实验一 PCA算法

**学号：20011453**

**姓名：罗仕松**

1. 实验目的

掌握PCA算法的原理

2. 实验要求

通过本实验应达到如下要求：

（1）理解数据降维过程

（2）熟练使用Python或其他工具实现PCA算法

3. 实验器材

（1）计算机一台

（2）Python或其他编程工具

4. 实验内容

**（1）阐述PCA降维过程：**

设有m条n维数据。

1）将原始数据按列组成n行m列矩阵X

2）将X的每一行（代表一个属性字段）进行零均值化，即减去这一行的均值

3）求出协方差矩阵

4）求出协方差矩阵的特征值及对应的特征向量r

5）将特征向量按对应特征值大小从上到下按行排列成矩阵，取前k行组成矩阵P

6）即为降维到k维后的数据

（2）数据集：可以使用sklearn库中的iris数据集，其中每个样本有4个特征参数，分别为花萼长度，花萼宽度，花瓣长度，花瓣宽度4个属性。（数据集不限）。

**数据集选取为：前150行**

**加载数据集代码：**iris = load\_iris()

（3）方法一：

（a）**数据标准化（去均值）代码：**mat = X  
  
# 测试数据的简单转换  
Mat = np.array(mat, dtype='float64')  
print('在PCA转换之前，数据是:\n', Mat)  
print('\n方法1：原始算法PCA:')  
p, n = np.shape(Mat) # shape of Mat  
t = np.mean(Mat, 0) # 每列平均值  
  
# 减去每列的平均值  
for i in range(p):  
 for j in range(n):  
 Mat[i, j] = float(Mat[i, j] - t[j])

**标准化后的数据（前五行）为：**

**[[-0.74333333 0.44266667 -2.358 -0.99933333]**

**[-0.94333333 -0.05733333 -2.358 -0.99933333]**

**[-1.14333333 0.14266667 -2.458 -0.99933333]**

**[-1.24333333 0.04266667 -2.258 -0.99933333]**

**[-0.84333333 0.54266667 -2.358 -0.99933333]]**

（b）**求协方差矩阵代码：**cov\_Mat = np.dot(Mat.T, Mat) / (p - 1)

**协方差矩阵结果为：**

**协方差**

**[[ 0.68569351 -0.042434 1.27431544 0.51627069]**

**[-0.042434 0.18997942 -0.32965638 -0.12163937]**

**[ 1.27431544 -0.32965638 3.11627785 1.2956094 ]**

**[ 0.51627069 -0.12163937 1.2956094 0.58100626]]**

U, V = np.linalg.eigh(cov\_Mat)  
# 重新排列特征向量和特征值  
U = U[::-1]  
for i in range(n):  
 V[i, :] = V[i, :][::-1]

（c）**求特征值和特征向量代码：**

**特征值及特征向量结果为：**

**[4.22824171 0.24267075 0.0782095 0.02383509]**

**[[-0.36138659 0.65658877 0.58202985 0.31548719]**

**[ 0.08452251 0.73016143 -0.59791083 -0.3197231 ]**

**[-0.85667061 -0.17337266 -0.07623608 -0.47983899]**

**[-0.3582892 -0.07548102 -0.54583143 0.75365743]]**

（d）将数据降到k维（k的值可以依据原数据集选取，如果使用iris数据集，k可以取值为2），按特征值大小排序，选取前k个特征值对应的特征向量，计算**降维后的数据（前五行）为：**

**[[ 2.68412563 0.31939725]**

**[ 2.71414169 -0.17700123]**

**[ 2.88899057 -0.14494943]**

**[ 2.74534286 -0.31829898]**

**[ 2.72871654 0.32675451]]**

（4）方法二

用sklearn进行PCA，from sklearn.decomposition import PCA

**降维后的数据（前五行）为：**

**[[-2.68412563 0.31939725]**

**[-2.71414169 -0.17700123]**

**[-2.88899057 -0.14494943]**

**[-2.74534286 -0.31829898]**

**[-2.72871654 0.32675451]]**

**5. 实验心得**

实验过程中了解了pca算法的流程，对pca算法以及机器学习有了一些新的认识，同时实验过程中也遇见了一些问题，在经过请教同学以及上网查询之后都解决了问题。