Distâncias

Patrícia de Siqueira Ramos

UNIFAL-MG, *campus* Varginha

2 de Outubro de 2018

Distância

- Distância é um conceito importante para algumas técnicas multivariadas (AA - análise de agrupamento - é uma delas)
- Dados os valores das u.a. i e j, as distâncias entre i e j são representadas em uma matriz $n \times n$:

$$\mathbf{D} = \begin{bmatrix} D_{11} & D_{12} & \dots & D_{1n} \\ D_{21} & D_{22} & \dots & D_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ D_{n1} & D_{n2} & \dots & D_{nn} \end{bmatrix},$$

Distância euclidiana

- Dados os valores das u.a. i e j, a medida mais comum de distância entre i e j é a distância euclidiana, dada por

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2},$$

em que X_{ik} e X_{jk}, $k = 1, \ldots, p$, são os valores das variáveis para as observações i e j.

Distância euclidiana padronizada

Se as variáveis estiverem em escalas diferentes, padroniza-se cada variável pela sua variância e tem-se a distância euclidiana padronizada:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \frac{(X_{ik} - X_{jk})^2}{S_{kk}}},$$

Distância de Mahalanobis

- Se as covariâncias também forem levadas em conta, essa deve ser a medida de distância adotada
- Apenas calcularemos da forma matricial

Cálculo das distâncias de forma matricial

- Considerar os vetores $\mathbf{X_{i.}}$ e $\mathbf{X_{j.}}$ os vetores colunas referentes às observações i e j.
- As formas matriciais das distâncias são:

euclidiana:

$$d_{ij}^2 = (\mathbf{X_{i.}} - \mathbf{X_{j.}})^T(\mathbf{X_{i.}} - \mathbf{X_{j.}})$$

euclidiana padronizada:

$$d_{ij}^2 = (\mathbf{X_{i.}} - \mathbf{X_{j.}})^T\mathbf{D}^{-1}(\mathbf{X_{i.}} - \mathbf{X_{j.}})$$

Mahalanobis:

$$d_{ij}^2 = (\mathbf{X_{i.}} - \mathbf{X_{j.}})^T\mathbf{S}^{-1}(\mathbf{X_{i.}} - \mathbf{X_{j.}})$$

Exercício

Obter as matrizes de distâncias $\mathbf{D}_{n\times n}$ euclidiana, euclidiana padronizada e de Mahalanobis da amostra aleatória ($n = 4$, $p = 3$):

$$\mathbf{X} = \begin{bmatrix} 7 & 3 & 9 \\ 4 & 6 & 11 \\ 4 & 2 & 5 \\ 5 & 5 & 7 \end{bmatrix}.$$