Entrega Ejercicios Cluster

Daniel Czarnievicz

2018

Ejercicio 1

En la clasificación jerárquica agregativa:

a. **Verdadero**. La metodología parte de tomar a cada individuo por separado. Define algoritmos para agregar individuos con individuos, individuos con grupos, y grupos con grupos. El proceso se itera en cada etapa hasta lograr un grupo con todos los individuos.

b. **Falso**. Por ejemplo, la distancia euclidea no tiene en cuenta las correlaciones entre las observaciones, mientras que la distancia de Mahalanobis sí. Pueden verse los laboratorios del curso para obtener contra-ejemplos de esta afirmación.

c. **Verdadero**. No dependen del algoritmo, dependen de la distancia.

d. **Falso**. El método une en la etapa 1 a $min\{A;\ B;\ C;\ D;\ ...\}$.

e. **Verdadero**. Esto se debe a que la distancia de Mahalanobis tiene en cuenta las correlaciones entre las variables en su fórmula.

$$d_{ij}^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})' \mathbf{\Sigma}^{-1} (x_{ik} - x_{jk})$$

f. **Falso**. El procedimiento comienza con cada observación como su propio grupo, y termina con un grupo con todas las observaciones.

En la clasificación no jerárquica:

a. **Falso**. En la clasificación no jerárquica se debe definir la cantidad de grupos de forma a priori.

b. **Verdadero.**. Ver Rencher sección 14.4.1a para una explicación del método.

c. **Falso**. El criterio no requiere especificar individuos representativos de forma a priori.

d. **Verdadero**. Este es el objetivo del método.

Ejercicio 2

a. **Falso**. Los primeros individuos que se unen son el A y el D dado que ellos son la solución a la ecuación:

$$Etapa\ \ 1 : min\big\{(A,B);\ (A,C);\ (A,D);\ (B,C);\ (B,D);\ (C,D)\big\}$$

b. **Verdadero**. Las etapas son las siguientes:

$$Etapa\ \ 1 : min\big\{(A,B);\ (A,C);\ (A,D);\ (B,C);\ (B,D);\ (C,D)\big\}$$

	A	**B**	**C**	**D**
A	0	6	3	2
B		0	5	8
C			0	4
D				0

$$Etapa\ \ 2 : min\big\{min(AD, B);\ min(AD, C);\ (B, C)\big\}$$

	(AD)	**B**	**C**
(AD)	0	6	3
B		0	5
C			0

$$Etapa\ \ 3 : min\big\{(ADC), B\big\}$$

	(ACD)	**B**
(ACD)	0	3
B		0

c. **Falso**, es 5.

d. **Falso**, es 5, lo que corresponde a la distancia entre C y B.

Ejercicio 3

1. Comenzaría con Simple Linkage ya que al ser el algoritmo más sensible a los valors atípicos permite ver claramente la distorsión que estos producen. Si este algoritmo tiene como resultado un dendogrrama alargado (es esperable que así sea), entonces se puede probar con Complete Linkage, y Ward. El problema de estos algoritmos es que podrían estar forzando un agrupamiento que no sea el que refleje la verdadera naturaleza de los datos (por ejemplo, Ward tiende a formar grupos esféricos).

   Dado que estas observaciones son atípicas en la gran mayoría de las variables, es esperable que ninguno de los algoritmos anteriores solucione el problema. Podría entonces separarse momentáneamente estas dos observaciones y repetir el proceso. Si existiera una estructura de datos que dichas observaciones no me permiten ver, se debería solucionar de esta forma. Las mismas deben luego ser incluidas en el grupo al que más se parezcan (por ejemplo, utilizando la mínima distancia al centroide del grupo). El principal problema con estas estrategias radíca en el concepto de dato atípico. Separar esas dos observaciones modificaría el rango intercuartilico de la distribución de las variables. Por tanto, luego de separadas, nuevas observaciones podrían caer dentro de la categoría de atípicas.

   En caso de que estas estrategias tampoco solucione el problema, puede deberse a que efectivamente no exista una estructura de grupos entre los datos (tal como sugería Simple Linkage con todas las observaciones). Si aún así es necesario formar tipologías, entonces sería recomendable utilizar métodos no jerárquicos (por ejemplo, k-medias) y combinarlos con fuzzy sets (la cual entrega el coeficiente de pertenencia a cada grupo. Por último, si la naturaleza de los datos lo permite, puede considerarse el armar un grupo solo con las dos observaciones que presentan valores atípicos.

2. Repetir el proceso del punto 1. Es probable que el método de Ward solucione el problema dado que la atipicidad está dada solo por una variable. Cuanto mayor número de variables, menos relevante será la atipicidad en una de ellas.

3. Lo importante en estos casos es utilizar los métodos correctos para medir distancias. Dos de estos posibles métodos son: Simple Matching Coefficient y, el coeficiente de Jaccard. Ambos se basan en el uso de tablas de contingencia. Utilizando estas distancias, todos los métodos de clustering vistos en el curso pueden ser utilizados para estos datos. Otra solución sería utilizar clustering basado en un modelo multinomial en caso de muestreos aleatorios con reposición, o en un modelo multihipergeométrico en caso de un muestreo sin reposición.

4. Comenzaría analizando las variables cuantitativas. Una vez formados estos grupos, analizaría la variable cualitativa. Si los niveles de la cualitativa quedan determinados por las otras variables, entonces las cuantitativas son suficientes para revelar la estructura de grupos (esto puede realizarse mediante una regresión Logit o Probit en caso de vaiables binarias, o sus extensiones para casos de variables con más de dos categorías). En caso de que al analizar la variable cualitativa, proporciones no despreciables

pertenezcan a los distintos grupos definidos por las cuantitativas, entonces deben emplearse métodos adicionales, por ejemplo los mencionados en la parte c.

Ejercicio 4

Análisis visual del dendograma: dado que las barras verticales del dendograma indican el nivel al que se unen los individuos/grupos, analizar visualmente la longitud de dichas barras da una idea de la estructura de grupos que presentan los datos. Este método es muy informal y solo permite una primera aproximación. En caso de tener una cantidad predeterminada de grupos, esto determinará el nivel en el cual cortar. En este caso donde dicha cantidad no está especificada a priori, sugiero 4 grupos.

Criterio del R^2: Lo que hace este indicador es relacionar la variación explicada y la variación total, dónde la variación explicada está representa por la estructura de grupos que se encuentra en cada nivel. Por lo tanto, el valor del R cuadrado se encuentra en el intervalo [0,1], valiendo 0 cuando todas las observaciones se encuentran en un mismo grupo (es decir, la variación explicada por la estructura de grupos es 0 ya que todos los individuos están en un mismo grupo), y valiendo 1 cuando cada individuo es un grupo (es decir, cuando la variación explicada es igual a la variación total).

En términos matemáticos:

$$R^2 = 1 - \frac{\sum_{k=1}^{g} \sum_{i=1}^{n_k} \sum_{j=1}^{p} \left(x_{ijk} - \bar{x}_{jk}\right)^2}{\sum_{i=1}^{n} \sum_{j=1}^{p} \left(x_{ij} - \bar{x}_{j}\right)^2}$$

donde p es la cantidad de variables, g la cantidad de grupos, n la cantidad total de observaciones, y n_k la cantidad de observaciones en el grupo k.

Para este ejemplo vemos que el mayor aumento del indicador se da cuando pasamos de 3 grupos ($R^2 = 0.3978791$) a 4 grupos ($R^2 = 0.5081371$) leyendo desde abajo hacia arriba. Dado que aquí es cuando se da el mayor aumento (en términos proporcionales), nos quedamos con 4 grupos.

Criterio pseudo-F: relaciona la suma de variaciones entre los grupos (variación explicada) con la suma de variaciones en los grupos (variación residual):

$$Pseudo\,F = \frac{\text{Variación entre grupos}/(k-1)}{\text{Variación en los grupos}/(n-k)}$$

La cual puede calcularse también como:

$$Pseudo\,F = \frac{tr(B)/(k-1)}{tr(W)/(n-k)}$$

O como cocientes del R^2:

$$Pseudo\,F = \frac{tr(B)/(k-1)}{tr(W)/(n-k)}$$

Reglas empíricas de utilización:

- Si el indicador crece de forma monótona al crecer el número de grupos, no se puede determinar una estructura de grupos.
- Si disminuye de forma monótona al crecer el número de grupos, no se puede determinar una estructura de grupos, pero se puede decir que existe jerarquía.
- Cuando se halla un máximo, la población presenta una estructura de grupos en ese máximo.

En este caso, en cuatro grupos también se observa un máximo local. Cabe destacar que este no es el único máximo local. Esto enfatiza la necesidad de no quedarse con un solo método, sino que combinarlos.

Nota: el indicador no es una variable aleatoria que se distribuye F.

Criterio pseudo-t^2: este indicador busca determinar la significación de juntar dos grupos. Busca determinar en cada paso, si la disminución de la SCR (variación intra-grupos) como resultado de pasar de k a $k+1$ grupos es significativa o no. Algebraicamente, este indicador también se basa en las tazas de las matrices de varianza en el grupo (W), antes y después de unir dos grupos:

$$\text{Pseudo-}t^2 = \frac{tr(W_{GL}) - \Big[tr\big(W_G + tr(W_L)\big)\Big]}{\Big[tr(W_G) + tr(W_L)\Big] / \big(n_G + n_L - 2\big)}$$

donde:

- n_i es la cantidad de observaciones en el grupo $i = G,\ L$.
- $tr(W_i)$ es la traza en el grupo $i = G,\ L,\ GL$

Si se analiza el vector de valores de pseudo-t^2 desde abajo (1 grupo con n observaciones) hacia arriba (n grupos con 1 observación cada uno), si en $k+1$ grupos presenta una caída "fuerte" respecto de k, entonces nos quedamos con $k+1$ grupos. En este caso, dicho comportamiento se observa al pasar desde 3 grupos (pseudo-$t^2 = 194.72$) a 4 grupos (pseudo-$t^2 = 44.84$), por lo que sugiere quedarse con 4 grupos.

Dado que tanto la inspección visual, como los tres índices sugieren quedarse con 4 grupos, esta es la opción por la que se opta. Como ya se adelantó, los diferentes métodos pueden no siempre coincidir, por lo que es necesario contemplar lo que todos ellos sugieren para tomar una decisión.