Taller de Análisis de datos Multivariados
INEGI junio 2016

Leticia Gracia Medrano
IIMAS Depto. Probabilidad y Estadística
lety@sigma.iimas.unam.mx

1. DISTANCIAS

1.1. Distancias entre individuos con datos numéricos

La distancia puede generalizarse introduciendo una matriz$A > 0$ es decir una matriz positiva definida con dimensiones $p \times p$ esto es: $(x-y)A(x-y)$, cuando $A = S^{-1}$ donde **S** es la matriz de varianzas y covarianzas de las variables que conforman a los vectores, esta es conocida como distancia de Mahalanobis .

Hay otras distancias como la *city block* que corresponde a:

$$d(x,y) = |(x_1 - y_1)| + \ldots + |(x_p - y_p)|$$

Y también existe la generalización:

$$d(x,y) = ((x_1 - y_1)^\alpha + \ldots + (x_p - y_p)^\alpha)^{1/\alpha}$$

con α un entero,conocida como distancia Minkowski.

Para que una función entre dos puntos sea considerada distancia debe cumplir con las siguientes propiedades:

$$d(x,y) > 0 \; \forall x \neq y$$

$$d(x,y) = 0 \Leftrightarrow x \equiv y$$

$$d(x,y) \leq d(x,z) + d(z,y) \; \forall x, y, z$$

1.2. Distancias entre individuos con datos binarios

A continuación se presentan varias medidas de disimilitud. Si se tienen dos sujetos, i y j y si se denota por

a = todas las coincidencias ++

b = todas las no coincidencias +-

c = todas las no coincidencias -+

d = todas las coincidencias -+

p es el total de características $a + b + c + d = p$.

disimilitud entre i y j es

i) $d_{ij} = 1 - \frac{coincidencias}{p} = 1 - \frac{a+d}{p} = \frac{b+c}{p}$

esta **disimilitud** (en ocasiones no cumple con las propiedades de las distancias) corresponde a la proporción de variables que no coinciden.

Hay quienes sostienen que no deben tomarse en cuenta las coincidencias en ausencias, entonces: ii) Coeficiente de Jaccard

$d_{ij} = \frac{b+c}{a+b+c}$

d no se toma en cuenta pues la ausencia de cierta característica no ayuda a decir si son o no parecidos.

iii)Coeficiente de Czekanowski

$d_{ij} = \frac{b+c}{2a+b+c}$ si se quiere compensar el echar fuera a d

1.3. Distancias entre individuos con datos cualitativos

$c_{ij} =$ coeficiente de coincidencias de Sneath

$c_{ij} = \frac{\text{número de atributos en los que las unidades coinciden}}{p}$

$d_{ij} = 1 - c_{ij}$

1.4. Distancias entre individuos con datos de distintas escalas

Se puede utilizar el índice de Gower para crear distancias cuando se consideran datos con varios tipos de escalas de medición.

Para cada variable x_k la similitud entre individuo i y el j se escribe como

$\delta_{ijk} = 1$ si puedo comparar a i contra j en la variable k o

$\delta_{ijk} = 0$ en otro caso.

y a $s_{ijk} = 1$ si son iguales o $s_{ijk} = 0$ si son diferentes.

La **similitud** entre i y j esta dada por

$$c_{ij} = \frac{\sum_{k=1}^{p} s_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}}$$

La disimilitud como $d_{ij} = 1 - c_{ij}$

Si las x_k son cuantitativas

$$s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{rango(x_k)}$$

.

Si x_k es cualitativa

$$s_{ijk} = 1$$

si los individuos concuerdan en la variable k

y

$$s_{ijk} = 0$$

si no.

1.5. Distancias entre variables con datos numéricos

Se puede tener otro enfoque, que tal que la disimilitud que quiere calcularse es entre (**columnas** de X), en este caso el coeficiente de correlación es una buena medida. Puede considerarse que si la correlación es cercana a 1 las variables en cuestión son cercanas. Ahora, antes de construir las medidas hay que decidir si

un coeficiente de correlación alto pero negativo significa un acercamiento grande entre las variables, o un total alejamiento.

Una disimilitud podría ser $d(V_1, V_2) = 1 - \rho(V_1, V_2)$ o también esta otra $d(V_1, V_2) = 1 - \rho(V_1, V_2)^2$

1.6. Distancias entre variables con datos cualitativos

Si se requiere construir disimilitud entre variables y se trabaja con datos cualitativos

Se construye una tabla de contingencias tamaño 2×2 donde $a+c+b+d = n$

$d_{kl} = \chi^2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(a+c)(c+d)(b+d)}$

pero depende del tamaño de la muestra $\chi^2 \leq n$.

Otra medida que puede usarse es

$d_{kl} = 1 - \sqrt{\frac{\chi^2}{a+b+c+d}}$.

2. Algunos conceptos de Estadística

La **media poblacional** se define como:

$E[X_i] = \int x_i dF(x_i) = \mu_i$

$$E(\mathbf{x}) = \begin{pmatrix} E(x_1) \\ E(x_2) \\ \vdots \\ E(x_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$

La **varianza poblacional**, cuando existe,se define como:

$Var[X_i] = \int (x_i - \mu_i)^2 f(x_i) dx_1 = \sigma_i^2 = \sigma_{ii}^2$

La **covarianza poblacional** se define como:

$cov(x_i, x_j) = \int \int (x_i - \mu_i)(x_j - \mu_j) f(x_i, x_j) dx_i dx_j = \sigma_{ij}$

Estos valores se presentan dentro de la matriz de varianzas y covarianzas Σ

$$\Sigma = E[(X - E[X])(X - E[X])'] = var(x)$$

$$= \begin{bmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & \cdots & E[(X_1 - \mu_1)(X_p - \mu_p)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & \cdots & E[(X_2 - \mu_2)(X_p - \mu_p)] \\ \vdots & \vdots & \vdots \\ E[(X_p - \mu_p)(X_1 - \mu_1)] & \cdots & E[(X_p - \mu_p)(X_p - \mu_p)] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \sigma_{21} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{bmatrix}$$

La **media muestral** de la j-ésima variable está dada por
$\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}$
Denotaremos al conjunto de las medias en un vector de medias muestrales
$\mathbf{x}' = (\bar{x}_1, \ldots, \bar{x}_p)$

La **varianza muestral** de la k-ésima variable se calcula como:
$S_{kk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2$

La **covarianza** entre la j-ésima variable y la k-ésima variable esta dada por
$S_{lk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{il} - \bar{x}_l)(x_{ik} - \bar{x}_k)$
La **matriz de covarianzas muestral** denotada por **S**, contiene a las varianzas y covarianzas.

$$\mathbf{S} = \frac{\mathbf{1}}{\mathbf{n-1}}\sum_{\mathbf{i=1}}^{\mathbf{n}}(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})' = \begin{bmatrix} {S_1}^2 & \ldots & S_{1p} \\ & \ddots & \vdots \\ & & {S_p}^2 \end{bmatrix}.$$

3. Análisis De Conglomerados

El objetivo de este análisis es formar grupos de observaciones, de manera que todas las unidades en un grupo sean similares entre ellas pero que sean diferentes a aquellas de otros grupos. La parte interesante es definir que es similar. Si por ejemplo el rango de las disimilitudes corre de 0 a C, podríamos definir que dos unidades son consideradas similares si su disimilitud es menor a $\frac{1}{2}C$.

Hay métodos **jerárquicos y no- jerárquicos**. Dentro de los jerárquicos están los aglomerativos y los divisivos.

Los métodos jerárquicos aglomerativos dan origen a un gráfico llamado **dendograma** . Estos método son iterativos y en cada paso debe recalcularse la matriz de distancias, en los casos en que se tienen muchas observaciones esto puede llevar mucho tiempo de cómputo.

A continuación se presentan varios de estos métodos:

- **Liga sencilla o del vecino más cercano**: La distancia entre dos conglomerados, es la mínima de todas las posibles distancias entre objetos que pertenecen a distintos conglomerados. Se unen los dos conglomerados más cercanos. Este método tiende a tener un buen desempeo cuando hay grupos de forma elongada, conocidos como tipo cadena.

- **Liga completa o del vecino más lejano**: La distancia entre dos conglomerados, es la máxima de todas las posibles distancias entre objetos que pertenecen a distintos conglomerados. Se unen los dos conglomerados más cercanos.Se desempea bien cuando los conglomerados son de forma circular.

- **Liga promedio**: La distancia entre dos conglomerados, es el promedio de todas las posibles distancias entre objetos que pertenecen a distintos conglomerados.Se unen los dos conglomerados más cercanos. Funciona bien para conglomerados tanto tipo cadena como circulares.

- **Liga centroide**: La distancia entre dos conglomerados está definida como la diferencia entre las medias (centroide) de cada conglomerado.Se unen los dos conglomerados más cercanos.

- **Liga mediana**: La distancia entre dos conglomerados está definida como la diferencia ponderada entre las medias (centroide) de cada conglomerado, la ponderación está dada por el tamao (número de unidades del conglomerado) de los conglomerados.Se unen los dos conglomerados más cercanos. Se espera un mejor rendimiento que el del centroide cuando los grupos varían mucho en tamao.

- **Método de Ward**: Se unen los dos conglomerdos que arrojen la menor varianza intragrupo (witih group) (con respecto al centroide de cada conglomerado).

Sólo las dos primeras ligas son invariantes ante transformaciones monótonas de las disimilitudes d_{ij} En algunos caso se puede trabajar con sólo la matriz de disimilitudes, en otros se requieren los datos originales.

Se construye el dendograma y se hace un corte a la altura máxima de la disimilitud que se fije para considerar a dos sujetos como similares.

En los métodos divisivos R tiene uno llamado DIANA.

Entre los no jerárquicos están:

- **k medias (k means)**: se requiere de antemano dar el número de grupos existente. Es un método iterativo, por lo que requiere de una solución inicial, y de allí se va optimizando una función objetivo. En cada paso se reasignan los elementos del conglomerado de manera tal que la suma de distancias al cuadrado de los puntos al centro de su conglomerado sea mínima.

- **basado en un modelo de mezcla de normales**: éste considera además diferentes estructuras de covarianza, es un método bayesiano, y usa estimación tipo EM.

- **de dos pasos**: se recomienda cuando se tienen muchos datos.

    1. **Formación de Preconglomerados**. La idea es reducir el tamaño de la matriz de distancias entre todos los posibles elementos. Los preconglomerados son conglomerados de los datos originales que se usan después en un método jerárquico. Se lee un caso, el algoritmo decide si se une a los conglomerados ya existentes, o forma un nuevo conglomerado. Una vez acabado el proceso de preaconglomeramiento, cada conglomerado es tratado como un elemento, la nueva matriz de distancias se reduce al número de preconglomerados.

2. **Agrupación de preconglomerados**. Se usa un método jerárquico estándar usando como elementos a los preconglomerados.

En general cada método puede llevar a una agrupación diferente. Se recomienda usar varios métodos para identificar los individuos que brincan de un grupo a otro según el método usado, tambien para definir los tamaños de los grupos.