# Métodos Probabilísticos en Inteligencia Artificial

7 Junio 2000

1. Diseñar un algoritmo genético (función de coste, representación de individuos, operadores de cruce y mutación) para el siguiente problema:

   Dados los $n^2$ primeros números enteros positivos, se trata de colocarlos en una matriz cuadrada $A = (a_{ij})$ con $i, j = 1, ..., n$ tratando de minimizar la expresión:

   $$| \prod_{i=1}^{n} a_{ii} - \sum_{i>j} a_{ij} | + | \prod_{i=1}^{n} a_{ii} - \sum_{i<j} a_{ij} | .$$

2. Enunciar los algoritmos de Forgy y McQueen para clasificación no supervisada particional, explicando las similitudess y diferencias entre los mismos. Apoyándote en un ejemplo de tu invención muestra que los resultados que se obtienen con los algoritmos anteriores no tienen porque coincidir.

3. Para determinar si un individuo está enfermo o sano en relación con una determinada enfermedad, a dicho individuo se le efectúan dos pruebas.

   Los resultados de la primera prueba se expresan en términos de positivo o negativo. Se sabe que si el individuo está enfermo, la probabilidad de que el resultado de la primera prueba sea positivo es de 0.95, mientras que si el individuo está sano, un resultado positivo en la primera prueba tiene asociada una probabilidad de 0.02.

   La segunda prueba tiene como posibles resultados: alto, normal y bajo. Se tiene que si el individuo está enfermo la probabilidad de que el resultado de esta segunda prueba sea alto es de 0.85, mientras que dicha probabilidad se reduce hasta 0.10 para resultados normales. Si el individuo está sano, la segunda prueba tiene un resultado alto con probabilidad 0.20 y normal con probabilidad 0.70.

   Los resultados proporcionados por las dos pruebas anteriores son, una vez conocido el status del individuo (enfermo o sano), independientes. Además la probabilidad de padecer la enfermedad anterior en la población estudiada es de 0.01.

   Supongamos que a un paciente susceptible de padecer la anterior enfermedad se le efectúan las dos pruebas citadas anteriormente, obteniéndose que en la primera de ellas el resultado es positivo, mientras que en la segunda el resultado es alto.

   Se solicita llevar a cabo un diagnóstico de dicho individuo, justificando el paradigma utilizado.

4. La tabla adjunta contiene 20 casos que constituyen el conjunto de entrenamiento para un clasificador basado en el algoritmo *id3*. Las variables $X_i$ con $i = 1, ..., 4$ corresponden a las 4 variables predictoras y la variable $C$ a la variable que tratamos de predecir.

   El problema de clasificación supervisada se relaciona con la predicción de si un alumno aprobará la asignatura $C$ = *Métodos Probabilísticos en Inteligencia Artificial* en su primera convocatoria, teniendo en cuenta si aprobó o no en primera convocatoria las asignaturas: $X_1$= *Análisis Matemático*, $X_2$= *Cálculo*, $X_3$= *Optimización* y $X_4$= *Probabilidad y Estadística*. El valor 1 en cualquiera de las 5 variables anteriores indica que la asignatura en cuestión fué aprobada en la primera convocatoria, mientras que el valor 2 hace alusión a que se necesitó de más de una convocatoria para aprobarla.

   Obtener el árbol de clasificación a partir del algoritmo *id3* justificando en cada paso la elección de la variable nodo por medio de la ganancia en información. No es necesario llevar a cabo el proceso de poda previa, aunque si es necesaria la explicación de esta característica del algoritmo.

5. Explicar los conceptos de muestras relacionadas y muestras independientes, asi como el tipo de test a aplicar en cada caso al tener en cuenta la normalidad o la no normalidad de las variables implicadas.

6. Se extrae una muestra aleatoria de tamaño $n$ de una distribución binomial $X$ con parámetros $m$ y $p$. Es decir $X \to B(m, p)$. La muestra se ha extraido con objeto de estimar el parámetro $p$ que permanece desconocido. Las variables aleatorias que forman la muestra anterior se denotan por $X_1, ..., X_n$.

| *Caso* | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $C$ |
|---|---|---|---|---|---|
| 1 | 1 | 2 | 1 | 1 | 1 |
| 2 | 2 | 1 | 1 | 1 | 1 |
| 3 | 1 | 2 | 1 | 1 | 1 |
| 4 | 2 | 2 | 1 | 1 | 1 |
| 5 | 1 | 1 | 1 | 1 | 1 |
| 6 | 2 | 2 | 1 | 1 | 1 |
| 7 | 1 | 1 | 1 | 1 | 1 |
| 8 | 2 | 2 | 1 | 1 | 1 |
| 9 | 1 | 1 | 2 | 2 | 1 |
| 10 | 2 | 1 | 2 | 2 | 1 |
| 11 | 1 | 2 | 1 | 2 | 2 |
| 12 | 2 | 2 | 1 | 2 | 2 |
| 13 | 1 | 2 | 2 | 2 | 2 |
| 14 | 2 | 2 | 2 | 2 | 2 |
| 15 | 1 | 2 | 1 | 2 | 2 |
| 16 | 2 | 2 | 2 | 2 | 2 |
| 17 | 1 | 2 | 1 | 2 | 2 |
| 18 | 2 | 1 | 2 | 1 | 2 |
| 19 | 1 | 1 | 2 | 1 | 2 |
| 20 | 2 | 2 | 2 | 1 | 2 |

Obtener el estimador máximo verosímil para el parámetro $p$. Demostrar que dicho estimador máximo verosímil es eficiente para $p$.

**Nota**. Los ejercicios 1, 2, 3 y 4 se puntúan sobre 1.5 puntos. Los ejercicios 5 y 6 se puntúan sobre 1 punto. Los 2 puntos restantes provienen de la calificación de la práctica.