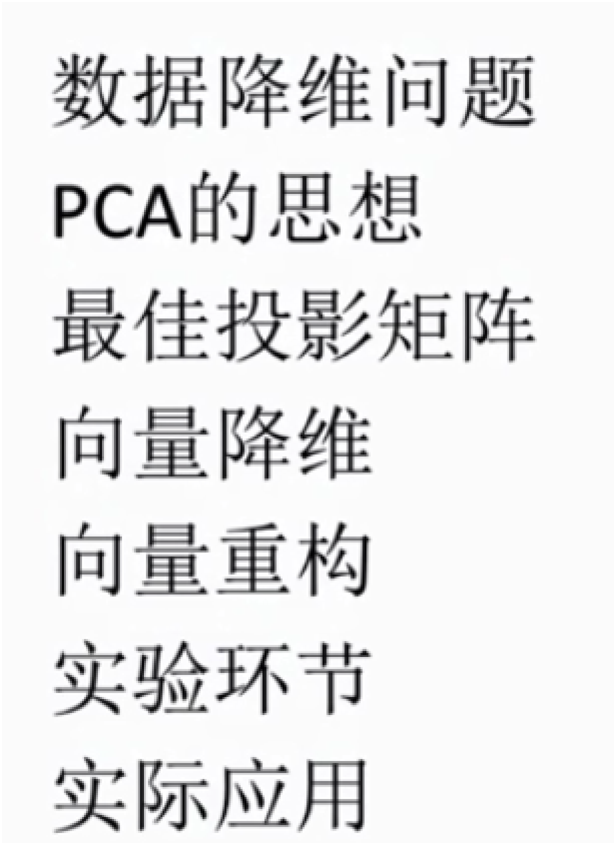
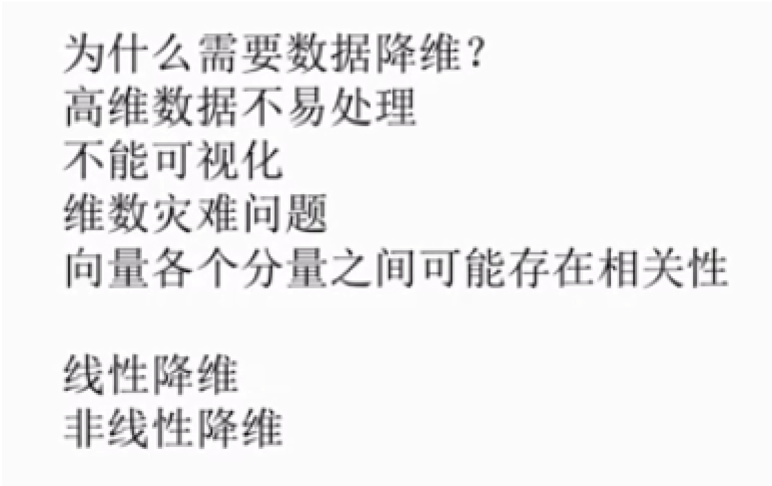
8 数据降维1 线性降维

8.1 本集内容简介



线性降维，典型的是PCA主成分分析

8.2 为什么需要数据降维

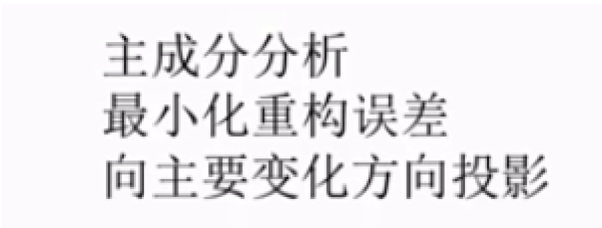


1. 高维的数组不容易处理
2. 数据无法可视化
3. 维数灾难
4. 特征各个分量之间可能存在相关性

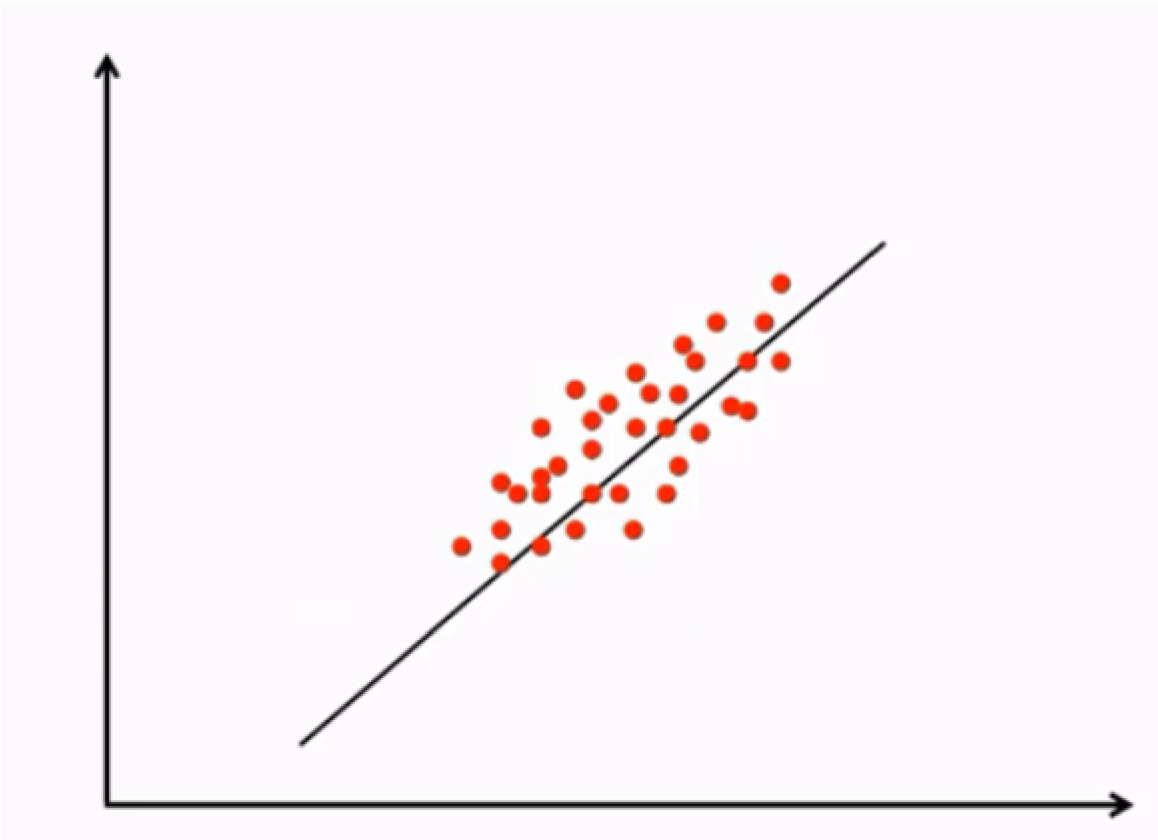
数据降维可以是有监督学习和无监督学习

1. 线性降维
2. 非线性降维

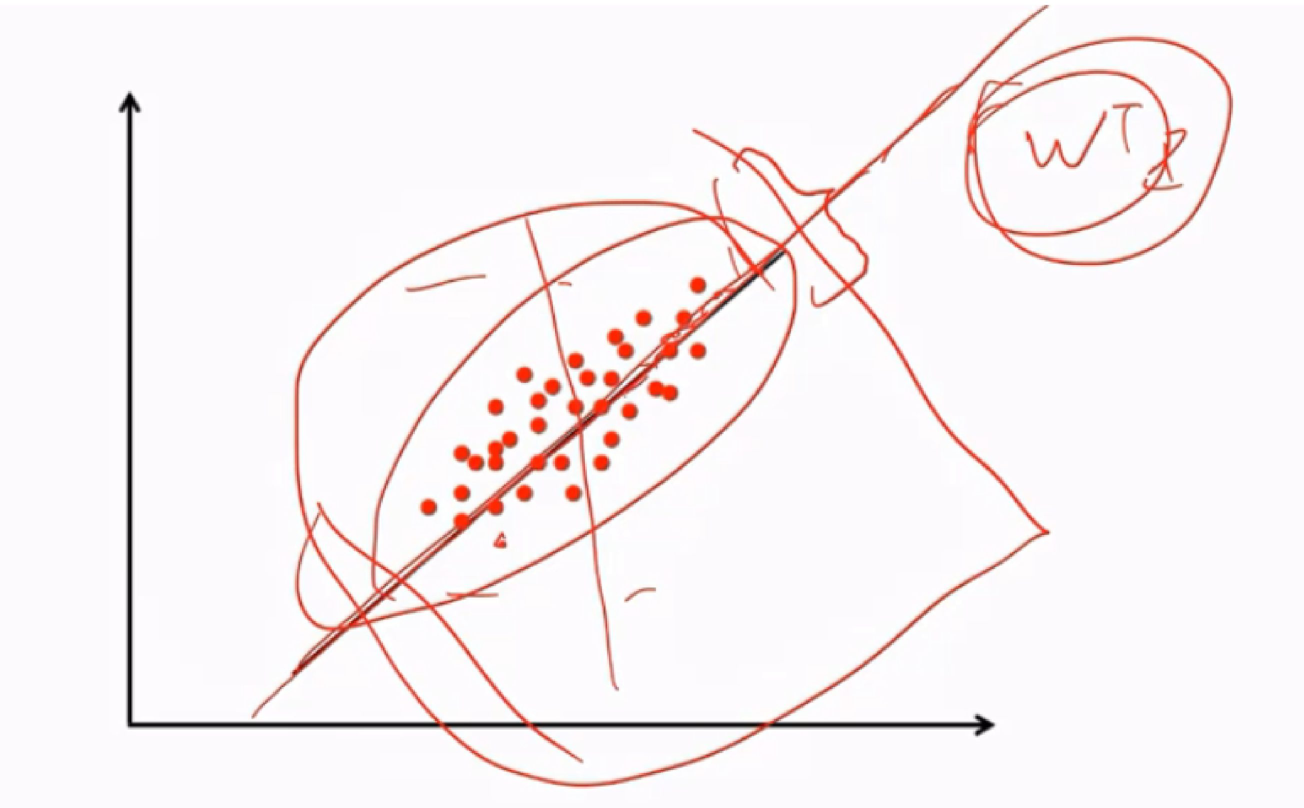
8.3 PCA简介



PCA 基于最小化重构误差



PCA主要思想，向数组主要变化方向投影。让W的转置和X做内积。



8.4 计算投影矩阵

$$\mathrm{x}_i = \mathrm{m} + a_i \mathrm{e}$$

$$L(a, \mathrm{e}) = \sum_{i=1}^{n} \left\| \mathrm{m} + a_i \mathrm{e} - \mathrm{x}_i \right\|^2$$

$$2\mathrm{e}^{\mathrm{T}} \left(\mathrm{m} + a_i \mathrm{e} - \mathrm{x}_i \right) = 0$$

$$a_i \mathrm{e}^{\mathrm{T}} \mathrm{e} = \mathrm{e}^{\mathrm{T}} \left(\mathrm{x}_i - \mathrm{m} \right)$$

$$a_i = \mathrm{e}^{\mathrm{T}} \left(\mathrm{x}_i - \mathrm{m} \right)$$

这里有一部分

投影到D维的空间。

$$\mathbf{x} = \mathbf{m} + \sum_{i=1}^{d'} a_i \mathbf{e}_i$$

$$L = \sum_{i=1}^{n} \left\| \mathbf{m} + \sum_{j=1}^{d'} a_{ij} \mathbf{e}_j - \mathbf{x}_i \right\|^2$$

$$\min_{\mathbf{W}} -\mathrm{tr}\left(\mathbf{W}^{\mathrm{T}} \mathbf{S} \mathbf{W}\right)$$
$$\mathbf{W}^{\mathrm{T}} \mathbf{W} = \mathbf{I}$$

8.5 完整的算法流程

计算投影矩阵的流程

1. 白化
2. 计算样本协方差矩阵
3. 计算协方差矩阵的特征值与特征向量

计算投影矩阵的流程：

1.计算所有样本向量的均值向量，并将所有向量减去均值向量

2.计算样本协方差矩阵

3.计算协方差矩阵的特征值与特征向量

4.将特征值从大到小排序，保留最大的一部分特征值和特征向量，构成投影矩阵

PCA投影的流程

PCA投影的流程：

1.计算所有样本的均值向量

2.所有样本减掉均值向量，然后再计算协方差矩阵

3.对协方差矩阵进行特征值分解，得到特征值和对应的特征向量

4.将减掉均值后的向量与特征向量矩阵相乘，得到投影结果

向量重构

让w 的转置 左乘y 再加上

计算投影矩阵的流程：

1.计算所有样本向量的均值向量，并将所有向量减去均值向量

2.计算样本协方差矩阵

3.计算协方差矩阵的特征值与特征向量

4.将特征值从大到小排序，保留最大的一部分特征值和特征向量，构成投影矩阵

PCA投影的流程：

1.计算所有样本的均值向量

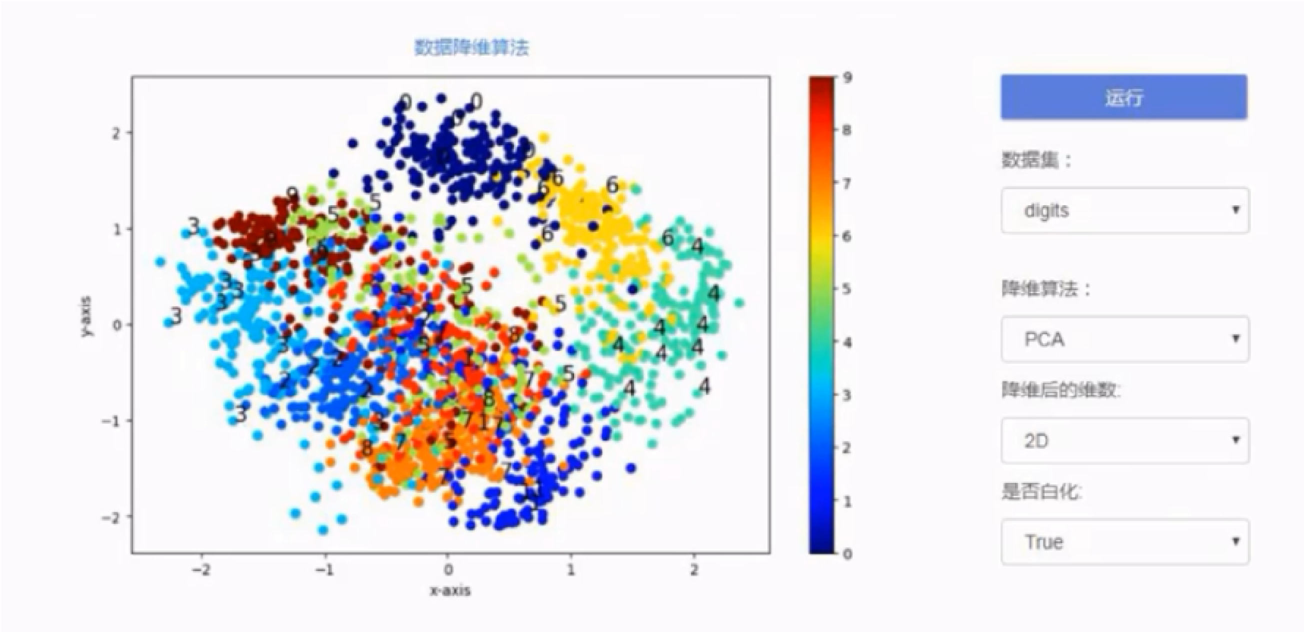
2.所有样本减掉均值向量，然后再计算协方差矩阵

3.对协方差矩阵进行特征值分解，得到特征值和对应的特征向量

4.将减掉均值后的向量与特征向量矩阵相乘，得到投影结果

8.6 实验环节

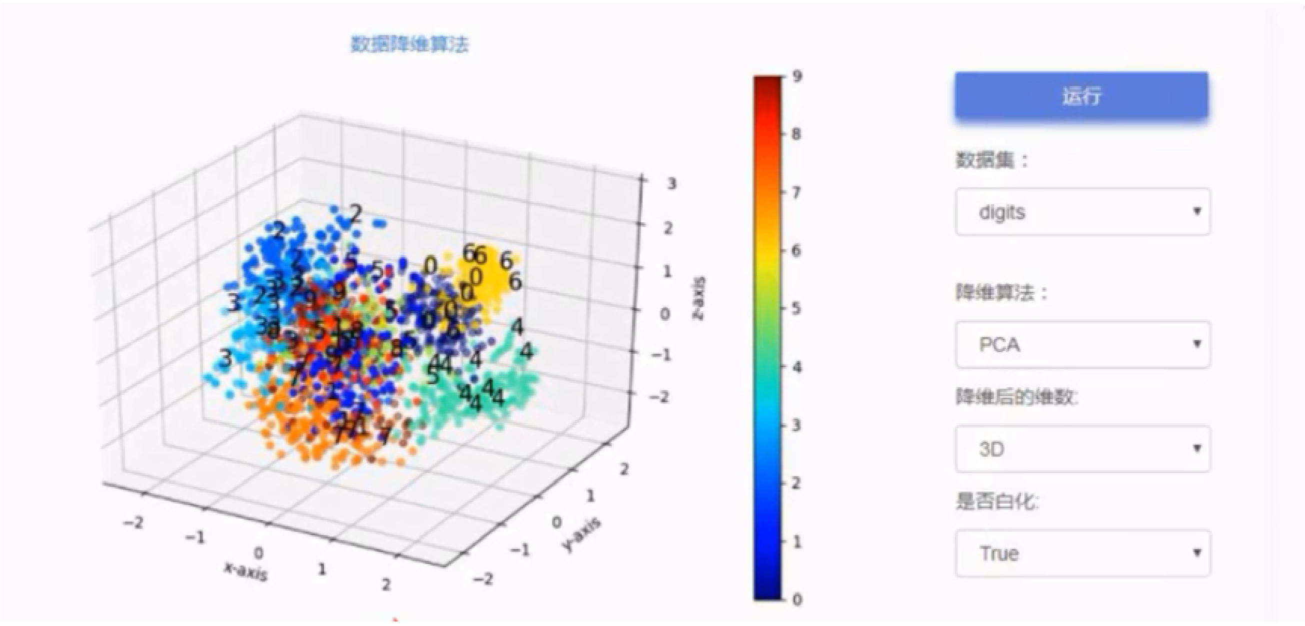
在云端实验室里面



- 0到9的数字
- 人脸

注意是否要白化

这里是投影到三维



8.8 降维总结

1. why 降维
2. PCA
3. 实验
4. 数字分类和人脸识别中的应用

主成分分析的优化目标是最小化重构误差，即用投影到低维空间中的向量近似重构原始向量，二者之间的误差要尽可能的小。最小化重构误差的目标为：

$$\min \sum_{i=1}^{n} \left\| \mathrm{m} + \sum_{j=1}^{d'} a_{ij} \mathrm{e}_j - \mathrm{x}_i \right\|^2$$