ANÁLISIS DE DATOS

Jesús García Herrero

ANALISIS DE DATOS

EJERCICIOS

Una empresa de seguros de automóviles quiere utilizar los datos sobre sus clientes para obtener reglas útiles que permita clasificar nuevos clientes en dos categorías, alto riesgo (**A**) y bajo riesgo (**B**), disponiéndose únicamente de dos atributos: la edad y el sexo. A partir de datos de siniestralidad referidos a un periodo de tiempo determinado, se dispone de la siguiente información acerca de los atributos:

- Se ha estimado la distribución de probabilidad del atributo edad como una variable normal con los siguientes parámetros para cada clase:
  - clientes tipo A: edad media: 21 años, desviación estándar: 2 años
  - clientes tipo B: edad media: 45 años, desviación estándar: 6 años

- En cuanto al atributo sexo, se dispone de la siguiente tabla resumen con todos los clientes existentes

Clase / **Sexo**	**A (alto riesgo)**	**B (bajo riesgo)**
Hombre	900	600
Mujer	300	900

Se pide:

1. Utilizando un clasificador de tipo bayesiano, con independencia entre atributos, determinar las reglas que permitan clasificar a un cliente nuevo en función de su sexo y edad.
2. Indicar las reglas para los casos de disponerse sólo de la edad y sólo del sexo del cliente
3. Determinar las nuevas reglas que se aplicarían en el primer caso (ambos atributos disponibles) si se considerara que es el doble de costoso clasificar incorrectamente a un cliente de tipo A como de tipo B que la situación de clasificar a uno de tipo B como de tipo A.

Ejercicio 2

Una empresa de venta telefónica dispone de una muestra de datos acerca de 1000 familias, y de una herramienta de análisis de datos que predice si una familia responderá o no de forma positiva a la oferta de un producto determinado. Se dispone de un sencillo modelo bastante fiable del comportamiento de las familias, que estima que únicamente el 10% de las familias estarán interesadas en el producto, y que además ha permitido, mediante simulación, evaluar la capacidad de predicción de la herramienta. La evaluación proporciona la matriz de confusión, definida de la forma siguiente:

Clase Predicha / **Clase Real**	**Sí**	**No**
Sí	TP	FN
No	FP	TN

Los valores de la matriz de confusión contienen el número de casos que aparecen en las posibles situaciones de: acierto en predicción positiva (TP), fallo en predicción positiva (FP), fallo en predicción negativa (FN), acierto en predicción negativa (TN).

El algoritmo de clasificación dispone además de un parámetro de ajuste, p, que permite variar los costes relativos de los dos tipos de error, FN y FP. Esto ha permitido llevar a cabo 5 simulaciones sobre los datos de las 1000 familias con diferentes valores de este parámetro, llegando a los resultados que se indican a continuación:

CP CR	Sí	No
Sí	40	60
No	160	740

p=1

CP CR	Sí	No
Sí	80	20
No	320	580

p=2

CP CR	Sí	No
Sí	90	10
No	510	390

p=3

CP CR	Sí	No
Sí	95	5
No	705	195

p=4

CP CR	Sí	No
Sí	100	0
No	900	0

p=5

Se pide:

1. Representar la gráfica con el "factor de elevado" (*lift chart*) en la que figure el número de familias que compran el producto frente al porcentaje de familias que es seleccionado para llamarlas realizar la oferta. Comentar como varían los dos tipos de error de clasificación a lo largo de la coordenada horizontal y superponer a esta gráfica los valores esperados en el caso de que la selección fuese aleatoria en lugar de utilizar la herramienta de análisis de datos.
2. Representar la ganancia de clasificación, definida como el cociente entre los valores de familias que compran el producto al utilizar la herramienta y al hacer selección aleatoria, para cada uno de los porcentajes de familias a los que se les hace la oferta. Determinar en qué situación se optimiza esta ganancia.
3. Representar en una gráfica, para cada valor de porcentaje de familias enviadas, la tasa de error global de clasificación de la herramienta, y el extremo superior del intervalo de confianza al 95%. Para ello puede suponerse que el error cometido al estimar una probabilidad con un valor P con N datos puede aproximarse como una variable estadística normal de media nula y varianza P(1-P)/N, y sabiendo que en una variable normal estandarizada $z_{2.5\%}=1.96$
4. Suponiendo que el coste de llamada es de 0.2 euros y que cada producto vendido supone un ingreso de 5 euros, representar el beneficio en función del porcentaje de familias llamadas y determinar a cuántas habría que llamar para maximizar el beneficio.

Ejercicio 3

Dada la hipotética muestra de datos de entrenamiento siguiente:

Atributos:

PATAS: {NINGUNA, DOS, MAYOR_QUE_DOS}
ACUÁTICO: {SI, NO}
PLUMAS: {SI, NO}

CLASE: {Ave, Pulpo, Humano, Pez}

Datos:

PATAS	ACUATICO	PLUMAS	CLASE
DOS	NO	SÍ	Ave
MAYOR_QUE_DOS	SI	NO	Pulpo
DOS	NO	NO	Humano
NINGUNA	SÍ	NO	Pez

1. Realizar un seguimiento detallado del algoritmo ID3 para construir el árbol de decisión para clasificar el animal según los tres atributos considerados.
2. Realizar un seguimiento detallado de un algoritmo de aprendizaje de reglas en escalada (de tipo "PRISM"), con criterios de preferencia de precisión y cobertura.
3. ¿Cómo puede mejorarse ID3 para tratar con atributos con múltiples valores, valores continuos y sobreajuste?
4. ¿Cómo se puede evitar el sobreajuste en un sistema de aprendizaje de reglas?

Ejercicio 4

Se tiene una situación con datos numéricos de dos atributos continuos, **x** e **y**, y con dos clases posibles, **A** y **B**. Se sabe que los datos proceden de distribuciones estadísticas con las siguientes características y parámetros:

- Probabilidades de clase: p(A)=0.6, p(B)=0.4
- Clase A: normal bivariada de parámetros $\mu_A = [0 \quad 0]^t$, $S_A = \begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix}$

$$f((x,y)\,|\,A) = \frac{1}{2\pi\sqrt{|S_A|}} \exp\left\{-\frac{1}{2}[x \quad y] S_A^{-1} \begin{bmatrix} x \\ y \end{bmatrix}\right\}$$

- Clase B: distribución uniforme sobre el atributo **x** en el intervalo [-1,1] e independiente del atributo **y**

$$f((x,y)\,|\,B) = \begin{cases} 1/2, & x \in [-1,1] \\ 0, & \text{en el resto} \end{cases}$$

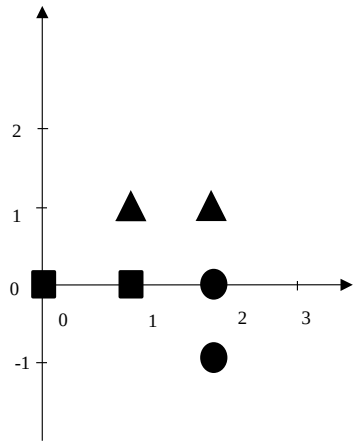
1. Diseñar el clasificador óptimo con un criterio bayesiano que a partir de un dato (x1,y1) determine la clase con mínimo error medio.
2. Modificar el clasificador anterior para que el criterio sea minimizar el coste medio si se penaliza el doble el error de clasificar un dato de la clase B como de la clase A que clasificar un dato de la clase A como de la clase B.
3. Si en lugar de conocer los parámetros de las distribuciones se dispusiera de una muestra representativa y un algoritmo de entrenamiento de tipo "Bayes ingenuo" con modelo de distribuciones normal, ¿en qué cambiarían los clasificadores anteriores?

4. ¿Cuál sería el criterio de clasificación con datos incompletos (sólo se dispone de **x**, de **y** o de ninguno)?

Ejercicio 5

Dada la siguiente muestra de entrenamiento de datos con dos atributos numéricos (x e y), perteneciente a tres clases posibles (representadas con cuadrados, círculos y triángulos), dibujar y razonar las fronteras de decisión si se utilizan los siguientes clasificadores

- clasificador de vecino más próximo (IB1)
- clasificación mediante regresión lineal (multiclase e interclase)
- C4.5 (sin poda del árbol)
- clasificador bayesiano ("Bayes Ingenuo")



Ejercicio 6

Una empresa de seguros de automóviles quiere utilizar los datos sobre sus clientes para obtener reglas útiles que permita clasificar nuevos clientes en dos categorías, alto riesgo (**A**) y bajo riesgo (**B**), disponiéndose únicamente de dos atributos: la edad y el sexo. A partir de datos de siniestralidad referidos a un periodo de tiempo determinado, se dispone de la siguiente información acerca de los atributos:

- Se ha estimado la distribución de probabilidad del atributo edad como una variable normal con los siguientes parámetros para cada clase:
  - clientes tipo A: edad media: 21 años, desviación estándar: 2 años
  - clientes tipo B: edad media: 45 años, desviación estándar: 6 años

- En cuanto al atributo sexo, se dispone de la siguiente tabla resumen con todos los clientes existentes

Clase / **Sexo**	**A (alto riesgo)**	**B (bajo riesgo)**
Hombre	900	600
Mujer	300	900

Se pide:

4. Utilizando un clasificador de tipo bayesiano, con independencia entre atributos, determinar las reglas que permitan clasificar a un cliente nuevo en función de su sexo y edad.
5. Indicar las reglas para los casos de disponerse sólo de la edad y sólo del sexo del cliente
6. Determinar las nuevas reglas que se aplicarían en el primer caso (ambos atributos disponibles) si se considerara que es el doble de costoso clasificar incorrectamente a un cliente de tipo A como de tipo B que la situación de clasificar a uno de tipo B como de tipo A.