SCC0276 - Aprendizado de Máquina

Aula PCA

Profa. Dra. Roseli Aparecida Francelin Romero
SCC - ICMC - USP

2019

Sumário

Tratamento de Dados

- Permite reduzir a dimensão dos dados:
  $x_1, x_2, ..., x_n \; -- > \; t_1, t_2, ..., t_p \; ; \; p <= n$
- $t_1, t_2, ..., t_p$ são ortogonais (sem correlação)
- Mantendo a variancia dos dados

Tratamento de Dados

- Requer uma transformação dos dados
- Os dados devem estar com media $= 0$ e variancia $=1$

Sumário

Média e Variância

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Dados centrados na média e variancia igual a 1:

$$x_{ij} = \frac{(x_{ij} - \bar{x}_j)}{\sigma_j}$$

Sumário

PCA

- Determinar a matriz de covariancia dos dados
- Determinar os auto-valores e respectivos auto-vetores
- os coeficientes do auto-vetor correspondente ao maior auto-valor formarão os coeficientes da componente principal Z_1.
- os coeficientes do auto-vetor correspondente ao segundo maior auto-valor formarão os coeficientes da 2a. componente principal, Z_2.
- e assim, sucessivamente.

PCA

- **Objetivo**: Dadas n variáveis, deseja-se achar combinações lineares dessas, para produzir índices que sejam não correlacionados, de tal forma que: Índices Zi: componentes principais.

PCA

- i-ésima componente principal
- $Z_i = a_{i1}X_1 + a_{i2}X_2 + ... + a_{in}X_n$
- com a seguinte propriedade: $a_{i1}^2 + a_{i2}^2 + ... + a_{in}^2 = 1$
- $Z_1, Z_2, ..., Z_p$ são não correlacionados

PCA

- PCA: resume-se em encontrar os autovalores e autovetores da matriz C de covariância dos dados
- Supondo que: os autovalores da matriz C estejam ordenados da seguinte forma:

$$\lambda_1 >= \lambda_2 >= ... = \lambda_n$$

- Os auto-vetores associados:

$$a_1, a_2, ..., a_n$$

Sumário

Propriedades

- $$a_i^T a_j = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$
- a soma dos auto-valores correspondem ao traço da matriz de covariância:
$$\lambda_1 + \lambda_2 + ... + \lambda_n = c_{11} + c_{22} + ... + c_{nn}$$
- $$var(Z_i) = \lambda_i$$

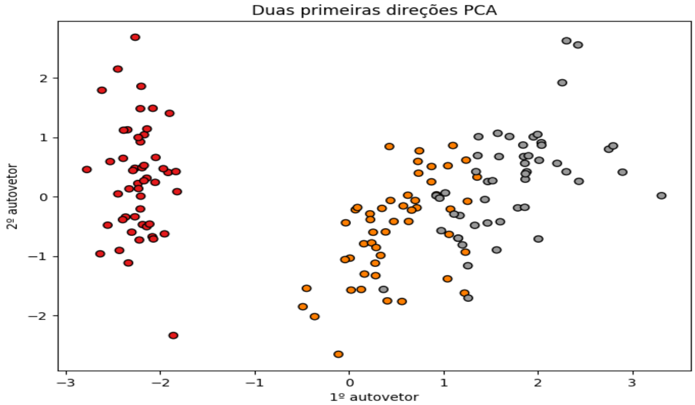

Figura 1: Duas componentes principais do iris.dat

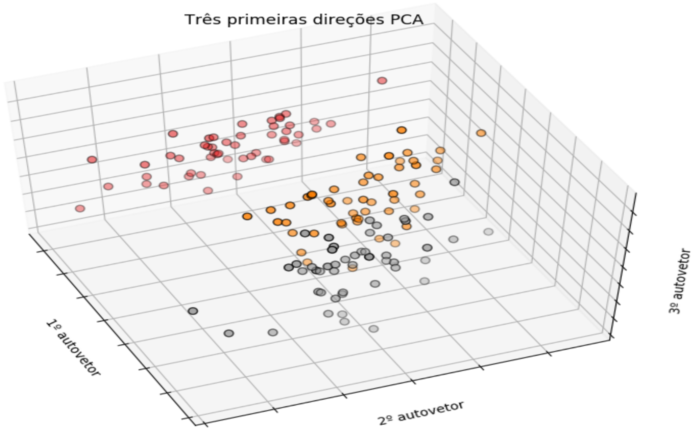

Figura 2: Duas componentes principais do iris.datr

Exercicio

- Encontrar as 2 componentes principais para o conjunto iris.dat. Verificar que as duas juntas representam 95.80% da variancia dos dados.