Análise Multivariada e Aprendizado Não-Supervisionado
Aula 0: Motivação - Análise de Componentes Principais

Prof. Cibele Russo

cibele@icmc.usp.br

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

Análise Multivariada

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

Aprendizado estatístico

Alguns exemplos de aprendizado estatístico:

- Predizer se um paciente, hospitalizado por um ataque cardíaco, terá um segundo ataque cardíaco, a partir de medições clínicas, dietéticas e demográficas do paciente.
- Predizer o preço de uma ação daqui a seis meses, com base na performance da companhia e de dados econômicos.
- Identificar perfis de clientes a partir de dados do seu relacionamento com uma empresa
- ... outros.

Aprendizado supervisionado

Aprendizado supervisionado

Problemas em que existem um conjunto de **variáveis explicativas** (*inputs*, preditores, variáveis independentes) e de **variáveis respostas** (*outputs*, variáveis dependentes) são chamados de problemas de **aprendizado supervisionado**.

Em problemas de aprendizado supervisionado, os dados possuem **rótulos** (aprendem com um professor).

Aprendizado supervisionado

Algumas técnicas de aprendizado supervisionado:

- Modelos de regressão linear,
- Regressão logística,
- Análise fatorial,
- Análise discriminante linear (LDA).

Aprendizado não-supervisionado

Aprendizado não-supervisionado

Problemas em que se deseja estudar a variabilidade dos dados de forma multivariada, possivelmente com **redução de dimensionalidade**, sem necessariamente atribuir valores para uma variável resposta, são chamados de problema de **aprendizado não-supervisionado**.

Em problemas de aprendizado não-supervisionado, os dados não possuem **rótulos** (não aprendem com um professor).

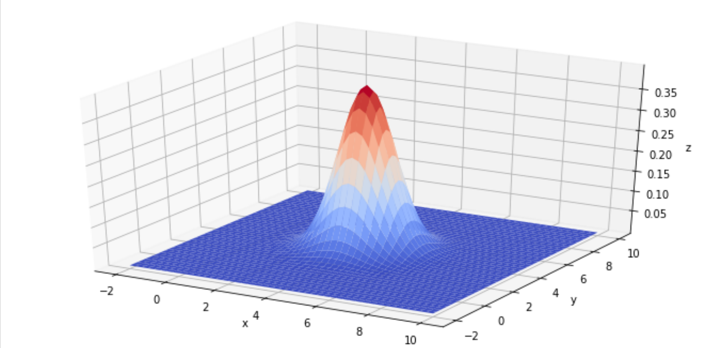
Aprendizado não-supervisionado

Algumas técnicas de aprendizado não-supervisionado:

- Análise de componentes principais (PCA),
- Análise de agrupamentos (Clustering): dendrograma, K-médias.
- Análise de correlações canônicas (CCA)
- Análise de correspondência (CA)

Análise multivariada

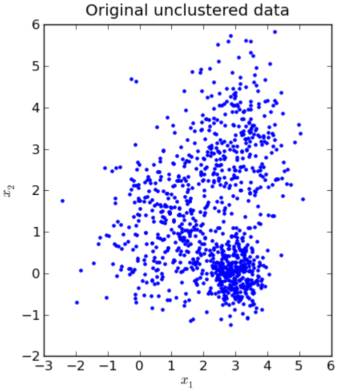
Distribuição normal bivariada

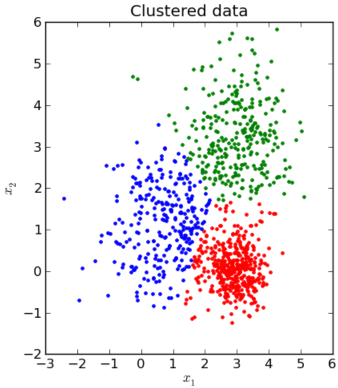


Fonte: https://towardsdatascience.com/a-python-tutorial-on-generating-and-plotting-a-3d-guassian-distribution-8c6ec6c41d03

Aprendizado não-supervisionado

Aprendizado não-supervisionado: exemplo (K-médias)

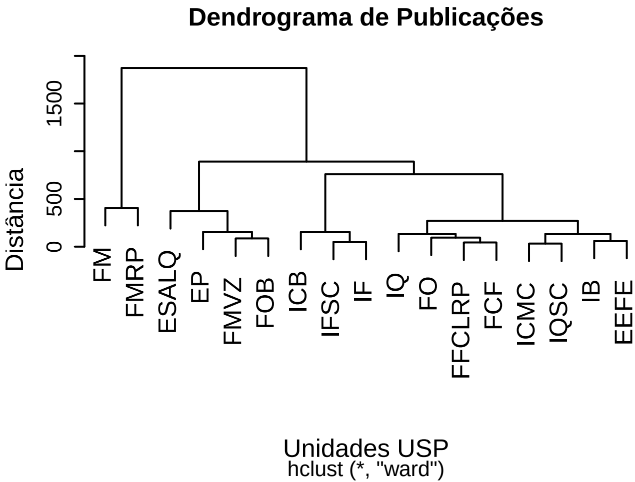




Fonte: https://mubaris.com/posts/kmeans-clustering/

Aprendizado não-supervisionado

Aprendizado não-supervisionado: exemplo (dendrograma)



Introdução - Dados multivariados

$\underset{\sim}{X}$ é um vetor aleatório de dimensão $p \times 1$, isto é,

$$\underset{\sim}{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix},$$

em que cada X_i é uma variável aleatória com média μ_i e variância σ_i^2, para $i = 1, \ldots, p$.

Introdução - Dados multivariados

A **covariância** e a **correlação** entre duas variáveis X_i e X_j, com $i, j = 1, \ldots, p$ e $i \neq j$ são dadas, respectivamente, por

$$\text{Cov}(X_i, X_j) = \sigma_{ij} \text{ e } \text{Cor}(X_i, X_j) = \rho_{ij}.$$

Obs: $\rho_{ij} = \dfrac{\sigma_{ij}}{\sqrt{\sigma_i^2}\sqrt{\sigma_j^2}}$.

Introdução - Dados multivariados

O **vetor de médias** (populacionais) de $\underset{\sim}{X}$ é $\underset{\sim}{\mu}_{p\times 1}$ e a matriz de variâncias e covariâncias (populacionais) de $\underset{\sim}{X}$ é $\Sigma_{p\times p}$, em que

$$\underset{\sim}{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \text{ e } \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ & \sigma_2^2 & \cdots & \sigma_{2p} \\ & & \ddots & \vdots \\ \text{sim.} & & & \sigma_p^2 \end{bmatrix}$$

Obs: Note que Σ é simétrica pois $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$, para $i, j = 1, \ldots, p$.

Introdução - Dados multivariados

Estamos particularmente interessados no caso em que as variáveis $X_1, \ldots, X_p$ estão correlacionadas, isto é, algumas (ou muitas) das covariâncias $\mathrm{Cov}(X_i, X_j), i, j = 1, \ldots, p$ e $i \neq j$ são não-nulas.

Quando isso ocorre, significa que existe **redundância entre dimensões**, e podemos ter interesse em **reduzir a dimensionalidade do problema**, construindo novas variáveis, não correlacionadas entre si, que sejam combinações lineares das variáveis originais.

Pode ser que poucas (k) novas variáveis expliquem grande parte da variabilidade existente nas (p) variáveis originais. Isso pode significar a **redução de custos** como tempo computacional e espaço para armazenamento de dados.

Introdução - Dados multivariados

Estamos particularmente interessados no caso em que as variáveis $X_1, \ldots, X_p$ estão correlacionadas, isto é, algumas (ou muitas) das covariâncias $\text{Cov}(X_i, X_j), i, j = 1, \ldots, p$ e $i \neq j$ são não-nulas.

Quando isso ocorre, significa que existe **redundância entre dimensões**, e podemos ter interesse em **reduzir a dimensionalidade do problema**, construindo novas variáveis, não correlacionadas entre si, que sejam combinações lineares das variáveis originais.

Pode ser que poucas (k) novas variáveis expliquem grande parte da variabilidade existente nas (p) variáveis originais. Isso pode significar a **redução de custos** como tempo computacional e espaço para armazenamento de dados.

Introdução - Dados multivariados

Estamos particularmente interessados no caso em que as variáveis $X_1, \ldots, X_p$ estão correlacionadas, isto é, algumas (ou muitas) das covariâncias $\text{Cov}(X_i, X_j), i, j = 1, \ldots, p$ e $i \neq j$ são não-nulas.

Quando isso ocorre, significa que existe **redundância entre dimensões**, e podemos ter interesse em **reduzir a dimensionalidade do problema**, construindo novas variáveis, não correlacionadas entre si, que sejam combinações lineares das variáveis originais.

Pode ser que poucas (k) novas variáveis expliquem grande parte da variabilidade existente nas (p) variáveis originais. Isso pode significar a **redução de custos** como tempo computacional e espaço para armazenamento de dados.

Análise de Componentes principais

Seja $\underset{\sim}{X}$ um vetor aleatório com vetor de médias $\underset{\sim}{\mu}$ e matriz de variâncias e covariâncias Σ.

Sejam $\lambda_1 \geq \ldots \geq \lambda_p$ os **autovalores** de Σ, com **autovetores** correspondentes $\underset{\sim}{e}_1, \ldots, \underset{\sim}{e}_p$, tais que

1. $\underset{\sim}{e}_i^\top \underset{\sim}{e}_j = 0$, para $i, j = 1, \ldots, p$ e $i \neq j$,
2. $\underset{\sim}{e}_i^\top \underset{\sim}{e}_i = 1$, para $i = 1, \ldots, p$,
3. $\Sigma \underset{\sim}{e}_i = \lambda_i \underset{\sim}{e}_i$, para $i = 1, \ldots, p$.

Considere a matriz ortogonal $O_{p \times p} = (\underset{\sim}{e}_1, \ldots, \underset{\sim}{e}_p)^\top$. Então

$\underset{\sim}{Y}_{p \times 1} = O^\top \underset{\sim}{X}$ é o **vetor de componentes principais** de Σ.

Análise de Componentes principais

Seja $\underset{\sim}{X}$ um vetor aleatório com vetor de médias $\underset{\sim}{\mu}$ e matriz de variâncias e covariâncias Σ.

Sejam $\lambda_1 \geq \ldots \geq \lambda_p$ os **autovalores** de Σ, com **autovetores** correspondentes $\underset{\sim}{e}_1, \ldots, \underset{\sim}{e}_p$, tais que

1. $\underset{\sim}{e}_i^\top \underset{\sim}{e}_j = 0$, para $i, j = 1, \ldots, p$ e $i \neq j$,
2. $\underset{\sim}{e}_i^\top \underset{\sim}{e}_i = 1$, para $i = 1, \ldots, p$,
3. $\Sigma \underset{\sim}{e}_i = \lambda_i \underset{\sim}{e}_i$, para $i = 1, \ldots, p$.

Considere a matriz ortogonal $O_{p \times p} = (\underset{\sim}{e}_1, \ldots, \underset{\sim}{e}_p)^\top$. Então

$\underset{\sim}{Y}_{p \times 1} = O^\top \underset{\sim}{X}$ é o **vetor de componentes principais** de Σ.

Análise de Componentes principais

Seja $\underset{\sim}{X}$ um vetor aleatório com vetor de médias $\underset{\sim}{\mu}$ e matriz de variâncias e covariâncias Σ.

Sejam $\lambda_1 \geq \ldots \geq \lambda_p$ os **autovalores** de Σ, com **autovetores** correspondentes $\underset{\sim}{e}_1, \ldots, \underset{\sim}{e}_p$, tais que

1. $\underset{\sim}{e}_i^\top \underset{\sim}{e}_j = 0$, para $i, j = 1, \ldots, p$ e $i \neq j$,
2. $\underset{\sim}{e}_i^\top \underset{\sim}{e}_i = 1$, para $i = 1, \ldots, p$,
3. $\Sigma \underset{\sim}{e}_i = \lambda_i \underset{\sim}{e}_i$, para $i = 1, \ldots, p$.

Considere a matriz ortogonal $O_{p \times p} = (\underset{\sim}{e}_1, \ldots, \underset{\sim}{e}_p)^\top$. Então

$\underset{\sim}{Y}_{p\times 1} = O^\top \underset{\sim}{X}$ é o **vetor de componentes principais** de Σ.

Análise de Componentes principais

Seja $\underset{\sim}{X}$ um vetor aleatório com vetor de médias $\underset{\sim}{\mu}$ e matriz de variâncias e covariâncias Σ.

Sejam $\lambda_1 \geq \ldots \geq \lambda_p$ os **autovalores** de Σ, com **autovetores** correspondentes $\underset{\sim}{e}_1, \ldots, \underset{\sim}{e}_p$, tais que

1. $\underset{\sim}{e}_i^\top \underset{\sim}{e}_j = 0$, para $i, j = 1, \ldots, p$ e $i \neq j$,
2. $\underset{\sim}{e}_i^\top \underset{\sim}{e}_i = 1$, para $i = 1, \ldots, p$,
3. $\Sigma \underset{\sim}{e}_i = \lambda_i \underset{\sim}{e}_i$, para $i = 1, \ldots, p$.

Considere a matriz ortogonal $O_{p\times p} = (\underset{\sim}{e}_1, \ldots, \underset{\sim}{e}_p)^\top$. Então

$\underset{\sim}{Y}_{p\times 1} = O^\top \underset{\sim}{X}$ é o **vetor de componentes principais** de Σ.

Componentes principais

Propriedades:

1. A j-ésima componente principal de Σ é dada por
$$Y_j = \underset{\sim}{e_j}^\top \underset{\sim}{X}.$$
2. $\mathrm{E}(Y_j) = \underset{\sim}{e_j}^\top \underset{\sim}{\mu}$.
3. $\mathrm{Var}(Y_j) = \underset{\sim}{e_j}^\top \Sigma \underset{\sim}{e_j} = \lambda_j$.
4. $\mathrm{Cov}(Y_i, Y_j) = \mathrm{Cor}(Y_i, Y_j) = 0$ para $i, j = 1, \ldots, p$ e $i \neq j$.
5. A proporção da variância total de $\underset{\sim}{X}$ que é explicada pela j-ésima componente principal é
$$\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}$$

Componentes Principais

Observações:

1. Com a matriz de variâncias e covariâncias amostrais, estima-se os componentes principais da matriz de variâncias e covariâncias populacionais.
2. O mesmo desenvolvimento pode ser feito para a matriz de correlações, com a vantagem de que as componentes principais são menos influenciadas pela magnitude das variâncias.

Componentes Principais

Observações:

1. Com a matriz de variâncias e covariâncias amostrais, estima-se os componentes principais da matriz de variâncias e covariâncias populacionais.

2. O mesmo desenvolvimento pode ser feito para a matriz de correlações, com a vantagem de que as componentes principais são menos influenciadas pela magnitude das variâncias.

Motivação 1

O conjunto de dados "Educação" (Fonte: Ipeadata) mostra dados educacionais por estado de acordo com a descrição a seguir.

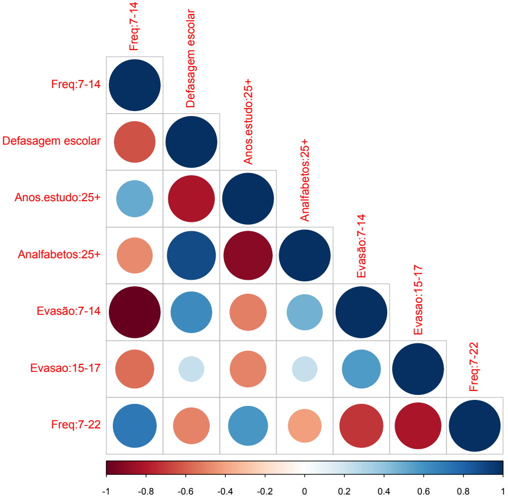
- Sigla (do estado)
- Estado
- X1: Frequência escolar de pessoas com 7 a 14 anos
- X2: Defasagem escolar em mais de 1 ano de atraso de pessoas com 7 a 14 anos (2000)
- X3: Anos de estudo de pessoas com 25 anos ou mais (2000)
- X4: Analfabetos com 25 anos ou mais (2000)
- X5: Evasão escolar de pessoas com 7 a 14 anos (2000)
- X6: Evasão escolar de pessoas com 15 a 17 anos (2000)
- X7: Frequência escolar de pessoas com 7 a 22 anos

Os comandos para análises se encontram no arquivo Motivação.ipynb

Matriz de correlações

	Freq:7-14	Def. escolar	Anos.est:25+	Analf.:25+	Ev:7-14	Ev:15-17	Freq:7-22
Freq:7-14	1.00	-0.64	0.51	-0.48	-1.00	-0.56	0.72
Defasagem escolar	-0.64	1.00	-0.82	0.89	0.64	0.24	-0.50
Anos.estudo:25+	0.51	-0.82	1.00	-0.90	-0.51	-0.50	0.59
Analfabetos:25+	-0.48	0.89	-0.90	1.00	0.48	0.23	-0.42
Evasão:7-14	-1.00	0.64	-0.51	0.48	1.00	0.56	-0.72
Evasao:15-17	-0.56	0.24	-0.50	0.23	0.56	1.00	-0.81
Freq:7-22	0.72	-0.50	0.59	-0.42	-0.72	-0.81	1.00

Gráfico de correlação



Perguntas

1. Deseja-se criar uma ordenação pela primeira componente principal da matriz de variâncias e covariâncias.
2. Quanto da variabilidade total dos dados é explicado pela primeira componente principai?

Motivação 2

Estudos mostram que grande parte de adultos e adolescentes norte-americanos usam regularmente substâncias psicoativas. Em um destes estudos (Huba et al. 1981, J. of Personality and Social Psychology), dados foram coletados de 1634 estudantes na área metropolitana de Los Angeles. Cada participante completou um questionário informando o número de vezes que cada item foi usado.

Os itens são os seguintes: cigarro, cerveja, vinho, licor, cocaína, tranquilizantes, medicamentos, heroína, maconha, haxixe, inalantes, alucinógenos e anfetaminas.

Ver arquivo Motivação_R.ipynb .

Motivação 2

Estudos mostram que grande parte de adultos e adolescentes norte-americanos usam regularmente substâncias psicoativas. Em um destes estudos (Huba et al. 1981, J. of Personality and Social Psychology), dados foram coletados de 1634 estudantes na área metropolitana de Los Angeles. Cada participante completou um questionário informando o número de vezes que cada item foi usado.

Os itens são os seguintes: cigarro, cerveja, vinho, licor, cocaína, tranquilizantes, medicamentos, heroína, maconha, haxixe, inalantes, alucinógenos e anfetaminas.
Ver arquivo Motivação_R.ipynb .