Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ciencias de la Computación

Semestre Otoño 2009
Inteligencia Artificial
CC52A

Profesor: Gonzalo Ríos
Auxiliar: Miguel Romero
Fecha: 30 de Abril

Control 1

1. **(1.5 pto)** El Banco Nacional de Pelotillehue ha tenido importantes pérdidas por haber ofrecido créditos a personas que dejaron de pagar antes de haber cumplido la deuda. El gerente de finanzas, Don Chuma, le ha entregado una gran cantidad de datos, con el fin de disminuir el riesgo del Banco al otorgar los créditos. El esquema de los datos se muestran a continuación:

Nombre Dato	**Valores**	**Descripción**
CLIENTE	INT	Un id único de cada cliente
NOMBRE	STRING	Nombre del cliente
EDAD	INT	Edad del cliente
SEXO	H,M	Sexo del cliente
ESTADO CIVIL	C,So,Se,Vi	Casado, Soltero, Separado, Viudo
CASA PROPIA	BOOL	True si el cliente tiene casa propia
INGRESO	INT	Ingreso mensual del cliente
INDEPENDENCIA	BOOL	True si el cliente es trabajador independiente
CREDITO	INT	Monto del crédito solicitado
CUOTAS	INT	Número de cuotas del crédito
INTERES	FLOAT	Interés anual del crédito
TIPO CREDITO	Hip,Con	Hipotecarios, Consumo
PAGOS	INT	Numero de meses que el cliente canceló

   Modele y describa cada una de las siguientes solicitudes que le ha hecho Don Chuma:

   (a) Conocer los perfiles principales de los clientes que solicitan créditos Hipotecarios, lo hayan pagado o no.

   (b) Predecir el Ingreso Mensual del cliente, a partir de su Nombre, Edad, Sexo, Estado Civil, Casa Propia e Independencia.

   (c) Determinar si otorgar el crédito a un cliente en particular. Para esto, debe observar que el banco desea aumentar su Beneficio, y al otorgar un crédito a una persona que no lo pagó, el banco pierde la diferencia, y no otorgar un crédito a una persona que si lo hubiera pagado, se pierde el interés del crédito.

2. **(1.5 pto)** Responda verdadero o falso, argumentando su respuesta de forma clara, definiendo las conceptos y dando ejemplos en cada caso. Las respuestas no fundamentadas no serán consideradas.

   (a) Las redes bayesianas son modelos sólo de clasificación.

   (b) Las suposiciones de independencia erróneas producen redes bayesianas más complejas para entrenar e inferir.

   (c) La regla de la cadena indica que el uso de independencia condicionales entre variables produce redes bayesinas más simples.

   (d) Al instanciar una variable de la red bayesina, la información debe propagarse por la red, siempre en el sentido de los arcos.

   (e) El puntaje de verosimilitud y el costo de descripción mínima no consideran la complejidad de la red, sino sólo la codicación de los datos en la red.

3. **(2 pto)** Considere la siguiente Red Bayesiana

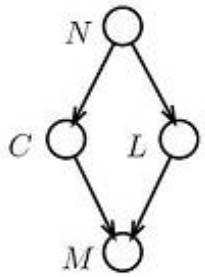

Figura 1

(a) Usando la Tabla 1 de probabilidades condicionales, calcule la probabilidad de que ocurra N dado que ocurrió M, y la probabilidad de que ocurra L dado que ocurrió C. Fundamente su desarrollo.

$P(n) = 0.5$	
$P(c \mid n) = 0.1$	$P(c \mid \neg n) = 0.5$
$P(l \mid n) = 0.8$	$P(l \mid \neg n) = 0.2$
$P(m \mid c, l) = 0.99$	$P(m \mid c, \neg l) = 0.9$
$P(m \mid \neg c, l) = 0.9$	$P(m \mid \neg c, \neg l) = 0$

Tabla 1

(b) Usando los datos de la Tabla 2, diga si conviene agregar un arco de C hacia L usando el puntaje de verosimilitud. Haga el mismo análisis considerando el costo de descripción mínima (para estimar las TPC use maxima verosimilitud).

N	C	L	M	Frecuencias
0	1	1	1	3
1	1	0	0	2
1	0	1	1	4
0	0	0	1	5

Tabla 2

4. **(1.5 pto)** Dado con conjunto de datos D=$\{(\vec{X}_i, \mathrm{C}_i)_{i=1}^n\}$, con $C_i \in \{0, 1\}$ se construyeron los siguientes modelos de clasificacción

- Una red neuronal con nucleos gaussianos
- Una red bayesiana naive, usando el estimador de laplace
- K-means, con K=5 y usando vecindad envolvente

A continuación, se muestra un resumen del resultado obtenido

N Datos	X_1	X_2	$\mathrm{C}_{\mathrm{real}}$	C_{neuro}	C_{naive}	C_{kmeans}
10	5	5	1	0	0	1
10	4	1	0	1	1	0
15	1	2	1	1	0	0
15	2	5	0	1	0	1

(a) Calcule el % de error de cada uno de los modelos

(b) Para cada dato $(\vec{X}_i, \mathrm{C}_i)$ escriba $(\vec{X}_i, M_i)$, donde $M_i \in \{C_{neuro}, C_{naive}, C_{kmeans}\}$ indica el modelo que hace la predicción correcta. Usando los datos $\{(\vec{X}_i, \mathrm{M}_i)_{i=1}^n\}$, construya un árbol de decisión, usando el índice *Gini* (En los splits, considere el umbral $u = 3$)

(c) Construya un modelo híbrido H tal que, a cada dado, se evalúa inicialmente en el árbol de decisión construido anteriormente, y al llegar a una hoja del árbol , se le aplica el modelo M_i a dicho dato, con $M_i \in \{C_{neuro}, C_{naive}, C_{kmeans}\}$, es decir, el modelo asociado a la clase asignada por el árbol. Calcule el % de error del modelo H. ¿Qué puede concluir de esto?

Tiempo: 3 horas