

Análisis Multivariado: Tarea 4

Métodos de Clasificación

Fecha de entrega: 26 de mayo.

Análisis de Conglomerados

1. (1 punto) El método flexible de Lance y Williams establece que podemos definir la distancia de un tercer cluster con dos cluster recién agrupados como

$$d(C_3, C_1 \cup C_2) = \alpha_1 d(C_3, C_1) + \alpha_2 d(C_3, C_2) + \beta d(C_1, C_2) + \gamma |d(C_3, C_1) - d(C_3, C_2)|.$$

Muestra que para los siguientes valores recuperamos las siguientes ligas.

Liga	α_i	β	γ
Simple	$\frac{1}{2}$	0	$-\frac{1}{2}$
Compuesta	$\frac{1}{2}$	0	$\frac{1}{2}$
Centroide	$\frac{n_i}{n_1+n_2}$	$-\frac{n_1 n_2}{(n_1+n_2)^2}$	0
Ward	$\frac{n_i+n_3}{n_1+n_2+n_3}$	$-\frac{n_3}{n_1+n_2+n_3}$	0
Mediana	$\frac{1}{2}$	$-\frac{1}{4}$	0

2. (1 punto) Considerando a \mathbf{T} la matriz de variación total dada por

$$\mathbf{T} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T = \mathbf{W} + \mathbf{B} = \sum_{i=1}^k \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i.})(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i.})^T + \sum_{i=1}^k n_i (\bar{\mathbf{x}}_{i.} - \bar{\mathbf{x}}..)(\bar{\mathbf{x}}_{i.} - \bar{\mathbf{x}}..)^T$$

donde \mathbf{W} es la matriz de variación dentro de los clusters (*within-cluster variation*) y \mathbf{B} es la matriz de variación entre clusters (*between-cluster variation*). Mostrar que para $k = 2$

- i. \mathbf{B} es una matriz simétrica de rango 1.

- ii. Usando el hecho de que para una matriz \mathbf{A} simétrica de rango 1 se cumple que

$$|\mathbf{I} - \mathbf{A}| = 1 - \text{tr}(\mathbf{A})$$

mostrar que los siguientes métodos para dividir n objetos p dimensionales son equivalentes

- (a) Minimizar $|\mathbf{W}|$
 - (b) Maximizar $\text{tr}(\mathbf{B}\mathbf{W}^{-1})$
 - (c) Minimizar $\text{tr}(\mathbf{W}\mathbf{T}^{-1})$
- (Hint: Mostrar $(a) \Leftrightarrow (c)$ y después $(b) \Leftrightarrow (c)$).

3. (3 puntos) El archivo *NCI.txt* contiene 6830 variables de 64 células cancerígenas de distintos tipos de cáncer. Realizar lo siguiente:
- a) Mediante un análisis de conglomerados (aglomerativo, divisivo y/o fuzzy) estudiar el posible agrupamiento de las células cancerígenas. ¿Qué valor de k parece más razonable?
 - b) Utilizar el algoritmo k -means con el valor de k elegido en el inciso anterior y realiza un análisis comparando ambos métodos, el valor de k elegido y los clusters creados.

Análisis de Discriminantes

4. (1 punto) En un problema de discriminación de dos grupos, asumir que

$$f_i(x) = \binom{n}{x} \theta_i^x (1 - \theta_i)^{n-x}$$

donde θ_1 y θ_2 son conocidas. Si π_1 y π_2 son las probabilidades a priori muestre que se obtiene una función discriminante lineal como regla óptima. Da la probabilidad de clasificación equivocada $P(1|2)$ asumiendo que $\theta_1 > \theta_2$.

5. (1 punto) El criterio de Fisher era encontrar una proyección

$$\mathbf{z} = \mathbf{X}\mathbf{a},$$

que maximizara la variación entre grupos relativo a la variación dentro de los grupos. Esto es,

$$\max_{\mathbf{a}} \frac{\mathbf{a}^T \mathbf{B} \mathbf{a}}{\mathbf{a}^T \mathbf{W} \mathbf{a}}.$$

Mostrar que el vector \mathbf{a} que resuelve el problema de optimización es el eigenvector más grande de $\mathbf{W}^{-1}\mathbf{B}$. Para esto:

- i. Notar que este problema de optimización es equivalente a

$$\begin{aligned} \max_{\mathbf{a}} \quad & \mathbf{a}^T \mathbf{B} \mathbf{a} \\ \text{s.a.} \quad & \mathbf{a}^T \mathbf{W} \mathbf{a} = 1. \end{aligned}$$

- ii. Considerando el cambio de variable $\mathbf{b} = \mathbf{W}^{\frac{1}{2}} \mathbf{a}$ resuelve el problema de optimización y encuentra \mathbf{a} .
- iii. Mostrar que $\mathbf{W}^{-1} \mathbf{B}$ y $\mathbf{W}^{-\frac{1}{2}} \mathbf{B} \mathbf{W}^{-\frac{1}{2}}$ tienen los mismos eigenvalores.

6. (3 puntos) La base de datos *SP500.txt* contiene el porcentaje de retornos desde inicios del 2001 a finales de 2005. Para cada fecha se tiene el porcentaje de retornos record para cada uno de los 5 días previos, el volumen de transacciones del día previo, el porcentaje de retorno del día actual y un indicador binario de si el mercado iba hacia arriba o hacia abajo en esa fecha. Realizar lo siguiente

- i. Analiza la correlación de las variables.
- ii. Separa los datos en una muestra para entrenar el modelo (del año 2001-2004) y una muestra para probar el modelo (año 2005).
- iii. El objetivo es predecir el comportamiento del mercado en el año 2005. Por lo que se pide:
- Realizar un análisis de discriminante lineal utilizando las variables Lag1 y Lag2 y el comando *lda()* de la paquetería *MASS* de *R*. Utilizando la función *partimat()* de la librería *klaR* de *R* o programando tu propia función, dibuja el hiper-plano que separa las regiones y predice a través de la función *predict()* el tipo de mercado para el año 2005. Obtén la matriz de confusión y analiza los resultados.
 - Repetir el inciso anterior pero esta vez utilizando un discriminante cuadrático utilizando la función *qda()*. ¿Qué método es mejor para estas dos variables?

6. (1 punto)¹ ¿Puedes encontrar un mejor análisis discriminante a los utilizados en el ejercicio anterior?

¹Ejercicio opcional.