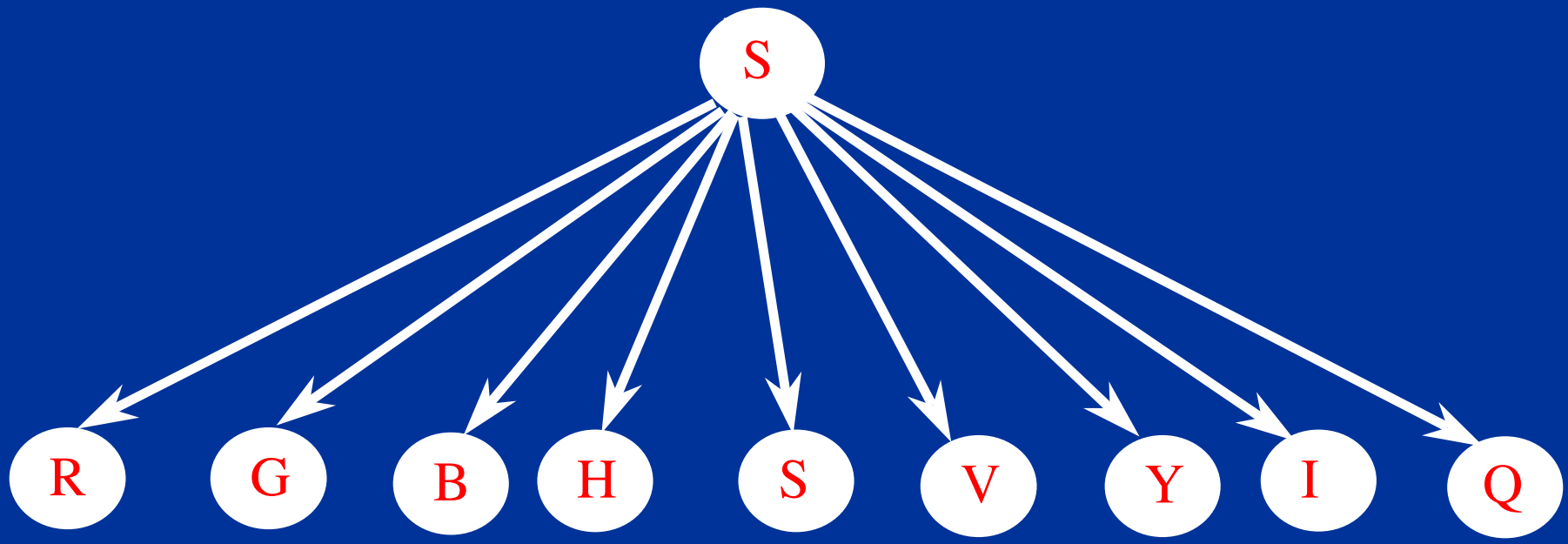
- Prueba de independencia de cada atributo dada la clase
- Información mutua condicional

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

- Eliminar, unir o (insertar) atributos

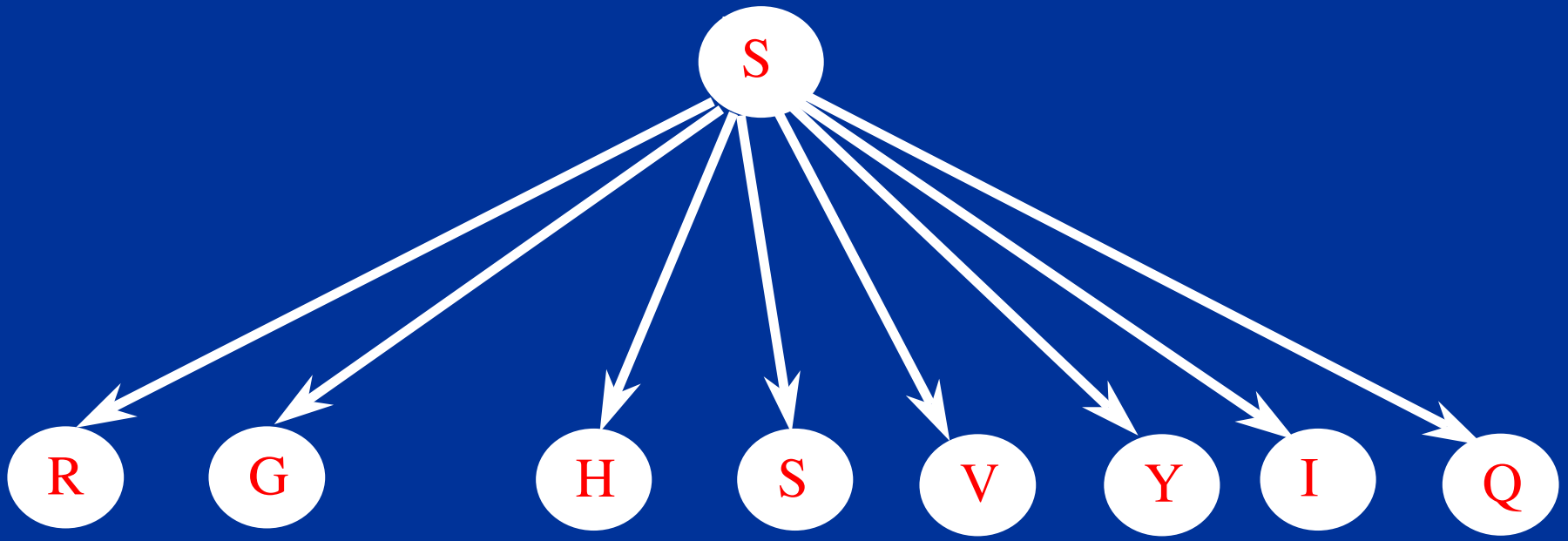
Ejemplo: clasificación de piel

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

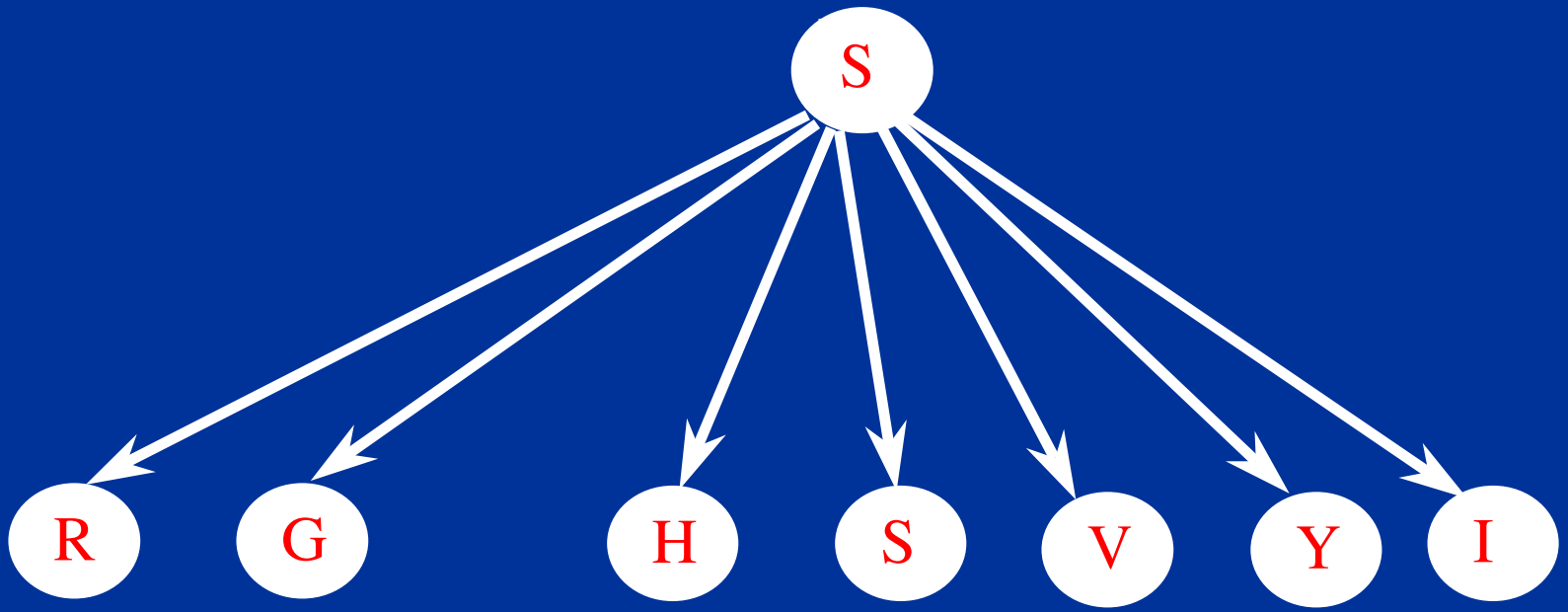


Mejora estructural

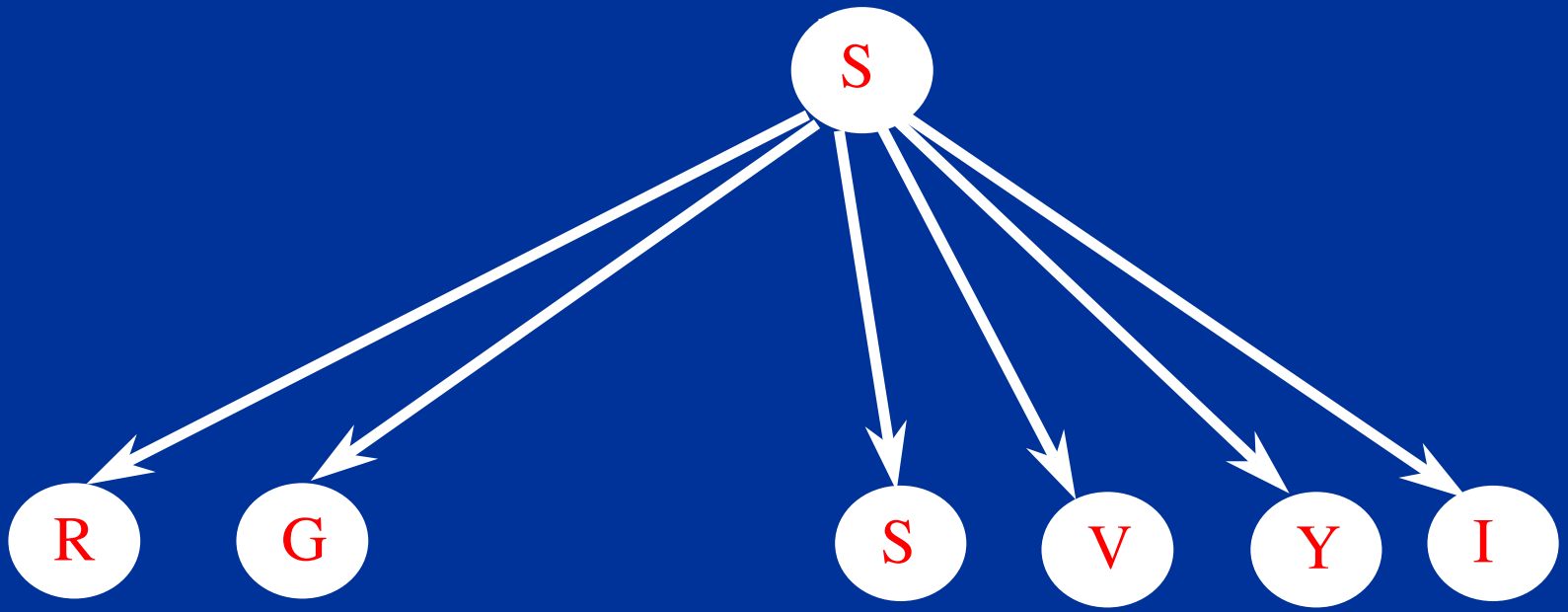
Elimina B



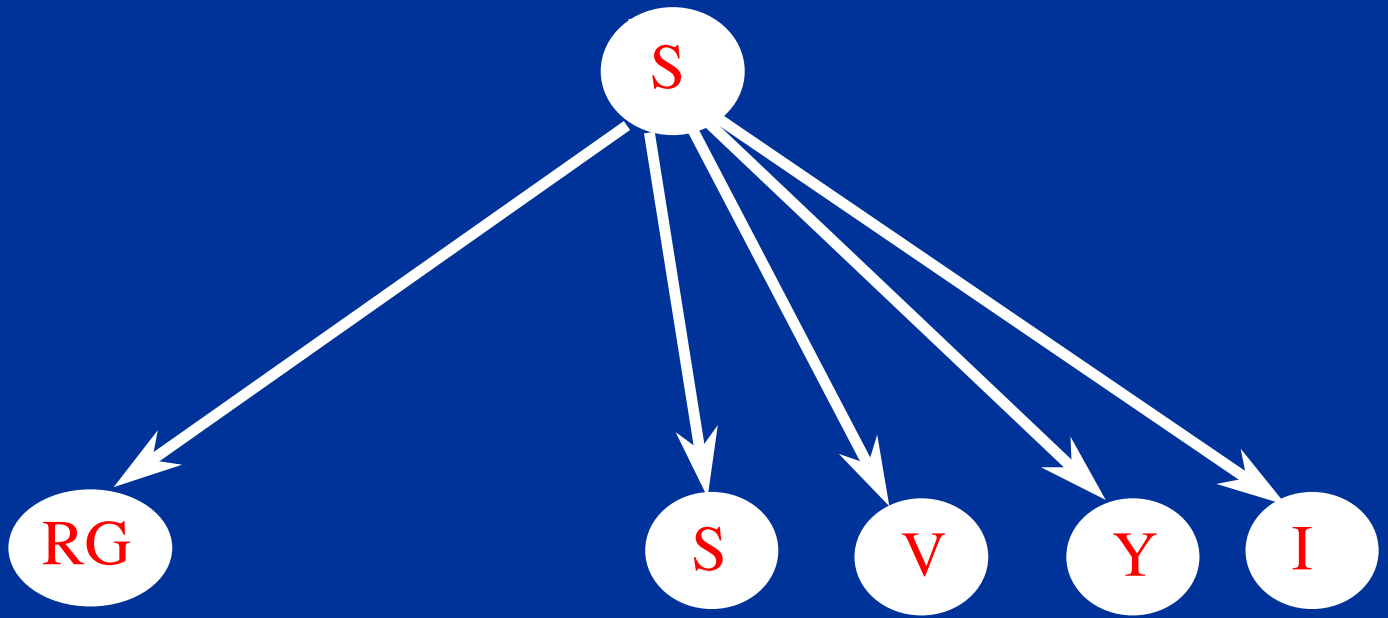
Elimina Q



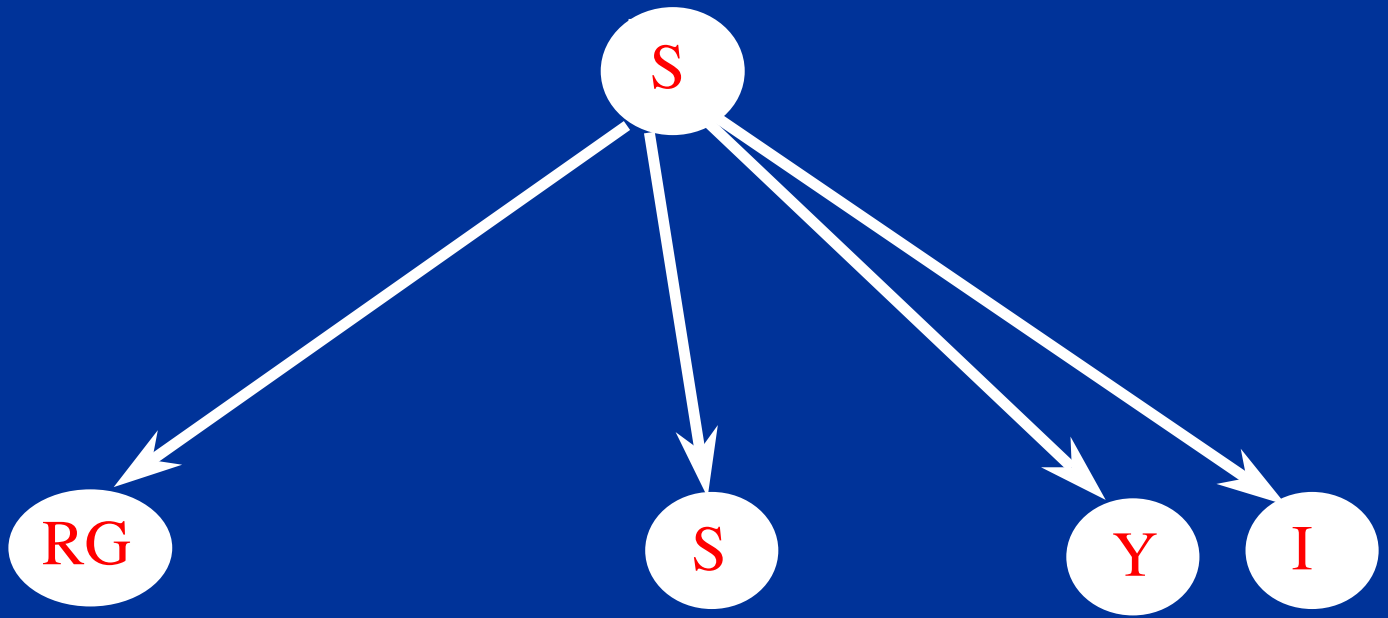
Elimina H



Unir RG

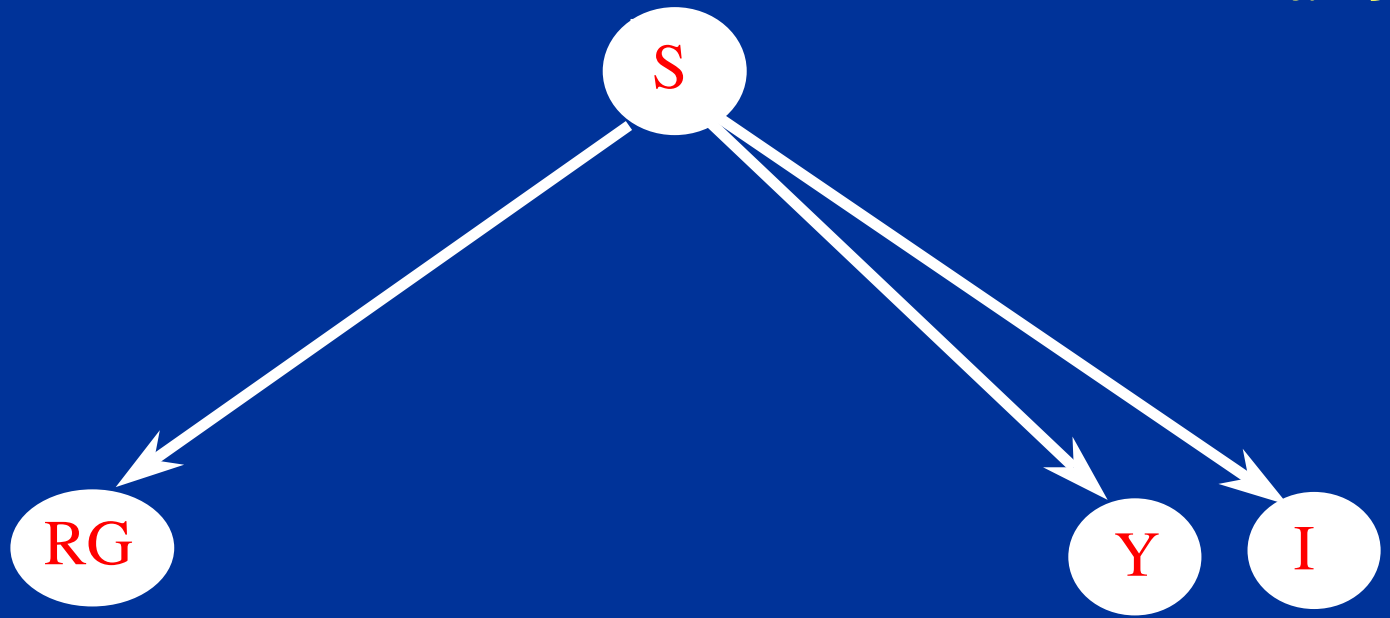


Elimina V



Elimina S

Exactitud: inicial 94%
final 98%



Clasificadores Multidimensionales

- A diferencia de los clasificadores tradicionales (de una dimensión), los clasificadores multidimensionales asignan cada instancia a varias clases a la vez
- Ejemplos:
  - Textos, un documento puede pertenecer a varias categorías
  - Genética, un gen puede tener varias funciones
  - Imágenes, una imagen puede tener varios objetos

Clasificadores Multidimensionales

- En un clasificador multidimensional se desea maximizar la probabilidad del conjunto posible de clases dados los atributos:

$$\text{Arg}_{\mathbf{H}} [\text{Max } P(H_1, H_2, \ldots H_m \mid E_1, E_2, \ldots E_N)]$$

$$\text{Arg}_{\mathbf{H}} [\text{Max } P(H_1, H_2, \ldots H_m \mid \mathbf{E})]$$

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

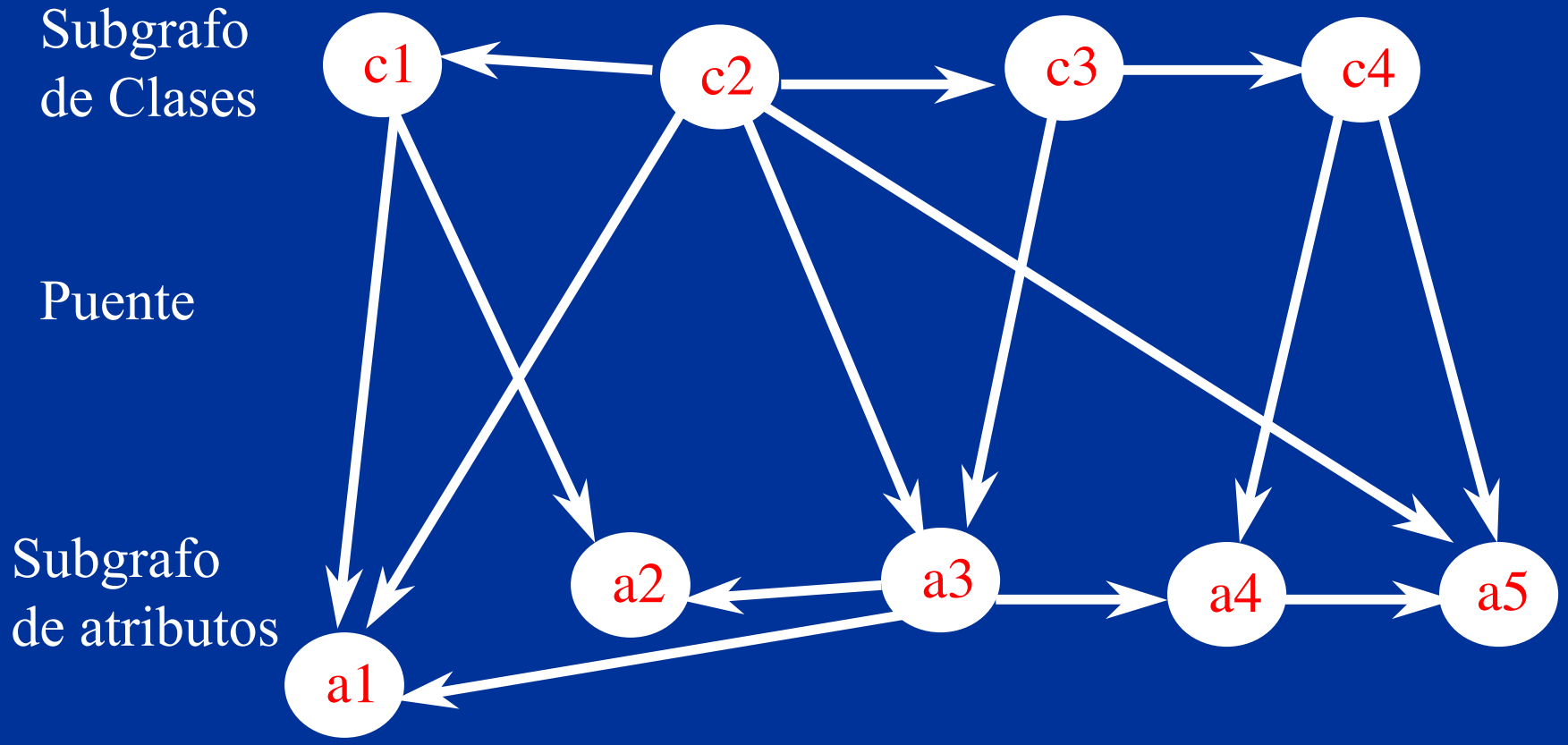
Enfoques Básicos

- Existen dos enfoques básicos para resolver este problema:
  - Binario: se separa en “m” clasificadores independientes y se obtiene un clasificador independiente para cada clase, y luego se concatena el resultado (no considera la dependencia entre clases)
  - Conjunto potencia: se forma una variable clase global que es la combinación de todas las clases individuales (complejidad computacional, aumenta exponencialmente con el número de clases base)

Clasificadores Bayesianos Multidimensionales

- Para considerar las dependencias entre clases y atributos, se construye un modelo gráfico de 3 capas:
  - Dependencias entre Clases
  - Dependencias entre Atributos
  - Dependencias entre Clases y Atributos
- Aprender y clasificar con este modelo es computacionalmente complejo (explicación más probable o *MAP*)

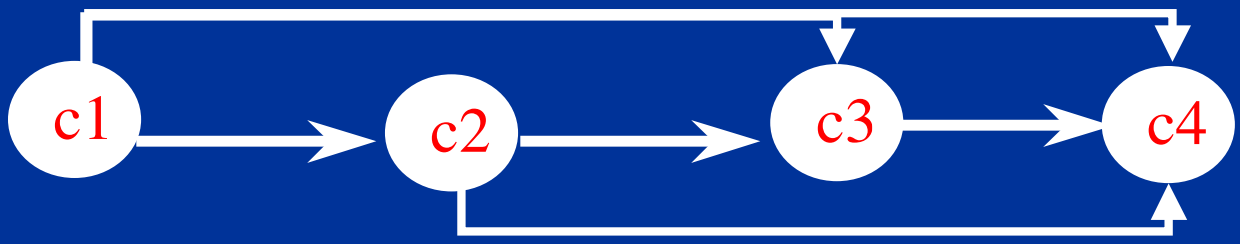
Clasificadores Bayesianos Multidimensionales



Clasificadores en Cadena

- Otra alternativa más sencilla es partir de clasificadores binarios y agregar otras clases como atributos para de cierta forma tomar en cuenta las dependencias entre clases
- Se forma una “cadena” entre las clases de forma que las clases previas en la cadena se incluyen como atributos de las siguientes clases

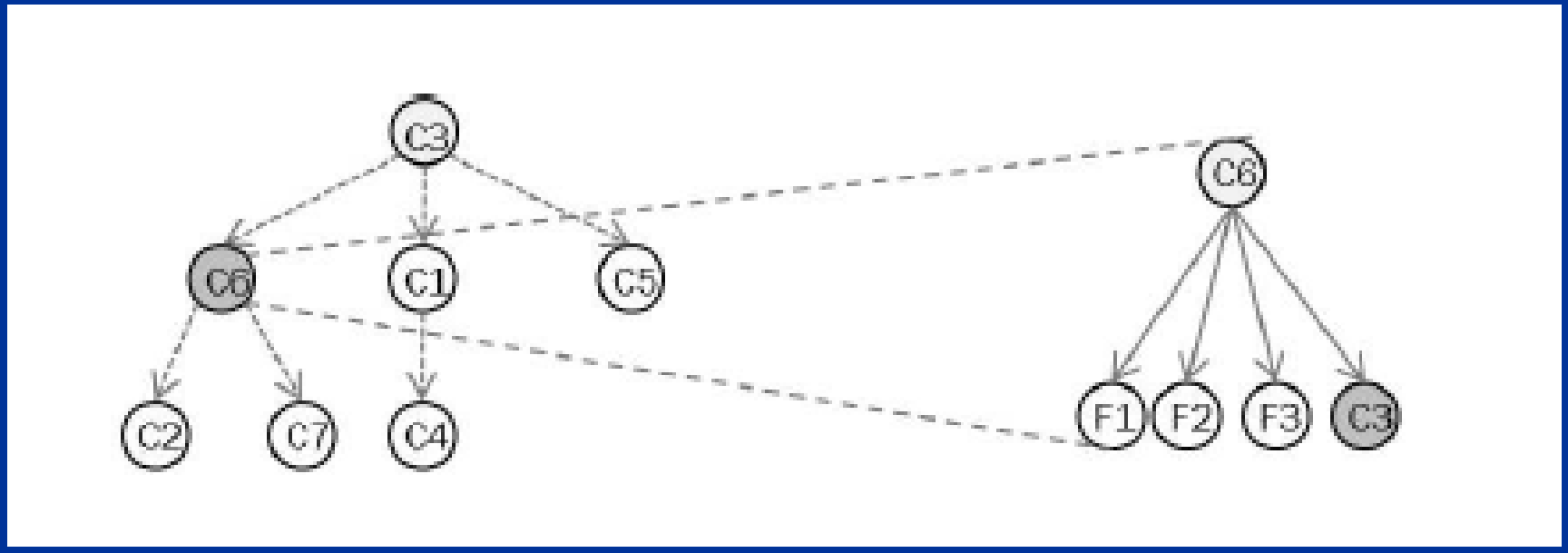
Clasificadores en Cadena



Clasificadores en Cadena Bayesianos

- Se aprende un modelo de dependencias entre clases (red bayesiana)
- El modelo de dependencias puede ser de diversos tipos: árbol, poliárbol, red multiconectada
- Dicho modelo se utiliza para definir el orden de la cadena y los atributos adicionales que se incluyen en cada clase

Clasificadores en Cadena Bayesianos



Descriminador lineal

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

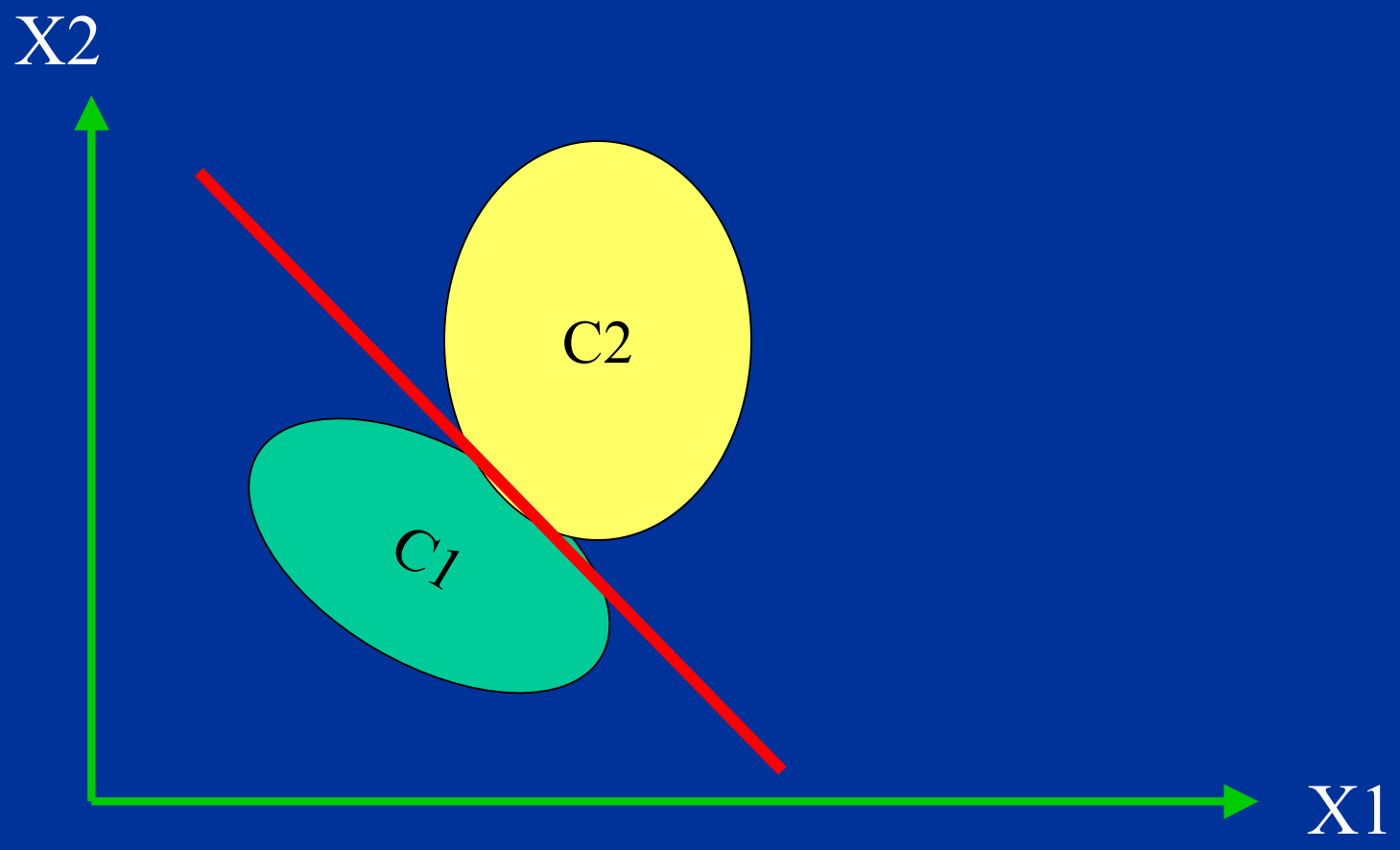
$$g(X) = \Sigma\, a_j\, x_j,$$

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

- Clasificadores multidimensionales:
  - Read, et al., “Classifier chains for multilabel classification”, ECML/PKDD, 2009
  - Bielza, Li, Larrañaga, “Multidimensional classification with Bayesian networks”, IJAR, 2011
  - Zaragoza, Sucar, Morales, Larrañaga, Bielza, “Bayesian chain classifiers for multidimensional classification”, IJCAI, 2011
- Evaluación:
  - C. Drummond, R. C. Holte, “Explicitly representing expected cost: an alternative to the ROC representation”.

Actividades

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