SME0822 Análise Multivariada e Aprendizado Não-Supervisionado
Aula 1: Organização dos dados e Análise descritiva

Prof. Cibele Russo

cibele@icmc.usp.br

http://www.icmc.usp.br/~cibele

Introdução

Análise Multivariada é um ramo da estatística que busca estudar e desenvolver métodos para **descrever** e **analisar** dados multivariados.

Alguns objetivos específicos das técnicas de análise multivariada são

- Redução ou simplificação dos dados
- Ordenação e agrupamento dos dados
- Investigação da dependência entre variáveis
- Predição
- Testes de hipóteses

Organização de dados

Suponha que são observadas $p \geq 1$ variáveis em n indivíduos, itens ou unidades experimentais (observações).

Notação

Seja x_{jk}: medição da k-ésima variável na j-ésima unidade experimental, com $j = 1, \ldots, n$ e $k = 1, \ldots, p$.

Podemos representar os dados construindo a matriz de dados

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}_{n \times p}$$

Medidas-resumo

A **média amostral** da k-ésima variável, $k = 1, \ldots, p$ é dada por

$$\bar{x}_k = \frac{1}{n} \sum_{j=1}^{n} x_{jk}$$

A **variância amostral** da k-ésima variável, $k = 1, \ldots, p$ é dada por

$$s_k^2 = \frac{1}{n} \sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2$$

Medidas-resumo

A **covariância amostral** entre a i-ésima e k-ésima variáveis, $i, k = 1, \ldots, p; i \neq k$, é dada por

$$s_{ik} = \frac{1}{n} \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

A **correlação amostral** entre a i-ésima e k-ésima variáveis, $i, k = 1, \ldots, p; i \neq k$, é dada por

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_i^2}\sqrt{s_k^2}} = \frac{\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)}\sqrt{\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)}}$$

Medidas-resumo

O **vetor de médias amostrais** é dado por

$$\underset{\sim}{\bar{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}$$

Medidas-resumo

A **matriz de variâncias e covariâncias amostrais** ou simplesmente matriz de covariâncias amostrais é dada por

$$S = \begin{pmatrix} s_1^2 & s_{12} & \dots & s_{1p} \\ s_{12} & s_2^2 & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{2p} & \dots & s_p^2 \end{pmatrix}$$

Medidas-resumo

A **matriz de correlações amostrais** é dada por

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{12} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \dots & 1 \end{pmatrix}$$

Obs: $-1 \leq r_{ik} \leq 1, \; \forall i, k = 1, \dots, p, i \neq k$.

Exemplo 1: vendas de livros

São coletadas informações a respeito de 4 registros de vendas de livros:

Variável 1 (valor da nota): 42, 52, 48, 58

Variável 2 (número de livros): 2, 3, 2, 3

Neste exemplo, temos $p = 2$ variáveis e $n = 4$ observações.

Como fica a matriz de dados neste caso?

$$X = \begin{pmatrix} 42 & 2 \\ 52 & 3 \\ 48 & 2 \\ 58 & 3 \end{pmatrix}_{4\times 2}$$

Exemplo

Temos interesse em resumir a informação dos dados em medidas-resumo. Para isso, utilizamos o vetor de médias amostrais e a matriz de variâncias e covariâncias amostrais de X.

Além disso, é de se esperar que, quanto maior o número de livros no pedido, maior será o valor da compra. Essa ideia pode ser representada por meio da matriz de correlações amostrais de X.

Exemplo 2: dados de flores - Iris

Dados originalmente apresentados em

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems.

Annals of Eugenics, 7, Part II, 179–188.

Table I

Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5·1	3·5	1·4	0·2	7·0	3·2	4·7	1·4	6·3	3·3	6·0	2·5
4·9	3·0	1·4	0·2	6·4	3·2	4·5	1·5	5·8	2·7	5·1	1·9
4·7	3·2	1·3	0·2	6·9	3·1	4·9	1·5	7·1	3·0	5·9	2·1
4·6	3·1	1·5	0·2	5·5	2·3	4·0	1·3	6·3	2·9	5·6	1·8
5·0	3·6	1·4	0·2	6·5	2·8	4·6	1·5	6·5	3·0	5·8	2·2
5·4	3·9	1·7	0·4	5·7	2·8	4·5	1·3	7·6	3·0	6·6	2·1
4·6	3·4	1·4	0·3	6·3	3·3	4·7	1·6	4·9	2·5	4·5	1·7
5·0	3·4	1·5	0·2	4·9	2·4	3·3	1·0	7·3	2·9	6·3	1·8
4·4	2·9	1·4	0·2	6·6	2·9	4·6	1·3	6·7	2·5	5·8	1·8
4·9	3·1	1·5	0·1	5·2	2·7	3·9	1·4	7·2	3·6	6·1	2·5
5·4	3·7	1·5	0·2	5·0	2·0	3·5	1·0	6·5	3·2	5·1	2·0
4·8	3·4	1·6	0·2	5·9	3·0	4·2	1·5	6·4	2·7	5·3	1·9
4·8	3·0	1·4	0·1	6·0	2·2	4·0	1·0	6·8	3·0	5·5	2·1
4·3	3·0	1·1	0·1	6·1	2·9	4·7	1·4	5·7	2·5	5·0	2·0
5·8	4·0	1·2	0·2	5·6	2·9	3·6	1·3	5·8	2·8	5·1	2·4
5·7	4·4	1·5	0·4	6·7	3·1	4·4	1·4	6·4	3·2	5·3	2·3
5·4	3·9	1·3	0·4	5·6	3·0	4·5	1·5	6·5	3·0	5·5	1·8
5·1	3·5	1·4	0·3	5·8	2·7	4·1	1·0	7·7	3·8	6·7	2·2
5·7	3·8	1·7	0·3	6·2	2·2	4·5	1·5	7·7	2·6	6·9	2·3
5·1	3·8	1·5	0·3	5·6	2·5	3·9	1·1	6·0	2·2	5·0	1·5
5·4	3·4	1·7	0·2	5·9	3·2	4·8	1·8	6·9	3·2	5·7	2·3
5·1	3·7	1·5	0·4	6·1	2·8	4·0	1·3	5·6	2·8	4·9	2·0