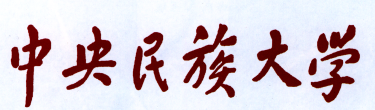
****

|  |  |
| --- | --- |
|  |  |

数据挖掘

实验一：主成分分析（PCA）算法

|  |  |
| --- | --- |
| 教 学 班： | 计算机科学与技术一班 |
| 学 号： | 19011402 |
| 姓 名： | 南佳霖 |
| 实验地点： | 理工楼913 |
| 指导教师： | 姜珊 |

2021

# 实验目的

掌握PCA算法的原理

# 实验要求

1. 理解数据降维过程
2. 熟练使用Python实现PCA算法

# 实验环境

1. PC机一台
2. Python或其他编程工具
3. Numpy、sklearn等第三方库

# 实验内容

1. **PCA降维过程：**

简单来说，就是将数据从原始的空间中转换到新的特征空间中，例如原始的空间是三维的(x,y,z)，x、y、z分别是原始空间的三个基，我们可以通过某种方法，用新的坐标系(a,b,c)来表示原始的数据，那么a、b、c就是新的基，它们组成新的特征空间。在新的特征空间中，可能所有的数据在c上的投影都接近于0，即可以忽略，那么我们就可以直接用(a,b)来表示数据，这样数据就从三维的(x,y,z)降到了二维的(a,b)。

1. **数据集**

* 数据集选择：

本次实验选择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。样本共分三类，用0、1、2表示

* 加载数据集代码：

*from sklearn.datasets import load\_iris*

*x, y = load\_iris(return\_X\_y=True)*

1. **数据标准化（去均值）代码：**

self.n\_features\_ = X.shape[1]

X = X - X.mean(axis=0) # 0均值化

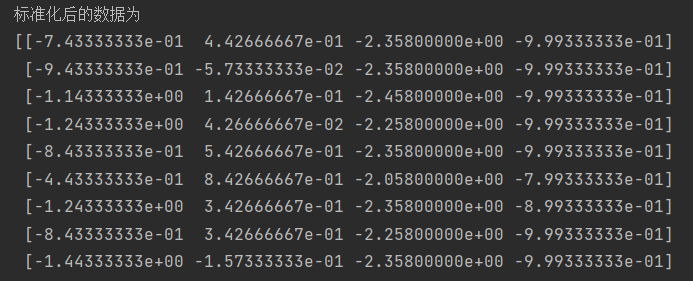


图1-1 标准化后的数据（部分）

1. **求协方差矩阵代码：**

self.covariance = X.T.dot(X) / X.shape[0]

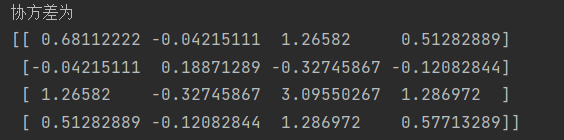


图1-2 协方差矩阵

1. **求特征值和特征向量代码：**

eig\_vals, eig\_vectors = np.linalg.eig(self.covariance)

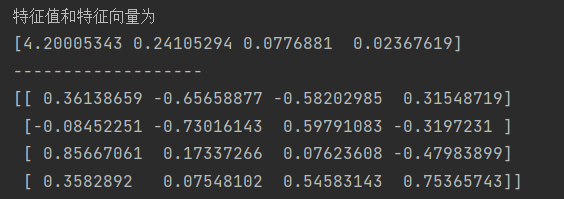
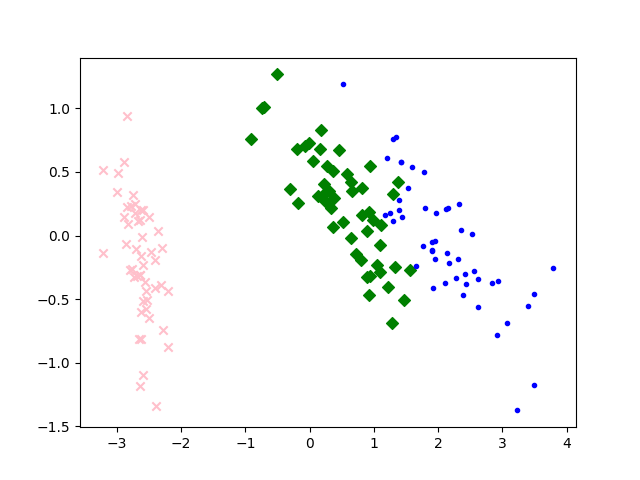


图1-3 特征值及特征向量：

1. 依据原数据集，将数据降到2维，并按特征值大小排序，选取前k个特征值对应的特征向量，计算**降维后的数据为：**



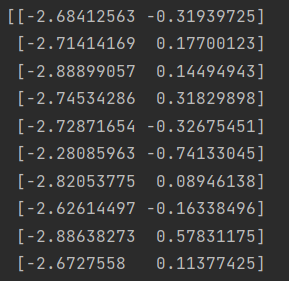


图1-4 （部分）降维后的数据（二维） 图1-5 二维数据的坐标表示

# 五．实验心得

* 通过这次实验、我更加深刻地了解了数据降维技术，对数据挖掘有了更加深入的认识。
* PCA算法属于无监督学习，在计算过程中完全不需要人为的设定参数或是根据任何经验模型对计算进行干预，最后的结果只与数据相关，与用户是独立的。
* 用PCA技术可以对数据进行降维，同时对新求出的“主元”向量的重要性进行排序，根据需要取前面最重要的部分，将后面的维数省去，可以达到降维从而简化模型或是对数据进行压缩的效果。同时最大程度的保持了原有数据的信息。