Tema 6: Clasificadores Bayesianos

Pedro Larrañaga, Iñaki Inza, Abdelmalik Moujahid

Departamento de Ciencias de la Computación e Inteligencia Artificial

Universidad del País Vasco

`http://www.sc.ehu.es/isg/`

Modelos básicos

- Naïve Bayes (Minsky, 1961)
- Seminaïve Bayes (Pazzani, 1997)
- Naïve Bayes aumentado a árbol (Friedman y col., 1997)
- Clasificador Bayesiano k-dependiente (Sahami, 1996)
- Red Bayesiana (Jensen, 2001)

Clasificadores Bayesianos

Clasificación Supervisada con Paradigmas Probabilistas

$$\gamma : (x_1, \ldots, x_n) \rightarrow \{1, 2, \ldots, r_0\}$$

- Matriz de costes: $co(r, s)$
- Minimización del coste total de errores

$$\gamma(\boldsymbol{x}) = \arg\min_k \sum_{c=1}^{r_0} co(k, c) p(c|x_1, \ldots, x_n)$$

  - Función de pérdida $0/1$

$$\gamma(\boldsymbol{x}) = \arg\max_c p(c|x_1, \ldots, x_n)$$

Naïve Bayes

Formulación clásica de un problema de diagnóstico

- m diagnósticos posibles no excluyentes

	X_1	$\ldots$	X_n	Y_1	$\ldots$	Y_m
$(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)})$	$x_1^{(1)}$	$\ldots$	$x_n^{(1)}$	$y_1^{(1)}$	$\ldots$	$y_m^{(1)}$
$(\boldsymbol{x}^{(2)}, \boldsymbol{y}^{(2)})$	$x_1^{(2)}$	$\ldots$	$x_n^{(2)}$	$y_1^{(2)}$	$\ldots$	$y_m^{(2)}$
$\ldots$		$\ldots$			$\ldots$	
$(\boldsymbol{x}^{(N)}, \boldsymbol{y}^{(N)})$	$x_1^{(N)}$	$\ldots$	$x_n^{(N)}$	$y_1^{(N)}$	$\ldots$	$y_m^{(N)}$

Naïve Bayes

Formulación clásica de un problema de diagnóstico

$$(y_1^*, \ldots, y_m^*) = arg \max_{(y_1, \ldots, y_m)} p(Y_1 = y_1, \ldots, Y_m = y_m | X_1 = x_1, \ldots, X_n = x_n)$$

$$p(Y_1 = y_1, \ldots, Y_m = y_m | X_1 = x_1, \ldots, X_n = x_n)$$

$$\propto p(Y_1 = y_1, \ldots, Y_m = y_m) p(X_1 = x_1, \ldots, X_n = x_n | Y_1 = y_1, \ldots, Y_m = y_m)$$

número de parámetros a estimar: $2^m - 1 + 2^m(2^n - 1)$

m	n		parámetros
3	10	$\simeq$	$8 \cdot 10^3$
5	20	$\simeq$	$33 \cdot 10^6$
10	50	$\simeq$	$11 \cdot 10^{17}$

Naïve Bayes

Diagnósticos excluyentes

$$c^* = arg \max_c p(C = c | X_1 = x_1, \ldots, X_n = x_n)$$

$$p(C = c | X_1 = x_1, \ldots, X_n = x_n) \;\propto\; p(C = c) p(X_1 = x_1, \ldots, X_n = x_n | C = c)$$

número de parámetros a estimar: $(r_0 - 1) + r_0(2^n - 1)$

r_0	n		parámetros
3	10	$\simeq$	$3 \cdot 10^3$
5	20	$\simeq$	$5 \cdot 10^6$
10	50	$\simeq$	$11 \cdot 10^{15}$

Naïve Bayes

Diagnósticos excluyentes y variables condicionalmente independientes dado el diagnóstico (naïve Bayes)

$$c^* = arg\max_c p(C = c | X_1 = x_1, \ldots, X_n = x_n)$$

$$= arg\max_c p(C = c) \prod_{i=1}^{n} p(X_i = x_i | C = c)$$

número de parámetros a estimar: $(r_0 - 1) + r_0 n$

r_0	n	parámetros
3	10	32
5	20	104
10	50	509

Naïve Bayes

Naïve Bayes (Minsky, 1961)

- Variables predictoras condicionalmente independientes dada C
  - Predictoras discretas

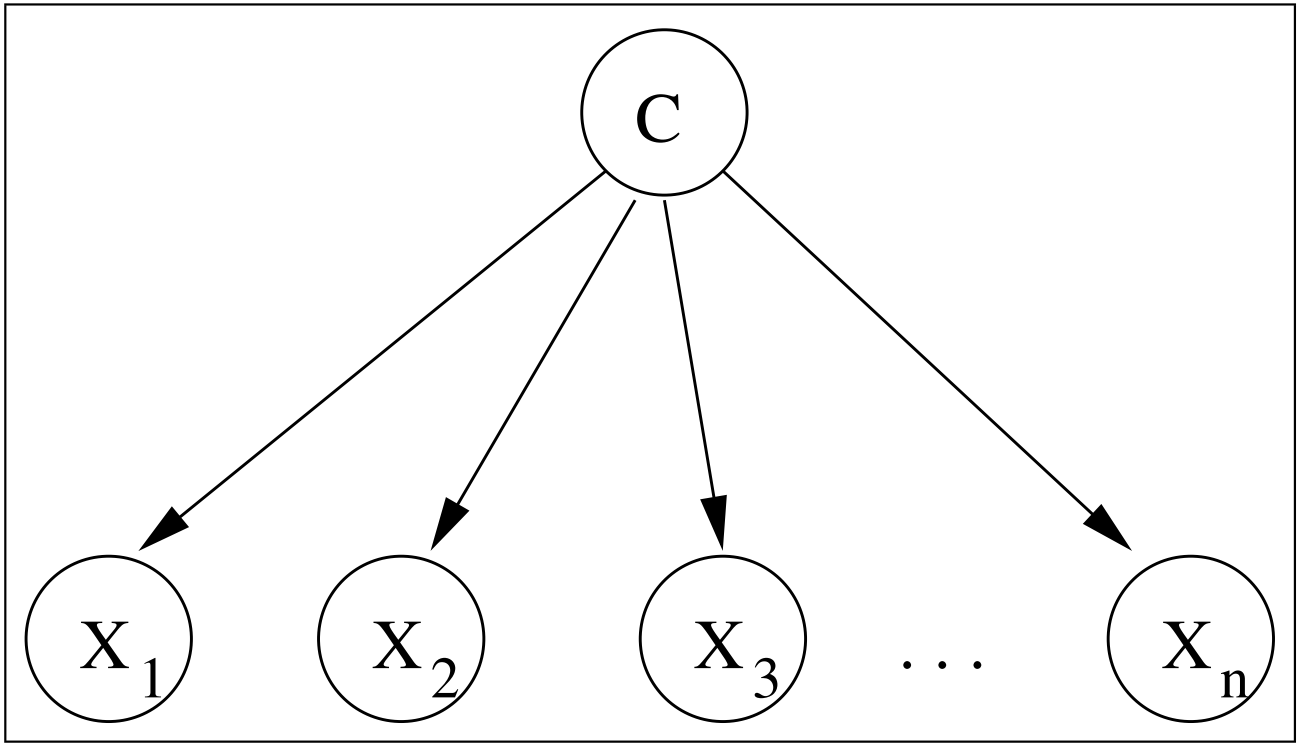
$$c^* = arg\max_c p(C=c)\prod_{i=1}^{n} p(X_i = x_i|C=c)$$

  - Predictoras continuas y normales

$$c^* = arg\max_c p(C=c)\prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi}\sigma_i^c} e^{\frac{-1}{2}\left(\frac{x_i-\mu_i^c}{\sigma_i^c}\right)^2}\right]$$

Naïve Bayes

Naïve Bayes



Estructura gráfica de un modelo naïve Bayes

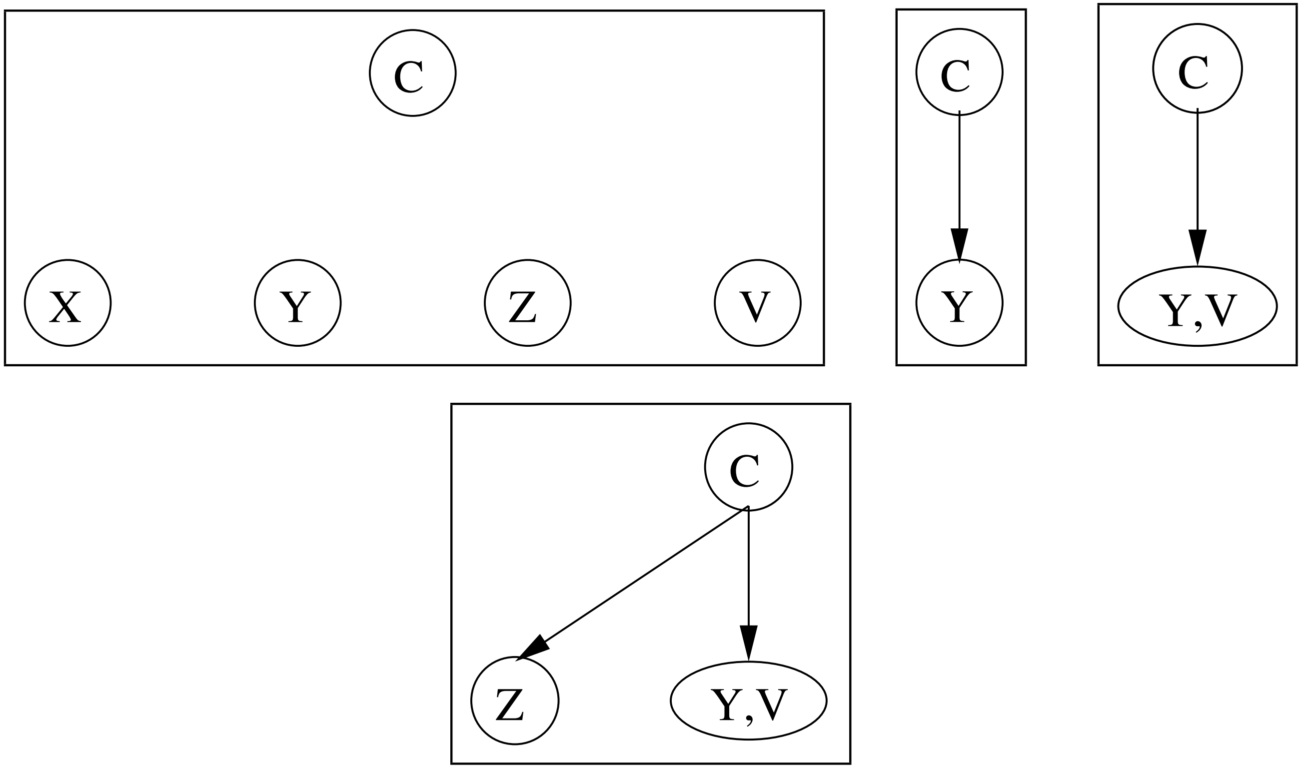
Seminaïve Bayes

```
Paso 1. Inicializar el conjunto de variables a utilizar a vacío.
     Clasificar todos los ejemplos como pertenecientes a la clase
     más frecuente
Paso 2. Repetir en cada paso la mejor opción entre:
     (a) Considerar cada variable que no está en el modelo como una
     variable a incluir en el modelo. Dicha variable debe incluirse
     condicionalmente independiente de las variables presentes en
     el modelo, dada la variable clase
     (b) Juntar cada variable no presente en el modelo con una
     variable que ya forme parte del mismo
     Evaluar cada posible opción por medio de la estimación del
     porcentaje de bien clasificados
     Hasta que ninguna opción produzca mejoras
```

Pseudocódigo del algoritmo *FSSJ* (Pazzani, 1997)

Seminaïve Bayes

Seminaïve Bayes



Proceso de construcción de un modelo seminaïve Bayes.

$$p(c|x, y, z, v) \propto p(c)p(z|c)p((y, v)|c)$$

Naïve Bayes aumentado a árbol

- Cantidad de información mutua entre X e Y

$$I(X,Y) = \sum_{i=1}^{r_X} \sum_{j=1}^{r_Y} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$

  mide la reducción de la incertidumbre de una de las variables cuando se conoce la otra

- Cantidad de información mutua entre X e Y condicionada a C

$$\begin{aligned} I(X,Y|C) &= \sum_c p(c) I(X,Y|C=c) \\ &= \sum_{i=1}^{r_X} \sum_{j=1}^{r_Y} \sum_{k=1}^{r_0} p(x_i, y_j, c_k) \log \frac{p(x_i, y_j|c_k)}{p(x_i|c_k)p(y_j|c_k)} \end{aligned}$$

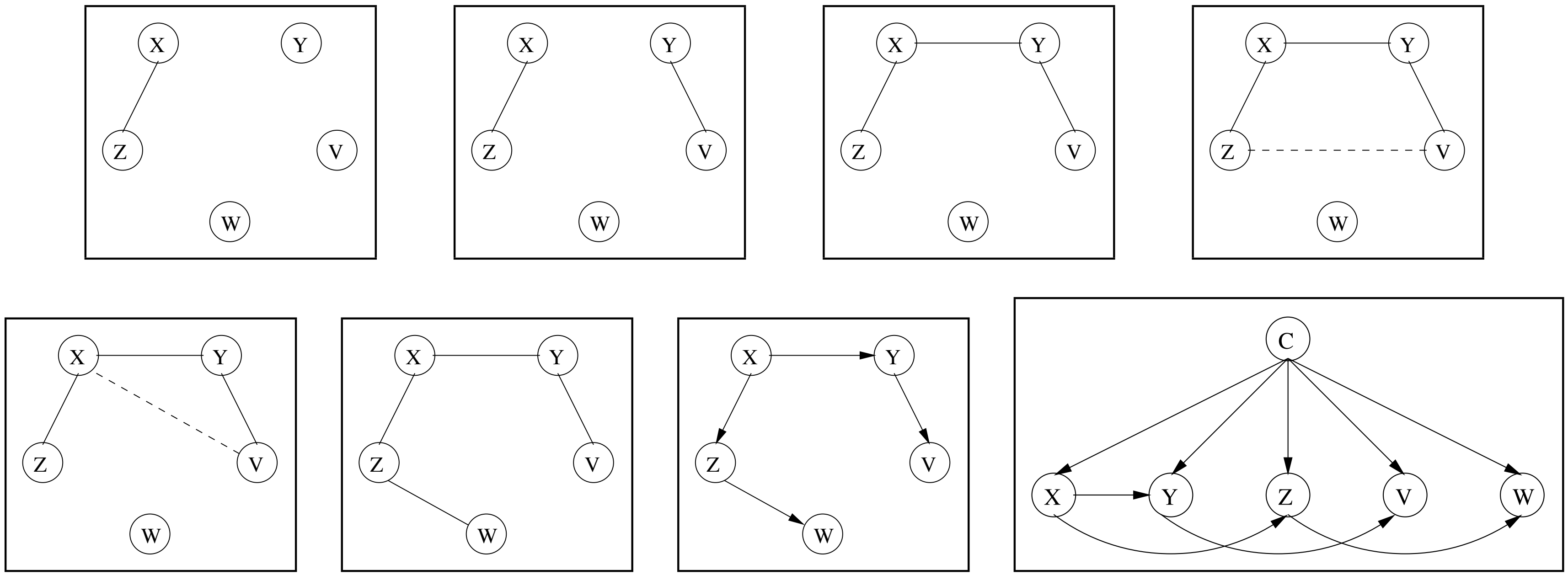
- Relación entre la cantidad de información mutua y la verosimilitud

Naïve Bayes aumentado a árbol

```
Paso 1. Calcular I(Xi, Xj | C) con i < j, i, j = 1, ..., n
Paso 2. Construir un grafo no dirigido completo cuyos nodos
     corresponden a las variables predictoras: X1, ..., Xn. Asignar
     a cada arista conectando las variables Xi y Xj un peso
     dado por I(Xi, Xj | C)
Paso 3. Asignar las dos aristas de mayor peso al árbol a construir
Paso 4. Examinar la siguiente arista de mayor peso, y añadirla al
     árbol a no ser que forme un ciclo, en cuyo caso se descarta
     y se examina la siguiente arista de mayor peso
Paso 5. Repetir el paso 4 hasta seleccionar n − 1 aristas
Paso 6. Transformar el árbol no dirigido resultante en uno dirigido
     escogiendo una variable como raiz, para a continuación
     direccionar el resto de aristas
Paso 7. Construir un modelo TAN añadiendo un nodo etiquetado
     como C y posteriormente un arco desde C a cada variable
     predictora Xi
```

Pseudocódigo del algoritmo *TAN* (Friedman y col., 1997)

Naïve Bayes aumentado a árbol



Proceso de construcción de TAN. $I(X,Z|C) > I(Y,V|C) > I(X,Y|C) > I(Z,V|C) >$

$$I(X,V|C) > I(Z,W|C) > I(X,W|C) > I(Y,Z|C) > I(Y,W|C) > I(V,W|C)$$

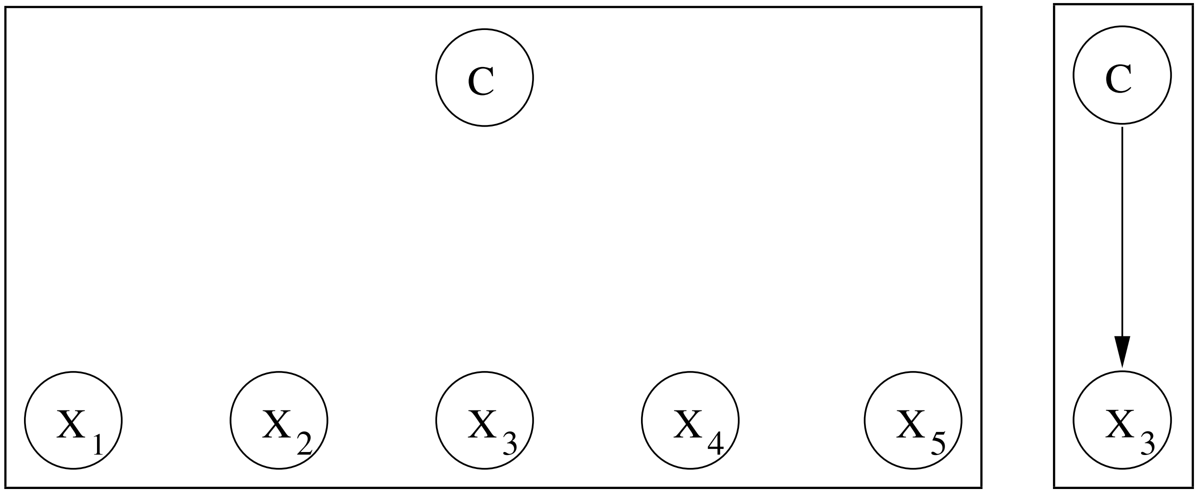
$$p(c|x,y,z,v,w) \propto p(c)p(x|c)p(y|x,c)p(z|x,c)p(v|y,c)p(w|z,c)$$

Clasificador Bayesiano k-dependiente

Clasificador Bayesiano k-dependiente (Sahami, 1996)

- Precalcula $I(X_i, C)$ y $I(X_i, X_j|C)$ para todo par de variables
- Añade en cada iteración, de entre las variables que no están en el modelo, aquella X_{max} que tenga mayor $I(X_i, C)$
- Asigna a la variable añadida como padres la variable C y aquellas k variables con mayor $I(X_j, X_{max}|C)$

Clasificador Bayesiano k-dependiente

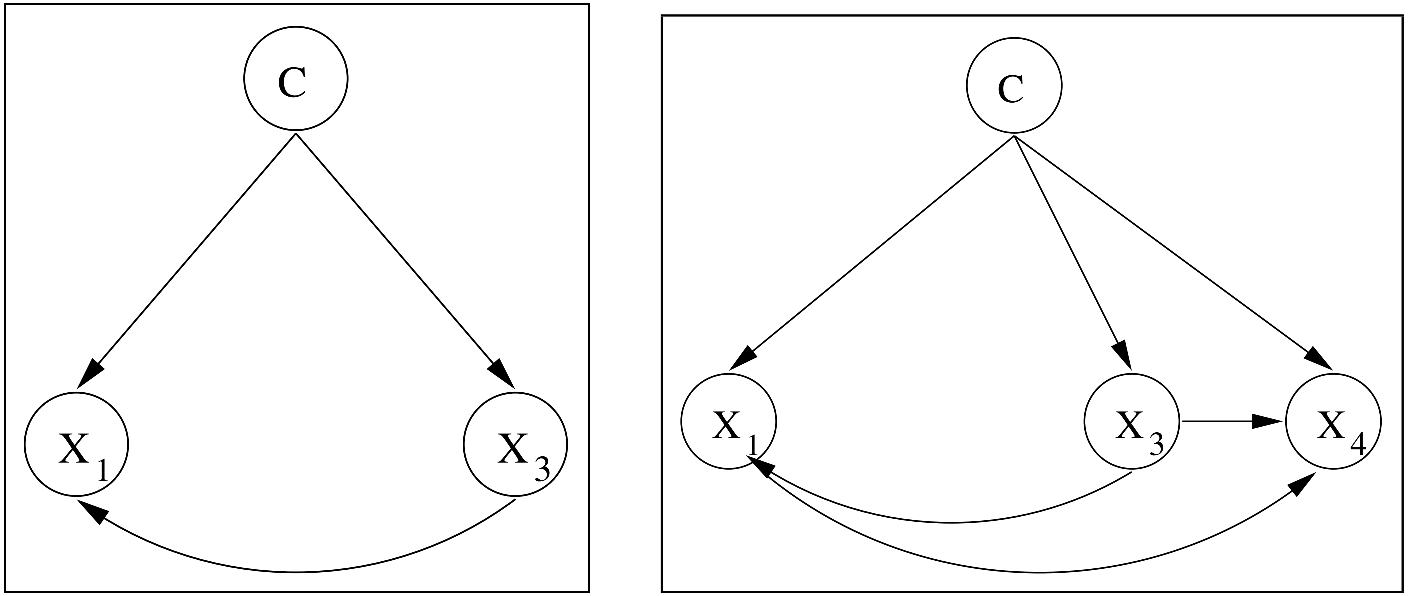


Proceso de construcción de kDB con $k = 2$.

$$I(X_3, C) > I(X_1, C) > I(X_4, C) > I(X_5, C) > I(X_2, C)$$

$$I(X_3, X_4|C) > I(X_2, X_5|C) > I(X_1, X_3|C) > I(X_1, X_2|C) > I(X_2, X_4|C) >$$
$$I(X_2, X_3|C) > I(X_1, X_4|C) > I(X_4, X_5|C) > I(X_1, X_5|C) > I(X_3, X_5|C)$$

Clasificador Bayesiano k-dependiente



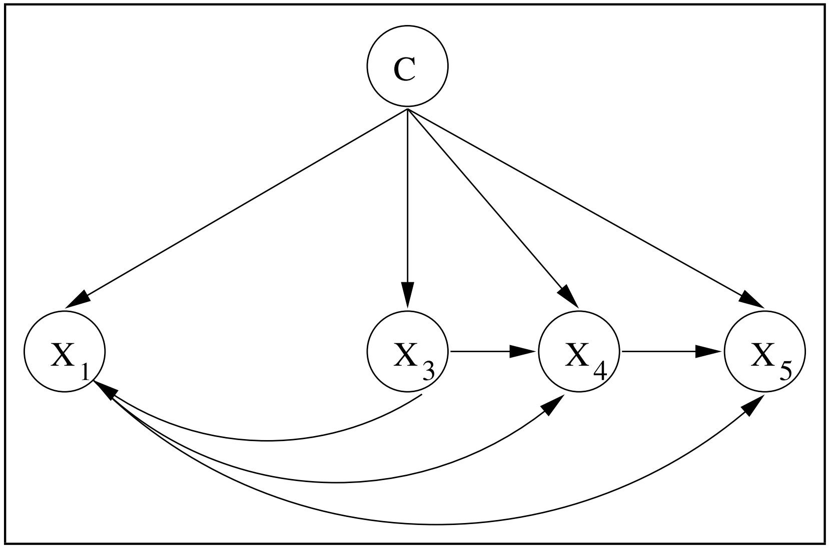
Proceso de construcción de kDB con $k = 2$.

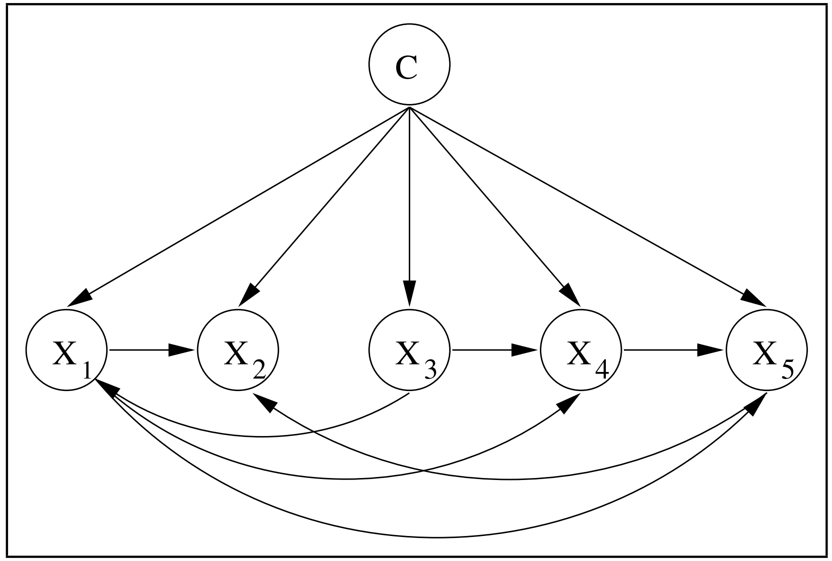
$$I(X_3, C) > I(X_1, C) > I(X_4, C) > I(X_5, C) > I(X_2, C)$$

$$I(X_3, X_4|C) > I(X_2, X_5|C) > I(X_1, X_3|C) > I(X_1, X_2|C) > I(X_2, X_4|C) >$$
$$I(X_2, X_3|C) > I(X_1, X_4|C) > I(X_4, X_5|C) > I(X_1, X_5|C) > I(X_3, X_5|C)$$

$$p(c|x_1, x_2, x_3, x_4, x_5)$$

Clasificador Bayesiano k-dependiente





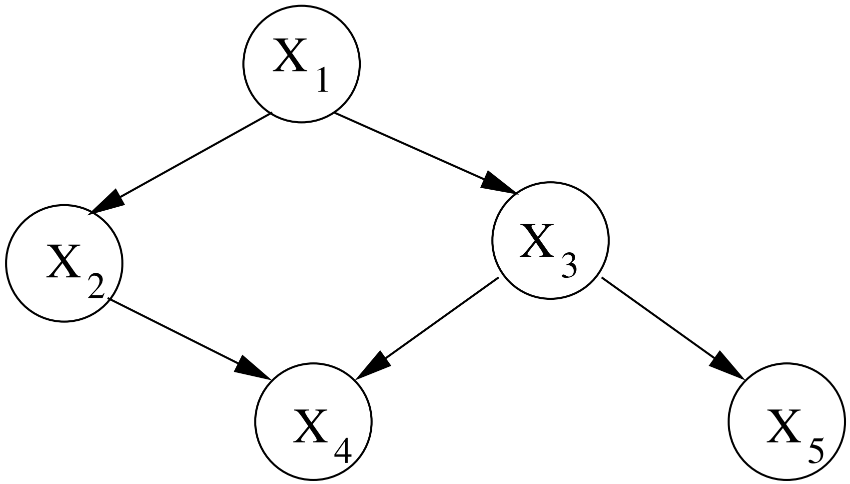
Proceso de construcción de kDB con $k = 2$.

$$I(X_3, C) > I(X_1, C) > I(X_4, C) > I(X_5, C) > I(X_2, C)$$

$$I(X_3, X_4|C) > I(X_2, X_5|C) > I(X_1, X_3|C) > I(X_1, X_2|C) > I(X_2, X_4|C) >$$
$$I(X_2, X_3|C) > I(X_1, X_4|C) > I(X_4, X_5|C) > I(X_1, X_5|C) > I(X_3, X_5|C)$$

$$p(c|x_1, x_2, x_3, x_4, x_5) \propto$$
$$p(c)p(x_1|x_3, c)p(x_2|x_1, x_5, c)p(x_3|c)p(x_4|x_1, x_3, c)p(x_5|x_1, x_4, c)$$

Red Bayesiana múltiplemente conectada



$$
\begin{array}{ll}
p(X_1 = 0) = 0{,}20 & p(X_4 = 0|X_2 = 0, X_3 = 0) = 0{,}80 \\
p(X_2 = 0|X_1 = 0) = 0{,}80 & p(X_4 = 0|X_2 = 1, X_3 = 0) = 0{,}80 \\
p(X_2 = 0|X_1 = 1) = 0{,}80 & p(X_4 = 0|X_2 = 0, X_3 = 1) = 0{,}80 \\
p(X_3 = 0|X_1 = 0) = 0{,}20 & p(X_4 = 0|X_2 = 1, X_3 = 1) = 0{,}05 \\
p(X_3 = 0|X_1 = 1) = 0{,}05 & p(X_5 = 0|X_3 = 0) = 0{,}80 \\
 & p(X_5 = 0|X_3 = 1) = 0{,}60
\end{array}
$$

Factorización de la distribución de probabilidad conjunta obtenida con la red Bayesiana adjunta