模式识别教案

邓世文 *

2017/8/31

Contents

* 哈尔滨师范大学数学科学学院 *(e-mail: dengswen@gmail.com).*

6 特征提取

6.1 引言

（1）降维的两种方法

特征选择 $\rightarrow$ 从 D 个特征中选取 $d < D$ 个特征。

特征提取 $\rightarrow$ 通过适当的变换将 D 特征转换成 $d < D$ 个新特征。

（2）特征提取的目的：

1) 降低特征空间的维度；

2) 消除特征间的相关性，减少特征中与分类无关的信号。

（3）特征变换

给定原始特征 $\mathbf{x} \in \mathbb{R}^D$，变换后的特征为 $\mathbf{y} \in \mathbb{R}^d$，可通过始下变换得到

$$\mathbf{y} = \mathbf{W}^T\mathbf{x} \tag{1}$$

其中 $\mathbf{W} \in \mathbb{R}^{D\times d}$ 称为变换矩阵。通常 $d < D$，实现降维；但也有通过非线性变换实现升维处理。

特征提取问题 $\rightarrow$ 依据训练样本求适当的 $\mathbf{W}$，使其满足某种特征变换的准则为最优。

6.2 基于类别可分性判据的特征提取

6.2.1 可分性判据与特征提取

以前一章的类别可分性判据 $J(\cdot)$ 作为度量新特征的准则，则特征提取问题可表述为如下优化问题

$$\mathbf{W}^* = arg\min_{\mathbf{W}} J(\mathbf{W}^T\mathbf{x}) \tag{2}$$

其中 $J(\cdot)$ 可是类内类间距离、概率距离或熵。

6.2.2 基于类内类间距离的特征提取

（1）类降维后的特征内、类间离散度矩阵

$$\bar{\mathbf{S}}_w = \mathbf{W}^T\mathbf{S}_w\mathbf{W} \tag{3}$$

$$\bar{\mathbf{S}}_b = \mathbf{W}^T\mathbf{S}_b\mathbf{W} \tag{4}$$

（2）J_1 准则下的降维问题

1) 目标函数

$$\mathbf{W}^* = arg\max_{\mathbf{W}} J_1(\mathbf{W}) = arg\max_{\mathbf{W}} Tr(\mathbf{W}^T(\mathbf{S}_w + \mathbf{S}_b)\mathbf{W}) \tag{5}$$

这是一个关于 $\mathbf{W}$ 的二次函数，其最大值为 $+\infty$。消除尺度影响，引入约束 $Tr(\mathbf{W}^T\mathbf{S}_w\mathbf{W}) = 1$，将优化问题转化为

$$arg\max J_1(\mathbf{W}) \tag{6}$$

$$\text{s.t } Tr(\mathbf{W}^T\mathbf{S}_w\mathbf{W}) = 1 \tag{7}$$

2) **构造** Lagrange **函数**

$$L(\mathbf{W},\Lambda) = J_1(\mathbf{W}) - Tr(\Lambda(\mathbf{W}^T\mathbf{S}_w\mathbf{W} - \mathbf{I})) \tag{8}$$

$$= Tr(\mathbf{W}^T(\mathbf{S}_w + \mathbf{S}_b)\mathbf{W}) - Tr(\Lambda(\mathbf{W}^T\mathbf{S}_w\mathbf{W} - \mathbf{I})) \tag{9}$$

其中 $\mathbf{I}$ **为单位阵，**Λ **为对角阵，其对解元素为** Lagrange **乘子。**

3) **求解**

Lagrange **函数的极值点** $\rightarrow \frac{\partial}{\partial \mathbf{W}} L(\mathbf{W},\Lambda) = \mathbf{0}$

求 Jacobian **矩阵** $\rightarrow$ **求关于** $\mathbf{W}$ **的微分**

$$dL(\mathbf{W},\Lambda) = 2Tr(\mathbf{W}^T(\mathbf{S}_w + \mathbf{S}_b)d\mathbf{W}) - 2Tr(\Lambda\mathbf{W}^T\mathbf{S}_w d\mathbf{W}) \tag{10}$$

Jacobina **矩阵** $\mathbf{J}$ **为**

$$\mathbf{J} = 2\mathbf{W}^T(\mathbf{S}_w + \mathbf{S}_b) - 2\Lambda\mathbf{W}^T\mathbf{S}_w \tag{11}$$

得到梯度

$$\nabla_{\mathbf{W}} L(\mathbf{W},\Lambda) = \mathbf{J}^T = 2(\mathbf{S}_w + \mathbf{S}_b)\mathbf{W} - 2\mathbf{S}_w\mathbf{W}\Lambda \tag{12}$$

令梯度为 0**，得到**

$$(\mathbf{S}_w + \mathbf{S}_b)\mathbf{W} = \mathbf{S}_w\mathbf{W}\Lambda \tag{13}$$

即

$$\mathbf{S}_w^{-1}(\mathbf{S}_w + \mathbf{S}_b)\mathbf{W} = \mathbf{W}\Lambda \tag{14}$$

$$\Rightarrow (\mathbf{I} + \mathbf{S}_w^{-1}\mathbf{S}_b)\mathbf{W} = \mathbf{W}\Lambda \tag{15}$$

$$\Rightarrow \mathbf{W} + \mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{W} = \mathbf{W}\Lambda \tag{16}$$

$$\Rightarrow \mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{W} = \mathbf{W}(\Lambda - \mathbf{I}) \tag{17}$$

由此可知，$\mathbf{W}$ **对应** $\mathbf{S}_w^{-1}\mathbf{S}_b$ **的特征向量，**$\Lambda - \mathbf{I}$ **为对角阵，其对角元素为** $\mathbf{S}_w^{-1}\mathbf{S}_b$ **的特征值，即**

$$\Lambda = \mathbf{I} + \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_D \end{bmatrix} = \begin{bmatrix} \lambda_1 + 1 & & \\ & \ddots & \\ & & \lambda_D + 1 \end{bmatrix} \tag{18}$$

由式 (13) **左乘** $\mathbf{W}^T$ **得到**

$$\mathbf{W}^T(\mathbf{S}_w + \mathbf{S}_b)\mathbf{W} = \mathbf{W}^T\mathbf{S}_w\mathbf{W}\Lambda \tag{19}$$

式 (9) **的优化问题的约束条件满足**

$$\mathbf{W}^T\mathbf{S}_w\mathbf{W} = \mathbf{I} \tag{20}$$

因此 J_1 准则为

$$J_1(\mathbf{W}) = Tr(\mathbf{W}^T(\mathbf{S}_w + \mathbf{S}_b)\mathbf{W}) = Tr(\mathbf{W}^T\mathbf{S}_w\mathbf{W}\Lambda) = Tr(\Lambda) = \sum_{i=1}^{d}(1+\lambda_i) \tag{21}$$

4) 结论

(i) 最优变换矩阵 $\mathbf{W}$ 就是由 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 的前 d 个特征向量构成。

(ii) 基于不同的类内类间距离准则所得到的变换矩阵相同。

(iii) 基于其它可分性准则大多难以找到闭式解，只能使用数值方法求解。

6.3 主成分分析 (PCA)

6.3.1 PCA 概述

（1）基本思想：由原始特征计算一组按重要性从大到小排序的且不相关的新特征，它们是原始特征的线性组合。

（2）问题描述

令 $\mathbf{x} = [x_1, \cdots, x_D]^T$ 表示原始特征向量，第 i 个新特征 ξ_i 可表示为

$$\xi_i = \sum_{j=1}^{D} a_{ij}x_j = \mathbf{a}_i^T\mathbf{x} \tag{22}$$

其中 $\|\mathbf{a}_i\|_2 = 1$。

令 $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_d] \in \mathbb{R}^{D\times d}$ 表示特征变换矩阵，$\boldsymbol{\xi} = [\xi_1, \cdots, \xi_d]^T \in$ 表示新的特征向量，则写成矩阵形式为

$$\boldsymbol{\xi} = \mathbf{A}^T\mathbf{x} \tag{23}$$

主成分分析的目的：就是求解最优正交矩阵 $\mathbf{A}$，使得新特征 ξ_i 的方差达到极值。

解释：为何需要新特征的方差最大化？$\rightarrow$ 正交变换能够新特征 ξ_i 间不相关，其方差越大，则样本在该方向上的差异就越大，因此这一特征就越重要。

6.3.2 求解方法

（1）求解第一个投影方向 $\mathbf{a}_1$

1) 第 1 个特征 ξ_1

$$\xi_1 = \mathbf{a}_1^T\mathbf{x} \tag{24}$$

其方差为

$$\mathrm{var}(\xi_1) = \mathbb{E}[\xi_1^2] - \mathbb{E}[\xi_1]^2 = \mathbb{E}[\mathbf{a}_1^T\mathbf{x}\mathbf{x}^T\mathbf{a}_1] - \mathbb{E}[\mathbf{a}_1^T\mathbf{x}]\mathbb{E}[\mathbf{a}_1^T\mathbf{x}] = \mathbf{a}_1^T\Sigma\mathbf{a}_1 \tag{25}$$

其中 Σ 是 $\mathbf{x}$ 的协方差矩阵，可基于训练样本估计得到。

2) 优化问题

$$\max_{\mathbf{a}_1} \mathbf{a}_1^T\Sigma\mathbf{a}_1 \tag{26}$$

$$\text{s.t. } \|\mathbf{a}_1\|_1 = 1 \tag{27}$$

3) Lagrange 函数

$$L(\mathbf{a}_1, \lambda_1) = \mathbf{a}_1^T \Sigma \mathbf{a}_1 - \lambda_1(\mathbf{a}_1^T \mathbf{a}_1 - 1) \tag{28}$$

其中 λ_1 为 Lagrange 乘子。

4) 利用一阶条件得到

$$\Sigma \mathbf{a}_1 = \lambda_1 \mathbf{a}_1 \tag{29}$$

可知 $\mathbf{a}_1$ 是协方差矩阵 Σ 的特征向量，其所对应的最大的特征值为 λ_1，其也是新特征 ξ_1 的方差，即

$$\mathrm{var}(\xi_1) = \mathbf{a}_1 \Sigma \mathbf{a}_1 = \lambda_1 \mathbf{a}_1^T \mathbf{a}_1 = \lambda_1 \tag{30}$$

其中 ξ_1 称为第一主成分，它在原始特征的所有线性组合时是方差最大的。

（2）其他的主成分

1) 第二个主成分 ξ_2 满足条件

要求在此方向上的特征方差最大，还要与前一个主成分不相关，即要求 ξ_1 和 ξ_2 的协方差 $\mathrm{cov}(\xi_1, \xi_2) = 0$，

$$\mathrm{cov}(\xi_1, \xi_2) = \mathbb{E}[\xi_1 \xi_2] - \mathbb{E}[\xi_1]\mathbb{E}[\xi_2] = 0 \tag{31}$$

从而得到

$$\mathbf{a}_2^T \Sigma \mathbf{a}_1 = 0 \tag{32}$$

由于 $\Sigma \mathbf{a}_1 = \lambda_1 \mathbf{a}_1$，因此要求 ξ_1, ξ_2 不相关等价于，要求 $\mathbf{a}_1, \mathbf{a}_2$ 正交，即

$$\mathbf{a}_1^T \mathbf{a}_2 = 0 \tag{33}$$

由此可知，满足 $\mathbf{a}_1^T \mathbf{a}_2 = 0$ 和 $\|\mathbf{a}\|_2 = 1$ 的 $\mathbf{a}_2$ 就是协方差矩阵 Σ 的第 2 大特征值 λ_2 所对应的特征向量。

2) 其他投影方向

此上述分析可知，所要寻找的投影方向就是协方差矩阵特征值分解按特征值大小排序后所对应的的特征向量。

3) 正交变换矩阵

选取前 d 个特征值所对应的特征向量作正交变换矩阵 $\mathbf{A}$ 的列向量，即 $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_d]$。

d 个主成分点数据全部方差的比例

$$r = \frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{D} \lambda_i} \tag{34}$$

因此可依据比例 r 来选取相应的降维维度 d。

6.4 K-L 变换

6.4.1 K-L 变换的基本原理

K-L 变换的基本原理与 PCA 相同，但 K-L 变换能够考虑到不同的分类信息，实现有监督的特征提取。

（1）正交基表示与表示误差

1) 原始特征空间的基

令 $\{\mathbf{u}_i\}_{i=1,\cdots,\infty}, \mathbf{u}_i \in \mathbb{R}^D$, 为 D 维原始特征空间的正交基，其满足 $\mathbf{u}_i^T\mathbf{u}_j = \delta_{ij}$。

2) 原始特征在基下的表示

任意原始特征 $\mathbf{x} \in \mathbb{R}^D$ 皆可表示为

$$\mathbf{x} = \sum_{i=1}^{\infty} c_i \mathbf{u}_i \tag{35}$$

其中 $c_i = \langle \mathbf{x}, \mathbf{u}_i \rangle$。

3) 原始特征近似

采用有限的 $d < D$ 项来逼近 $\mathbf{x}$

$$\hat{\mathbf{x}} = \sum_{i=1}^{d} c_i \mathbf{u}_i \tag{36}$$

4) 近似误差

$$e = \mathbb{E}\left[\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2\right] = \mathbb{E}\left[\left\|\sum_{i=d+1}^{\infty} c_i \mathbf{u}_i\right\|_2^2\right] = \mathbb{E}\left[\sum_{i=d+1}^{\infty} c_i^2\right] = \mathbb{E}\left[\sum_{i=d+1}^{\infty} \langle \mathbf{x}, \mathbf{u}_i \rangle^2\right] \tag{37}$$

$$= \mathbb{E}\left[\sum_{i=d+1}^{\infty} \mathbf{u}_i^T \mathbf{x}\mathbf{x}^T \mathbf{u}_i\right] = \sum_{i=d+1}^{\infty} \mathbf{u}_i^T \mathbb{E}\left[\mathbf{x}\mathbf{x}^T\right] \mathbf{u}_i \tag{38}$$

$$= \sum_{i=d+1}^{\infty} \mathbf{u}_i^T \mathbf{\Psi} \mathbf{u}_i \tag{39}$$

其中 $\mathbf{\Psi} = \mathbb{E}[\mathbf{x}\mathbf{x}^T]$ 为二阶矩阵，定义在有限训练集 $X = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$ 上的二阶矩阵为

$$\mathbf{\Psi} = \mathbb{E}[\mathbf{x}\mathbf{x}^T] = XX^T \tag{40}$$

自为自相关矩阵（产生矩阵）。

5) 基本思想：通过最小化误差 e 来求得降维变换矩阵

（2）优化问题

1) 目标函数

$$\min_{\mathbf{u}} e = \sum_{i=d+1}^{\infty} \mathbf{u}_i^T \mathbf{\Psi} \mathbf{u}_i \text{ s.t. } \|\mathbf{u}_i\|_2^2 = 1, \forall i \tag{41}$$

2) Lagrange 函数

$$L(\mathbf{u}, \lambda) = \sum_{i=d+1}^{\infty} \mathbf{u}_i^T \mathbf{\Psi} \mathbf{u}_i - \sum_{i=d+1}^{\infty} \lambda_i (\mathbf{u}_i^T \mathbf{u}_i - 1) \tag{42}$$

3) 求解

令关于 $\mathbf{u}_i$ 的偏导为 0，得到

$$(\mathbf{\Psi} - \lambda_i \mathbf{I})\mathbf{u}_i = 0 \tag{43}$$

这表明 λ_i 和 $\mathbf{u}_i$ 分别是 $\mathbf{\Psi}$ 的特征值和特征向量，即

$$\mathbf{\Psi}\mathbf{u}_i = \lambda_i \mathbf{u}_i \tag{44}$$

4) 最小误差

因此最小误差为

$$e = \sum_{d+1}^{\infty} \lambda_i \tag{45}$$

如果令 $d = 0$，则上式对所有 $i = 1, \cdots, \infty$ 皆成立。

（2）实现 K-L 变换

1) K-L 变换的过程

给定训练集 $X = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$，实现 K-L 变换的过程：

(a) 计算自相关矩阵（产生矩阵）$\mathbf{\Psi} = XX^T$；

(b) 对 $\mathbf{\Psi}$ 进行特征值分解；

(c) 将 $\mathbf{\Psi}$ 的特征向量 $\mathbf{u}_i$ 按其所对应的特征值 λ_i 由大到小排序；

(d) 将前 d 个特征向量构成的矩阵 $U = [\mathbf{u}_1, \cdots, \mathbf{u}_d] \in \mathbb{R}^{D \times d}$ 作为特征提取的正交变换矩阵；

2) 新特征生成过程

新的 d 维特征表示为 $\mathbf{y} = U^T\mathbf{x}$，其满足错误最小。

（3）K-L 变换与 PCA 的区别

K-L 变换使用的是自相关矩阵，PCA 使用的是协方差矩阵。

在 Bayesian 框架下，这些模型皆可统一到一起，其区别仅是对错误赋与不同的先验，从而实现一系列更为复杂的模型，如 PPCA、ICA、FA 等。

（4）K-L 变换的优点

1) 最佳压缩表示；

2) 新特征是不相关的；

3) 表示熵最小；

4) 总体熵最小。

6.4.2 有监督的 K-L 变换

（1）K-L 变换的产生矩阵 $\mathbf{\Psi}$ 可采用多种形式

1) 基本定义 $\rightarrow \mathbf{\Psi} = \mathbb{E}[\mathbf{x}\mathbf{x}^T] \approx XX^T \rightarrow$ 自相关矩阵

2) 协方差矩阵 $\rightarrow \mathbf{\Psi} = \mathbb{E}[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T]$，其中 μ 为样本均值，$\rightarrow$ 主成分分析 (PCA)

引入类别信息

3) 总类内离散度矩阵 $\rightarrow \mathbf{\Psi} = S_w = \sum_{i=1}^{C} P_i \Sigma_i$，其中 $P_i = p(\omega_i)$ 是类 ω_i 的先验概率，Σ_i 是 ω_i 类的协方差矩阵 $\Sigma_i = \mathbb{E}[(\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^T]$

（2）如果对样本的分类信息有特定的先验知识，可设计特定的 K-L 变换进行特征进取。

6.4.3 从类均值中提取判别信息

（1）先验：类别的主要分类信息包含在均值中。

（2）基本思想：优先提取均值信息

1) 以总类内离散度矩阵 S_w 作为产生矩阵，即 $\mathbf{\Psi} = S_w$，以消除特征间的相关性；

2) 在变换后的特征中，选择方差小、类均值与总体均值判别大的特征作为新特征。

（3）实现过程

1) 计算生成矩阵（总类离散度矩阵）：$\mathbf{\Psi} = S_w$；

2) 对 $\mathbf{\Psi}$ 特征值分解 $\rightarrow\{\lambda_i, \mathbf{u}_i\}_{i=1,\cdots,D}$。第 i 个变换后特征 $y_i = \mathbf{u}_i^T\mathbf{x}$，$i = 1, \cdots, D$，其中每维特征的方差为 λ_i；

3) 计算变换后特征的分类性能指标

$$J(y_i) = \frac{\mathbf{u}_i^T S_b \mathbf{u}_i}{\lambda_i}, \text{ for } i = 1, \cdots, D \tag{46}$$

其中 $S_b = \sum_{j=1}^{C} P_j(\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^T$ 为原始特征空间的类间离散度矩阵，$\boldsymbol{\mu}_j$ 和 $\boldsymbol{\mu}$ 分别第 j 个类的均值和总体均值。

4) 构造变换矩阵 $U \in \mathbb{R}^{D\times d}$

依据性能指标对变换后的特征排序

$$J(y_1) \geq J(y_2) \geq \cdots \geq J(y_D) \tag{47}$$

其中第 i 个特征 y_i 对应第 i 个特征向量 $\mathbf{u}_i$，选择前 d 个特征所对应的特征向量构成变换矩阵 $U = [\mathbf{u}_1, \cdots, \mathbf{u}_d]$。

（4）例

先验概率相同 $(P1 = P2 = 0.5)$ 的两类 ω_1 和 ω_2 分类问题，两个类别均值分别为

$$\mu_1 = [4, 2]^T \text{ 和} \mu_2 = [-4, -2]^T$$

协方差矩阵分别为

$$\Sigma_1 = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \text{ 和} \Sigma_2 = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}$$

（注：已知矩阵

$$A = \begin{bmatrix} 3.5 & 1.5 \\ 1.5 & 3.5 \end{bmatrix}$$

的特征值分解：特征值分别为 $5, 2$，与其相对应的特征向量分别为 $[0.707, 0.707]^T$ 和 $[0.707, -0.707]^T$。）

要求：

1) 采用 K-L 变换计算提取一维均值判别信息变换矩阵；

2) 画出新的坐标轴；

3) 给定样本点 $(3.5, 2.4), (-4.2, -2.4), (2.7, 2.5), (-4.1, -1.9)$，分别计算它们的一维变换特征 (4.1713, -4.6662, 3.6764, -4.2420)，并指出它们的类别 (ω_1 或 ω_2)。

求解：

1) 基于类均值判别信息的 K-L 变换

i) 计算生成矩阵

$$\mathbf{\Psi} = S_w = 0.5 * \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} + 0.5 * \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix} = \begin{bmatrix} 3.5 & 1.5 \\ 1.5 & 3.5 \end{bmatrix}$$

ii) 对 $\mathbf{\Psi}$ 特征值分解

利用已知条件可知此矩阵特征值分解的特征值为 $5, 2$，和特征向量为 $[0.707, 0.707]^T$ 和 $[0.707, -0.707]^T$。

iii) 计算分类性能指标

$$S_b = \begin{bmatrix} 16 & 8 \\ 8 & 4 \end{bmatrix}$$

得到 $J(y_1) = 3.6,\ J(y_2) = 1$。

v) 变换矩阵 $U = [\mathbf{u}_1] = [0.707\ 0.707]^T$。

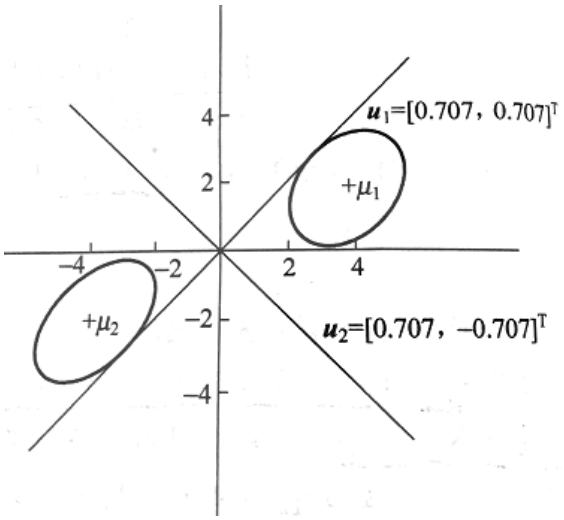

图 1: 基于类均值判别信息的 K-L 变换.

6.5 **核** PCA

给定数据集 $X = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$，假定数据已去均值，其协方差矩阵为 $\mathbf{\Psi} = XX^T$。PCA 的主成分由 $\mathbf{\Psi}$ 的特征向量 $U = [\mathbf{u}_1, \cdots, \mathbf{u}_d]$ 决定。

令 $\{\lambda, \mathbf{u}\}$ 为 $\mathbf{\Psi}$ 的特征值和特征向量。

$$\begin{aligned} &\mathbf{\Psi}\mathbf{u} = \lambda\mathbf{u} \\ \Rightarrow &\mathbf{u} = \frac{1}{\lambda}\mathbf{\Psi}\mathbf{u} = \frac{1}{\lambda}\left(XX^T\right)\mathbf{u} = \frac{1}{\lambda}\left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{u} = \frac{1}{\lambda}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T\mathbf{u} = \sum_{i=1}^{N}\frac{\langle\mathbf{x}_i, \mathbf{u}\rangle}{\lambda}\mathbf{x}_i \\ &= \sum_{i=1}^{N}\alpha_i\mathbf{x}_i = X\boldsymbol{\alpha} \end{aligned}$$

其中 $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_N]$, $\alpha_i = \frac{1}{\lambda}\langle \mathbf{x}_i, \mathbf{u} \rangle$。

因此

$$
\begin{aligned}
&\boldsymbol{\Psi}\mathbf{u} = \lambda \mathbf{u} \\
\Rightarrow &\boldsymbol{\Psi} X\boldsymbol{\alpha} = \lambda X\boldsymbol{\alpha} \\
\Rightarrow &XX^T X\boldsymbol{\alpha} = \lambda X\boldsymbol{\alpha} \\
\overset{\text{左乘}X^T}{\Longrightarrow} &X^T XX^T X\boldsymbol{\alpha} = \lambda X^T X\boldsymbol{\alpha} \\
\Rightarrow &KK\boldsymbol{\alpha} = \lambda K\boldsymbol{\alpha} \\
\Rightarrow &K\boldsymbol{\alpha} = \lambda \boldsymbol{\alpha}
\end{aligned}
$$

其中 $K = XX^T$ 为核矩阵 (Gram) 矩阵，

$$K_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle \tag{48}$$

其中可将内积 $\langle \cdot, \cdot \rangle$ 替换为核函数 $k(\cdot, \cdot)$。

6.6 局部线性嵌入 (locally linear embedding, LLE)

（1）LLE 能够保持数据内部的全局的非线性结构。

（2）LLE 基于简单的几何直觉（假设）：样本点位于或接近某个潜在的低维流形，流形的任何小的局部可用小平面近似。

6.6.1 提取数据内部几何结构

基本思想：基于局部平面假定，以样本的邻域来近似样本 (回归)，从而提取回归系数。

（1）数据和定义

给定数据集 $\Omega = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$，$\mathbf{x}_i \in \mathbb{R}^D$。

给定数据集中任一点 $\mathbf{x} \in \Omega$, 以 B 表示样本 $\mathbf{x}$ 的邻域点的下标集合。

（2）局部线性模型

假定样本可由其邻域中的点来表示

$$\mathbf{x} \approx \sum_{i=1}^{N} w_i \mathbf{x}_i \tag{49}$$

其中 $w_i = 0$, if $i \notin B$。

(3) 近似误差

$$e(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{x} - \sum_{i=1}^{N} w_i \mathbf{x}_i\right\|^2 = \frac{1}{2}\left\|\sum_{i=1}^{N} w_i(\mathbf{x} - \mathbf{x}_i)\right\|^2 \tag{50}$$

$$= \frac{1}{2}\left(\sum_{i=1}^{N} w_i(\mathbf{x} - \mathbf{x}_i)\right)^T \left(\sum_{i=1}^{N} w_i(\mathbf{x} - \mathbf{x}_i)\right) \tag{51}$$

$$= \frac{1}{2}\left(\sum_{i=1}^{N} w_i(\mathbf{x} - \mathbf{x}_i)^T\right) \left(\sum_{i=1}^{N} w_i(\mathbf{x} - \mathbf{x}_i)\right) \tag{52}$$

$$= \sum_{ij} w_i w_j \langle(\mathbf{x} - \mathbf{x}_i, \mathbf{x} - \mathbf{x}_j)\rangle \tag{53}$$

$$= \frac{1}{2}\mathbf{w}^T C \mathbf{w} \tag{54}$$

其中 $\mathbf{w} = [w_1, \cdots, w_N]^T$，$C_{ij} = \langle(\mathbf{x} - \mathbf{x}_i, \mathbf{x} - \mathbf{x}_j)\rangle$ 是以 $\mathbf{x}$ 为均值邻域样本的相关矩阵 (Gram 矩阵)。

(4) 优化问题

$$\min_{\mathbf{w}} e(\mathbf{w}) = \frac{1}{2}\mathbf{w} C \mathbf{w} \tag{55}$$

$$\text{s.t.} \mathbf{1}^T \mathbf{w} = 1 \tag{56}$$

(5) 求解

Lagrange 函数

$$L(\mathbf{w}, \lambda) = \frac{1}{2}\mathbf{w}^T C \mathbf{w} - \lambda(\mathbf{1}^T \mathbf{w} - 1) \tag{57}$$

令 $\frac{\partial L}{\partial \mathbf{w}} = 0$，得到

$$C\mathbf{w} - \lambda \mathbf{1} = 0 \tag{58}$$

$$\Rightarrow \mathbf{w} = \lambda C^{-1} \mathbf{1} \tag{59}$$

对 $\mathbf{w}$ 标准化

$$w_i = \begin{cases} w_i, & \text{if } i \in B \\ 0, & \text{else} \end{cases} \tag{60}$$

$$\mathbf{w} = \frac{C^{-1}\mathbf{1}}{\mathbf{1}^T C^{-1} \mathbf{1}} \tag{61}$$

(6) 对 N 个样本点可求得矩阵 $W = [\mathbf{w}_1, \cdots, \mathbf{w}_N] \in \mathbb{R}^{N \times N}$，此矩阵表征了数据 X 内部的几何特征 (结构)，即流形结构。

6.6.2 流形嵌入表示

基本思想：依据数据集的几何关系实现数据在低维空间的嵌入表示，新的嵌入表示能够保持数据集中原有的几何关系，其中数据集中的几何关系由矩阵 W 所刻画。

假定数据集中每个样本 $\mathbf{x} \in \mathbb{R}^D$ 所对应的嵌入表示为 $\mathbf{y} \in \mathbb{R}^D$，注意这里的 $\mathbf{y}$ 是嵌入表示但还没有降维。令 $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_N] \in \mathbb{R}^{D \times N}$ 为对应于数据集 $X = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$ 的嵌入表示。

（1）总体数据集的嵌入损失函数

$$e(\mathbf{Y}) = \sum_i \left\| \mathbf{y}_i - \sum_j W_{ij}\mathbf{y}_j \right\|^2 \tag{62}$$

$$= \sum_i \left(\mathbf{y}_i - \sum_j W_{ij}\mathbf{y}_j \right)^T \left(\mathbf{y}_i - \sum_k W_{ik}\mathbf{y}_k \right) \tag{63}$$

$$= \sum_i \left(\mathbf{y}_i^T - \sum_j W_{ij}\mathbf{y}_j^T \right) \left(\mathbf{y}_i - \sum_k W_{ik}\mathbf{y}_k \right) \tag{64}$$

$$= \sum_i \left(\mathbf{y}_i^T\mathbf{y}_i - \sum_j W_{ij}\mathbf{y}_j^T\mathbf{y}_i - \sum_k W_{ik}\mathbf{y}_i^T\mathbf{y}_k + \left(\sum\nolimits_j W_{ij}\mathbf{y}_j^T\right)\left(\sum\nolimits_k W_{ik}\mathbf{y}_k\right) \right) \tag{65}$$

$$= \sum_i \left(\mathbf{y}_i^T\mathbf{y}_i - \sum_j W_{ij}\mathbf{y}_j^T\mathbf{y}_i - \sum_k W_{ik}\mathbf{y}_i^T\mathbf{y}_k + \sum\nolimits_{jk} W_{ij}W_{ik}\mathbf{y}_j^T\mathbf{y}_k \right) \tag{66}$$

$$= \sum_i \mathbf{y}_i^T\mathbf{y}_i - \sum_i\sum_j W_{ij}\mathbf{y}_j^T\mathbf{y}_i - \sum_i\sum_k W_{ik}\mathbf{y}_i^T\mathbf{y}_k + \sum_i\sum\nolimits_{jk} W_{ij}W_{ik}\mathbf{y}_j^T\mathbf{y}_k \tag{67}$$

$$= \sum_{ij} \mathbf{y}_i^T\mathbf{y}_j\delta_{ij} - \sum_{ij} W_{ij}\mathbf{y}_j^T\mathbf{y}_i - \sum_{ik} W_{ik}\mathbf{y}_i^T\mathbf{y}_k + \sum_i\sum_{jk} W_{ij}W_{ik}\mathbf{y}_j^T\mathbf{y}_k \tag{68}$$

$$= \sum_{ij} \mathbf{y}_i^T\mathbf{y}_j\delta_{ij} - \sum_{ij} W_{ij}\mathbf{y}_j^T\mathbf{y}_i - \sum_{ij} W_{ji}\mathbf{y}_i^T\mathbf{y}_j + \sum_m\sum_{ij} W_{mj}W_{mi}\mathbf{y}_i^T\mathbf{y}_j \tag{69}$$

$$= \sum_{ij} \left(\mathbf{y}_i^T\mathbf{y}_j\delta_{ij} - W_{ij}\mathbf{y}_j^T\mathbf{y}_i - W_{ji}\mathbf{y}_i^T\mathbf{y}_j + \sum_m W_{mj}W_{mi}\mathbf{y}_i^T\mathbf{y}_j \right) \tag{70}$$

$$= \sum_{ij} \mathbf{y}_i^T\mathbf{y}_j \left(\delta_{ij} - W_{ij} - W_{ji} + \sum_m W_{mj}W_{mi} \right) \tag{71}$$

$$= \sum_{ij} \mathbf{y}_i^T\mathbf{y}_j M_{ij} \tag{72}$$

$$= Tr(YMY^T) \tag{73}$$

其中 $M = (\mathbf{I} - W)^T(\mathbf{I} - W) \in \mathbb{R}^{N\times N}$ 为一对称矩阵（Gram 矩阵），其元素定义为 $M_{ij} = \delta_{ij} - W_{ij} - W_{ji} + \sum_m W_{mj}W_{mi}$。

（2）优化问题

令 $P = Y^T$，并加入约束 $P^TP = \mathbf{I}$ 以使每一维的特征为单位方差,，得到如下优化问题

$$\min_P Tr(P^TMY) \tag{74}$$

$$\text{s.t. } P^TP = \mathbf{I} \tag{75}$$

（3）求解 → 特征值分解

（4）流形嵌入

特征向量矩阵 P 的列为特征向量，行对应数据样本 $\mathbf{x}$ 的流形嵌入 $\mathbf{y}$。

（5）降维表示

包含 N 个特征向量的矩阵 $P = [\mathbf{p}_1, \cdots, \mathbf{p}_N]$，其所对应的特征值为 $\lambda_1 \geq \cdots \geq \lambda_N$。令 P_d 为特征值最

小的 d 个特征向量，则矩阵 P_d 的行向量或 P_d^T 的列向量，就是样本 $\mathbf{x} \in \Omega$ 的降维嵌入特征 $\mathbf{y} \in \mathbb{R}^d$。

注：如果数据集 X 是未去均值的矩阵，则存一个为 0 的最小特征值，在进行特征降维表示时，应该去从特征向量矩阵 P 中，除特征值为 0 的特征向量后，再保存最小的 d 个特征向量。