机器学习实验报告

主成分分析（Principal Component Analysis）

姓名：秦海滨

学号：1180300523

2020 年 11 月 10 号

目录

1 实验目的

实现一个 PCA 模型，能够对给定数据进行降维（即找到其中的主成分）。

2 实验要求及环境

2.1 实验要求

1. 人工生成一些数据（如三维数据），让它们主要分布在低维空间中。如首先让某个维度的方差远小于其它维度，然后对这些数据旋转。生成这些数据后，用你的 PCA 方法进行主成分提取。

2. 找一个人脸数据（小点样本量），用你实现 PCA 方法对该数据降维，找出数据主成分，然后用这些主成分对每一副人脸图像进行重建，比较一些它们与原图像有多大差别（用信噪比衡量）。

2.2 实验环境

操作系统：Windows10
开发环境：Spider 4.1.4，Python 3.8

3 基本思想

PCA（Principal Component Analysis）是一种常见的数据分析方式，一般用于高维数据的降维，可用于提取数据的主要特征分量。其数学推导一般从最大可分性和最近重构性进行，前者的优化条件表示需要在降维后数据间的方差最大，后者的优化条件表示划分出的点与划分平面（投影面）的距离最小。

我们可以认为 PCA 完成了将高维的数据点投影到低维的超平面上，其投影结果既要满足样本点在这个超平面上的投影尽可能的分开（方差较大），也要满足样本点到这个超平面的距离足够近。

首先，我们对于需要进行 PCA 的数据进行中心化，即保证其均值为 0，公式如下：

$$x_i = x_i - \frac{1}{m}\sum_{j=1}^{m} x_j$$

即对每一个样本点减去其样本均值，这样能保证对经过中心化后的数据常规的线性变换即是绕着原点的旋转变化。而且这样能够保证以一种较为简明的

形式表示样本的协方差矩阵，如下：

$$\Sigma = E[(X - E[X])(X - E[X])^T] = \frac{1}{m}XX^T$$

设使用的投影坐标系的标准正交向量基为 $W = \{w_1, w_2, \ldots, w_d\}, d < n$，这样每个样本点 x_i 降维后得到的坐标为 $z_i = \{z_i1; z_i2; \ldots; z_id\}$。其中，$z_{ij} = w_j^T x_i$，表示 x_i 在 d 维坐标系下第 j 维的坐标。若基于 z_i 来重构 x_i，我们会得到 $\hat{x} = Wz$。则考虑整个训练集上，原样本点 x_i 与基于投影重构的样本点 $\hat{x}_i$ 之间的距离为：

$$\sum_{i=1}^{m} || \sum_{j=1}^{d} z_{ij} w_j - x_i ||_2^2$$

根据最近重构性，该式应被最小化。考虑到 w_j 是标准正交基，$\sum_i x_i x_i^T$ 是协方差矩阵，我们可以得到上述距离公式与 $W^T XX^T W$ 的迹成反比，这就是我们需要优化的目标。我们要使可重构性最小，则要使矩阵 $W^T XX^T W$ 的迹最大。

同样的，从最大可分性出发。样本点 x_i 在降维平面上的投影是 $W_T x_i$，样本点的协方差矩阵为 $\sum_i W^T x_i x_i^T W$，这与最近重构性得出的优化目标相同！

我们对于优化目标式采用拉格朗日乘子法，有：

$$XX^T w_i = \lambda_i w_i$$

于是，我们只需要对协方差矩阵 XX^T 进行特征值分解，将所求的的特征值排序：$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$，从中取出前 d 个特征值对应的特征向量构成 $W^* = (w_1, w_2, ..., w_d)$，这就是主成分分析的解。

4 数据的生成

4.1 三维数据的生成

为了能够明显地可视化，我们选择了三维数据进行数据分析。能够在给定了数据在三个特征维度上的均值和协方差矩阵，利用库中的生成函数生成出需要降维的数据。

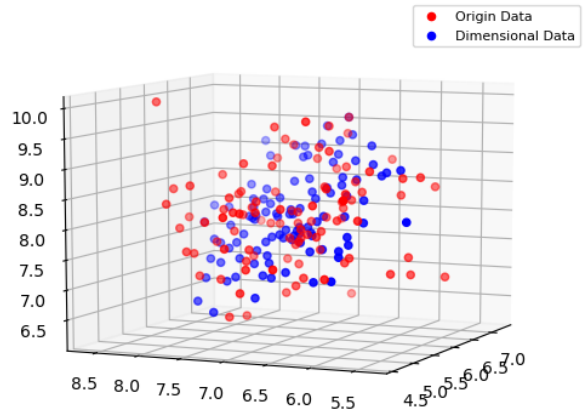
4.2 人脸数据读取

我从一些热门电影的演员中截取的人脸图片，将其裁剪为正方形便于后续计算。

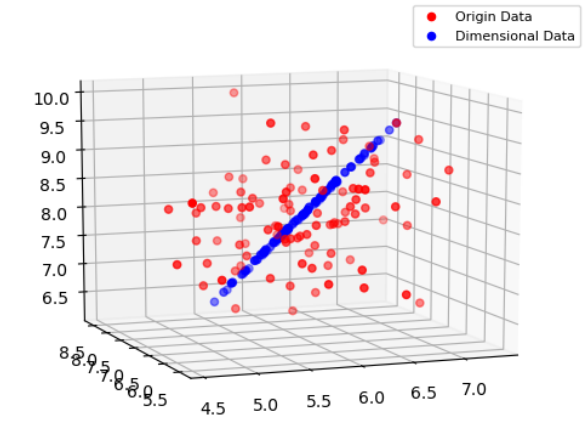
5 实验结果分析

5.1 三维数据降维后的结果

三维数据降维到二维之后会投影到一个平面上，程序运行结果如下：

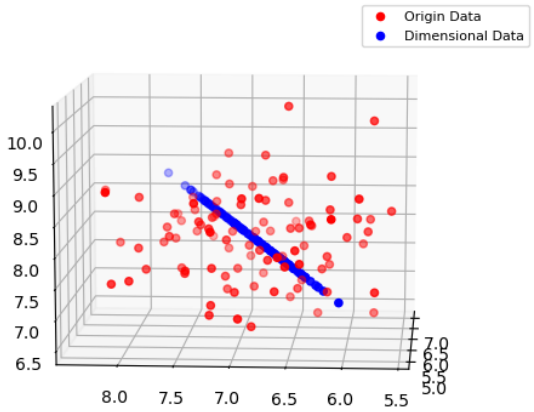


(a) 三维数据降维到二维的结果（平面正视图）

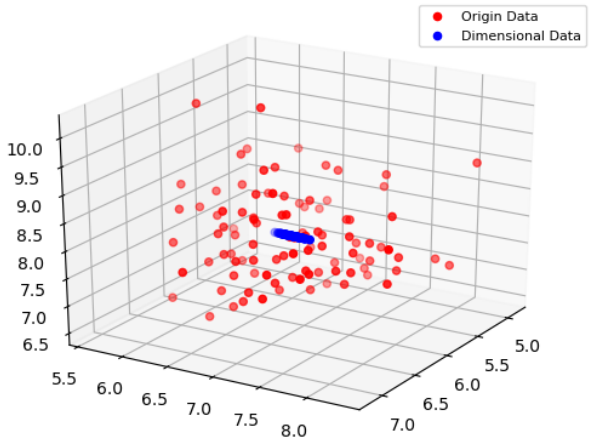


(b) 三维数据降维到二维的结果（平面侧视图）

我们也可以尝试将数据降维到一维，数据应该被投影到一条直线上，程序运行结果如下：



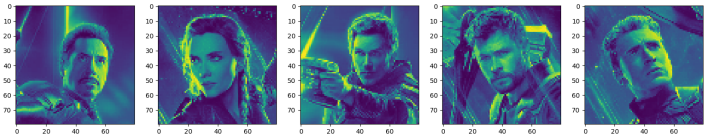
(c) 三维数据降维到一维的结果（平面正视图）



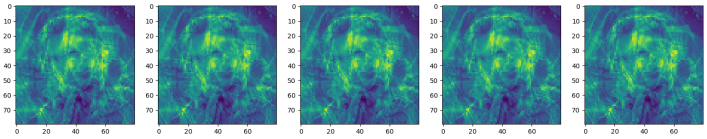
(d) 三维数据降维到一维的结果（平面侧视图）

5.2 人脸数据降维结果

我们截取的人脸数据共有 3600 个维度，我们分别对其降低到 1 维、2 维、三 3 维、4 维和 5 维以及 100 维观察输出结果以及信噪比，原始的图像如下：

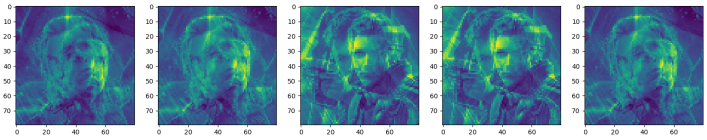


(e) 原始数据



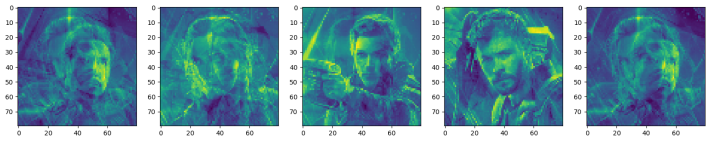
(f) 降低到 1 个维度的结果

由上图可以看出，降低到一个维度后的图像已经基本丧失了所有人脸特征，信噪比为 14.92，难以辨认出人脸归属。



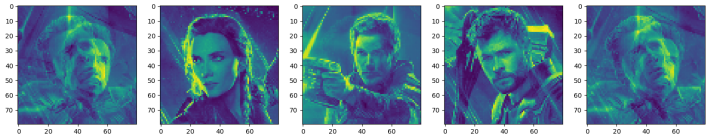
(g) 降低到 2 个维度的结果

降低到 2 个维度的数据与原数据相比也有着较大的差别，但相比于维度为 1 的的数据而言可以在一些地方辨认出人脸特征，信噪比为 16.51。



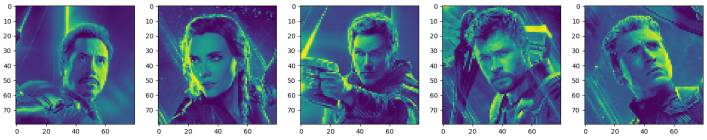
(h) 降低到 3 个维度的结果

降低到 3 个维度后，人脸特征也在一步步提升，但依旧较为模糊，信噪比为 18.59。



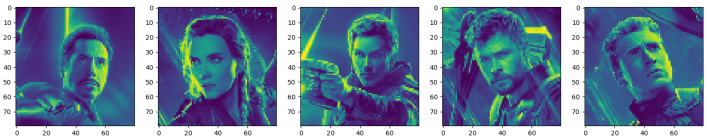
(i) 降低到 4 个维度的结果

降低到 4 个维度后，我们已经能较为完全地辨认出人脸特征，此时信噪比为 22.60。



(j) 降低到 5 个维度的结果

降低 5 个维度后，我们惊奇地发现，人脸数据与原数据相比已经恢复地十分全面，没有什么太大的差别，信噪比为 154.17。这大概使我们确认，5 个维度就能较为完全地保存数据特征，作为比较，我们输出降低到 100 维的结果进行比对：



(k) 降低到 100 个维度的结果

此时的信噪比为 154.17，与降低到 5 维相比差距并不大，这也印证了我们之前的推论。

由以上几个结果对比而言，我们可以发现降低的维度越多，信噪比越低，但当维度到达一个值的时候，就足以刻画大多的数据特征，更多的提供维度也不能明显提高信噪比了。

6 结论

在很多时候高维的数据对于数据的分析会带来很多的不便，复杂的模型与漫长的分析时间极大降低了效率。由此提出的主成分分析方法，能够在所有维度

中选出能够最完全表示原来数据集的维度，通过低维度的表达尽可能的模拟表示原来的信息。通过最近重构性和最大可分性选出主成分表示数，尽可能地减少数据丢失。

显然，低维空间和高维空间必有不同，因为对应于最小的 n-d 个特征值的特征向量被舍弃了，这是降维所导致的必然结果，但舍弃这部分信息能够在其他方面带来更多的优势。一方面，舍弃这部分信息能使样本的采样密度增大，这是降维的重要动机！另一方面，当数据收到噪声影响时，最小的特征值对应的特征向量往往与噪声有关，舍弃这部分向量在一定程度上起到了降噪的作用。