Análisis de Clasificación

Juan Sosa, PhD

Universidad Externado de Colombia

I - 2018

Análisis de Clasificación (Clustering)

Objetivo

Identificar grupos (*clusters*) de individuos que sean homogéneos dentro de los grupos y heterogéneos entre grupos.

Observaciones

- Diferente del análisis discriminante!
- Los grupos no están definidos a priori (o si incluso existen grupos).
- Es necesario considerar una distancia.
- Cuántos grupos?
- Como asignar los individuos?

Métodos

- Clasificación jerárquica.
- Agrupamiento de K-medias.
- Agrupamiento basado en el modelo (estructura probabilística).

Agrupamiento jerárquico

Algoritmo

1. *Start:* $C_1, C_2, \ldots, C_n$ (singletones), i.e., $K = n$.
2. Encontrar y unir el par de clusters más *cercanos*, C_i y C_j.
3. Decrecer el numero de clusters en 1, i.e., $K \leftarrow K - 1$.
4. *Stop* si $K = 1$. De lo contrario, volver al paso 2.

Distancias (similaridad) entre individuos

- $d_{ij} = \sqrt{\sum_\ell (x_{i,\ell} - x_{j,\ell})^2}$: Euclidiana.
- $d_{ij} = \max_\ell |x_{i,\ell} - x_{j,\ell}|$: Norma máxima.
- $d_{ij} = \sum_\ell |x_{i,\ell} - x_{j,\ell}|$: Manhattan.
- $d_{ij} = \left(\sum_\ell (x_{i,\ell} - x_{j,\ell})^p\right)^{1/p}$: Minkowski.

Distancias entre clusters A y B

- $d_{AB} = \min_{i \in A, j \in B}\{d_{ij}\}$: *single linkage clustering.*
- $d_{AB} = \max_{i \in A, j \in B}\{d_{ij}\}$: *complete linkage clustering.*
- $d_{AB} = \frac{1}{n_a\, n_B} \sum_{i \in A} \sum_{i \in B} d_{ij}$: *group average clustering.*

Agrupamiento jerárquico (cont.)

Cómo elegir K?

Examinar los tamaños de los cambios de altura en el dendrograma y tomar un "gran" cambio para indicar el número apropiado de clusters para los datos.

Cons: *chaining*

Tendencia juntar puntos intermedios a un cluster ya establecido en lugar de inicializar un nuevo cluster.

Observaciones

- Se debe ser cuidadoso al elegir las distancias (conmesurabilidad).
- No tiene estructura probabilística.

Agrupamiento de K-medias

Método

Encontrar la partición de n individuos en K grupos que minimicen el *within-group sum of squares* (WGSS):

$$\text{WGSS} = \sum_{j=1}^{p} \sum_{\ell=1}^{K} \sum_{i \in G_\ell} (x_{ij} - \bar{x}_j^{(\ell)})^2$$

donde $\bar{x}_j^{(\ell)} = \sum_{i \in G_\ell} x_{ij}$ es la media en el grupo G_ℓ con la variable j.

Algoritmo (Steinley, 2008)

- Partición inicial (agrupamiento jerárquico).
- Calcular el cambio en el criterio de agrupamiento moviendo cada individuo de su propio cluster a otro cluster.
- Ejecutar el cambio que conlleve al mayor cambio en el valor del criterio de agrupamiento.
- Repetir pasos 2. y 3. hasta que el movimiento de ningún individuo cause una mejora.

Agrupamiento de K-medias (cont.)

Propiedad

$$\begin{aligned} \text{SC TOTAL} &= \text{SC ENTRE} + \text{SC DENTRO} \\ &= \text{SC EXPLICADA} + \text{SC NO EXPLICADA} \end{aligned}$$

$$\text{Porcentaje de variabilidad explicado} = \frac{\text{SC EXPLICADA}}{\text{SC TOTAL}}$$

Cómo elegir K?

Graficar el WGSS frente a K y elegir aquel valor de K donde el decrecimiento del WGSS no sea "significativo".

Cons

- No es invariante a la escala de medición (conmensurabilidad).
- Tiende a construir grupos con estructura "esferica".

Observaciones

- No tiene estructura probabilística.

Agrupamiento basado en el modelo

Objetivo

Desarrollar un modelo estadístico que caracterice el mecanismo aleatorio que explique como se genera la data.

Modelo

$$f(\boldsymbol{x}; \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{j=1}^{K} \pi_j f_j(\boldsymbol{x}, \boldsymbol{\theta}_j)$$

donde $\sum_{j=1}^{J} \pi_j = 1$ (probabilidades de la mezcla).

Objetivo

Crear los grupos a partir de las probabilidades a posteriori:

$$\mathbb{P}\mathrm{r}\left[\text{Grupo } k \mid \mathbb{X}\right] = \frac{\pi_k \, f(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma})}{\sum_{j=1}^{K} \pi_j \, f(\boldsymbol{x} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma})}$$

Agrupamiento basado en el modelo (cont.)

Diagnósticos

Establecer la estabilidad del agrupamiento:

- Fuerza de predicción: Cross-Validation (OK si superior a 0.8 o 0.9).
- Indice de Jaccard: Jittering o Bootstraping (OK si superior a 0.75).