题目2：特征降维和特征学习

2018211568号 2018211316班 杜明欣

任务定义

> MINST 是一个手写数字数据集，使用 PCA 对数据进行降维。观察前两个特征向量所对应的图像，即
>
> 将数据嵌入到 R2 空间。绘制降维后的数据，并分析二维特征是否能够足以完成对输入的分类，对结果进行分析和评价

输入输出

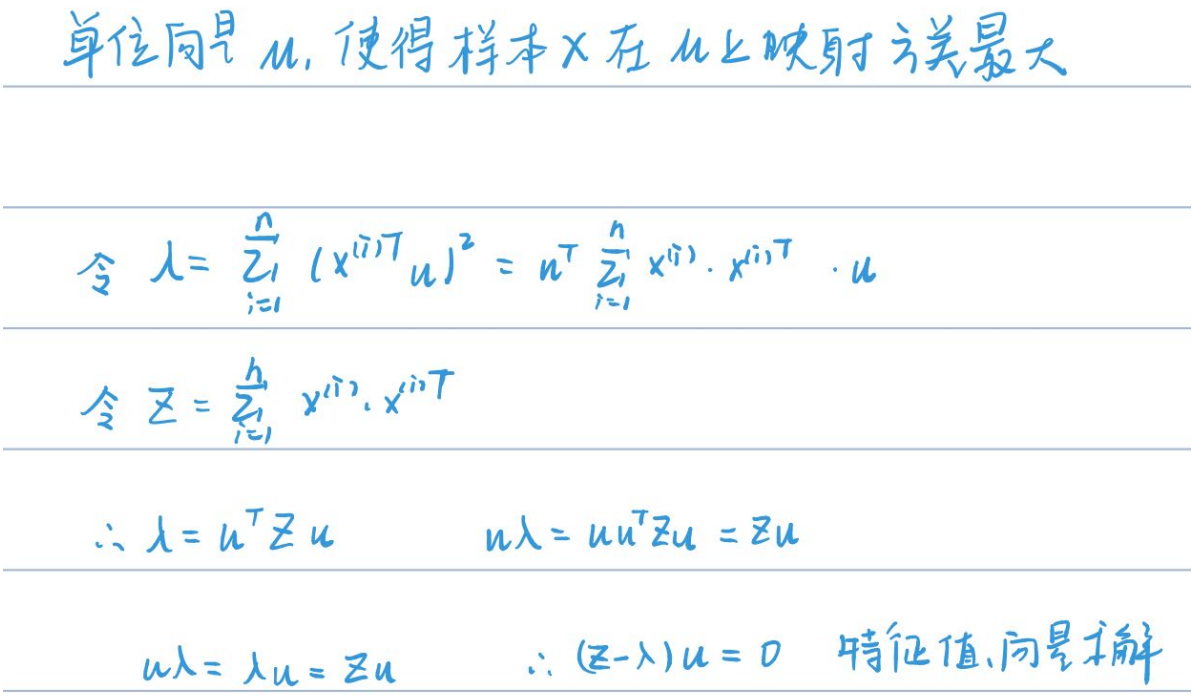
> 输入：MINST 是一个手写数字数据集

方法描述

PCA主成分分析法

核心思想：寻找样本**方差最大**的方向，方差越大越体现样本的不同性，更能体现样本特征

公式推导



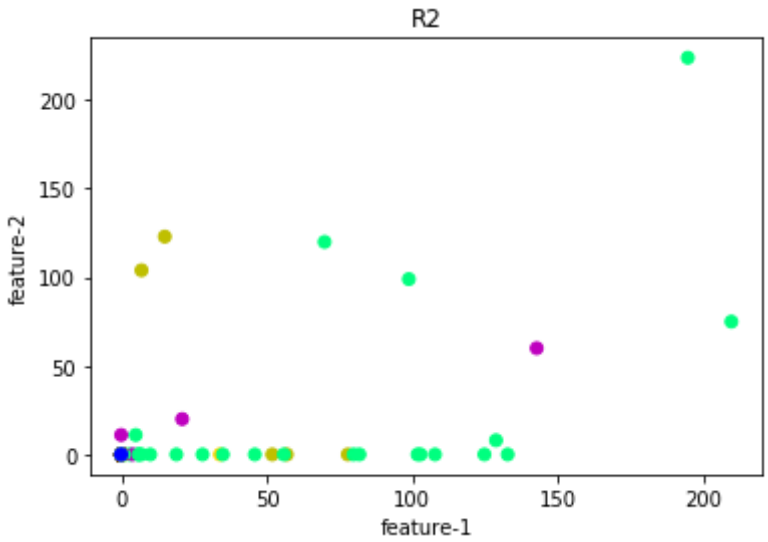
代码实现

```
def down(data):
    #去中心化 uncen_data 减均值后shape 10000*784
    ave=np.sum(data,axis=0)/(data.shape[0])
    uncen_data=data-ave

    #计算协方差 cov_data 784*784
```

```
    cov_data=np.cov(uncen_data.transpose())

    #计算特征值向量
    evals, evecs = linalg.eig(cov_data)
    sumvals=evals.sum()
    maxd=max2(evals)
    max2vals=evals[maxd].sum()
    print("贡献率：",max2vals/sumvals)
    #挑选最大2个特征值对应特征向量
    down_vector=evecs[maxd]
    #计算降维后矩阵
    down_matrix = np.dot(uncen_data, down_vector.transpose())
    return down_matrix
```

结果分析

PCA降维结果图



降维效果分析

结果图形

PCA降维至2维效果并不好，大部分样本集中于feature-2=0区域，未能区分出差别

每类数字用不同颜色标记，从绘制图形看降维后结果未体现出全部数字类别，故降维效果有限

评价指标

采用**累积方差贡献率**衡量**提取信息**能力，累积方差贡献率越**大**说明提取信息能力越**强**

计算公式：协方差矩阵最大2个特征值之和与所有特征值之和的比值

```
贡献率：  (0.1759214994467868+0j)
```

本实验降维后累积方差贡献率仅为17.6%不足以代表样本主体信息，因此二维特征**不足**以完成对输入的分类