实现PCA主要成分分析

一、实验题目

对给定的数据ColorHistogram.asc进行PCA主成分分析，并输出以下三个内容：PCA之前数据方差、PCA之后的数据（降至5维）、PCA之后数据方差。

二、实验原理

PCA的主要思想是将n维特征映射到k维上，这k维是全新的正交特征也被称为主成分，是在原有n维特征的基础上重新构造出来的k维特征。

PCA的工作就是从原始的空间中顺序地找一组相互正交的坐标轴，新的坐标轴的选择与数据本身是密切相关的。其中，第一个新坐标轴选择是原始数据中方差最大的方向，第二个新坐标轴选取是与第一个坐标轴正交的平面中使得方差最大的，第三个轴是与第1,2个轴正交的平面中方差最大的。依次类推，可以得到n个这样的坐标轴。

通过这种方式获得的新的坐标轴，我们发现，大部分方差都包含在前面k个坐标轴中，后面的坐标轴所含的方差几乎为0。于是，我们可以忽略余下的坐标轴，只保留前面k个含有绝大部分方差的坐标轴。事实上，这相当于只保留包含绝大部分方差的维度特征，而忽略包含方差几乎为0的特征维度，实现对数据特征的降维处理。

三、实验步骤

基于特征值分解协方差矩阵实现PCA算法

输入：数据集X需要降到k维。

1)去平均值(即去中心化)，即每一位特征减去各自的平均值。

|  |
| --- |
|  |

2)计算协方差矩阵，注：这里除或不除样本数量n或n-1,其实对求出的特征向量没有影响。

|  |
| --- |
|  |

3)用特征值分解方法求协方差矩阵的特征值与特征向量。

|  |
| --- |
|  |

4)对特征值从大到小排序，选择其中最大的k个。然后将其对应的k个特征向量分别作为行向量组成特征向量矩阵P。

5)将数据转换到k个特征向量构建的新空间中，即Y=PX。

|  |
| --- |
|  |

四、实验结果

|  |
| --- |
|  |
|  |