# 案例三 书写数字降维

## 【1】案例背景

本题由 **顾翔 921127940122**负责

本案例提供了一个手写数字的图片数据集，整个数据集共有1797张分辨率为8×8的手写数字照片(从0~9的10类数字图片)，并且每张图片有64个像素，即有64维特征，请利用本章所示的降维方法，将图片数据映射到二维空间并进行可视化展示

## 【2】方法陈述

首先使用pandas对数据进行处理，接着，使用主成分分析算法（PCA）对数据进行降维，最后将数据展示出来，分析结果。

运行环境，python3安装有pandas、scikit-learn、numpy库，运行软件：Anaconda、jupyter notebook

## 【3】实验代码

### 步骤一、数据处理

1.数据处理

import **pandas** as **pd**

import **numpy** as **np**

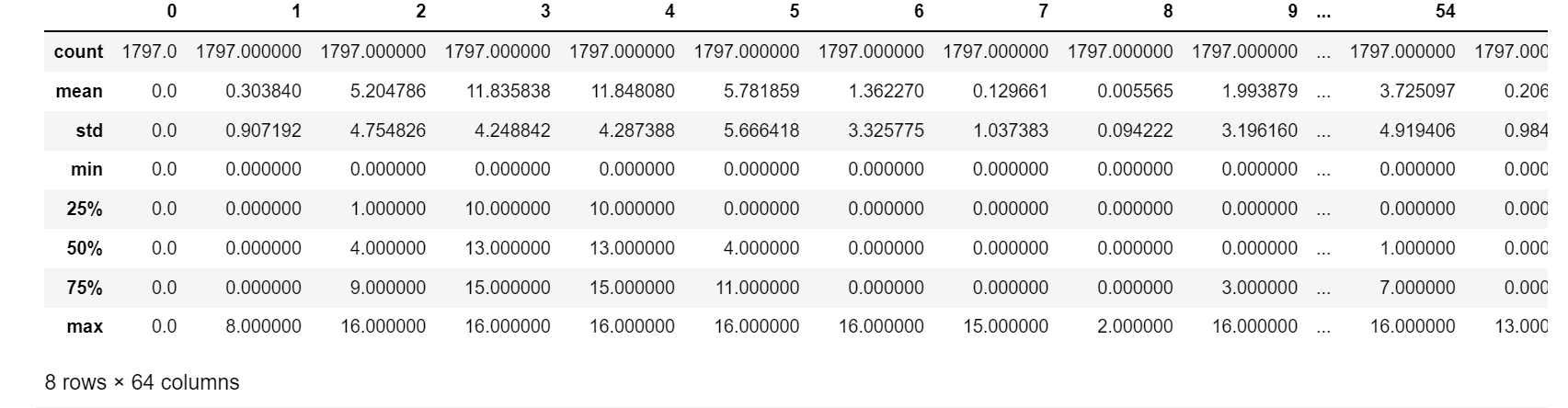
df=**pd**.**read\_table**('optdigits.tes',sep=',',header=None)

df.**head**()

tag=df[64]

x=df.loc[:,:63]

x.**describe**()



### 步骤二、降维——主成分分析（PCA）

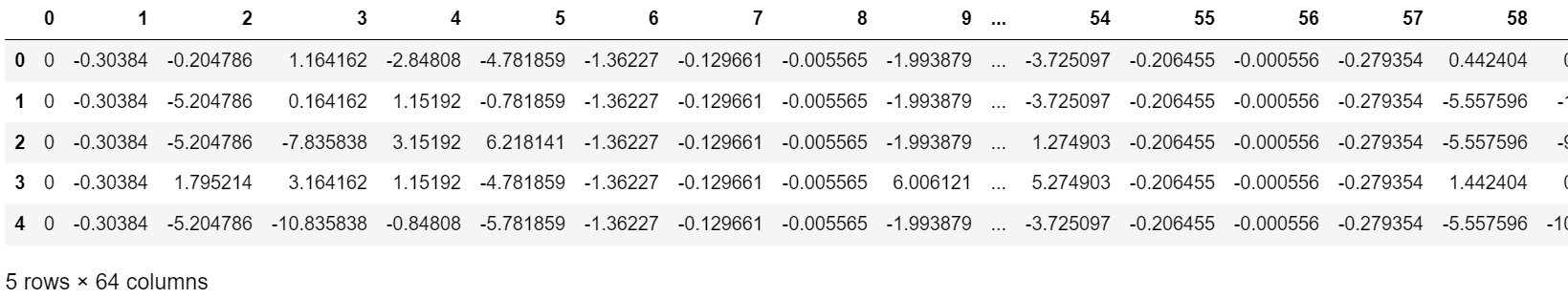
1. 对X中的样本进行中心化处理

m=x.**mean**()

for i in **range**(x.shape[0]):

    x.loc[i,:]=x.loc[i,:].values-m

x.**head**()

****

1. 计算协方差矩阵 ∑

pov=**np**.**dot**(x.T,x)/(x.shape[0]-1)

1. 对协方差矩阵做特征值分解并将特征值排序，按题目要求，取最大的二个特征值对应的特征向量组成转换矩阵w

e\_vals,e\_vecs=**np**.**linalg**.**eig**(pov)

sort\_evals=e\_vals

sort\_evals.sort()

lambda1=sort\_evals[-1]

lambda2=sort\_evals[-2]

for i in **range**(e\_vals.shape[0]):

    if(e\_vals[i]==lambda1):

        vecs1=e\_vecs[i]

    elif(e\_vals[i]==lambda2):

        vecs2=e\_vecs[i]

w=**np**.**array**([vecs1,vecs2]).T

1. Y=xw

y=**np**.**dot**(x,w)

## 结果分析

1. 将降维至二维的数据展现出来
2. 图中，不同的数据产生了较为明显的区分，但也有部分区分不明显，主要为6（蓝色）和8（紫色），推测原因为6，8在写法上相近，因此在降维损失更多的细节后，变得更加难以区分

import **matplotlib**.**pyplot** as **plt**

res=**pd**.**DataFrame**(y)

res.**insert**(res.shape[1],'tag',tag)

def **draw**(res,i,colour):

    temp=res.loc[res['tag']==i]

**plt**.**scatter**(temp[0],temp[1],s=5,c=colour)

**draw**(res,0,'red')

**draw**(res,1,'pink')

**draw**(res,2,'orange')

**draw**(res,3,'lightyellow')

**draw**(res,4,'lawngreen')

**draw**(res,5,'turquoise')

**draw**(res,6,'skyblue')

**draw**(res,7,'slategrey')

**draw**(res,8,'darkviolet')

**draw**(res,9,'gold')

**plt**.**xlim**([-15,20])

**plt**.**ylim**([-15,15])

**plt**.**show**()

