Guía de Ejercicios: Redes Bayesianas

Profesor: Carlos Hurtado

25 de agosto de 2008

1. Clasificador Bayesiano Naive

Problema 1 *Considere el siguiente registro de edad para un grupo de personas clasificadas en dos clases: A, B:*

Edad	*Clase*
20	*A*
15	*A*
35	*B*
40	*B*
25	*A*
24	*B*
50	*B*
10	*A*
30	*B*
18	*A*

(i) Induzca un clasificador Bayesiano a partir de estos datos, suponiendo que las edades distribuyen normal para cada clase. El modelo no sólo debe clasificar, sino estimar un puntaje discriminante. Es decir, dada una persona para la cual sólo conocemos la edad, el clasificador debe estimar un puntaje discriminante, como probabilidad, para cada clase.
(ii) Grafique el puntaje discriminante para cada clase en función de la edad.

Problema 2 *Al clasificar un objeto $\vec{x} = (x_1, \ldots, x_n)$ en alguna clase c en un clasificador Bayesiano Naive se asume que $P(\vec{x} \mid c) = P(x_1 \mid c) \times \ldots \times P(x_n \mid c)$. Dé un ejemplo (distinto al dado en los apuntes del curso) de un clasificador Bayesiano Naive, lo más realista posible, en que la suposición de independencia condicional no funciona. Defina en su ejemplo las variables de entrada, la variable de clases y las probabilidades necesarias.*

Problema 3 *Muestre en su ejemplo del problema 2, cómo la mala suposición de independencia condicional lleva a una clasificación errónea. Es decir, muestre cómo los puntajes discriminantes hacen que la clase seleccionada sea distinta a la clase que se seleccionaría sin la mala suposición de independencia.*

2. Modelación de Redes Bayesianas

Problema 4 *Consider la red Bayesiana de detección de fraude dada en los apuntes del curso, con las variables: $F, E, S, G, , J$. Encuentre un ordenamiento de estas variables que genere una red con un número de parámetros significativamente mayor que la red dada en los apuntes. Justifique (usando la noción de caminos d-conectados) por qué los arcos de su red no se pueden simplificar.*

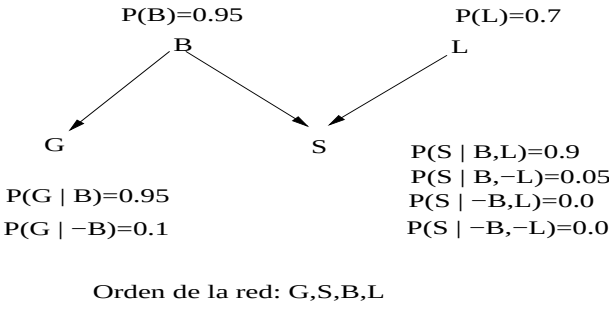

Figura 1: Red Bayesiana

G	S	B	L	Num. Datos
v	v	v	v	5
v	f	v	f	3
f	f	f	v	2

Cuadro 1: Datos

Problema 5 *Considere la red Bayesiana R de la figura 2. Sea D un conjunto de datos que distribuyen de acuerdo a la probabilidad conjunta codificada por R. Supongamos que eliminamos el arco (L, S) de R, obteniendo una red R'. Estime las TPCs de R' a partir de D usando máxima verosimilitud (es decir, cada parámetro de una distribución discreta se estima como una frecuencia).*

Problema 6 *Considere la estructura de la red Bayesiana "Family Out" dada en la figura 2. Construya una estructura de red equivalente dado el ordenamiento de variables: do, lo, hb, bp, fo.*

Hint: *Genere una red "completa" y luego razone sobre caminos d-conectados para simplificar arcos.*

Problema 7 *Investigue cómo se denominan los distintos tipos de inferencia probabilística que se realizan usando redes Bayesianas. Dé ejemplos de cada uno.*

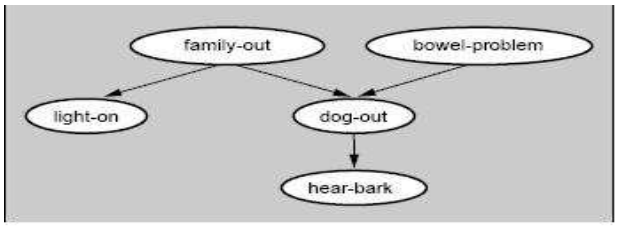

Figura 2: Red Bayesiana "Family Out"

3. Aprendizaje de Redes Bayesianas

Problema 8 *Considere sólo la estructura de la Red Bayesiana de la Figura 2 (olvídese de sus TPCs) y los datos de la tabla 1. Diga si conviene eliminar el arco* (L, S) *de la red basado en el puntaje de descripción mínima.*

Hint: *Recuerde que el puntaje de descripción mínima de una estructura* G *con datos* D *es:*

$$PuntajeMDL(G, D) = n \sum_{x_i \in dom(X)} \Pr_D(x_i) \log \Pr_{G,D}(x_i) - \frac{dim(G) log n}{2},$$

donde n *es el número de datos,* X *es el vector de variables de la red,* $dim(G)$ *es el número de parámetros de la red,* $\Pr_D$ *es la probabilidad conjunta observada en los datos y* $\Pr_{G,D}$ *es la probabilidad conjunta codificada por "para la cual"la red (estimada a base de* D*).*

4. Problemas Resultos

Resolución Problema 5. Si D distribuye de acuerdo a la red R, entonces debemos estimar las TPCs de R usando la distrubución codificada por R. Luego, para los nodos B, L, G las TPC's de R y R' son las mismas.

Para el nodo S, sea debemos calcular la TPCs de R', que denotamos $P'(S|B)$ y $P'(S| - B)$, en base a la probabilidad codificada por R, sea esta última denotada por P.

$P'(S|B) = P(S|B) = \frac{P(S,B)}{P(B)}$

$= \frac{P(S,B,L)}{P(B)} + \frac{P(S,B,-L)}{P(B)}$

$= P(S|B,L)P(L|B) + P(S|B,-L)P(-L|B)$

Pero del gráfo de R se puede ver que B y L son independientes (no existe un camino d-conectado de B a L dada evidencia vacía). Luego:

$P'(S|B) = P(S|B,L)P(L) + P(S|B,-L)P(-L)$

$P'(S|B) = 0{,}9 \times 0{,}7 + 0{,}05 \times 0{,}3 = 0{,}63 + 0{,}015 = 0{,}645$

En forma análoga obtenemos:

$P'(S|-B) = P(S|-B,L)P(L) + P(S|-B,-L)P(-L)$

$P'(S|-B) = 0 + 0 = 0.$

Resolución Problema 8. Sea G_1 la red inicial.

Primero estimamos las TPCs de G_1 en base a los datos de la tabla 1. Estimamos en base a frecuencia y obtenemos:

$\Pr(B) = 0{,}8$

$\Pr(L) = 0{,}7$

$\Pr(G|B) = 1$, $\Pr(G|-B) = 0.$

$\Pr(S|B,L) = 1$, $\Pr(S|B,-L) = 0$, $\Pr(S|-B,L) = 0$, $\Pr(S|-B,-L) =$ no definida

En base a estas probabilidades, obtenemos la probabilidad conjunta estimada para G_1, para los valores de la tabla 1:

G	S	B	L	$\Pr_D$	$\Pr_{G_1,D}$
v	v	v	v	0.5	$0{,}8 \times 0{,}7 \times 1 \times 1 = 0{,}56$
v	f	v	f	0.3	$0{,}8 \times 0{,}3 \times 1 \times 1 = 0{,}24$
f	f	f	v	0.2	$0{,}2 \times 0{,}7 \times 1 \times 1 = 0{,}14$

Luego el puntaje MDL para G_1 con los datos D es:

$PuntajeMDL(G_1, D) = 10 \times (0{,}5 \log 0{,}56 + 0{,}3 \log 0{,}24 + 0{,}2 \log 0{,}7) - \frac{16 \log 10}{2}$

$PuntajeMDL(G_1, D) = -16{,}03 - 26{,}58 = -42{,}61$

Ahora, repetimo el procedimiento con el grafo G_2, donde se ha eliminado el arco (L, S).

Las TPCs son:

$\Pr(B) = 0{,}8$

$\Pr(L) = 0{,}7$

$\Pr(G|B) = 1$, $\Pr(G|-B) = 0.$

$\Pr(S|B) = 5/8 = 0{,}63$, $\Pr(S|-B) = 0$

En base a estas probabilidades, obtenemos la probabilidad conjunta estimada para G_2, para los valores de la tabla 1:

G	S	B	L	$\Pr_D$	$\Pr_{G_2,D}$
v	v	v	v	0.5	$0{,}8 \times 0{,}7 \times 1 \times 0{,}63 = 0{,}35$
v	f	v	f	0.3	$0{,}8 \times 0{,}3 \times 1 \times 0{,}37 = 0{,}09$
f	f	f	v	0.2	$0{,}2 \times 0{,}7 \times 1 \times 1 = 0{,}14$

$PuntajeMDL(G_2, D) = -23{,}67 - \frac{12 \log 10}{2}$

$PuntajeMDL(G_2, D) = -23{,}67 - 19{,}93 = -43{,}6.$

Luego no conviene eliminar el arco (L, S).