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：1160300312
姓名：靳贺霖

一、 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），可以利用已有的矩阵特征向量提取方法。

二、实验要求及实验环境

实验要求：

（1） 首先人工生成一些数据（如三维数据），让他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他维度，然后对这些数据进行旋转。生成这些数据后，用你的 PCA 方法进行主成份提取。

（2） 利用手写体数字数据 mnist，用你实现的 PCA 方法对该数据进行降维，找出一些主成分，然后用这些主成分对每一幅图像进行重建，比较一些它们与原图像有多大差别（可以用信噪比衡量）

实验环境：

操作系统：windows 7
编译环境：python3.7,
编译器：PyCharm

三、设计思想（本程序中的用到的主要算法及数据结构）

1．算法原理

这次我们的任务是要对数据进行降维，用的是主成分分析（PCA）算法。主成分分析法希望通过降维能将多个维度的指标转化为几个具有代表性的综合指标，从而实现数据的有损失压缩。

PCA 考虑这样一个问题：对于在正交坐标系上的样本点，如何将其用少于原坐标系维数的坐标系来表达这些样本点。算法主要是通过两方面来考虑的。一是最近重构性，即样本点到新坐标系面的距离足够近；二是最大可分性，即样本点在这个新的坐标系面上的投影能尽可能分开（方差最大）。我们接下来通过这两个方面来讨论 PCA 算法的原理。

首先来考虑最近重构性。首先为了方便向量的表示，我们使数据样本中心化，即使所有的样本向量之和为零向量。然后定义我们要求的低维新坐标系为$\mathrm{W}=\{w_1,w_2,\cdots,w_{d'}\}$，其中每个$w_i$都是标准正交基向量，$d'<\mathrm{d}$，d为原高维坐标系。这样，我们就可以将我们所有的数据点投影到低维坐标系中。设样本点集合为$\mathrm{X}=\{x_1,x_2,\cdots x_m\}$，则我们就可以得到所有的样本点到其投影的点的距离和为

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

这样，我们的问题就转化成了将这个式子最小化。其中$\boldsymbol{z_i}=\{z_{i1},z_{i2},\cdots,z_{id'}\}$，为样本点$x_{\mathrm{i}}$在新的低维坐标系下的投影，而且我们易知有$z_{i1}=w_j{}^T x_{\mathrm{i}}$。所以我们对上式子进行化简，我们的问题就转变成了下面这样的问题

$$\min_W -tr(W^TXX^TW)$$
$$\mathrm{s.t.}\quad W^TW=I$$

然后我们再考虑最大可分性。由于我们的样本已经中心化，所以对于投影的向量即为向

量和均值的差，这样，我们就可以得到方差为$\sum_i W^T x_i x_i^T W$，因此，我们的问题就可以转化维这样的问题

$$\max_W \quad tr(W^T X X^T W)$$
$$\text{s.t.} \quad W^T W = I$$

可以看出对于考虑最大可分性和最近重构性，二者转化为的目标问题其实是相同的。对于这个问题的求解我们可以使用拉格朗日乘子法，并进行求导，得到的

$$XX^T W = \lambda W$$

这样，对于这个问题的求解，我们只需要对XX^T进行特征值分解，然后取大的d'个特征值对应的特征向量为 PCA 的解即可。

2. 算法的实现

首先考虑数据的生成。因为要考虑旋转前的某个维度的方差小于其他维度，可以选择直接生成多维高斯分布。通过调整协方差矩阵来调整分布的“形状”（即使某个维度的方差较小）。而我们知道协方差矩阵非对角线的元素来决定分布的形状，只需要调整协方差矩阵非对角线元素即可。

然后对于 PCA 算法的实现，首先我们要对数据进行处理，让所有数据向量都减去所有数据向量的均值向量，从而实现中心化。然后计算协方差矩阵，并求出协方差特征值和特征向量。然后对选择出大的d'个特征值对应的特征向量作为新的基向量。然后对于低维的数据在高维中的表示，我们可以通过先对数据进行降维再进行升维，然后再加上所有数据向量的均值向量来实现，即

$$X_{\text{d\_reduction}} = (X - \mu)W^T W + \mu$$

这样，我们就完成了 PCA 算法的实现。

接下来考虑信噪比计算的实现。对于 mnist 数据降维以后的结果，我们使用峰值信噪比（Peak Signal to Noise Ratio）来实现。PSNR 通常用来评价一幅图像压缩后和原图像相比质量的好坏，计算公式如下所示

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}||I(i,j) - K(i,j)||^2$$

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) = 20 \cdot \log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right)$$

其中计算 MSE 为均方差，PSNR 即为峰值信噪比，其中的 MAX 通常取 255，为图像的灰度值。PSNR 越高说明压缩后失真越少。

四、实验结果与分析

1. 对生成的数据进行 PCA 主成分提取

首先考虑使用 PCA 将二维的数据降为一维的数据。这里生成的二维数据服从二维高斯分布，其中均值为[2, 10]，协方差阵为[[10, 9], [9, 10]]，数据点有 100 个，得到的降维结果如 figure1 所示。可以从图中很明显得看出数据降维后是满足最近重构性和最大可分性的。

然后考虑使用 PCA 将三维数据降为二维的数据和一维的数据。这里生成的三维数据服

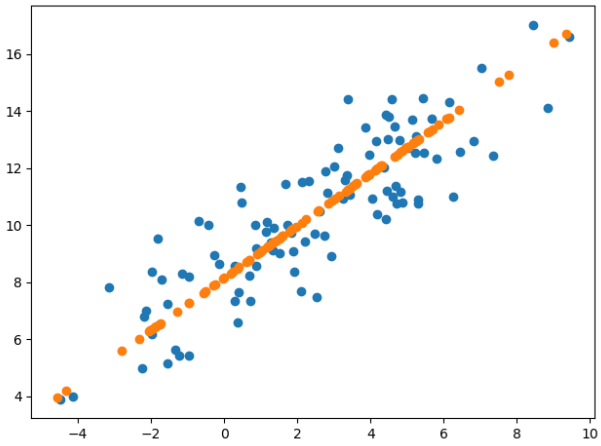

figure1：使用 CPA 将二维数据降为一维的结果

从三维高斯分布，其中均值为[2, 10, 4]，协方差阵为[[2, 2, 2], [2, 10, 2], [2, 2, 10]]，一共有 200 个样本点。得到的降维结果如 figure2 所示，从图中可以看出，虽然三维数据看起来比

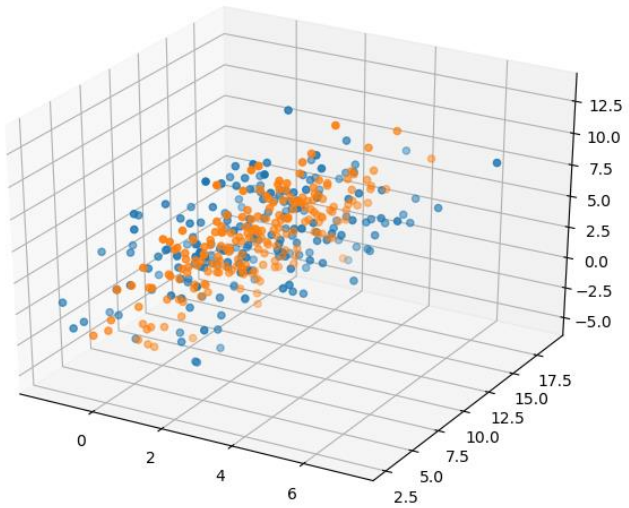

figure2：使用 CPA 将三维数据降为二维的结果

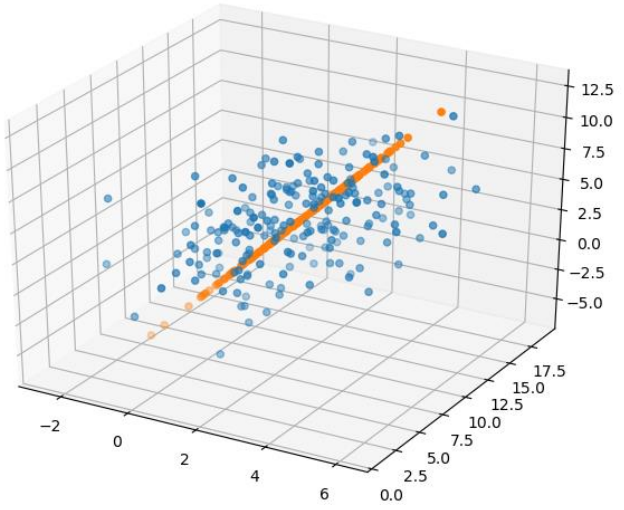

figure3：使用 CPA 将三维数据降为一维的结果

较不容易分辨，但是仍然可以辨别出结果是朝着保留信息最多的方向降维的。然后考虑三维数据到一维数据的降维。这里仍沿用上一部分得到的三维高斯分布，得到的结果如figure3 所示，可以看出三维降为一维的结果虽然不那么好，但是也是理论上应该得到的结果。由于三维以上的图像难以表示，不在这里进行讨论。

2. 对 mnist 数据使用 PCA 降维

这里选择从官网上下载一个手写体 mnist 数据集合。对其进行处理为一个 28*28 的矩阵，然后使用 CPA 对其找到主成分进行降维。我们可以将 28*28 的矩阵视作 28 个 28 维的样本向量，然后对其进行降维。首先将其降到 14 维，效果图如 figure4 所示，可以看出降

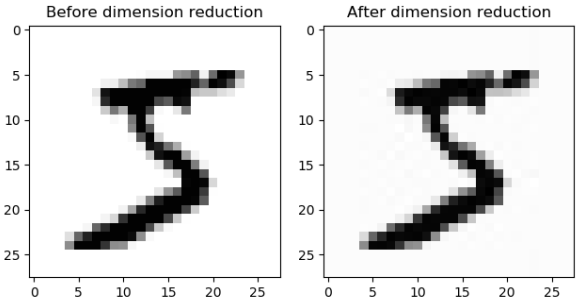

figure4：将一个 mnist 数据从 28 维度降到 14 维的结果

虽然维度降低了，但是其实从直观看二者差别几乎没有。其实这个也比较容易理解，因为图片中有大量的留白，使用 CPA 是保留主成分的降维，所以丢失的维度大概率都是留白的部分。然后考虑将其降到 7 维，效果图如 figure5 所示。这次就可以看出二者明显的区别，但是可以看出主成分是没有明显的改变的，可以明显看出显示的数字为 5。然后考虑

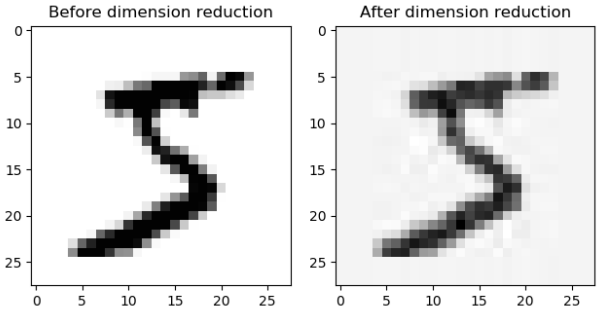

figure5：将一个 mnist 数据从 28 维降到 7 维的结果

将其降维到 2 维，得到的效果图如 figure6 所示，可以明显看出数据出现了很大程度的丢失，

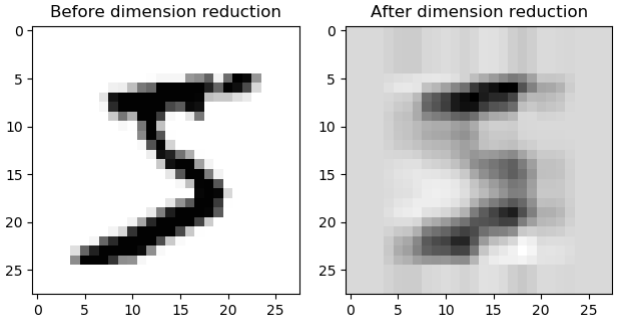

figure6：将一个 mnist 数据从 28 维降到 2 维的结果

但是仍然可以依稀辨别出要表达的数字为 5，然后我们考虑最极端的情况，将数据从 28 维降到 1 维，得到的结果图如 figure7 所示。这次得到的结果就已经完全无法由人眼判别出

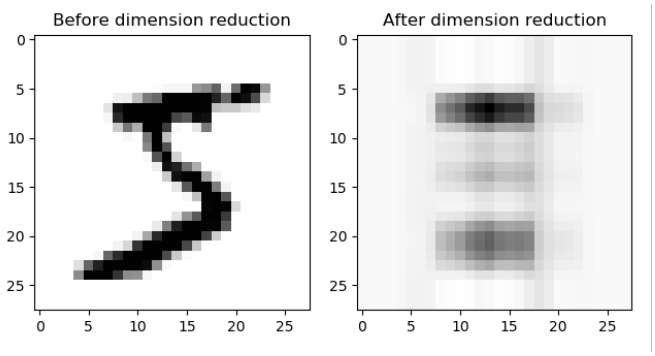

figure7：极端情况，将一个 mnist 数据从 28 维降到 1 维的结果

结果，虽然保留了一部分的数据，但是对人来说，数据已经丢失，无法辨别。然后考虑计算得到的图像和原图像的信噪比，截取部分得到的结果如下表所示：

降低到维度	21	14	7	2	1
信噪比	118.89	-5.74	-14.44	-20.98	-22.05

然后以降低以后的维度为横坐标，对应和原图像的信噪比维纵坐标，得到的折线图如 figure8

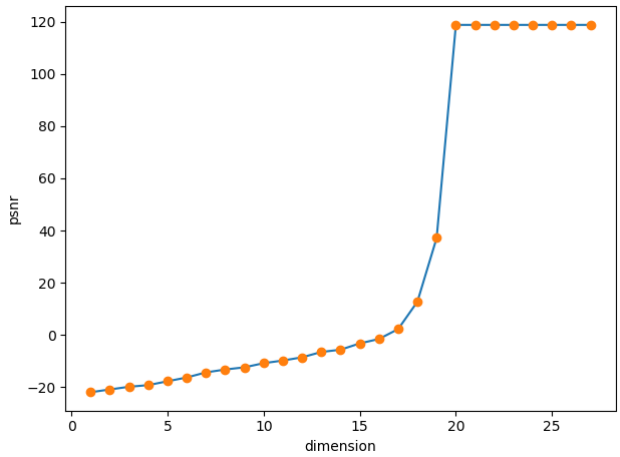

figure8：降维后的维度和信噪比的关系折线图

所示。可以看出，信噪比和降低以后的维度是大致成正相关的。而且这种趋势在 17 维之前几乎为线性，17-20 维出现指数级增长，到达 20 维以后，不再变化。这种趋势和数据的表达形式和数据含量都是有关系的，但是大致趋势几乎就是这样。

五、结论

1. 使用 PCA 对给定数据进行降会对数据造成一定程度的损失，但是会最大程度得保留数据的特征，能够实现主成分提取；
2. 使用 PCA 对 mnist 上的实际手写体数据进行降维时，可以看出降低以后的维度较大时几乎和原图无差别，但是如果降低维度以后的维度较低时，图像有会有明显的噪声出现，但一定程度上不影响辨别。但是当维度降到很低甚至 1 时，已经无法辨别实际要表达的数据；
3. 使用 PCA 对 mnist 上的数据降维后和原图进行信噪比测量可以得到，信噪比和降低以后的维度大致是成正相关的，但是增长速率在不同的维度不同。维度高时维度变化对应的信噪比甚至几乎没有变化。

六、参考文献

[1] http://huaxiaozhuan.com/%E7%BB%9F%E8%AE%A1%E5%AD%A6%E4%B9%A0/chapters/10_PCA.html
[2] https://en.wikipedia.org/wiki/Principal_component_analysis
[3] http://deeplearning.net/tutorial/gettingstarted.html
[4] https://en.wikipedia.org/wiki/MNIST_database
[5] https://es.wikipedia.org/wiki/PSNR

七、附录：源代码（带注释）

将源代码作为附件发送