Tratamento nos dados

Redução de Dimensionalidade

Transformação dos dados

- Dados centrados na média e variância 1
  - $xm = 1/n \Sigma_{i=1}^{n} x_i$
  - $x_i = (x_i - xm)/\sigma$ (centrados na média)
- Escalamento pela variância (quando temos muito dominante em rel. as demais)
  - $var_j = 1/ (n-1)\Sigma_{i=1}^{n} (x_{ij} - xm_j)^2$
  - $x_{ij\ (sv)} = x_{ij} / var$ desvio= σ=sqrt(var)

Transformação nos dados

- Pre-processamento dos dados
  - log 10 (deve ser usado quando existem grandes disparidades de magnitude nos dados)
  - Normalização

    $|| x || = \Sigma x_{ij}^2$ para i, j = 1, ..., ; $x_{ij\ (norm)} = x_{ij} / || x ||$

    não aconselhável para dados com menos de 10 variáveis
  - Transformação de dados do intervalo [a,b] para [-1,1]:

    $y = (2x – b –a) / (b –a)$

Analise de Componentes Principais - PCA

- Reduz o no. de variáveis
  - $x_1, x_2, …, x_p$ ==> $t_1, t_2,…., t_m$ (m<p)
  - os erros propagados nos valores de w são menores, porque $t_1, t_2,…., t_m$ são ORTOGONAIS (sem correlação)

Covariância

- Como é a matriz de covariância?
- Medir c/ var. sobre a média de seus n valores $x_i = 1/n \Sigma_{i=1} x_{ij}$
- $c_{ij} = 1/n \Sigma_{k=1} (x_{ik} - x_i) (x_{jk} - x_j)$

onde c_{ij} é a covariância de x_i e x_j.

- C (pxp)
- A covariância é por vezes chamada de medida de dependência linear entre as duas variáveis aleatórias.

Covariância

$$\begin{pmatrix} 138.32 & -102.12 & -36.20 & -0.94 & -0.14 \\ & 79.74 & 22.38 & 153 & 0.11 \\ & & 13.82 & -0.58 & 0.02 \\ & & & 0.64 & 0.03 \\ & & & & 0.26 \end{pmatrix}$$

PCA

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

PCA

i-ésima componente principal

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + ... + a_{ip}X_p$$

Com restrição: $a_{i1}^2 + a_{i2}^2 + \cdots + a_{ip}^2 = 1$

$Z_1, Z_2, ..., Z_{i-1}, Z_i$ Não correlacionados

PCA

- PCA: resume-se em encontrar os autovalores e autovetores da matriz *C* de covariância dos dados
- C= $\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pp} \end{bmatrix}$

PCA

- Supondo que: os autovalores da matriz *C* estejam ordenados da seguinte forma:

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_j \geq ... \geq \lambda_p$$

- Os auto-vetores associados:

$$a_1, a_2, \ldots, a_j, \ldots, a_p$$

PCA

- Propriedades
- Para $a_i^T a_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + ... + a_{ip}X_p$$

$a_{i1}, a_{i2}, \ldots, a_{ip}$

- são os elementos do *i*-ésimo autovetor correspondente

PCA

- a soma dos auto-valores corresponde ao **traço da matriz covariância C**

$$\lambda_1 + \lambda_2 + ... + \lambda_p = c_{11} + c_{22} + ... + c_{pp}$$

- **Var(Z_i) = λ_i**

Teste com o conj. Iris

- Os dados da Íris coletados por Fisher em 1936 (*Understanding the Data*, 2005), existem quatro (p=4) medidas em cm das pétalas e sépalas de três tipos (g=3) de flores Íris (*setosa*, *versicolor* e *virginica*), com cinqüenta observações de cada classe (cada tipo de flor) – n=150;

Exemplo

Íris Setosa				Íris Versicolor				Íris Virginica			
X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
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
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.1	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.1	1.5	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

Conjunto de dados Iris

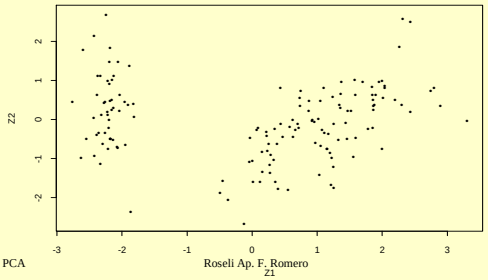
Exemplo

		Autovetores (coeficientes)			
Componente	Autovalor	X_1	X_2	X_3	X_4
1	2.91082	0.522371	-0.263356	0.581254	0.565611
2	0.92122	0.372320	0.925556	0.021094	0.065417
3	0.14735	-0.721015	0.242033	0.140889	0.633804
4	0.02061	-0.261998	0.124137	0.801155	-0.523543

PCA

- Conclusões:
- Z1 é responsável por 72.77% do total da variância.
- Z2 é responsável por 23.03% do total da variância.
- Z3 é responsável por 3.68% do total da variância.
- Z4 é responsável por 0.52% do total da variância.

PCA



PCA – Conceitos Básicos

- Polinômio Característico
  $P(\lambda) = \det (A - \lambda I) = 0$

  porque $Av = \lambda v \Rightarrow (A - \lambda I)v = 0 \Rightarrow$
  se v não nulo $\Rightarrow \det(A - \lambda I) = 0$

Exemplo:

- Determinar os auto-valores e auto-vetores de:

$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$

$\det(A - \lambda I) = \begin{pmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{pmatrix} = (2-\lambda)^2 - 3 = 0$

- Sol: $\lambda_1=3$ e $\lambda_2=1$

Exemplo

- Determinar o auto-vetor assoc. a $\lambda_1=3$

$-v_1 + v_2 = 0$

$v_1 - v_2 = 0$

Sol: $v_1 = v_2$

Portanto $v = [1\ 1]^T$

Analogamente, $v = [1\ -1]^T$ para $\lambda_2=1$

Operações em PCA

- Det. dos auto-valores e auto-vetores:
  calculo do determinante e
  sol. de um sistema de eq. lineares.

Exercicio

- Implementar a técnica PCA (use o Matlab ou algum pacote estatístico, ou 1 ling. de progr.) e verifique as 2 comp. Principais obtidas para o conj. iris

Reconstrução dos dados originais

$$Z = [\,Z_1, Z_2, \ldots, Z_p\,]^T$$

$$= [\,X^T a_1, X^T a_2, \ldots, X^T a_{p-1}\,]^T$$

$= A^T X$

$X = A\,Z = \sum_{i=1}^{p} Z_i a_i$ $\quad A^T = A^{-1}$

Redução da Dimensionalidade

- Sejam $\lambda_1, \lambda_2, \ldots, \lambda_m$
- os m auto-valores da matriz C
- Entao X' ~ X onde:

$$X' = \sum_{i=1}^{m} Z_i a_i \qquad m < p$$

- o erro: e = X - X' onde:

$$e = \sum_{i=m+1}^{p} Z_i a_i$$

- O vetor de erro e é ortogonal ao vetor X', que aproxima X.

  Verifique isto!!!

  A equação: $e^T X' = 0$: princípio da ortogonalidade
- Existe uma rede neural que implementa PCA, proposta por RUBN-89 (vamos ver!!!)

•Algoritmo da PCA

1. Padronize os dados correspondentes às variáveis para que estes tenham média igual a 0 e variância igual a 1.
2. Calcule a matriz de covariância *C*.
3. Encontre os autovalores da matriz *C* e seus correspondentes auto-vetores. Os coeficientes da *i*-ésima componente principal são dados pelo auto-vetor associado ao i-ésimo auto-valor

4. Descarte as componentes que acumulem uma pequena proporção da variação dos dados. Por exemplo, se os dados originais tiverem 20 variáveis e as três primeiras componentes principais forem responsáveis por 90% do total da variância, as outras 17 componentes principais podem ser ignoradas.

Algoritmo PCA - Matlab

```
%%%%%%%%%%%%%%%%%%
%  Componentes Principais     %
%     PCA  - Roseli Romero    %
%%%%%%%%%%%%%%%%%%
% carregamento doa arquivo iris.dat
load iris.txt
% verificando a dimensao do conjunto
[n,p] = size(iris);
% tirando a primeira coluna da matriz de dados
iris(:,1)=[ ];
p= p-1;
X = iris;   % matriz X contem os dados
% centrando os dados na media 0
S = std(X);  % S armazena os desvios-padrao de cada coluna de X
M = mean(X); % M armazena as medias de cada coluna de X
X = X - ones(n,1) * M;
% transformando os dados para ter variancia 1
X = X./(ones(n,1)*S);
% calculando a matriz de covariancia C
C = (X'*X)/n;
% auto-valores (A) e auto-vetores (V) da matriz de covariancia
[V,A] = eig(C);
A = diag(A);
```



```
% ordenando auto-valores e auto-vetores por ordem crescente de auto-valores
V = V';   %coloca auto-vetores nas linhas de V
A= [A,V]; % concatena A e V (cada linha de A contem um auto-valor e auto-vetor)
A=sortrows(A,1);  %ordena os auto-valores em ordem crescente
A=A(:,1) %separa auto-valores em A e imprimi na tela
V       %imprimi V na tela
%calculando as componentes principais
Z=[];
for i=1:p
  Vt = V(p-i+1,:); %pega os auto-vetores em ordem decrescente dos auto-valores
  Z(:,i) = X * Vt';  % otem-se a i-esima componente principal
end
% salvando as componentes principais em um arquivo
save('componentes_iris.txt')
Z1=Z(:, 1:2)
L= Z1(1:50,:);
L1=Z1(51:100,:);
L2=Z1(101:150,:);
% fazendo o grafico das duas primeiras componentes
%plot(Z1,Z2,'-')
%fazendo o grafico das tres classes separadas
plot(L(:,1),L(:,2),'.',L1(:,1),L1(:,2),'+',L2(:,1),L2(:,2),'*')
```


- Implementar a tecnica PCA (use o Matlab ou algum pacote estatistico, ou 1 ling. de progr.) e verifique as 4 comp. Principais obtidas para o conj. iris

- Dada a matriz:

$$X = \begin{pmatrix} 3 & 0 & 1 & 5 \\ 4 & 7 & 4 & 3 \\ 3 & 2 & 6 & 2 \\ 5 & 3 & 6 & 4 \end{pmatrix}$$

- transformar essa matriz:
  dados centrados pela média
  escalamento pela variância