

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：必修

实验题目：PCA模型实验

学号：1160300314

姓名：朱明彦

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

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. 利用手写体数字数据mnist，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

- OS: Ubuntu 16.04.5 LTS
- python 3.7.0

三、设计思想(本程序中用到的主要算法及数据结构)

1.算法原理

PCA(主成分分析, Principal Component Analysis)是最常用的一种降维方法。在周志华老师的机器学习书中给出了有关于两种有关PCA的推导，分别从最近重构性和最大可分性两种方面进行。

如果超平面可以对正交属性空间的所有样本进行恰当表达，就要具有下面两个性质

- 最近重构性：样本点到这个超平面的距离都足够近
- 最大可分性：样本点在这个超平面上的投影尽可能分开

1.1 中心化

在PCA开始时都假设数据集进行了中心化，即：对于数据集 $\mathbf{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ ，其中 $\mathbf{x}_i \in \mathbb{R}^n$ 。对每个样本均进行如下操作：

$$\mathbf{x}_i \leftarrow \mathbf{x}_i - \frac{1}{m} \sum_{j=1}^m \mathbf{x}_j$$

其中 $\mu = \frac{1}{m} \sum_{j=1}^m \mathbf{x}_j$ 称为样本集的 D 的中心向量。之所以进行中心化，是因为经过中心化之后的常规的线性变换就是绕原点的旋转变换，也就是坐标变换；以及 $\sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i^T = \mathbf{X}^T \mathbf{X}$ 就是样本集的协方差矩阵。

经过中心化后的数据，有 $\sum_{j=1}^m \mathbf{x}_j = \mathbf{0}$ 。设使用的投影坐标系的标准正交向量基为 $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d\}$ ， $d < n$ ，每个样本降维后得到的坐标为：

$$\mathbf{z} = \{z_1, z_2, \dots, z_d\} = \mathbf{W}^T \mathbf{x} \quad (1)$$

因此，样本集与降维后的样本集表示为：

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_m^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{bmatrix}, \mathbf{Z} = \begin{bmatrix} \mathbf{z}_1^T \\ \vdots \\ \mathbf{z}_m^T \end{bmatrix} = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,d} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ z_{m,1} & z_{m,2} & \cdots & z_{m,d} \end{bmatrix}$$

1.2 从最近重构性原理解释

在得到 \mathbf{z} 后，需要对其进行重构，重构后的样本设为

$$\hat{\mathbf{x}} = \mathbf{W} \mathbf{z} \quad (2)$$

将式(1)(2)代入，那么对于整个数据集上的所有样本与重构后的样本之间的误差为：

$$\sum_{i=1}^m \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2 = \sum_{i=1}^m \|\mathbf{W} \mathbf{W}^T \mathbf{x}_i - \mathbf{x}_i\|_2^2 \quad (3)$$

根据定义，可以有：

$$\mathbf{W} \mathbf{W}^T \mathbf{x}_i = \mathbf{W} (\mathbf{W}^T \mathbf{x}_i) = \sum_{j=1}^d \mathbf{w}_j (\mathbf{w}_j^T \mathbf{x}_i) \quad (4)$$

由于 $\mathbf{w}_j^T \mathbf{x}_i$ 是标量，有 $\mathbf{w}_j^T \mathbf{x}_i = (\mathbf{w}_j^T \mathbf{x}_i)^T = \mathbf{x}_i^T \mathbf{w}_j$ ，从而式(4)变为：

$$\begin{aligned} \sum_{i=1}^m \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2 &= \sum_{i=1}^m \|\mathbf{W} \mathbf{W}^T \mathbf{x}_i - \mathbf{x}_i\|_2^2 \\ &= \sum_{i=1}^m \left\| \sum_{j=1}^d (\mathbf{x}_i^T \mathbf{w}_j) \mathbf{w}_j - \mathbf{x}_i \right\|_2^2 \\ &= \sum_{i=1}^m \left\| \mathbf{x}_i - \sum_{j=1}^d (\mathbf{x}_i^T \mathbf{w}_j) \mathbf{w}_j \right\|_2^2 \end{aligned} \quad (5)$$

此外，根据 \mathbf{X} 的定义有：

$$\begin{aligned} \|\mathbf{X} - \mathbf{X} \mathbf{W} \mathbf{W}^T\|_F^2 &= \sum_{i=1}^m \sum_{j=1}^n \left[x_{i,j} - \left(\sum_{k=1}^d w_{k,j} \times \mathbf{x}_i^T \mathbf{w}_k \right) \right]^2 \\ &= \sum_{i=1}^m \left\| \mathbf{x}_i - \sum_{k=1}^d (\mathbf{x}_i^T \mathbf{w}_k) \mathbf{w}_k \right\|_2^2 \\ &= \sum_{i=1}^m \left\| \mathbf{x}_i - \sum_{j=1}^d (\mathbf{x}_i^T \mathbf{w}_j) \mathbf{w}_j \right\|_2^2 \end{aligned} \quad (6)$$

结合式(5)(6)可以化简优化目标：

$$\begin{aligned}
\mathbf{W}^* &= \arg \min_{\mathbf{W}} \sum_{i=1}^m \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2 = \arg \min_{\mathbf{W}} \|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{W}^T\|_F^2 \\
&= \arg \min_{\mathbf{W}} \text{tr}[(\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{W}^T)^T(\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{W}^T)] \\
&= \arg \min_{\mathbf{W}} \text{tr}[\mathbf{X}^T\mathbf{X} - \mathbf{X}^T\mathbf{X}\mathbf{W}\mathbf{W}^T - \mathbf{W}\mathbf{W}^T\mathbf{X}^T\mathbf{X} + \mathbf{W}\mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W}\mathbf{W}^T] \\
&= \arg \min_{\mathbf{W}} [\text{tr}(\mathbf{X}^T\mathbf{X}) - \text{tr}(\mathbf{X}^T\mathbf{X}\mathbf{W}\mathbf{W}^T) - \text{tr}(\mathbf{W}\mathbf{W}^T\mathbf{X}^T\mathbf{X}) + \text{tr}(\mathbf{W}\mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W}\mathbf{W}^T)] \\
&= \arg \min_{\mathbf{W}} [\text{tr}(\mathbf{X}^T\mathbf{X}) - \text{tr}(\mathbf{X}^T\mathbf{X}\mathbf{W}\mathbf{W}^T) - \text{tr}(\mathbf{X}^T\mathbf{X}\mathbf{W}\mathbf{W}^T) + \text{tr}(\mathbf{X}^T\mathbf{X}\mathbf{W}\mathbf{W}^T\mathbf{W}\mathbf{W}^T)] \\
&= \arg \min_{\mathbf{W}} [\text{tr}(\mathbf{X}^T\mathbf{X}) - \text{tr}(\mathbf{X}^T\mathbf{X}\mathbf{W}\mathbf{W}^T) - \text{tr}(\mathbf{X}^T\mathbf{X}\mathbf{W}\mathbf{W}^T) + \text{tr}(\mathbf{X}^T\mathbf{X}\mathbf{W}\mathbf{W}^T)] \\
&= \arg \min_{\mathbf{W}} [\text{tr}(\mathbf{X}^T\mathbf{X}) - \text{tr}(\mathbf{X}^T\mathbf{X}\mathbf{W}\mathbf{W}^T)] \\
&= \arg \min_{\mathbf{W}} [-\text{tr}(\mathbf{X}^T\mathbf{X}\mathbf{W}\mathbf{W}^T)] \\
&= \arg \max_{\mathbf{W}} [\text{tr}(\mathbf{X}^T\mathbf{X}\mathbf{W}\mathbf{W}^T)] \\
&= \arg \max_{\mathbf{W}} [\text{tr}(\mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W})]
\end{aligned} \tag{7}$$

从而优化目标为 $\mathbf{W}^* = \arg \max_{\mathbf{W}} [\text{tr}(\mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W})]$ ，约束为 $\mathbf{W}^T\mathbf{W} = \mathbf{I}_{d \times d}$

1.3 从最大可分性原理解释

对于原始数据样本点 \mathbf{x}_i 在降维后在新空间的超平面上的投影为 $\mathbf{W}^T\mathbf{x}_i$ 。若使样本点的投影尽可能分开，应该使样本点在投影后的方差最大化，即使下式最大化：

$$\begin{aligned}
\arg \max_{\mathbf{W}} &= \arg \max_{\mathbf{W}} \sum_{i=1}^m \mathbf{W}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{W} \\
&= \arg \max_{\mathbf{W}} \text{tr}(\mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W}) \\
\text{s.t. } &\mathbf{W}^T\mathbf{W} = \mathbf{I}
\end{aligned} \tag{8}$$

可以看到式(7)与(8)等价。PCA的优化问题就是要求解 $\mathbf{X}^T\mathbf{X}$ 的特征值。

只需将 $\mathbf{X}^T\mathbf{X}$ 进行特征值分解，将得到的特征值进行排序： $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ ，提取前 d 大的特征值对应的单位特征向量即可构成变化矩阵 \mathbf{W} 。

2.算法的实现

给定样本集 $\mathbf{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ 和低维空间的维数 d

1. 对所有的样本进行中心化操作：

$$1. \text{ 计算样本均值 } \mu = \frac{1}{m} \sum_{j=1}^m \mathbf{x}_j$$

$$2. \text{ 所有样本减去均值 } \mathbf{x}_j = \mathbf{x}_j - \mu, j \in \{1, 2, \dots, m\}$$

2. 计算样本的协方差矩阵 $\mathbf{X}^T\mathbf{X}$

3. 对协方差矩阵 $\mathbf{X}^T\mathbf{X}$ 进行特征值分解

4. 取最大的 d 个特征值对应的单位特征向量 $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d$ ，构造投影矩阵 $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d)$

5. 输出投影矩阵 \mathbf{W} 与样本均值 μ

四、实验结果分析

1.生成数据的测试

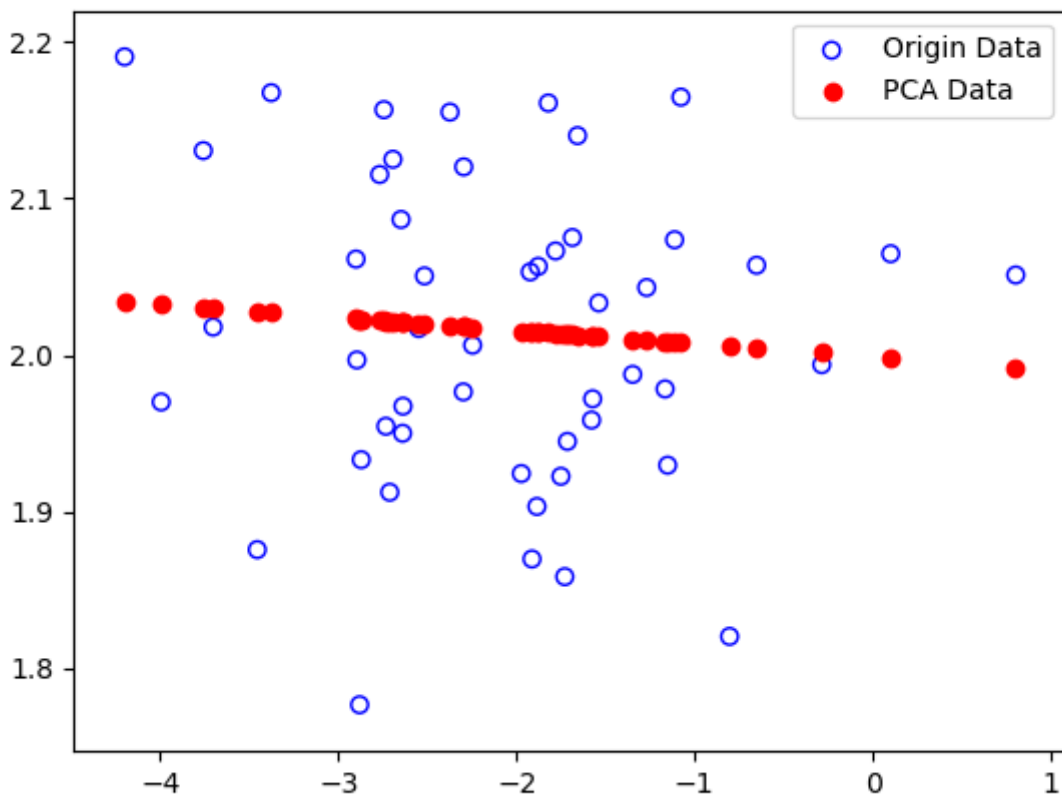
为了方便进行数据可视化，在这里只进行了2维数据和3维数据的在PCA前后的对比实验。

2维数据的测试

在2维数据的测试中，选择使用2维高斯分布产生样本，使用的参数为：

$$\mathbf{mean} = [-2, 2], \mathbf{cov} = \begin{bmatrix} 0.01 & 0 \\ 0 & 1 \end{bmatrix}$$

可以看到第1维的方差远小于第2维的方差($0.01 \ll 1$)，因此有直观感觉在第2维包含了更多的信息，所以直接进行PCA，得到的结果如下：



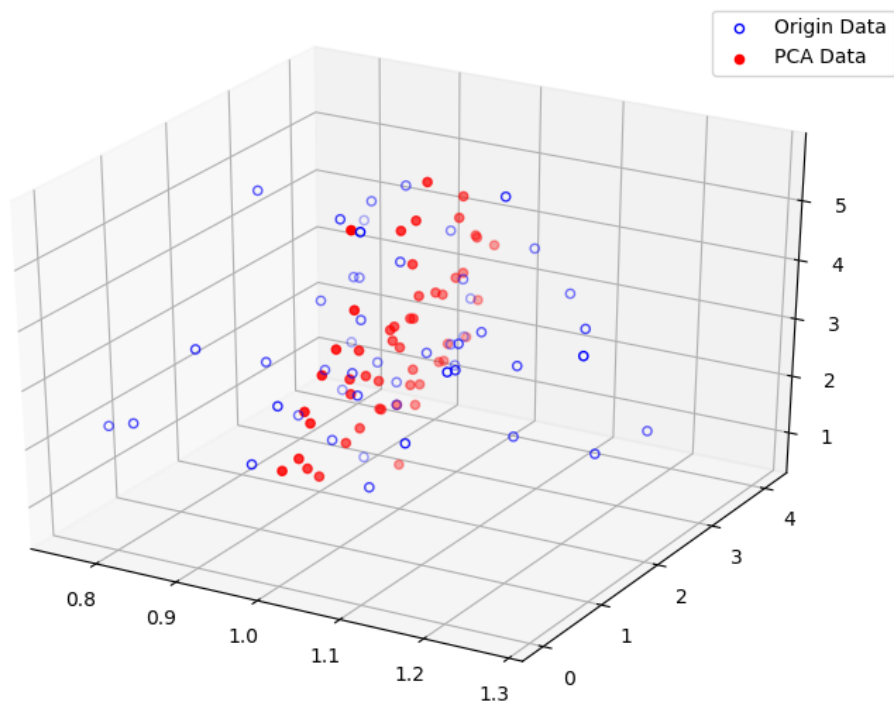
可以看到在PCA之后的数据分布在直线(1维)上，另外其在横轴上的方差更大，纵轴上的方差更小(注意横轴纵轴在单位长度上表示的大小不同)，所以在进行PCA之后得到的直线与横轴接近。

3维数据的测试

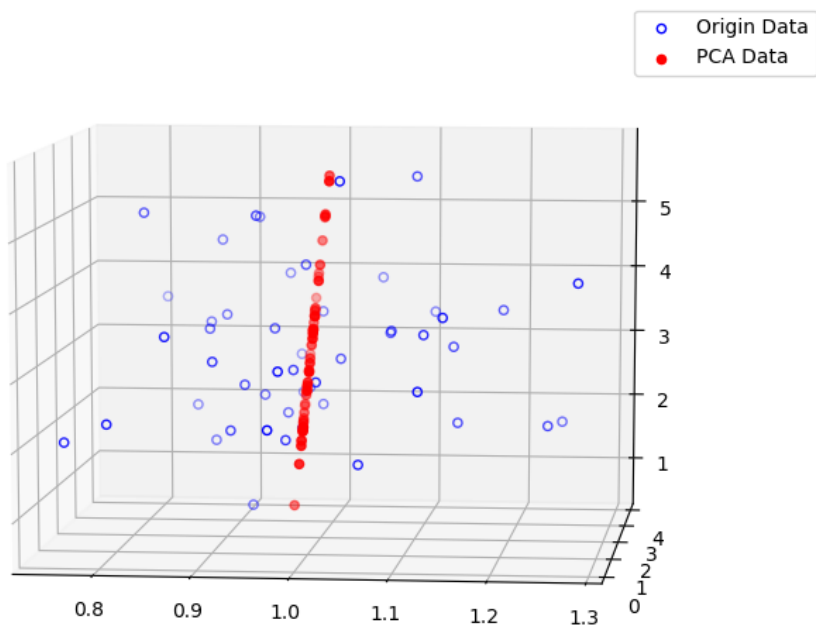
在3维数据的测试中，使用3维高斯分布随机产生样本，使用的参数为：

$$\mathbf{mean} = [1, 2, 3], \mathbf{cov} = \begin{bmatrix} 0.01 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

同样，可以看到第1维的方差是远小于其余两个维度的，所以在第1维相较于其他两维信息更少，进行PCA得到的结果如下：

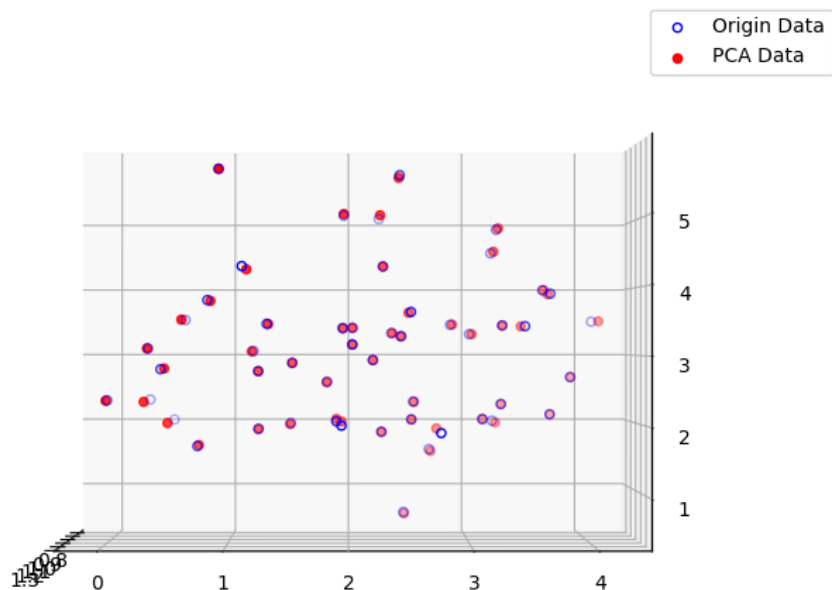


可以看到在底面的一个轴的单位长度表示的长度更小，即在原始数据上的第1维数据，对上面的图片进行旋转，我们可以看到：



降维后的数据分布在一个平面(2维)上, 并且与方差最小的1维相垂直。

对比其他方向, 可以看到经过PCA将样本数据进行了投影, 投影在了一个平面上, 如下图所示。



2.mnist手写数据集测试

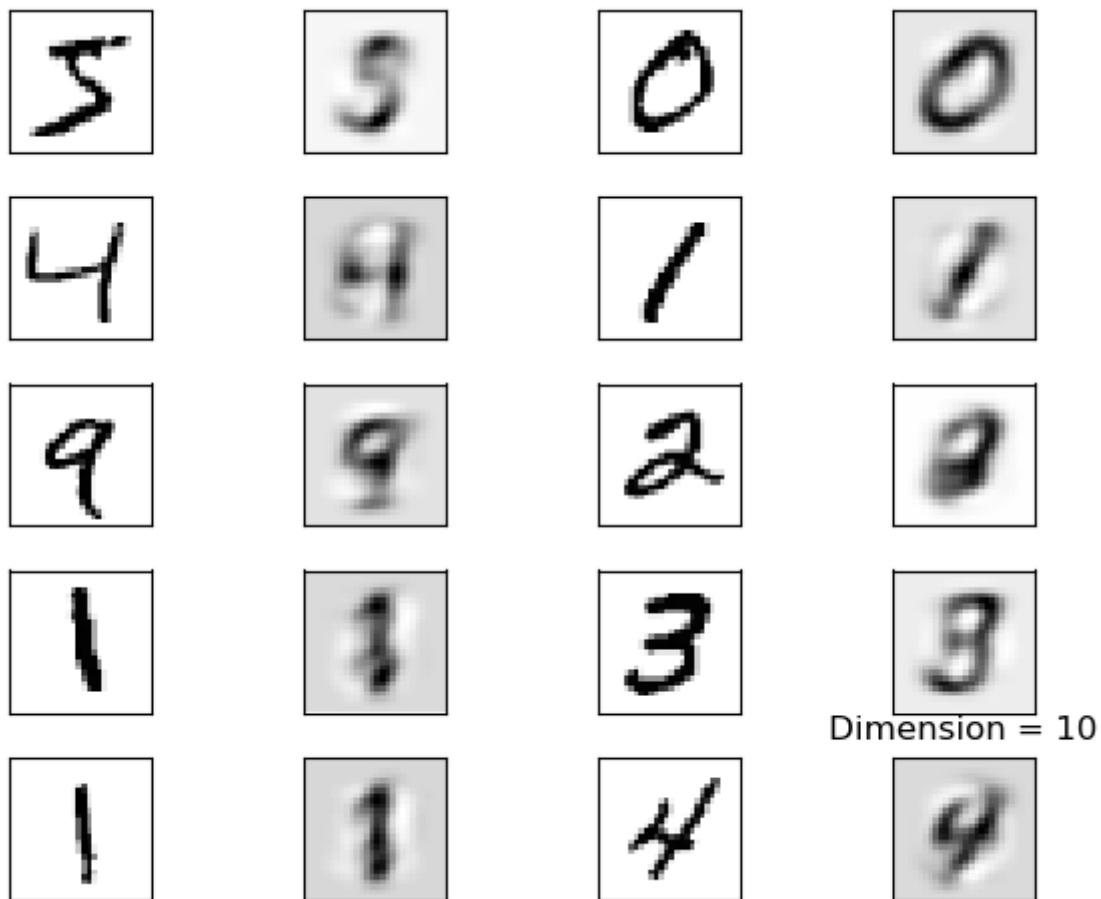
MNIST数据集来自美国国家标准与技术研究所(National Institute of Standards and Technology), 在本次实验中仅使用了其中的训练集(training set)部分, 来自250个不同人手写的数组构成, 其中50%为高中生, 50%来自人口普查局的工作人员。

图片是以字节的形式进行存储的, 训练集包括60000个样本。每张图片由 28×28 个像素点组成, 每个像素点用一个灰度值表示, 总的来说每个样本有784个属性。

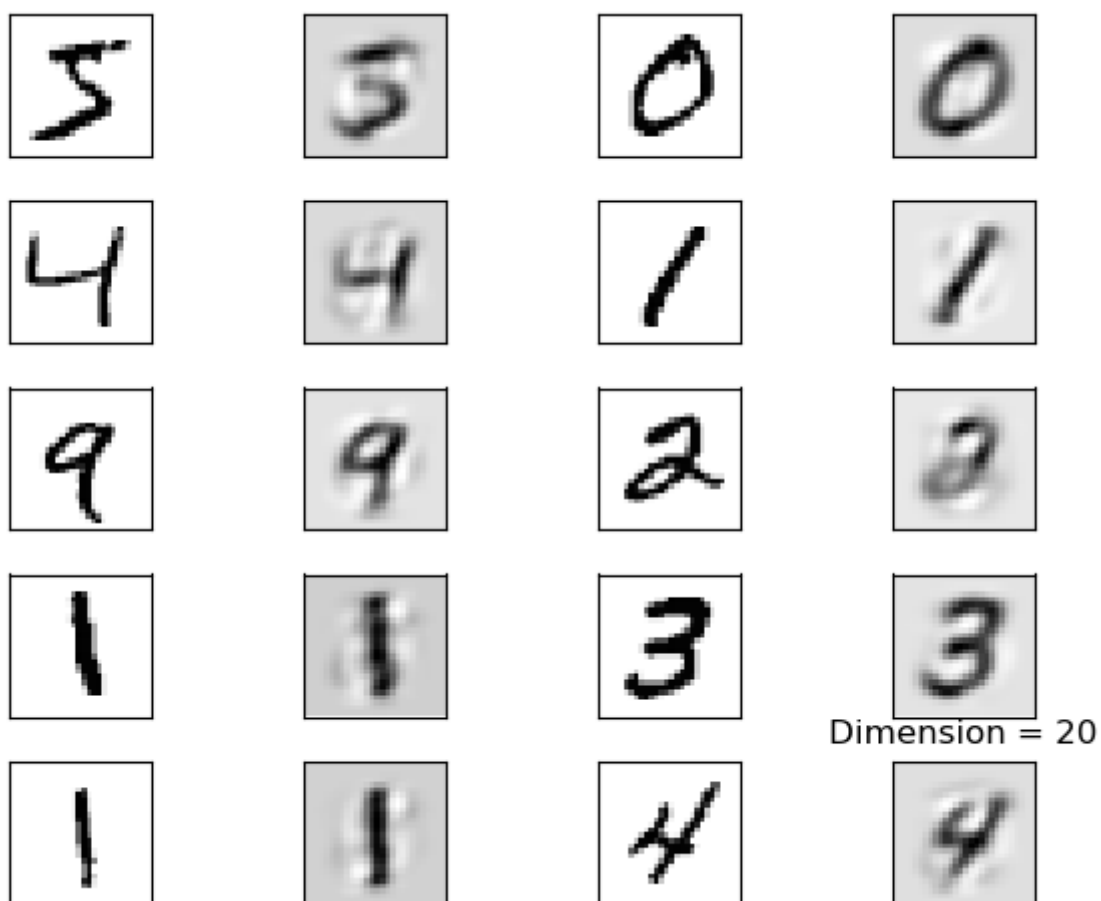
在读取时, 我们使用训练集, 分别得到训练集矩阵(60000×784), 每一行代表一张图片, 训练集对应的label(60000×1), 每一行为0~9, 表示对应行的图片代表的数字。

在训练时, 我们将784维的数据分别降维到10, 20, 30, 60, 100维, 并对其对应的信噪比进行对比, 得到下面的结果。

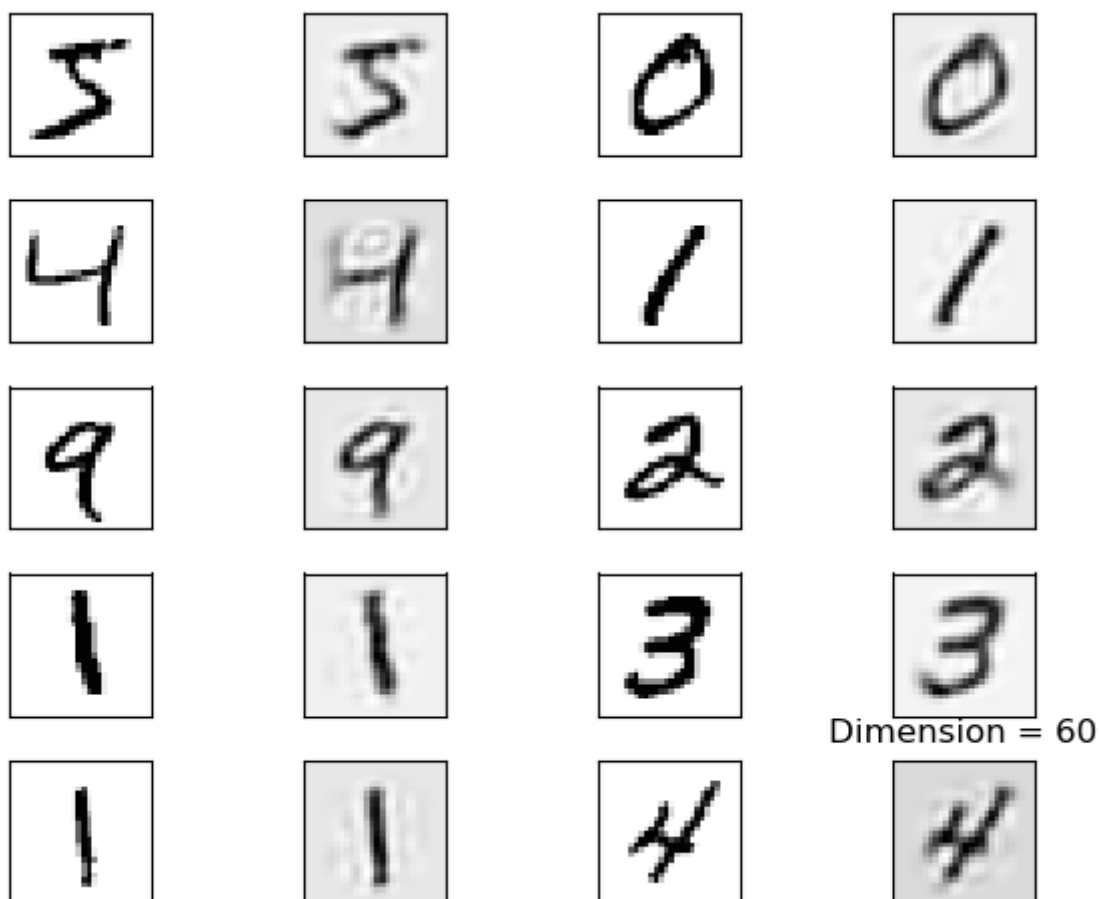
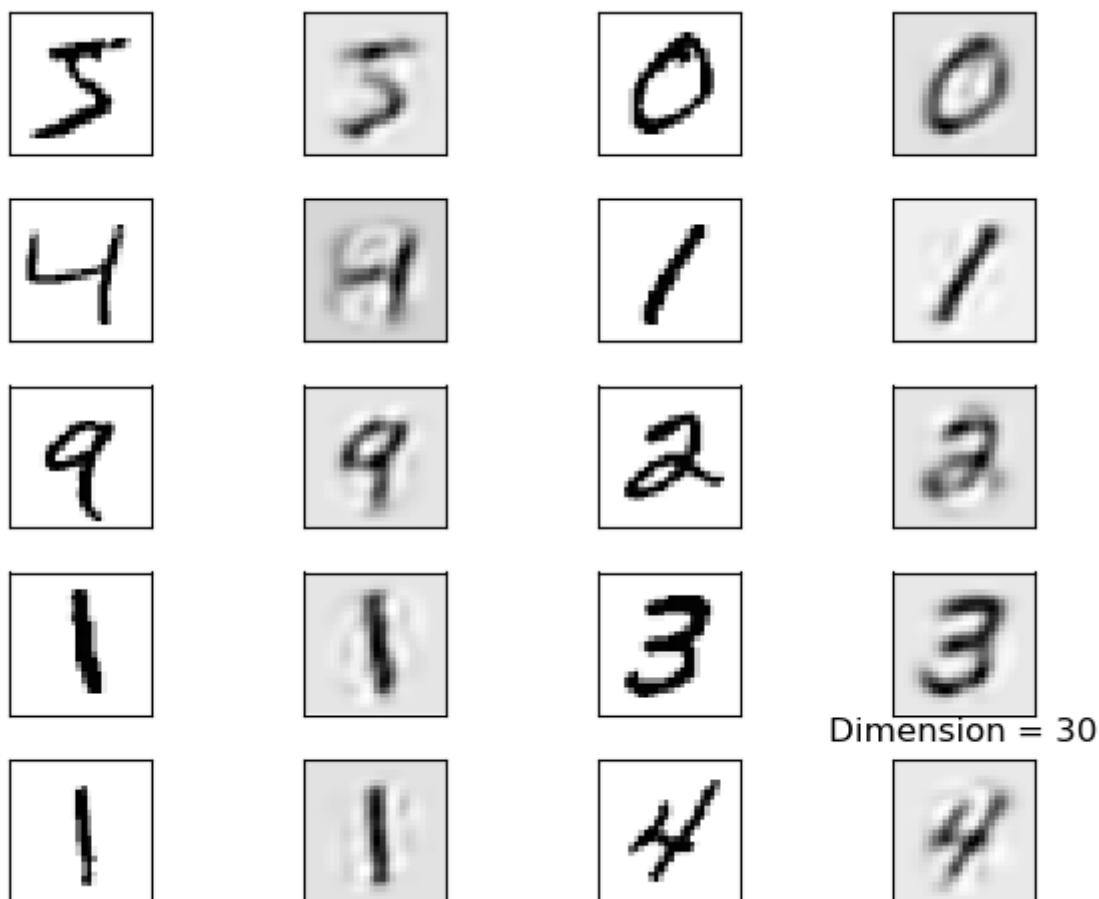
每张图片左侧为**784**维原始数据显示结果, 右侧为对应的**PCA**之后的图像。

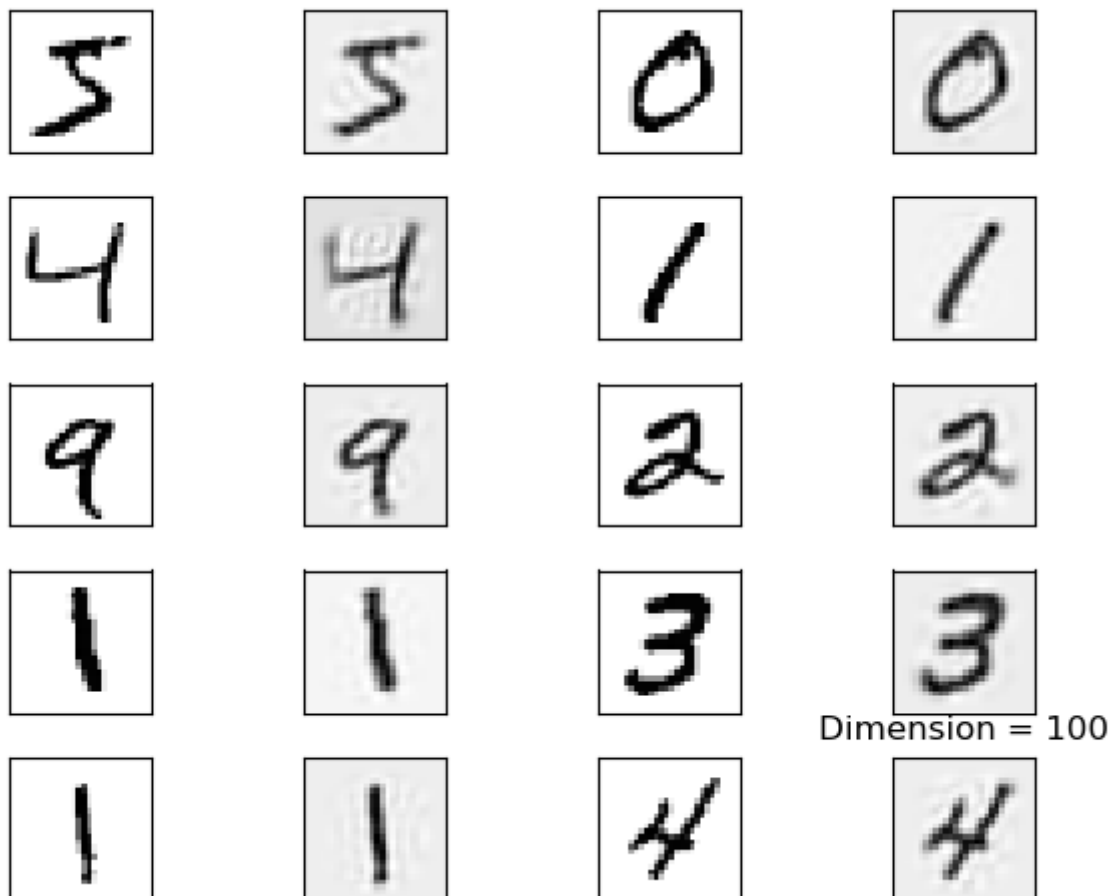


降维到10的时候，有些数字已经可以大致分，如0，1，但是对于其余数字还不能区分。



降维到20维时，可以看到又有一些数字，比如数字9，3已经可以分辨了，但是仍然有些数字比较模糊，特别是这里的2，进一步提高低维空间的维数。





可以看到，随着低维空间的维数提高，对于源数据的信息保留的更加全面。

使用的信噪比的公式为：

$$MSE = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} ||I(i, j) - K(i, j)||^2$$

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

下面是不同维数下信噪比的记录，可以观察到随着低维空间的维数升高，信噪比在下降。

Dimension	10	20	30	60	100
PSNR	33.25	31.60	30.37	27.74	25.36

五、结论

- PCA算法中舍弃了 $n - d$ 个最小的特征值对应的特征向量，一定会导致低维空间与高维空间不同，但是通过这种方式有效提高了样本的采样密度；并且由于较小特征值对应的往往与噪声相关，通过PCA在一定程度上起到了降噪的效果。
- PCA降低了训练数据的维度同事保留了主要信息，但在训练集上的主要信息未必是重要信息，被舍弃掉的信息未必无用，只是在训练数据上没有表现，因此PCA也有可能加重了过拟合。
- 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。
- 保留均值向量，能够通过向量减法将新样本进行中心化。

六、参考文献

- [THE MNIST DATABASE of handwritten digits](#)
- [Christopher Bishop. Pattern Recognition and Machine Learning.](#)
- 周志华 著. 机器学习, 北京: 清华大学出版社, 2016.1
- [AI算法工程师手册 数据降维](#)

七、附录:源代码(带注释)

仅有lab4.py文件，见压缩包