哈尔滨工业大学计算机科学与技术学院
实验报告

课程名称: 机器学习

课程类型: 选修

实验题目: PCA 模型实验

学号: 1190202110

姓名: 田雪洋

2021 年 11 月 22 日

一、实验目的

目标：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）

二、实验要求及实验环境

1. 实验要求

目标：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）

测试：（1）首先人工生成一些数据（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唯独，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，找出一些主成分，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. 实验环境

Windows10; python3.8.6;Pycharm

三、设计思想（本程序中的用到的主要算法及数据结构）

PCA(Principal Component Analysis)，即主成分分析方法，是一种使用最广泛的数据降维算法。其基本原理是利用正交变换把由线性相关变量表示的观测数据转换为少数几个由线性无关变量表示的数据，这些线性无关变量称为主成分，且主成分的个数通常小于原始变量的个数。下面从两个不同的角度实现 PCA

1. 最大投影方差

对于一组观测数据 $x = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x_n})$,$x_i$ 是 m 维随机变量，我们首先对 x_i 进行中心化，中心化可以给后面的计算带来极大的便利，因为中心化之后的常规线性变换就是绕原点的旋转变化，也就是坐标变换。然后，为了方便推导，我们定义: 样本均值:

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i \tag{1}$$

样本投影均值:

$$\mu = \frac{1}{n}\sum_{i}^{n} w^T x_i \tag{2}$$

样本的协方差矩阵:

$$\mathrm{cov}(X) = \Sigma = \frac{1}{n}\sum_{i}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \tag{3}$$

样本在 w 上的投影后的方差为:

$$\begin{aligned}
\sigma^2 &= \frac{1}{n}\sum_{i}^{n}\left(w^T x_i - \mu\right)^2 \\
&= \frac{1}{n}\sum_{i}^{n}\left(w^T x_i - \frac{1}{n}\sum_{i}^{n} w^T x_i\right)^2 \\
&= \frac{1}{n}\sum_{i}^{n}\left(w^T x_i - w^T\left(\frac{1}{n}\sum_{i}^{n} x_i\right)\right)^2 \\
&= \frac{1}{n}\sum_{i}^{n}\left(w^T (x_i - \bar{x})\right)^2 \\
&= \frac{1}{n}\sum_{i}^{n}\left(w^T (x_i - \bar{x})\right)\left(w^T (x_i - \bar{x})\right)^T \\
&= \frac{1}{n}\sum_{i}^{n} w^T (x_i - \bar{x})(x_i - \bar{x})^T w \\
&= w^T\left[\frac{1}{n}\sum_{i}^{n}(x_i - \bar{x})(x_i - \bar{x})^T\right] w \\
&= w^T \Sigma w
\end{aligned} \tag{4}$$

为了在降维的同时，尽可能保留原样本的信息，因此，我们需要让样本点在投影的超平面内尽可能的散开，这里使用方差衡量投影后的离散程度。这

就是最大投影方差的主要思想。而要求方差最大的投影面，即，

$$\hat{w} = \arg\max_{w} w^T \Sigma w \tag{5}$$

而另一方面，为了表示投影方便，我们规定 w 为单位向量, 即 $ww^T = 1$。综上，根据拉格朗日乘数法，得到优化的目标函数为:

$$L(w, \lambda) = w^T \Sigma \mathbf{w} + \lambda \left(1 - w^T w\right) \tag{6}$$

对目标函数求导，得到

$$\frac{\partial L(w, \lambda)}{\partial w} = 2\Sigma w - 2\lambda w \tag{7}$$

令该式为 0，求得:

$$\Sigma w = \lambda w \tag{8}$$

由特征值和特征向量的定义可以看出，λ 是 Σ 的特征值，w 是 Σ 的特征向量。

对于 (4) 式，我们将 (8) 式中的结果代入得到:

$$\sigma^2 = \lambda \tag{9}$$

所以，我们仅需要求出前 k 大的特征值，即可求出前 k 大的特征向量，即投影面。

此外由于我们对数据进行了中心化处理，所以 (1) 式的均值实际上就是 0，则协方差矩阵可以写成:

$$\mathrm{cov}(X) = \Sigma = \frac{1}{n} X X^T \tag{10}$$

(10) 式得到的协方差矩阵为对称矩阵，所以我们可以得到如下特征分解:

$$\Sigma = G \Lambda G^T \tag{11}$$

其中 G 是 Σ 的特征向量组成的矩阵，Λ 是相对应的特征值组成的对角矩阵，我们仅需要使该对角矩阵中的元素取前 k 大的特征值即可，求出最终的答案。具体算法过程如下:

(1) 设有 n 个 m 维的随机变量 $x = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{\boldsymbol{n}})$

(2) 将 x 的每一行进行中心化，即减去这一行的均值，是均值为 0;

(3) 求出协方差矩阵 $\mathrm{cov}(X) = \Sigma = \frac{1}{n} X X^T$;

(4) 求出协方差矩阵对应的特征值和特征向量;

(5) 将特征向量按对应特征值大小从上到下按列排列成矩阵，取前 k 列组成矩阵 w;

(6) 则 $y = w^T x$ 即为降维到 k 维后的数据。

2. 最小投影距离

除了最大投影方差，PCA 也可以从最小投影距离的角度进行推导。显然，对于中心化后的数据样本点，坐标原点的距离不变，而投影面的方差最大，也就等价于这些样本点到投影面的距离最小。所以，下面从最小投影距离的角度推导 PCA。

对于中心化后的数据 $x = \left(x^{(1)}, x^{(2)}, \ldots, x^{(n)}\right)$，我们定义 $z^{(i)} = \left(z_1^{(i)}, z_2^{(i)}, \ldots, z_k^{(i)}\right)^T$ 为经过投影变换后新坐标系中 $x^{(i)}$ 的 k 维坐标。其中，$z_j^{(i)} = w_j^T x^{(i)}$

那么，当我们采用 z_j 去恢复 $x^{(i)}$ 时会得到恢复数据 $\bar{x}^{(i)}$，最小投影距离也就是求得：

$$\operatorname{argmin} \sum_{i=1}^{n} \left\| \bar{x}^{(i)} - x^{(i)} \right\|_2^2 \tag{12}$$

对该式化简得到：

$$\begin{aligned}
\sum_{i=1}^{n} \left\| \bar{x}^{(i)} - x^{(i)} \right\|_2^2 &= \sum_{i=1}^{m} \left\| W z^{(i)} - x^{(i)} \right\|_2^2 \\
&= \sum_{i=1}^{m} \left(W z^{(i)} \right)^T W z^{(i)} - 2 \sum_{i=1}^{m} \left(W z^{(i)} \right)^T x^{(i)} + \sum_{i=1}^{m} x^{(i)T} x^{(i)} \\
&= \sum_{i=1}^{m} \left(z^{(i)T} \right) z^{(i)} - 2 \sum_{i=1}^{m} \left(z^{(i)T} \right) W^T x^{(i)} + \sum_{i=1}^{m} x^{(i)T} x^{(i)} \\
&= \sum_{i=1}^{m} \left(z^{(i)T} \right) z^{(i)} - 2 \sum_{i=1}^{m} \left(z^{(i)T} \right) z^{(i)} + \sum_{i=1}^{m} x^{(i)T} x^{(i)} \\
&= - \sum_{i=1}^{m} \left(z^{(i)T} \right) z^{(i)} + \sum_{i=1}^{m} x^{(i)T} x^{(i)} \\
&= - \operatorname{tr} \left(W^T \sum_{i=1}^{m} \left(x^{(i)T} \right) x^{(i)} \right) W \Bigg) + \sum_{i=1}^{m} x^{(i)T} x^{(i)} \\
&= - \operatorname{tr} \left(W^T X X^T W \right) + \Sigma
\end{aligned} \tag{13}$$

由于协方差矩阵 Σ 是不变的，所以最小化 (13) 式等于最小化:

$$\begin{aligned} \text{argmax} \quad & \text{tr}\left(W^T X X^T W\right) \\ \text{s.t.} \quad & W^T W = I \end{aligned} \tag{14}$$

这个结果与前面的最大投影误差法得到的结果相同。

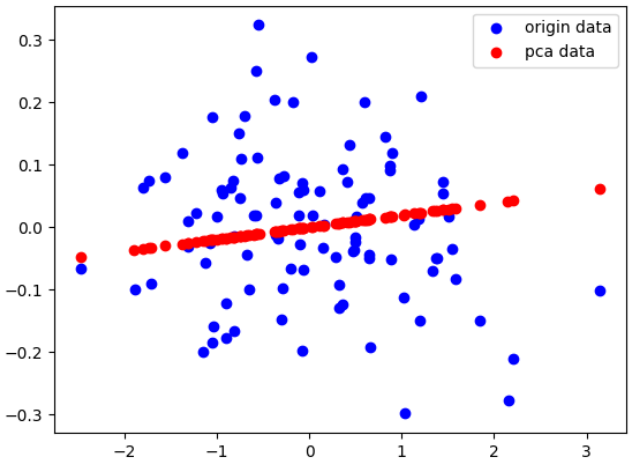
四、实验结果与分析

1. 生成数据

为了方便进行数据可视化，在这里只使用了 2 维数据和 3 维数据的在 PCA 前后的对比实验。

(1). 二维降到一维

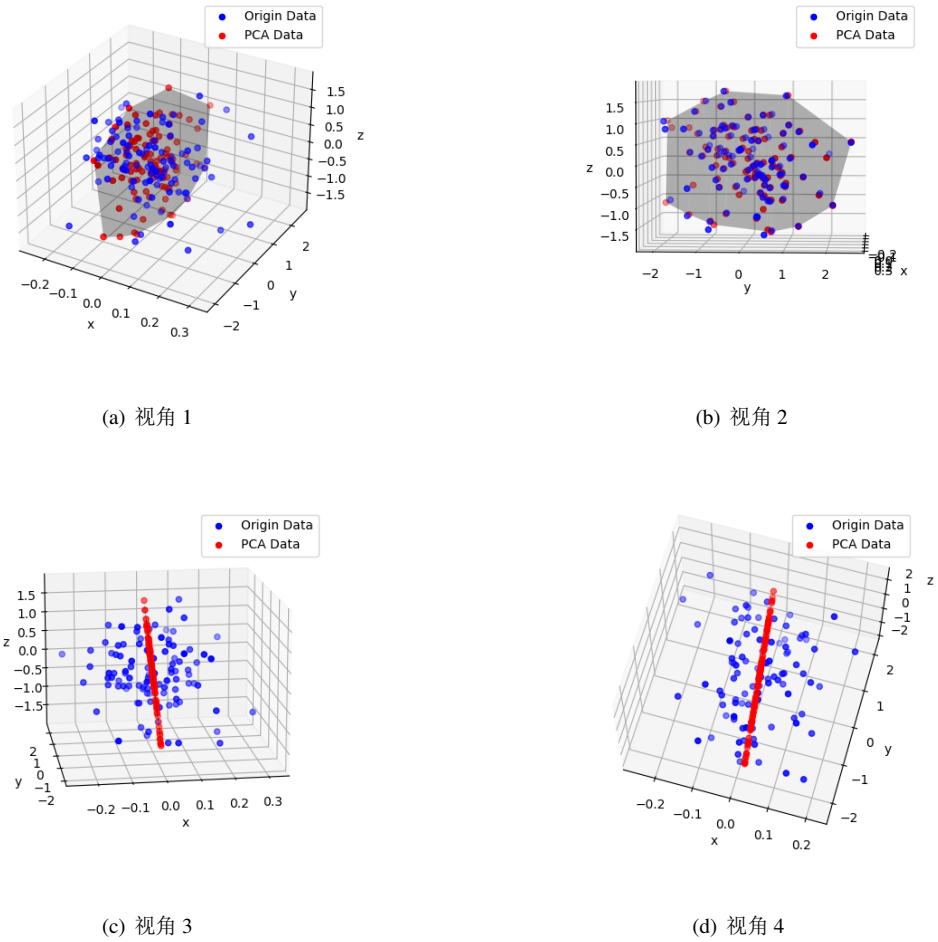
在这里，生成 $\mu = [-2, 2]$ 和 $\Sigma = [[1, 0], [0, 0.01]]$ 的二维高斯分布，可以看到在第二维的方差远小于第一维的方差，因此，第二维包含了更多的信息，进行 PCA 降维的结果如下:



可以看到 PCA 降维的结果基本上分布在与第一维平行的直线上，且第一维的方差极小，第二维的方差较大，符合预期。

(1). 三维降到二维

在这里，生成 $\mu = [1, 2, 3]$ 和 $\Sigma = [[0.01, 0, 0], [0, 1, 0], [0, 0, 1]]$ 的三维高斯分布，可以看到在第一维的方差远小于第二维和第三维的方差，因此，第二维和第三维包含了更多的信息，进行 PCA 降维的结果如下:



(a) 视角 1

(b) 视角 2

(c) 视角 3

(d) 视角 4

图 1: 不同视角下的 PCA 结果

可以看到 PCA 降维的结果基本上分布在第二维第三维组成的平面，符合预期。

2. 人脸数据测试

图像的信噪比和图像的清晰度一样，都是衡量图像质量高低的重要指标。图像的信噪比是指视频信号的大小与噪波信号大小的比值，其公式为:

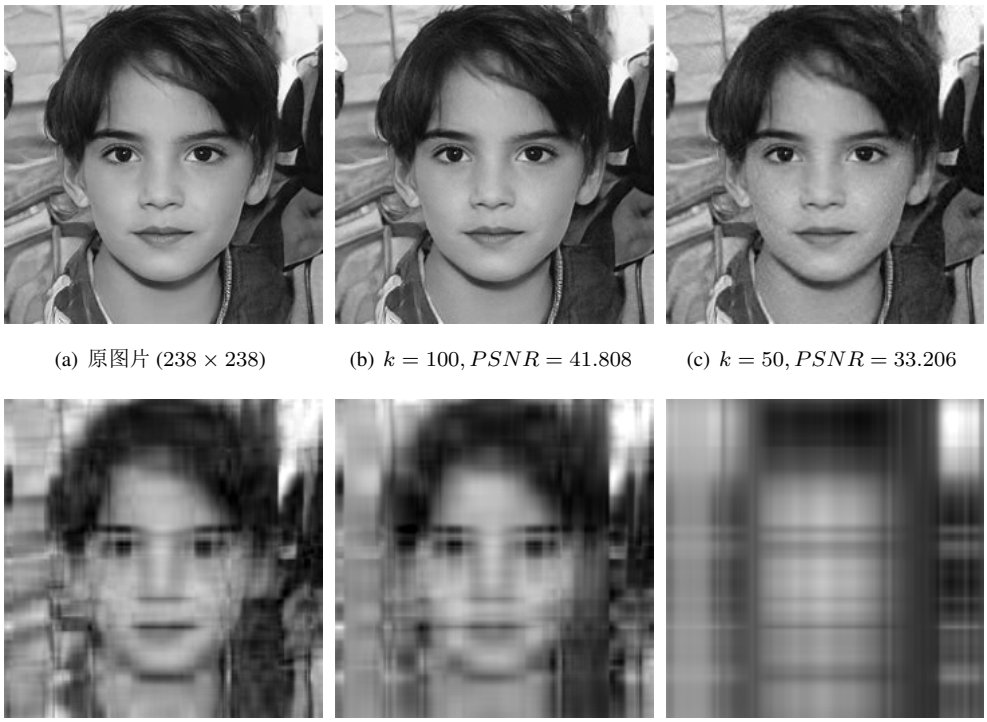
$$
\begin{aligned}
MSE &= \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\|I(i,j)-K(i,j)\|^2 \\
PSNR &= 10\cdot\log_{10}\left(\frac{MAX_I^2}{MSE}\right) = 20\cdot\log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right)
\end{aligned}
\tag{15}
$$

下面分别使用一男一女两组人脸图片利用 PCA 进行降维处理，并计算信噪比：

(a) 原图片 (227×305)　(b) $k = 100, PSNR = 44.985$　(c) $k = 50, PSNR = 34.780$

(d) $k = 10, PSNR = 24.571$　(e) $k = 5, PSNR = 21.515$　(f) $k = 1, PSNR = 14.716$

图 2: 女性人脸图像降维结果



(a) 原图片 (238 × 238) (b) $k = 100, PSNR = 41.808$ (c) $k = 50, PSNR = 33.206$

(d) $k = 10, PSNR = 23.646$ (e) $k = 5, PSNR = 20.896$ (f) $k = 1, PSNR = 16.466$

图 3: 男性人脸图像降维结果

五、结论

- PCA 算法中舍弃了较小的几个特征值对应的特征向量，有效提高了样本的采样密度；而且由于特征值对较小的几个特征向量往往与噪声有关，因此，PCA 同时也起到了降噪的效果
- PCA 不仅将数据压缩到低维，并且将降维之后的各维特征相互独立。

六、参考文献

(1) 周志华著. 机器学习, 北京: 清华大学出版社, 2016.1

(2) 李航著. 统计学习方法，北京：清华大学出版社，2020.6

七、附录：源代码（带注释）

源代码见相关文件

(1) bfacetest.py: 测试男性人脸图片使用 PCA 降维效果

(2) gfacetest.py: 测试女性人脸图片使用 PCA 降维效果

(3) data.py：数据预处理及 PCA 算法实现

(4) test2D.py: 利用 PCA 将二维数据降维成一维数据的测试

(5) test3D.py: 利用 PCA 将三维数据降维成二维数据的测试