降维

在特征向量所处的高维空间中，包含很多的冗余和噪音，我们希望通过降维的方式来寻找数据内部的特征，从而提升特征的表达能力，降低训练复杂度。

常见的降维方法：

- 主成分分析（Principal Component Analysis）
- 线性判别分析（Linear Discriminant Analysis）
- 等距映射（Isometry）
- 局部先行嵌入（Locally Linear Embedding）
- 拉普拉斯特征映射（Laplacian Eigenmaps）

PCA主成分分析

1. **什么是PCA？什么是主成分？**
    - PCA是一种线性、非监督、全局的降维算法
    - PCA旨在找到数据的主成分，通过主成分来表征原始数据
    - 比如在三维空间中的数据点分布在一个过原点的平面上，那么我们可以通过旋转坐标轴，使得xy轴与平面重合，就可以只通过两个维度来表示这些点，旋转后的坐标轴包含的信息就是主成分
    - 主成分是特征向量变化最大的方向（the direction of largest variation），也可以理解为数据在这个方向上的投影分布得更加分散（方差更大），所以PCA其实是找最佳投影方向
    - 主成分可以使数据的信噪比最大化，因为信号的方差越大，噪声的方差越小

2. **如何通过最大方差理论找到主成分？PCA如何求解？**
   - 先将数据进行去中心化：$x_1, x_2, ..., x_n = v_1 - \mu, v_2 - \mu, ..., v_n - \mu$
   - 向量内积是第一个向量投影到第二个向量上的长度，所以向量x_i在ω上的投影坐标为$x_i^T \omega$
   - PCA的目标是找到一个投影方向ω使得投影后的点$x_i^T \omega$的方差最大
   - 因为去中心化后投影点的均值为0，所以投影点的方差为：

$$D(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{\omega})^2 = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{\omega})^{\mathrm{T}}(\boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{\omega})$$

$$= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}_i\,\boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{\omega}$$

$$= \boldsymbol{\omega}^{\mathrm{T}}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\,\boldsymbol{x}_i{}^{\mathrm{T}}\right)\boldsymbol{\omega}.$$

     - 括号中的部分其实是数据的协方差矩阵（covariance matrix）
   - 因此PCA的优化问题变成了找到方向ω使得：

$$\begin{cases} \max\{\boldsymbol{\omega}^{\mathrm{T}}\Sigma\boldsymbol{\omega}\}, \\ s.t. \quad \boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega} = 1. \end{cases}$$

     - 限制：ω是归一化后的单位方向向量，不然max函数只会让ω越来越大
   - 对ω求导后可以知道，x投影后的方差就是协方差矩阵的特征值（eigine value），所以求投影点的最大方差，就是求x协方差矩阵的的最大特征值
   - 次佳的投影方向则位于最佳投影方向的正交空间（orthogonal）
   - 我们将x协方差矩阵的特征值从大到小排序，取前d大的特征值对应的特征向量，即为我们的主成分$w_1, w_2, ..., w_d$
   - 最后通过向量内积，将n维的样本映射到d维中：

$$\boldsymbol{x}_i' = \begin{bmatrix} \boldsymbol{\omega}_1{}^{\mathrm{T}}\boldsymbol{x}_i \\ \boldsymbol{\omega}_2{}^{\mathrm{T}}\boldsymbol{x}_i \\ \vdots \\ \boldsymbol{\omega}_d{}^{\mathrm{T}}\boldsymbol{x}_i \end{bmatrix}.$$

   - 降维后的信息占比为：

$$\eta = \sqrt{\frac{\sum_{i=1}^{d}\lambda_i{}^2}{\sum_{i=1}^{n}\lambda_i{}^2}}.$$

- PCA是线性降维法，可以通过核方法来应对非线性可投影的数据

3. **如何从回归的角度（最小平方误差和理论）定义PCA的目标并求解？**
    - PCA可以理解为在高维空间中，找到一个d维超平面，使得样本点到这个超平面的距离平方和最小，点x_k到超平面D的距离为：

    $$\text{distance}(\boldsymbol{x}_k, \boldsymbol{D}) = \| \boldsymbol{x}_k - \widetilde{\boldsymbol{x}_k} \|_2,$$

    - 最佳投影超平面由d个正交基构成$W = \{w_1, w_2, ..., w_d\}$，点$x_k$在超平面$D$上的投影是：

    $$\widetilde{\boldsymbol{x}_k} = \sum_{i=1}^{d} (\boldsymbol{\omega}_i^{\mathrm{T}} \boldsymbol{x}_k) \boldsymbol{\omega}_i,$$

    - 从这个角度出发，PCA的优化目标是：

    $$\begin{cases} \underset{\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_d}{\arg\min} \sum_{k=1}^{n} \| \boldsymbol{x}_k - \widetilde{\boldsymbol{x}_k} \|_2^2, \\ s.t. \quad \underset{\forall i,j}{\boldsymbol{\omega}_i^{\mathrm{T}} \boldsymbol{\omega}_j} = \delta_{ij} = \begin{cases} 1, \ i = j; \\ 0, \ i \neq j. \end{cases} \end{cases}$$

    - 可以简化为求矩阵的迹（对角线元素之和）：

    $$\begin{cases} \underset{\boldsymbol{W}}{\arg\max} \ tr(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{W}), \\ s.t. \quad \boldsymbol{W}^{\mathrm{T}} \boldsymbol{W} = I. \end{cases}$$

LDA线性判别分析

PCA没有考虑数据的标签，只是把原数据映射到方差较大的方向上，映射后不同类别的数据会完全混合在一起，很难被区分开。

1. **对于有类别标签的数据，如何设计目标函数使得降维的过程中不损失类别信息？**
   - 找到一个投影方向w，使得投影后的样本尽可能按照原始分类分开
   - 例如二分类任务，我们希望最大化投影点的均值差：

     $$D(C_1,C_2)=\|\widetilde{\boldsymbol{\mu}_1}-\widetilde{\boldsymbol{\mu}_2}\|_2^2$$

     $$\boldsymbol{\mu}_1=\frac{1}{N_1}\sum_{x\in C_1}\boldsymbol{x},\quad \boldsymbol{\mu}_2=\frac{1}{N_2}\sum_{x\in C_2}\boldsymbol{x}$$

     $$\widetilde{\boldsymbol{\mu}_1}=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\mu}_1,\quad \widetilde{\boldsymbol{\mu}_2}=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\mu}_2,$$

   - 因此，LDA的优化目标是：

     $$\begin{cases}\max\limits_{\boldsymbol{\omega}}\|\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\|_2^2, \\ s.t.\quad \boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega}=1.\end{cases}$$

   - 但只是最大化两类点投影后的中心距离，投影后的点会有重叠的部分，所以我们还需要最小化两类点投影后的各自的内部方差，即：最大化类间距离和最小化类内距离
   - 所以LDA的目标函数是类间距离和类内距离的比值：

     $$\max_{\boldsymbol{\omega}} J(\boldsymbol{\omega})=\frac{\|\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\|_2^2}{D_1+D_2},$$

     - D_1和D_2是两类点投影后的内部方差：

       $$D_1=\sum_{\boldsymbol{x}\in \boldsymbol{C}_1}(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}-\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\mu}_1)^2=$$

       $$\sum_{\boldsymbol{x}\in \boldsymbol{C}_1}\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\mu}_1)(\boldsymbol{x}-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{\omega},$$

   - 定义类内散度矩阵S_B和类内散度矩阵S_w，则目标函数可以被简化为：

     $$J(\boldsymbol{\omega})=\frac{\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\omega}}{\sum_{\boldsymbol{x}\in \boldsymbol{C}_i}\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\mu}_i)(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\omega}}.$$

     $$J(\boldsymbol{\omega})=\frac{\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_B\boldsymbol{\omega}}{\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{\omega}}$$

   - 求解LDA，只需要针对w求偏导，并令导数等于0，得：

     $$\boldsymbol{S}_w^{-1}\boldsymbol{S}_B\boldsymbol{\omega}=\lambda\boldsymbol{\omega}$$

     - λ是目标函数$J(w)$s
   - 所以求解LDA只需要求出样本的均值差和类内方差：

     $$\omega=\boldsymbol{S}_w^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

- LDA比PCA更适合降维有类别信息的数据，但是LDA对数据做了很强的假设：
    - 每个类的数据都是高斯分布
    - 各个类的协方差相等
- 线性模型对噪声的鲁棒性很好，但是表达能力有限，可以引入核方法进行扩展

2. **多类别LDA的求解过程：**
    - 多类别情况下的LDA，类间散度是全局散度和类内散度的差，即每个类别的中心经过投影后和全局中心的距离
    - 多类别LDA的求解步骤：
        - 计算每个类别的均值μ_j和全局均值μ
        - 计算类内散度矩阵S_w，全局散度矩阵S_t：

          $$\boldsymbol{S}_w = \sum_{\boldsymbol{x} \in C_i} (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^{\mathrm{T}}$$

          $$\boldsymbol{S}_t = \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathrm{T}}.$$

        - 计算类间散度矩阵S_b：

          $$\begin{aligned}\boldsymbol{S}_b &= \boldsymbol{S}_t - \boldsymbol{S}_w \\ &= \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathrm{T}} - \sum_{\boldsymbol{x} \in C_i} (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^{\mathrm{T}} \\ &= \sum_{j=1}^{N} \left(\sum_{\boldsymbol{x} \in C_j} (\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} - \sum_{\boldsymbol{x} \in C_j} (\boldsymbol{x} - \boldsymbol{\mu}_j)(\boldsymbol{x} - \boldsymbol{\mu}_j)^{\mathrm{T}} \right) \\ &= \sum_{j=1}^{N} \boldsymbol{m}_j (\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^{\mathrm{T}},\end{aligned}$$

        - 定义目标函数：

          $$J(\boldsymbol{W}) = \frac{tr(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{S}_b \boldsymbol{W})}{tr(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{S}_w \boldsymbol{W})},$$

        - 对矩阵$S_w{}^{-1}S_b$进行特征值分解，将特征值从大到小排序
        - 取特征值前d大的特征值对应的特征向量$w_1, w_2, ..., w_d$，将n维样本映射到d维空间：

          $$\boldsymbol{x}_i' = \begin{bmatrix} \boldsymbol{\omega}_1^{\mathrm{T}} \boldsymbol{x}_i \\ \boldsymbol{\omega}_2^{\mathrm{T}} \boldsymbol{x}_i \\ \vdots \\ \boldsymbol{\omega}_d^{\mathrm{T}} \boldsymbol{x}_i \end{bmatrix}.$$

3. **LDA和PCA从原理和应用上有什么异同?**
    - 无监督任务使用PCA（无监督算法），有监督任务使用LDA（监督算法）
    - 从求解过程看，相似度很高，但是原理有所区别：
        - PCA旨在最大化投影点的方差（假设方差越大，信息量越多），用主成分来表示原始数据可以去除冗余的维度
        - LDA旨在最小化类内方差和最大化类间方差，利用数据标签找到数据中最具判别性的维度，使得投影后的数据更容易被区分
    - 应用场景区别，例如语音识别：
        - 可以使用PCA过滤掉固定频率（方差较小）的背景噪音
        - 但是目标是识别人声的话，就需要通过LDA来考虑标签，使降维后的数据有区分性
    - PCA和LDA在人脸识别的应用：
        - 基于PCA的方法叫特征脸（Eigenface）方法，对人脸特征的协方差矩阵做特征分解，较大的特征值对应的特征向量具有与人脸相似的形状