Métodos Matemáticos en Ciencias de la Computación

03 Septiembre, 2007. Parte teórica. 5 puntos.

1. El método de Máxima Verosimilitud es el método más común para estimar los parámetros de un modelo paramétrico. Sea $X_1, ..., X_n$ n variables independientes con una función de distribución de probabilidad dada por $f(x;\theta)$ donde θ es el vector de los parámetros. La función verosímil es definida por: $L_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$. La función log-verosímil es definida por $l_n(\theta) = logL_n(\theta)$.

   Suponemos que las variables $X_1, ..., X_n$ siguen una distribución de Bernoulli de parámetro p. La función de probabilidad es $f(x;\theta) = p^x(1-p)^{1-x}$. Demostrar que el estimador de máxima verosimilitud para esta función, denotado por $\hat{p}$, viene dado por: $\hat{p} = \frac{S}{n}$ donde $S = \Sigma_{i=1}^{n} X_i$.

2. Escribir el pseudocódigo del algoritmo de inducción de árboles de clasificación denominado TDIDT *(Top Down Induction of Decision Trees)* explicando los criterios de selección de la variable informativa utilizados en los algoritmos *ID3* y *C4.5*.

3. Escribir el pseudocódigo del algoritmo IREP (Incremental Reduced Error Pruning) explicando las ideas fundamentales en la que se basa dicho algoritmo considerando el caso en el que la variable clase C toma dos posibles valores.

4. La tabla adjunta contiene 5 casos que constituyen el conjunto de datos a partir del cual se pretende aplicar el Clustering ascendente jerarquico.

	X_1	X_2	X_3
O_1	4	1	2
O_2	2	5	4
O_3	3	4	7
O_4	2	6	4
O_5	2	7	3

   La medida utilizada para definir la similitud entre dos objetos es la distancia euclidea: $d(O_i, O_j) = \Sigma_{k=1}^{3}(o_i^k - o_j^k)^2$ mientras que entre dos conglomerados utilizamos el enlace medio entre conglomerados.

5. Supongamos que se conoce que la probabilidad a priori de recidiva para un enfermo que sufre por vez primera un ataque de corazón es de 0,30. Supongamos que en el modelo que tenemos construido para tratar de predecir la recidiva, tenemos dos variables predictoras que se suponen independientes dado el valor de la clase. Las dos variables predictivas son la realización de ejercicio moderado de manera asidua y la edad del paciente. Mientras que la primera es discreta con dos posibles valores (si,

no), la segunda es continua y está para cada valor de la clase distribuida normalmente.

Entre los que no sufren de recidiva, el 80 por ciento efectúa ejercicio moderado de manera asidua, mientras que entre los que tienen recidiva dicha probabilidad se reduce al 10 por ciento. La edad de los pacientes que no tienen recidiva sigue un modelo normal con esperanza matemática de 60 años y desviación típica de 10 años, mientras que en el caso de los pacientes con recidiva los parámetros del modelo normal son 75 años de esperanza matemática y 15 de desviación típica.

Efectuar el diagnostico para un paciente de 65 años que efectúa ejercicio moderado de manera asidua.