Análisis Multivariado: Tarea 4

Técnicas de reducción de dimensión II

Fecha de entrega: 2 de mayo

Escalamiento multidimensional

1. (0.5 puntos) Mostrar que para la matriz doblemente centrada $\mathbf{B}$ se tiene que

$$b_{ij} = a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot}$$

2. (0.5 puntos) Sea $\mathbf{S}$ una matriz de similitudes, esto es, $s(i,j) = s(j,i)$ y $s(i,j) \leq s(i,i)$ para todo i y j. Recordar que se puede definir una matriz de distancias $\mathbf{D}$ a partir de la transformación

$$d(i,j) = \sqrt{s(i,i) - 2s(i,j) + s(j,j)}.$$

   Mostrar que si $\mathbf{S}$ es semi-positiva definida, entonces $\mathbf{D}$ es euclidiana con matriz doblemente centrada $\mathbf{B} = \mathbf{HSH}$. Para esto

   i. Mostrar que $\mathbf{D}$ en efecto es una matriz de distancias.

   ii. Mostrar que $\mathbf{HSH}$ es semi-definida positiva.

   iii. Mostrar que $\mathbf{B} = \mathbf{HSH}$ y concluir que $\mathbf{D}$ es euclidiana.

3. (2 puntos) El archivo *dist_ciudades_RU.txt* contiene la distancia en carretera de 14 ciudades del Reino Unido.

   i. Mediante un escalamiento multidimensional clásico reconstruye el mapa del Reino Unido. ¿Puedes identificar las ciudades?

   ii. ¿La matriz de distancias es Euclidiana? De no serlo, ajusta un escalamiento multidimensional añadiendo la constante aditiva. ¿Existe una mejora al momento de reconstruir el mapa e identificar las ciudades?

   iii. Realiza un escalamiento multidimensional no métrico y compara tus resultados con los incisos anteriores.

4. (2 puntos) En este ejercicio se hará un análisis de un *bowl* de puntos en $\mathbb{R}^3$. Para esto se debe realizar lo siguiente:

   i. Construir el bowl dado por las coordenadas

   $$x_i = \frac{1}{2}\sqrt{v_i}\cos(2\pi u_i)$$
   $$y_i = \frac{1}{2}\sqrt{v_i}\sin(2\pi u_i)$$
   $$z_i = v_i - \frac{1}{2},$$

   donde u_i y v_i son variables aleatorias uniformes en el intervalo $(0, 1)$ (considerar e.g. $n = 1000$) y graficar la figura resultante coloreando los puntos usando un mapeo apropiado (e.g., utilizando la tercer coordenada).

   ii. Obtener la matriz de distancias Euclidianas.

   iii. Realizar un escalamiento multidimensional métrico con $k = 2$ y graficar los resultados coloreando los puntos utilizando el mismo mapeo que en el primer inciso. ¿Es el resultado esperado? Comenta en los resultados haciendo énfasis en los supuestos del escalamiento multidimensional métrico.

   iv. ¿Puedes encontrar una mejor solución con alguna otra técnica vista en clase?

Análisis de correspondencias

5. (1 puntos) El archivo *disciplinas_recursos* contiene los datos de una organización de investigación que clasificó a 796 investigadores de acuerdo con su disciplina científica, así como de acuerdo a los recursos que se les habían otorgado, A, B, C, D y E, siendo los que más recibieron los del grupo A, los que menos los del grupo D y los que fueron rechazados en el grupo E.

   i. Realiza un análisis de correspondencias y comenta en los resultados, justificando de manera adecuada todos los pasos del análisis.

   ii. Para 53 investigadores que laboran en museos y no en universidades se tiene que 4 de ellos pertenecen al grupo A, 12 al grupo B, 11 al grupo C, 19 al grupo D y 7 al grupo E. ¿Cómo se comparan con los 796 investigadores restantes?

6. (2 puntos) En este ejercicio se analizarán los resultados a la pregunta "Una mujer con un niño en edad escolar en casa, ¿debe trabajar a tiempo completo, a tiempo parcial, o debe permanecer en casa?", realizada a 33,590 individuos de 24 países como parte de la encuesta

de 1994 del Programa Internacional de Investigación sobre la familia y los cambios de rol de género.

(a) Realiza un análisis de correspondencias para la base de datos *women.xls* y comenta en los resultados.

(b) Realiza el mismo procedimiento que en el inciso anterior pero ahora con la base *women2.xls*, la cual tiene desagregadas las repuestas por sexo. Comenta en tus resultados.

7. (2 puntos) Considera la base de datos *survival.txt* la cual tiene resultados de supervivencia de 764 individuos de 3 ciudades diferentes y desagregados por grupo de edad y realiza lo siguiente.

(a) Construye la matriz indicadora que constará de 764 renglones y 8 columnas.

(b) Construye la matriz de Burt.

(c) Realiza un análisis de correspondencias múltiple en cualquiera de las dos matrices anteriores y comenta en los resultados.