K-medias

Fundamentos

- Supongamos que tenemos una muestra de n elementos con p variables. El objetivo es dividir esta muestra en un número de grupos prefijado, K.
- El algoritmo de K −medias requiere las cuatro etapas siguientes:

1. Seleccionar K puntos como centros de los grupos iniciales. Esto puede hacerse:
   a) asignando aleatoriamente los objetos a los grupos y tomando los centros de los grupos formados
   b) tomando como centros los K puntos más alejados entre sí
   c) construyendo los grupos con información a priori, o bien seleccionando los centros a priori.

Fundamentos

2. Calcular las distancias euclídeas de cada elemento al centro de los K grupos, y asignar cada elemento al grupo más próximo. La asignación se realiza secuencialmente y al introducir un nuevo elemento en un grupo se recalculan las coordenadas de la nueva media de grupo.

3. Definir un criterio de optimalidad y comprobar si reasignando uno a uno cada elemento de un grupo a otro mejora el criterio.

4. Si no es posible mejorar el criterio de optimalidad, terminar el proceso.

Implementación del algoritmo

- El criterio de homogeneidad que se utiliza en el algoritmo de K −medias es la *suma de cuadrados dentro de los grupos (SCDG)* para todas las variables, que es equivalente a la suma ponderada de las varianzas de las variables en los grupos:

$$SCDG = \sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{i=1}^{n_k} \left(x_{ijk} - \bar{\mathbf{x}}_{jk}\right)^2$$

donde x_{ijk} es el valor de la variable j en el elemento i del grupo k, y $\bar{\mathbf{x}}_{jk}$ es la media de esta variable en el grupo.

Implementación del algoritmo

- El criterio se escribe:

$$\min SCDG = \min \sum_{k=1}^{K} \sum_{j=1}^{p} n_k s_{jk}^2$$

donde n_k es el número de elementos del grupo k y s_{jk}^2 es la varianza de la variable j en dicho grupo.

- La varianza de cada variable en cada grupo es claramente una medida de la heterogeneidad del grupo y al minimizar las varianzas de todas las variables en los grupos obtendremos grupos más homogéneos.

Implementación del algoritmo

- Un posible criterio alternativo de homogeneidad sería minimizar las distancias al cuadrado entre los centros de los grupos y los puntos que pertenecen a ese grupo. Si medimos las distancias con la norma euclídea, este criterio se escribe:

$$\min \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \bar{\mathbf{x}}_k)^t (x_{ik} - \bar{\mathbf{x}}_k) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} d^2(i,g)$$

donde $d^2(i,g)$ es el cuadrado de la distancia euclídea entre el elemento i del grupo g y su media de grupo.

Implementación del algoritmo

- Es fácil comprobar que ambos criterios son idénticos. Sabemos que un escalar es igual a su traza, entonces podemos escribir el último criterio como:

$$\min \sum_{k=1}^{K} \sum_{i=1}^{n_k} traza[d^2(i,g)] = traza\left[\sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \bar{\mathbf{x}}_k)^t (x_{ik} - \bar{\mathbf{x}}_k)^t\right]$$

Y llamando W a la matriz de SCDG:

$$W = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \bar{\mathbf{x}}_k)(x_{ik} - \bar{\mathbf{x}}_k)^t$$

Tenemos que

$$\min traza(W) = \min SCDG$$

Como la traza es la suma de los elementos de la diagonal principal ambos criterios coinciden.

Implementación del algoritmo

- Este criterio se denomina criterio de la traza, y fue propuesto por Ward (1963).

- El criterio de la traza tiene dos propiedades importantes. La primera es que no es invariante ante cambios de medida en las variables. Cuando las variables vayan en unidades distintas conviene estandarizarlas, para evitar que el resultado del algoritmo dependa de cambios irrelevantes en la escala de medida.

- Cuando vayan en las mismas unidades suele ser mejor no estandarizar, ya que es posible que una varianza mucho mayor que el resto sea precisamente debida a que existen dos grupos de observaciones en esa variable, y si estandarizamos podemos ocultar la presencia de los grupos.

Implementación del algoritmo

- La segunda propiedad del criterio de la traza es que minimizar la distancia euclídea produce grupos aproximadamente esféricos.
- Por otro lado este criterio está pensado para variables cuantitativas.
- Aunque puede aplicarse si existe un pequeño número de variables binarias, si una parte importante de las variables son atributos, es mejor utilizar los métodos jerárquicos.

Implementación del algoritmo

- La maximización de este criterio requeriría calcularlo para todas las posibles particiones en el número de grupos especificado.
- Esto es computacionalmente muy costoso, salvo para valores de n muy pequeños.
- Por eso, sólo encontraremos mínimos locales de SCDG, con lo cual, se recomienda aplicar el algoritmo usando diferentes configuraciones iniciales.

Implementación del algoritmo

- El algoritmo de k−medias busca la partición óptima con la restricción de que en cada iteración sólo se permite mover un elemento de un grupo a otro.

- El algoritmo funciona como sigue:
  1. Partir de una asignación inicial.
  2. Comprobar si moviendo algún elemento se reduce W.
  3. Si es posible reducir W, mover el elemento, recalcular las medias de los dos grupos afectados por el cambio y volver al punto 2. Si no es posible reducir W terminar.

- En consecuencia, el resultado del algoritmo puede depender de la asignación inicial y del orden de los elementos. Por eso siempre conviene repetir el algoritmo con distintos valores iniciales y permutando los elemento de la muestra.

Número de grupos

- En la aplicación habitual del algoritmo de K −medias hay que fijar el número de grupos, K.
- Obviamente, este número no puede estimarse con un criterio de homogeneidad ya que la forma de conseguir grupos muy homogéneos y minimizar la $SCDG$ es hacer tantos grupos como observaciones, con lo que siempre $SCDG = 0$.
- Se han propuesto distintos métodos para seleccionar el número de grupos.

Número de grupos

- Un procedimiento aproximado que se utiliza bastante es realizar un test F aproximado de reducción de variabilidad.

- Consiste en comparar la $SCDG$ de K grupos con la de $K+1$, y calcular la reducción proporcional de variabilidad que se obtiene aumentando un grupo adicional.

- El test es:

$$F = \frac{SCDG(K) - SCDG(K+1)}{SCDG(K+1)/(n-K-1)}$$

Número de grupos

- Se está comparando la disminución de variabilidad al aumentar un grupo con la varianza promedio.
- El valor de F suele compararse con una distribución $F_{p,\,p(n-K-1)}$
- Esta regla no esta muy justificada porque los datos no tienen porque verificar las hipótesis necesarias para aplicar la distribución F.
- Una regla empírica que da resultados razonables, sugerida por Hartigan (1975), e implantada en algunos programas informáticos, es introducir un grupo más si este cociente es mayor que 10.