PCA 降维以及维数的确定

作者：vcjmhg
原文链接：https://ld246.com/article/1592444387242
来源网站：链滴

概述

PCA（principal components analysis）即主成分分析技术，又称为主分量分析，旨在利用降维的想，把多个指标转换为少数的几个综合指标。

主成分分析是一种简化数据集的技术，它是一个线性变换。这个线性变化把数据变换到一个新的坐标统中，使得任何数据投影的第一大方差在第一个坐标上（称为第一主成分），第二个大的方差在第二坐标上（称为第二主成分），以此类推。主成分分析经常用于减少数据集的维数，同时保持数据集的方差贡献最大的特征。这是通过保留低阶主成分，忽略高阶主成分做到的。这样低阶成分往往能够保住数据的最重要方面。

PCA的原理就是将原来的样本数据投影到一个新的空间中。其中就是将原来的样本数据空间经过坐标换矩阵变到新空间坐标下，新空间坐标由其原数据样本中不同维度之间的协方差矩阵中几个最大特征对应的前几个特征向量组成。较小的特征值向量作为非主要成分去掉，从而可以达到提取主要成分代原始数据的特征，降低数据复杂度的目的。

算法步骤

1. 将n次采样的m维数据组织成矩阵形式 $X\in R^{n\times m}$。具体形式如下示：

$$\left(\begin{matrix}\begin{matrix}x_{11}&x_{12}\\x_{21}&x_{22}\\\end{matrix}&\begin{matrix}\dots&x_{1m}\\\cdots&x_{2m}\\\end{matrix}\\\begin{matrix}\vdots&\vdots\\x_{n1}&x_{n2}\\\end{matrix}&\begin{matrix}\ddots&\vdots\\\cdots&x_{nm}\\\end{matrix}\\\end{matrix}\right)$$

2. 将样本矩阵 X的每一列零均值化得新矩阵 $X^{\prime}$。

$$\boldsymbol{x}_{i} \leftarrow \boldsymbol{x}_{i}-\frac{1}{m} \sum_{i=1}^{m} \boldsymbol{x}_{i}$$

3. 计算其样本数据维度之间的相关度，此处使用协方差矩阵 C：

$$cov=\frac{1}{m}X^\prime{X^\prime}^T$$

4. 计算协方差矩阵 C的特征值及其对应的特征向量，并特征值按照从大到小排列。

$$\left(\lambda_1,\lambda_2,\cdots,\lambda_t\right)=\left(\begin{matrix}\begin{matrix}p_{11}&_{12}\\p_{21}&p_{22}\\\end{matrix}&\begin{matrix}\cdots&p_{1t}\\\cdots&p_{2t}\\\end{matrix\\\begin{matrix}\vdots&\vdots\\p_{n1}&p_{n2}\\\end{matrix}&\begin{matrix}\ddots&\vdots\\\cdots&p_{nt}\\\end{matrix}\\\end{matrix}\right)=\left(\boldsymbol{P}_{1}, \boldsymbol{P}_{2}, \ldots, \boldsymbol{P}_{i}\right)\ ,\ (其中\lambda_1>\lambda_2>\cdots>\lambda_t)$$

5. 根据降维要求，比如此处降到 k维，取其前个k向量组成降维矩阵P，如下所示：

$$P=\left(\boldsymbol{P}_{1}, \boldsymbol{P}_{2}, \ldots, \boldsymbol{P}_{k}\right)^T ,\ P\in R^{\times n}$$

6. 通过变换矩阵P对原样本数据 X进行坐标变换，从而达到数据降维与主成分提取的目的。

$$Y=X\bullet P\ ,\ Y\in R^{k\times m}$$

重建误差的计算

在投影完成之后，需要对投影的误差进行重建，从而计算数据降维之后信息的损失，一般来说通过以

公式来计算。

{error}_1=\frac{1}{k}\sum_{i=1}^{k}{||x^{\left(i\right)}}-x_{approx}^{\left(i\right)}||^2

{error}_2=\frac{1}{m}\sum_{i=1}^{m}{||x^{\left(i\right)}}||^2

其中：

- mm个样本表示为(x^{\left(1\right)},x^{\left(2\right)},\cdots,x^{\left(m\right)})(x^{\left(1\right),x^{\left(2\right)},\cdots,x^{\left(m\right)})
- 对应投影后的数据表示为 (x_{approx}^{\left(1\right)},x_{approx}^{\left(2\right)},\cdots,x_{approx}^{\left(m\right)})(x_{approx}^{\left(1\right)},x_{approx}^{\left(2\right)},\cdots,x_{approx}^{\left(m\right)})。

则其比率\eta\eta为

\eta=\frac{{error}_1}{{error}_2}

通过\eta\eta来衡量数据降维之后信息的损失。

算法描述

进而我们总结出算法描述如下：

输入： 样本集D=\left\{\boldsymbol{x}_{1}, \boldsymbol{x}_{2}, \ldots, \boldsymbol{x}_{m}\right}D=\left\{\boldsymbol{x}_{1}, \boldsymbol{x}_{2}, \ldots, \boldsymbol{x}_{m}\right\}；

低维空间维数

kk

过程：

1. 对所有样本进行零均值化： \boldsymbol{x}_{i} \leftarrow \boldsymbol{x}_{i}-\frac{1}{m} \su_{i=1}^{m} \boldsymbol{x}_{i}\boldsymbol{x}_{i} \leftarrow \boldsymbol{x}_{i}-\frac{1}{m} \sum{i=1}^{m} \boldsymbol{x}_{i}；
2. 计算样本的协方差矩阵 \mathbf{X X}^{\mathrm{T}}\mathbf{X X}^{\mathrm{T}}；
3. 对协方差矩阵 \mathbf{X X}^{\mathrm{T}}\mathbf{X X}^{\mathrm{T}}做特征值分解；
4. 取最大的 kk个特征值所对应的特征向量\left(\boldsymbol{P}_{1}, \boldsymbol{P}_{2}, \ldots, \oldsymbol{P}_{k}\right)\left(\boldsymbol{P}_{1}, \boldsymbol{P}_{2}, \ldots, \boldsymbol{P}_{k} right)；
5. 进行矩阵变换 Y=P\bullet X\ ,\ Y\in R^{k\times m}Y=P\bullet X\ ,\ Y\in R^{k\times m}

输出： 变换后的矩阵Y=X\bullet P\ ,\ Y\in R^{k\times m}Y=X\bullet P\ ,\ Y\in R^{k\times m}

算法实现

选用的数据集

使用数据集为：Imported Analog EMG – Voltage下的EMG1、EMG2、...、EMG8部分的数据

实验代码展示

```
fileName = 'c:\Users\Administrator\Desktop\机器学习作业\PCA\pcaData1.csv';
X = csvread(fileName);
m = size(X,1);
meanLine = mean(X,2);
R = size(X ,2);
%对原始数据做均值化处理，每一列都减去均值
A = [];
for i = 1:R
    temp = X(:,i) - meanLine;
    A = [A temp];
end
%求其协方差矩阵
C = A'*A/R;
%求协方差矩阵的特征值及其特征向量
[U,S,V] = svd(C);
%设置降维的维度数k，从1维计算到R-1维
k=8;
%计算投影后的样本数据Y
P=[];
for x = 1:k
    P = [P U(:,x)];
end
Y = X*P;
%计算数据重建误差以及比率
err1 = 0;
%获取样本X重建后的矩阵XR
XR= Y * pinv(P);
for i = 1:m
    err1 = norm(X(i,:)-XR(i,:))+err1;
end
%计算数据方差
err2 = 0;
for i=1:m
    err2 = norm(X(i,:))+err2;
end
eta = err1/err2
```

结果展示与分析

通过计算我们发现对应的特征值以及其对应的投影方向如下：

\lambda_1\lambda_1=1.8493对应的投影方向为(-0.0164,0.0300,-0.2376,0.4247,-0.6717,0.2356,-0.2196,0.4551)(-0.0164,0.0300,-0.2376,0.4247,-0.6717,0.2356,-0.2196,0.4551)

\lambda_2\lambda_2=1.3836对应的投影方向为(0.0910,0.1724,-0.0097,-0.8267,-0.1464,0.3599,0.0025,0.3570)(0.0910,0.1724,-0.0097,-0.8267,-0.1464,0.3599,0.0025,0.3570)

\lambda_3\lambda_3=0.5480对应的投影方向为(-0.1396,-0.4457,-0.1668,0.0870,0.2812,0.7696,-0.1742,-0.2115)(-0.1396,-0.4457,-0.1668,0.0870,0.2812,0.7696,-0.1742,-0.2115)

\lambda_4\lambda_4=0.4135对应的投影方向为(0.0622,0.1782,0.3136,-0.0080,-0.5387,0.2841,0.300,-0.6214)(0.0622,0.1782,0.3136,-0.0080,-0.5387,0.2841,0.3300,-0.6214)

\lambda_5\lambda_5=0.3218对应的投影方向为(0.2126,-0.7813,0.3136,-0.0080,-0.5387,0.2841,03300,-0.6214)(0.2126,-0.7813,0.3136,-0.0080,-0.5387,0.2841,0.3300,-0.6214)

\lambda_6\lambda_6=0.1322对应的投影方向为(-0.0959,0.0340,-0.6943,0.0068,0.0269,0.0042,0.119,0.0064)(-0.0959,0.0340,-0.6943,0.0068,0.0269,0.0042,0.7119,0.0064)

\lambda_7\lambda_7=0.0620对应的投影方向为(0.8881,-0.0497,-0.3407,-0.0198,-0.0103,-0.0424-0.2075,-0.2176)(0.8881,-0.0497,-0.3407,-0.0198,-0.0103,-0.0424,-0.2075,-0.2176)

\lambda_8=9.5959\times 10^{-17}\lambda_8=9.5959\times 10^{-17}对应的投影方向为(0.3536,.3536,0.3536,0.3536,0.3536,0.3536,0.3536,0.3536)(0.3536,0.3536,0.3536,0.3536,0.3536,0.3536,03536,0.3536)

k取不同值时对应的误差比率如下所示：

k的取值	**数据重建误差eat**
1	0.8265
2	0.7105
3	0.6499
4	0.5940
5	0.5521
6	0.5294
7	0.5162

参考

1. PCA主成分数量（降维维度）选择

2. Imported Analog EMG – Voltage下的EMG1、EMG2、…、EMG8部分进行PCA/NMF降维