Homework 02

## 一、 实验目标：

对数据集中的所有数字 3 进行主成分分析

## 二、 实验过程：

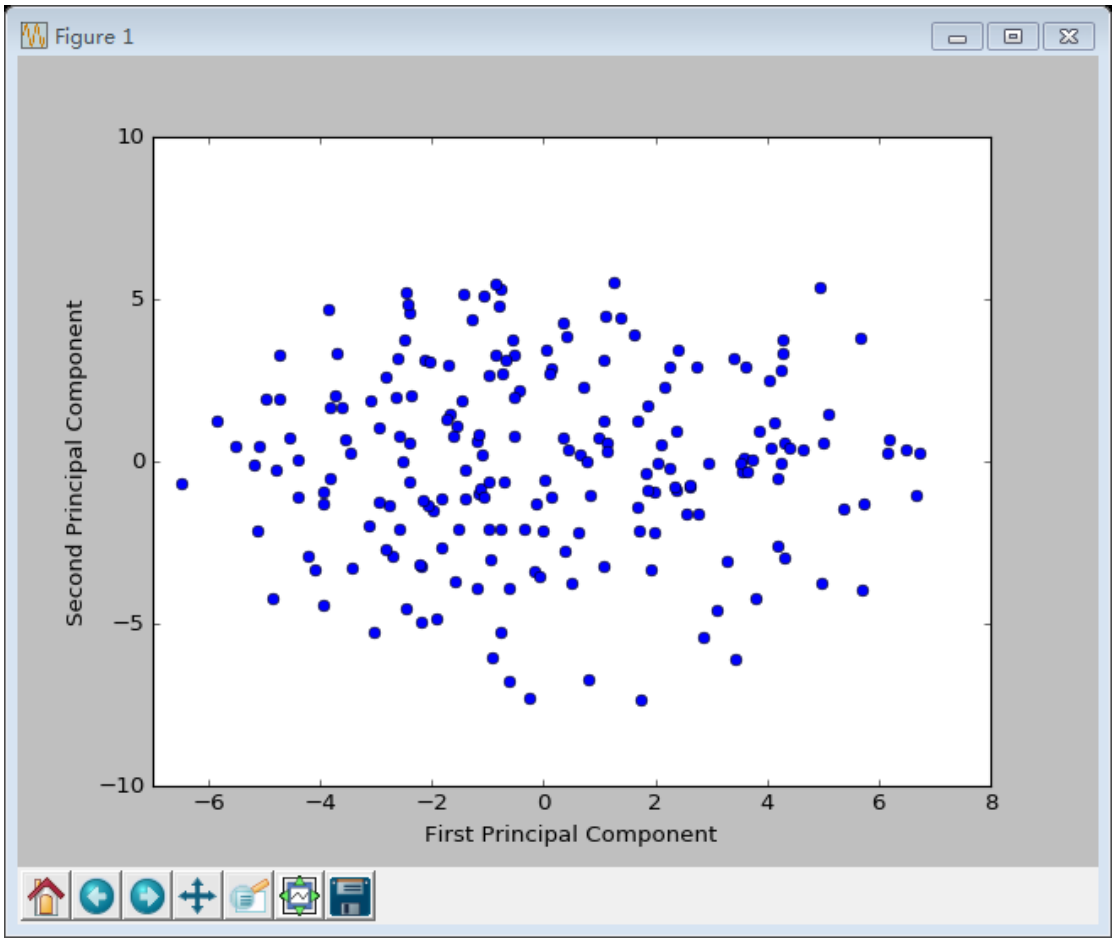
第一步，打开数据集文件“optdigits-orig.tra”，读取每个数字的向量，维度为 1024, 如果最后读取到该数字为 3, 则作为一次成功采样记录到 numVecMatrix 中，最终获得 199 行，1024 列的数字 3 的采样集，每一行为一个样本。

第二步，对每一列计算平均值，即 1024 个特征每个特征的平均值，然后每个特征减去其平均值，获得矩阵 metrixAve。

第三步，对 metrixAve 求协方差矩阵，对协方差矩阵求特征值后进行排序，选择最大的特征值作为 First Principal Component 和 Second Principal Component，获取两个特征值的特征向量。

第四步，将 metrixAve 映射到以两个特征值的特征向量生成的平面上，绘制图像。

## 三、 实验结果：




## 四、 实验代码及注释：

```
from pca_module import *
import matplotlib.pyplot as plt
import numpy as np
import numpy.linalg as lg
numVec = []
numVecMatrix = []
#Open file
fileHandle = open ('optdigits-orig.tra')
fileList = fileHandle.readlines()
lineNum = 0
flag = False
#Get all vectors whose number is 3
for fileLine in fileList:
    lineNum = lineNum + 1
    line = fileLine.rstrip()
```

```
        if lineNum < 22:
            continue
        if (lineNum - 21) % 33 != 0 :
            numVec += [int(x) for x in line]
        else:
            #If number is 3, store the vector. Then clear the numVec
            if(line == ' 3'):
                numVecMatrix.append(numVec)
            numVec = []
fileHandle.close()

#Minus mean
matrix = np.array(numVecMatrix)
meanVals = np.mean(matrix, axis=0)
metrixAve = matrix - meanVals

#Calculate covariance matrix
cov = np.cov(metrixAve, rowvar=0)
evals,evecs = np.linalg.eig(np.mat(cov))
indices = np.argsort(evals)
indices = indices[-1:-3:-1]
evecs = evecs[:,indices]
#Projection
lowDDataMat = metrixAve * evecs

#Plot
dataArr1 = np.array(lowDDataMat)
m = np.shape(dataArr1)[0]
axis_x1 = []
axis_y1 = []
for i in range(m):
    axis_x1.append(dataArr1[i,0])
    axis_y1.append(dataArr1[i,1])
plt.xlabel('First Principal Component')
plt.ylabel('Second Principal Component')
plt.axis([-7, 8, -10, 10])
plt.plot(axis_x1,axis_y1,"o")
plt.show()
```