云南大学数学与统计学院
《数据挖掘与决策支持实验》上机实践报告

课程名称： 数据挖掘与决策支持实验	**年级：** 2015 级	**上机实践成绩：**
指导教师： 彭程	**姓名：** 刘鹏	**专业：** 信息与计算科学
上机实践名称： 实现基于主成分分析的特征提取	**学号：** 20151910042	**上机实践日期：** 2018-07-05
上机实践编号： 03	**组号：**	

一、实验目的

学习使用 R 语言进行数据离散化。

二、实验内容

初始数据集为 Iris 鸢尾花数据。实现基于主成分分析的特征提取。

三、实验平台

Windows 10 Pro 1803；
Microsoft© Visual Studio 2017 Enterprise。
Version 1.1.442 – © 2009-2018 RStudio, Inc.

四、算法设计

在多元统计分析中，主成分分析（英语：Principal components analysis，PCA）是一种分析、简化数据集的技术。主成分分析经常用于减少数据集的维数，同时保持数据集中的对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留住数据的最重要方面。但是，这也不是一定的，要视具体应用而定。由于主成分分析依赖所给数据，所以数据的准确性对分析结果影响很大。

主成分分析由卡尔·皮尔逊于 1901 年发明，用于分析数据及建立数理模型。其方法主要是通过对协方差矩阵进行特征分解，以得出数据的主成分（即特征向量）与它们的权值（即特征值。PCA 是最简单的以特征量分析多元统计分布的方法。其结果可以理解为对原数据中的方差做出解释：哪一个方向上的数据值对方差的影响最大？换而言之，PCA 提供了一种降低数据维度的有效办法；如果分析者在原数据中除掉最小的特征值所对应的成分，那么所得的低维度数据必定是最优化的（也即，这样降低维度必定是失去讯息最少的方法）。主成分分析在分析复杂数据时尤为有用，比如人脸识别。

PCA 是最简单的以特征量分析多元统计分布的方法。通常情况下，这种运算可以被看作是揭露数据的内部结构，从而更好的解释数据的变量的方法。如果一个多元数据集能够在一个高维数据空间坐标系中被

显现出来，那么 PCA 就能够提供一幅比较低维度的图像，这幅图像即为在讯息最多的点上原对象的一个‘投影’。这样就可以利用少量的主成分使得数据的维度降低了。

PCA 跟因子分析密切相关，并且已经有很多混合这两种分析的统计包。而真实要素分析则是假定底层结构，求得微小差异矩阵的特征向量。

PCA 的数学定义是：一个正交化线性变换，把数据变换到一个新的坐标系统中，使得这一数据的任何投影的第一大方差在第一个坐标（称为第一主成分）上，第二大方差在第二个坐标（第二主成分）上，以此类推。

定义一个$n \times m$的矩阵，$\boldsymbol{X}^T$为去平均值（以平均值为中心移动到远点）的数据，其行为数据样本，列为数据类别。则$\boldsymbol{X}$的奇异值分解为$\boldsymbol{X} = \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{V}^T$，其中$n \times m$矩阵$\boldsymbol{W}$是$\boldsymbol{X}\boldsymbol{X}^T$的本征矢量矩阵，$\Sigma$为$n \times m$的非负矩形对角矩阵，$\boldsymbol{V}$是n × n的本征矢量矩阵。据此：

$$\begin{aligned}\boldsymbol{Y}^{\mathrm{T}} &= \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W} \\ &= \boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} \\ &= \boldsymbol{V}\boldsymbol{\Sigma}^T\end{aligned}$$

当$m < n - 1$时，$\boldsymbol{V}$在通常情况下不是唯一定义的，而$\boldsymbol{Y}$则是唯一定义的。$\boldsymbol{W}$是一个正交矩阵，$\boldsymbol{Y}^T\boldsymbol{W}^T = \boldsymbol{X}^T$，且$\boldsymbol{Y}^{\mathrm{T}}$的第一列由第一主成分组成，第二列由第二主成分组成，依此类推。

为了得到一种降低数据维度的有效办法，我们可以利用$\mathbf{W}_L$把$\boldsymbol{X}$映射到一个只应用前面L个向量的低维空间中去：

$$\boldsymbol{Y} = \boldsymbol{W}_{\mathbf{L}}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{\Sigma}_{\mathrm{L}}\boldsymbol{V}^{\mathrm{T}}$$

其中$\boldsymbol{\Sigma}_L = \boldsymbol{I}_{L\times m}$，且$\boldsymbol{I}_{L\times m}$为$L \times m$的单位矩阵。

X 的单向量矩阵 W 相当于协方差矩阵的本征矢量 C = X XT,

$$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} = \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{\Sigma}\mathbf{W}^{\mathrm{T}}$$

在欧几里得空间给定一组点数，第一主成分对应于通过多维空间平均点的一条线，同时保证各个点到这条直线距离的平方和最小。去除掉第一主成分后，用同样的方法得到第二主成分。依此类推。在$\boldsymbol{\Sigma}$中的奇异值均为矩阵$\boldsymbol{X}\boldsymbol{X}^T$的本征值的平方根。每一个本征值都与跟它们相关的方差是成正比的，而且所有本征值的总和等于所有点到它们的多维空间平均点距离的平方和。PCA 提供了一种降低维度的有效办法，本质上，它利用正交变换将围绕平均点的点集中尽可能多的变量投影到第一维中去，因此，降低维度必定是失去讯息最少的方法。PCA 具有保持子空间拥有最大方差的最优正交变换的特性。然而，当与离散余弦变换相比时，它需要更大的计算需求代价。非线性降维技术相对于 PCA 来说则需要更高的计算要求。

PCA 对变量的缩放很敏感。如果我们只有两个变量，而且它们具有相同的样本方差，并且成正相关，那么 PCA 将涉及两个变量的主成分的旋转。但是，如果把第一个变量的所有值都乘以 100，那么第一主成分就几乎和这个变量一样，另一个变量只提供了很小的贡献，第二主成分也将和第二个原始变量几乎一致。这就意味着当不同的变量代表不同的单位（如温度和质量）时，PCA 是一种比较武断的分析方法。但是在 Pearson 的题为 "On Lines and Planes of Closest Fit to Systems of Points in Space"的原始文件里，是假设在欧几里得空间里不考虑这些。一种使 PCA 不那么武断的方法是使用变量缩放以得到单位方差。

五、程序代码

5.1 程序描述

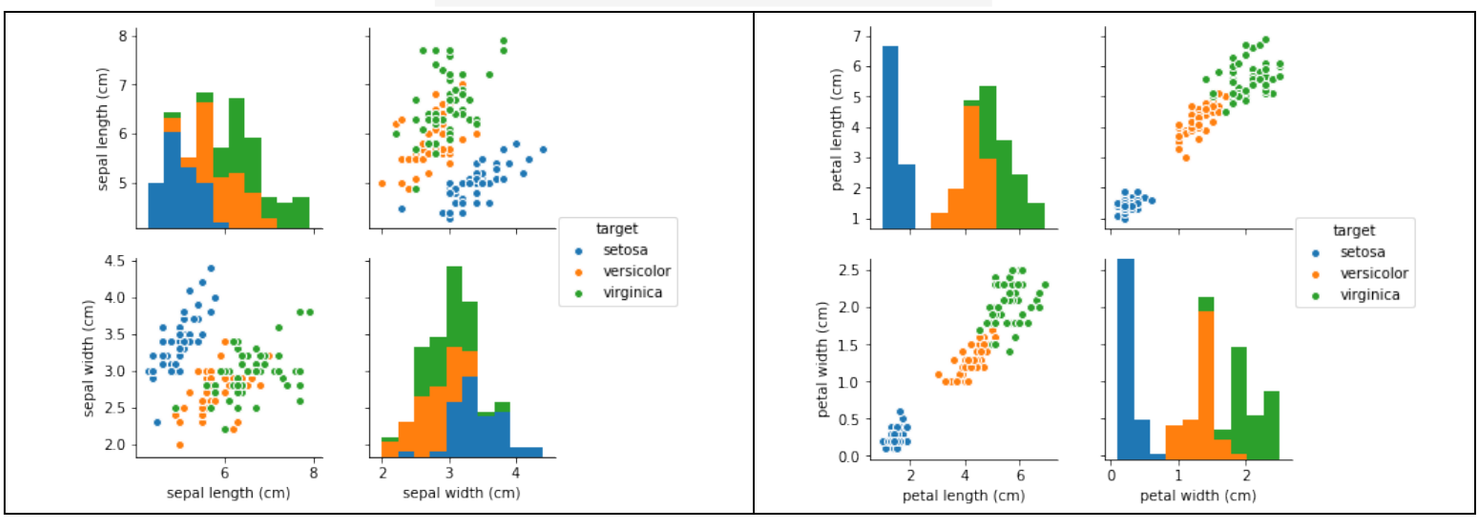
5.2 程序代码

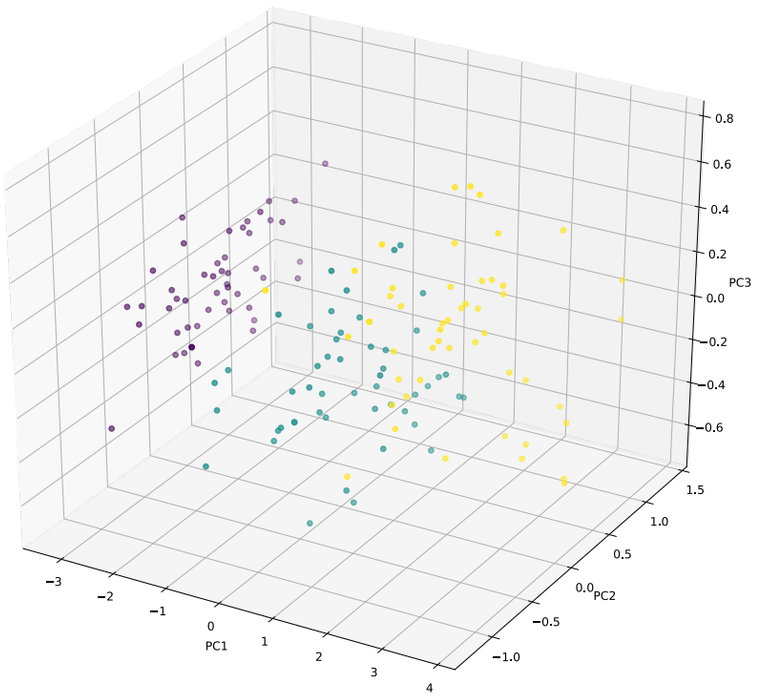
```
from sklearn import datasets
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
iris = datasets.load_iris()

data1=pd.DataFrame(np.concatenate((iris.data,iris.target.reshape(150,1)),axis=1),columns=np.append(iris.feature_names,'target'))
data=pd.DataFrame(np.concatenate((iris.data,np.repeat(iris.target_names,50).reshape(150,1)),axis=1), columns=np.append(iris.feature_names,'target'))
data=data.apply(pd.to_numeric,errors='ignore')

sns.pairplot(data.iloc[:,[0,1,4]],hue='target')
sns.pairplot(data.iloc[:,2:5],hue='target')

plt.scatter(data1.iloc[:,0],data1.iloc[:,1],c=data1.target)
plt.xlabel('sepal length (cm)')
plt.ylabel('sepal width (cm)')

from sklearn.decomposition import PCA
x_reduced = PCA(n_components=3).fit_transform(data.iloc[:,:4])
x_reduced

fig=plt.figure()
ax=Axes3D(fig)
ax.scatter(x_reduced[:,0],x_reduced[:,1],x_reduced[:,2],c=data1.iloc[:,4])
ax.set_xlabel('PC1')
ax.set_ylabel('PC2')
ax.set_zlabel('PC3')
```

```
29
30  PC=PCA(n_components=4).fit(data.iloc[:,:4])
31  PC.explained_variance_ratio_
```

程序代码 1

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	0.0
1	4.9	3.0	1.4	0.2	0.0
2	4.7	3.2	1.3	0.2	0.0
3	4.6	3.1	1.5	0.2	0.0
4	5.0	3.6	1.4	0.2	0.0
5	5.4	3.9	1.7	0.4	0.0
6	4.6	3.4	1.4	0.3	0.0
7	5.0	3.4	1.5	0.2	0.0
8	4.4	2.9	1.4	0.2	0.0
9	4.9	3.1	1.5	0.1	0.0
10	5.4	3.7	1.5	0.2	0.0
11	4.8	3.4	1.6	0.2	0.0
12	4.8	3.0	1.4	0.1	0.0
13	4.3	3.0	1.1	0.1	0.0
14	5.8	4.0	1.2	0.2	0.0
15	5.7	4.4	1.5	0.4	0.0
16	5.4	3.9	1.3	0.4	0.0
17	5.1	3.5	1.4	0.3	0.0
18	5.7	3.8	1.7	0.3	0.0
19	5.1	3.8	1.5	0.3	0.0
20	5.4	3.4	1.7	0.2	0.0
21	5.1	3.7	1.5	0.4	0.0
22	4.6	3.6	1.0	0.2	0.0
23	5.1	3.3	1.7	0.5	0.0
24	4.8	3.4	1.9	0.2	0.0
25	5.0	3.0	1.6	0.2	0.0
26	5.0	3.4	1.6	0.4	0.0
27	5.2	3.5	1.5	0.2	0.0
28	5.2	3.4	1.4	0.2	0.0
29	4.7	3.2	1.6	0.2	0.0
...	...	...	...	...	...




PC1
PC2
PC3
−3
−2
−1
0
1
2
3
4
−1.0
−0.5
0.0
0.5
1.0
1.5
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8

七、实验体会

八、参考文献

[1] https://zh.wikipedia.org/wiki/%E4%B8%BB%E6%88%90%E5%88%86%E5%88%86%E6%9E%90