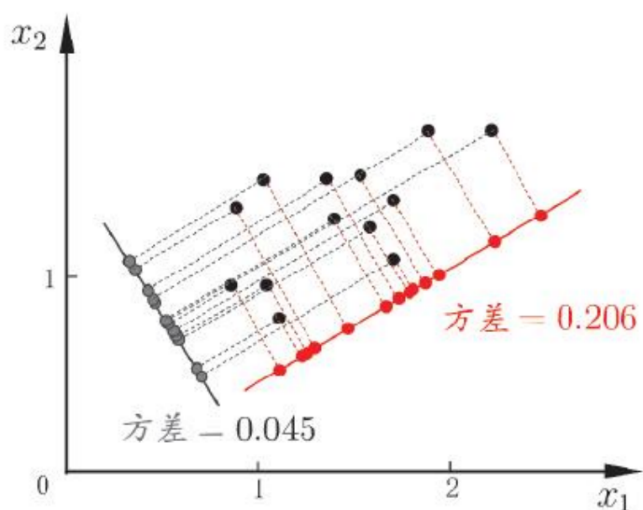


PCA是无监督方法：只有x没有y，没有标记，没有类别信息先验
对高维中的冗余、噪声去除，提取本真属性

搞清楚目的：不同后续任务导致不同降维方法
把高维空间中的低维结构恢复出来

一、两种基础思路

- 最大可分性：样本点在超平面上的投影尽可能分开
- 最近重构性：样本点到超平面距离都足够近



$$\begin{aligned} \min_{\mathbf{W}} \quad & -\text{tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W}) \\ \text{s.t.} \quad & \mathbf{W}^T \mathbf{W} = \mathbf{I}. \end{aligned}$$

- 求tr：把矩阵变成标量可优化
- 给W正交的限制，避免无穷大

对式(10.15)或(10.16)使用拉格朗日乘子法可得

$$\mathbf{X} \mathbf{X}^T \mathbf{W} = \lambda \mathbf{W}, \quad (10.17)$$

于是，只需对协方差矩阵 $\mathbf{X} \mathbf{X}^T$ 进行特征值分解，将求得特征值排序： $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$ ，再取前 d' 个特征值对应的特征向量构成 $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{d'})$ 。这就是主成分分析的解。PCA 算法描述如图 10.5 所示。

最大可分性

投影方差最大化

$$\begin{aligned} \max_{\mathbf{W}} \quad & \text{tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W}) \\ \text{s.t.} \quad & \mathbf{W}^T \mathbf{W} = \mathbf{I}, \end{aligned}$$

最近重构性

丢掉维度后变化最小

假定投影变换后得到的新坐标系为 $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d\}$, 其中 \mathbf{w}_i 是标准正交基向量

$$\|\mathbf{w}_i\|_2 = 1, \mathbf{w}_i^T \mathbf{w}_j = 0 (i \neq j).$$

若丢弃新坐标系中的部分坐标, 即将维度降低到 $d' < d$, 则样本点在低维坐标系中的投影是 $\mathbf{z}_i = (z_{i1}; z_{i2}; \dots; z_{id'})$ $z_{ij} = \mathbf{w}_j^T \mathbf{x}_i$

若基于 \mathbf{z}_i 来重构 \mathbf{x}_i , 则会得到 $\hat{\mathbf{x}}_i = \sum_{j=1}^{d'} z_{ij} \mathbf{w}_j$.

- 重构误差用平方损失

考虑整个训练集, 原样本点 \mathbf{x}_i 与基于投影重构的样本点 $\hat{\mathbf{x}}_i$ 之间的距离为

$$\begin{aligned} \sum_{i=1}^m \left\| \sum_{j=1}^{d'} z_{ij} \mathbf{w}_j - \mathbf{x}_i \right\|_2^2 &= \sum_{i=1}^m \mathbf{z}_i^T \mathbf{z}_i - 2 \sum_{i=1}^m \mathbf{z}_i^T \mathbf{W}^T \mathbf{x}_i + \text{const} \\ &\propto -\text{tr} \left(\mathbf{W}^T \left(\sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{W} \right). \end{aligned} \quad (10.14)$$

\mathbf{w}_j 是正交基, $\sum_i \mathbf{x}_i \mathbf{x}_i^T$ 是协方差矩阵

方差构成了对角线上的元素, 协方差构成了非对角线上的元素

$$\begin{aligned} \min_{\mathbf{W}} \quad & -\text{tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W}) \\ \text{s.t.} \quad & \mathbf{W}^T \mathbf{W} = \mathbf{I}. \end{aligned}$$

二、计算方法

基于特征值分解协方差矩阵

输入：样本集 $D = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$;
低维空间维数 d' .

过程:

- 1: 对所有样本进行中心化: $\mathbf{x}_i \leftarrow \mathbf{x}_i - \frac{1}{m} \sum_{i=1}^m \mathbf{x}_i$;
- 2: 计算样本的协方差矩阵 \mathbf{XX}^T ;
- 3: 对协方差矩阵 \mathbf{XX}^T 做特征值分解;
- 4: 取最大的 d' 个特征值所对应的特征向量 $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{d'}$.

输出：投影矩阵 $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{d'})$.

图 10.5 PCA 算法

1、先中心化：样本均值为0，每一位特征减去各自的平均值

算方差时不用减去均值

对验证集、测试集执行零均值化操作时，均值必须从训练集计算而来

2、从原数据集X直接得到协方差矩阵（均值已消去）

假设我们只有 a 和 b 两个变量，那么我们将它们按行组成矩阵 X：

$$X = \begin{pmatrix} a_1 & a_2 & \cdots & a_m \\ b_1 & b_2 & \cdots & b_m \end{pmatrix}$$

然后：

$$\frac{1}{m} X X^T = \begin{pmatrix} \frac{1}{m} \sum_{i=1}^m a_i^2 & \frac{1}{m} \sum_{i=1}^m a_i b_i \\ \frac{1}{m} \sum_{i=1}^m a_i b_i & \frac{1}{m} \sum_{i=1}^m b_i^2 \end{pmatrix} = \begin{pmatrix} \text{Cov}(a, a) & \text{Cov}(a, b) \\ \text{Cov}(b, a) & \text{Cov}(b, b) \end{pmatrix}$$

我们可以看到这个矩阵对角线上的分别是两个变量的方差，而其它元素是 a 和 b 的协方差。两者被统一到了一个矩阵里。

这里除或不除样本数量n或n-1,其实对求出的特征向量没有影响

3、特征值分解求协方差矩阵的特征值和特征向量

4、求解W：选最大的d'个特征值对应的特征向量，作为行向量组成W

工具包指令了前d'可有快速方法

5、Y=WX得到降维后结果

逐一选取最大方差方向

PCA 也可看作是逐一选取方差最大方向, 即先对协方差矩阵 $\sum_i \mathbf{x}_i \mathbf{x}_i^T$ 做特征值分解, 取最大特征值对应的特征向量 \mathbf{w}_1 ; 再对 $\sum_i \mathbf{x}_i \mathbf{x}_i^T - \lambda_1 \mathbf{w}_1 \mathbf{w}_1^T$ 做特征值分解, 取最大特征值对应的特征向量 \mathbf{w}_2 ; 由 \mathbf{W} 各分量正交及

$$\sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i^T = \sum_{j=1}^d \lambda_j \mathbf{w}_j \mathbf{w}_j^T$$

可知, 上述逐一选取方差最大方向的做法与直接选取最大 d' 个特征值等价.

基于SVD分解协方差矩阵

特征值和特征向量是针对方阵才有的, 任意形状矩阵都可以做奇异值分解

数据标准化

协方差矩阵易受特征尺度影响 (取值范围)

超参数 d' 确定

- 用户指定
- 交叉验证确定, 重构误差可能是U型曲线
- 设置重构阈值: 占用95%的特征值大小, 选取下式成立最小的 d'

$$\frac{\sum_{i=1}^{d'} \lambda_i}{\sum_{i=1}^d \lambda_i} \geq t .$$

三、例题&分析

算法示例

$$X = \begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}$$

以X为例，我们用PCA方法将这两行数据降到一行。

1)因为X矩阵的每行已经是零均值，所以不需要去平均值。

2)求协方差矩阵：

$$C = \frac{1}{5} \begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} -1 & -2 \\ -1 & 0 \\ 0 & 0 \\ 2 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{6}{5} & \frac{4}{5} \\ \frac{4}{5} & \frac{6}{5} \end{pmatrix}$$

3)求协方差矩阵的特征值与特征向量。

求解后的特征值为：

$$\lambda_1 = 2, \quad \lambda_2 = \frac{2}{5}$$

对应的特征向量为：

$$c_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad c_2 \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

其中对应的特征向量分别是一个通解， c_1 和 c_2 可以取任意实数。那么标准化后的特征向量为：

$$\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \quad \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

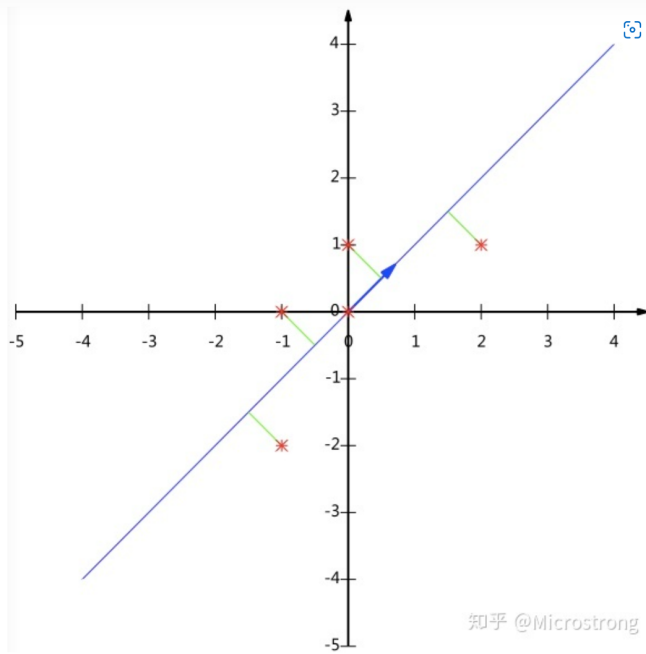
4)矩阵P为：

$$P = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

5)最后我们用P的第一行乘以数据矩阵X，就得到了降维后的表示：

$$Y = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} -\frac{3}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \frac{3}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}$$

• 示意图：



数据矩阵X降维投影结果

注意：如果我们通过特征值分解协方差矩阵，那么我们只能得到一个方向的PCA降维。这个方向就是对数据矩阵X从行(或列)方向上压缩降维。

有关协方差

设原始数据矩阵 X 对应的协方差矩阵为 C ，而 P 是一组基按行组成的矩阵，设 $Y=PX$ ，则 Y 为 X 对 P 做基变换后的数据。设 Y 的协方差矩阵为 D ，我们推导一下 D 与 C 的关系：

$$\begin{aligned}
 D &= \frac{1}{m} Y Y^T \\
 &= \frac{1}{m} (P X) (P X)^T \\
 &= \frac{1}{m} P X X^T P^T \\
 &= P \left(\frac{1}{m} X X^T \right) P^T \\
 &= P C P^T
 \end{aligned}$$

这样我们就看清楚了，我们要找的 P 是能让原始协方差矩阵对角化的 P 。换句话说，优化目标变成了寻找一个矩阵 P ，满足 $P C P^T$ 是一个对角矩阵，并且对角元素按从大到小依次排列，那么 P 的前 K 行就是要寻找的基，用 P 的前 K 行组成的矩阵乘以 X 就使得 X 从 N 维降到了 K 维并满足上述优化条件。

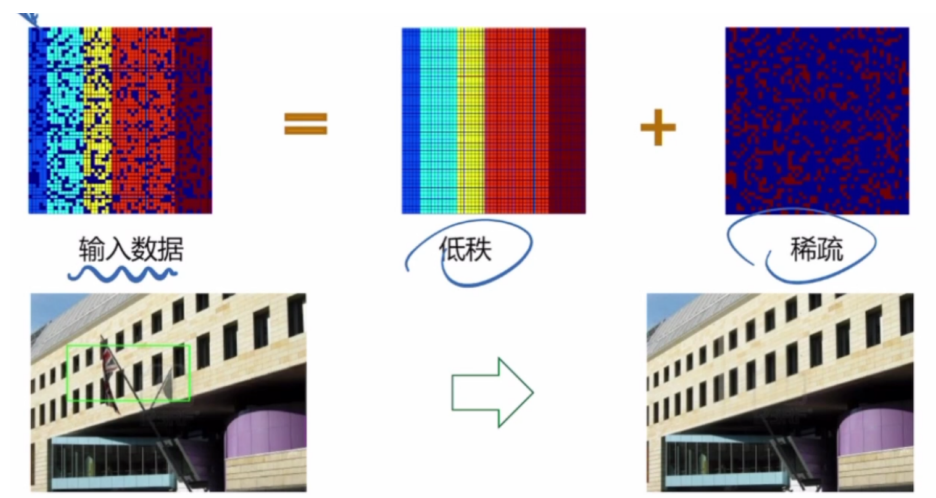
四、PCA应用

- 人脸识别中的特征脸，是降维后的特征向量
- 识别图片：使用均值和几个特征的线性组合即可表示所有可能的

Robust PCA:

不但要找到低秩重构，还要重构误差是稀疏的（少量非0）

用于图片去噪，人脸去墨镜



函数推广

找到函数空间的基

傅里叶变换：轻量级处理大规模核函数