

# Métodos Matemáticos en Ciencias de la Computación

04 Junio 2003. Parte Teórica. 6 puntos

1. Diseñar un algoritmo genético –especificando la representación de los individuos, la función de coste, al igual que los operadores de cruce y mutación– para resolver el siguiente problema de optimización combinatoria: Dados  $n$  números reales  $a_1, \dots, a_n$ , encontrar la partición de los mismos en cuatro grupos de tal manera que la suma de los números que pertenecen al grupo  $i$ -ésimo sea la mitad de la suma de los números pertenecientes al grupo  $i + 1$ -ésimo, con  $i = 1, 2, 3$ .

2. Se ha implementado un algoritmo genético para encontrar el óptimo en un problema de optimización combinatoria, en el cual se conoce de antemano dicho óptimo. La condición de parada del algoritmo genético se fija en cuanto se realiza un número determinado de evaluaciones. El algoritmo genético se ejecuta las veces necesarias hasta obtener el óptimo. De esta manera se obtiene que la primera vez que se consigue el óptimo ha sido en la tercera ejecución. A continuación se han necesitado siete ejecuciones para obtener de nuevo el óptimo. Finalmente el óptimo se ha alcanzado por tercera vez después de cinco ejecuciones. Con los datos anteriores, y utilizando el modelo probabilístico adecuado, obtener tanto el estimador máximo verosímil como la estimación máximo verosímil para el parámetro  $p$  el cual denota la probabilidad de obtener el óptimo con una ejecución del algoritmo genético implementado.

3. Aplicar el lema de Neyman–Pearson para obtener la región de rechazo óptima para el siguiente test de hipótesis:

$$H_0 : \lambda = \lambda_0$$

$$H_1 : \lambda = \lambda_1$$

a partir de una muestra de tamaño  $n$  extraída de una población  $X$  con distribución de probabilidad de Poisson de parámetro  $\lambda$ .

4. Dadas tres variables aleatorias discretas  $X, Y$  y  $Z$ . Demostrar que se verifica la siguiente igualdad entre la cantidad de información mutua de las variables  $X$  e  $Y$  condicionada a  $Z$  y las entropías condicionadas de  $X$  a  $Z$ , de  $Y$  a  $Z$  y de la conjunta  $(X, Y)$  a  $Z$ :

$$I(X, Y|Z) = H(X|Z) + H(Y|Z) - H(X, Y|Z)$$

5. Supongamos que el 20 por ciento de la población masculina con edades superiores a 60 años tiene cáncer de próstata. Para tratar de diagnosticarlo nos fijamos en los niveles de expresión de dos genes concretos que denotamos por  $G_1$  y  $G_2$ . Ambas variables se van a considerar discretizadas a dos valores.

Se conoce que la probabilidad de que un individuo con cáncer de próstata tenga un valor alto de  $G_1$  es de 0.75, mientras que para individuos sin cáncer de próstata el valor alto de  $G_1$  tan solo se da con probabilidad 0.20.

Por lo que respecta a  $G_2$ , se tiene que entre los que presentan cáncer de próstata, el valor alto de  $G_2$  se da con probabilidad 0.10, y entre los que no presentan cáncer de próstata dicho valor alto se da con probabilidad 0.60.

Efectuar el diagnóstico para un individuo con valor de expresión alta en  $G_1$  y baja en  $G_2$ . Explicar las premisas sobre las que se basa el paradigma utilizado para llevar a cabo el diagnóstico.

6. Diseñar un algoritmo de estimación de distribuciones en su versión UMDA –especificando la representación de los individuos, la función de coste, al igual que las posibles dificultades derivadas de la esencia propia del UMDA– para el problema del clustering particional.