

特征提取：把D个特征通过适当变换转换为d(D>d)个新特征，实现特征空间降维



2.6.1 基于类别可分离性判据 指标函数 J(W)

$$\begin{aligned} J_1(W) &= \text{tr}[W^T(S_w + S_b)W] \\ J_2(W) &= \text{tr}[(W^T S_w W)^{-1} (W^T S_b W)] \\ J_3(W) &= \ln \frac{|W^T S_b W|}{|W^T S_w W|} \\ J_4(W) &= \frac{\text{tr}(W^T S_b W)}{\text{tr}(W^T S_w W)} \\ J_5(W) &= \frac{|W^T S_b W|}{|W^T S_w W|} (\Sigma = S_w + S_b) \end{aligned}$$

→ 变换后特征空间
→ 原始特征空间的类内离差矩阵
→ 原始特征空间的类间离差矩阵

目标 $\rightarrow J(y) = \max_{\{W\}} J(W)$

阵，可知

$$\begin{aligned} S_w &= W^T S_w W \\ S_b &= W^T S_b W \\ W^T S_w W &= W^T \sum_{i=1}^c P_i \frac{1}{N} \sum_{k=1}^N (\vec{x}_k^{(i)} - \vec{m}^{(i)}) (\vec{x}_k^{(i)} - \vec{m}^{(i)})^T \\ S_b &= \sum_{i=1}^c P_i (\vec{m}^{(i)} - \vec{m}) (\vec{m}^{(i)} - \vec{m})^T \\ W^T S_w W &= W^T \left[\sum_{i=1}^c P_i \frac{1}{N} \sum_{k=1}^N (\vec{x}_k^{(i)} - \vec{m}^{(i)}) (\vec{x}_k^{(i)} - \vec{m}^{(i)})^T \right] W \\ &= \sum_{i=1}^c P_i \frac{1}{N} \sum_{k=1}^N [W^T \vec{x}_k^{(i)} - W^T \vec{m}^{(i)}] (\vec{x}_k^{(i)} - \vec{m}^{(i)})^T W \\ &= \sum_{i=1}^c P_i \frac{1}{N} \sum_{k=1}^N [\vec{y}_k^{(i)} - \vec{m}_y^{(i)}] (\vec{y}_k^{(i)} - \vec{m}_y^{(i)})^T W \\ &= S_y^* \end{aligned}$$

$$\text{cov}: \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

以小为例，问题表示为

$$\max J_1(W)$$

$$\text{st. } \text{tr}(W^T S_b W) \leq C \quad \text{调整尺度, 针对}$$

利用 Lagrange 乘子法求解

2.6.2 PCA

• PCA 的思想

将 n 维特征映射到 k 维上 ($k < n$)，这 k 维是全新的正交特征，称为主元，这一组按重要性从大到小排列的新特征，它们是原有特征的线性组合，并且相互之间是不相关的。

• PCA 方法

Example:

Two Dimensional Data—One Dimensional Data

Rows represent samples. Columns represent features. There are ten samples, and each sample has two features.

	x	y
2.5	2.4	
0.5	0.7	
2.2	2.9	
1.9	2.2	
3.1	3.0	
2.1	2.7	
2	1.6	
1	1.1	
1.5	1.6	
1.1	0.9	

PCA 聚类

(1.2) 对特征做方差归一化

① 求均值归零

② 求协方差矩阵 $\text{cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$

③ 求 $\text{cov}(X, Y)$ 的特征值与特征向量 $(|\lambda - C| = 0, \lambda_i P_i = CP)$

④ 将特征值按从大到小排序，选最大的 k 个，并取对应的 P_i

特征向量分作为列向量组成特征向量矩阵

⑤ 投影

Fifth Step: 将样本点投影到选取的特征向量上。假设样例数为 m ，特征数为 n ，减去均值后的样本矩阵为 $\text{DataAdjust}(m \times n)$ ，协方差矩阵是 $n \times n$ ，选取的 k 个特征向量组成的矩阵为 $\text{EigenVectors}(n \times k)$ 。那么投影后的数据 FinalData 为

$$\text{FinalData}(m \times k) = \text{DataAdjust}(m \times n) \times \text{EigenVectors}(n \times k)$$

• 为什么协方差矩阵特征向量就是 k 维理想特征？

在信号处理中认为信号具有较大的方差，噪声有较小的方差。

如图，样本在横轴上的投影方差较大，在纵轴上的投影方差较小，那么认为纵轴上的投影是由噪声引起的。

因此我们认为，最好的 k 维特征是将 n 维样本点转换为 k 维后，每一维上的样本方差都很大，即最大化方差。

特征处理 \Rightarrow 将原点移到样本点中心

我们要求的是最佳的 u ，使得投影后的样本点本方差最大。由投影后均值为 0 ，因此方差为： $\frac{1}{m} \sum_{i=1}^m (x_i^T u)^2 = \frac{1}{m} \sum_{i=1}^m x_i^T x_i u^T u$ ， u 是特征向量， u 是特征向量，最佳的投影直线是特征值 λ 最大时对应的特征向量，其次是 λ 第二大对应的特征向量，依次类推。

因此，我们只需要对协方差矩阵进行特征值分解，得到的前 k 大特征向量就是最佳的 k 维新特征，而且这 k 维新特征是正交的。得到前 k 个 u 以后，样本可以通过以下变换可以得到新的样本。

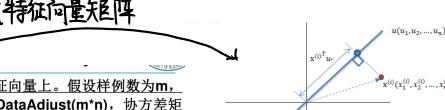
用 λ 表示 $\frac{1}{m} \sum_{i=1}^m (x_i^T u)^2$ ， u 表示 $\frac{1}{m} \sum_{i=1}^m x_i x_i^T u^T$ ，那么上式写成

$$u = u^T \Sigma u$$

由于 u 是单位向量，故 $u^T u = 1$ ，上式两边左乘 u^T ， $u^T u = \lambda u^T \Sigma u = \lambda u$

$$\Sigma u = \lambda u$$

即 λ 是 u 对应的特征值。



红色点表示 $x_i^{(i)}$ ，蓝色点表示 $x_i^{(j)}$ 在 u 上的投影， u 垂直的斜率也是直线的方向向量，而且是单位向量。蓝色点 $x_i^{(j)}$ 上的投影点，黑色点表示的是 $x_i^{(i)}$ > $x_i^{(j)}$ (即 $x_i^{(i)}$ 比 $x_i^{(j)}$ 更远于它在 u 上的投影) 的每一点特征值都为 1，因此投影到 u 上的样本点 (只一个到原点的距離) 的均值仍然是 0。

2.6.3 K-L 变换 (相关性好)

1. 最优描述的 K-L 变换 (沿类间距离大的方向降维)

设共 C 个类别，各类出现的先验概率为 $P(o_i) \quad i=1, \dots, C$

以 x_i 表示来自第 i 类的向量，则第 i 类簇群的自相关矩阵为： $R_i = E(x_i x_i^T)$

混合分布的自相关矩阵 R 是： $R = \sum_i P(o_i) R_i = \sum_i P(o_i) E(x_i x_i^T)$

然后求出 R 的特征值和特征向量： $\Lambda = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{bmatrix}$

$\Phi = (\Phi_1, \Phi_2, \dots, \Phi_p)$

将特征值降序排列，为了降到 d 维，取前 d 个特征向量，构成变换矩阵 A

$$A = \begin{bmatrix} \Phi_1^T \\ \Phi_2^T \\ \vdots \\ \Phi_d^T \end{bmatrix} \Rightarrow A_{d \times p}$$

为什么 K-L 变换是均方误差 (MSE: Mean Square Error) 定义下的最佳变换？

其中， $y^{(i)}$ 表示 D 维向量 y 的第 i 个特征分量。 Φ_i 表示第 i 个特征分量。

且 $y^{(i)} = \Phi_i^T x$ ， $x = \Phi^{-1} y^{(i)} = \sum_{j=1}^D \Phi_j y^{(j)}$

则 $y^{(i)} = E[\Phi_i^T x] = E[\Phi_i^T \sum_{j=1}^D \Phi_j y^{(j)}] = \sum_{j=1}^D E[\Phi_i^T \Phi_j] y^{(j)}$

$= \sum_{j=1}^D E[y^{(j)} \Phi_j^T \Phi_i]$ $= \sum_{j=1}^D \Phi_j^T \Phi_i y^{(j)}$ $= \sum_{j=1}^D \lambda_j \Phi_j y^{(j)}$

从上式可知 $y^{(i)}$ 是 y 的第 i 个特征分量， Φ_i 是第 i 个特征分量的单位向量。

以上方法称为最优描述的 K-L 变换。沿类间距离大的方向降维，从而均方误差最小。

最优描述的 K-L 变换扔掉了最不显着的特征，然而，显着的特征其实并不一定对分类有帮助。还是要找出对分类作用大的特征。

2. 运用 DKL 消除两类问题的特征相关性方法

$y^{(1)} = U^T x^{(1)} \quad x^{(1)} \in \omega_1$ U 是 S_{ω_1} (第一类的协方差) 的特征矢量矩阵

$U^T S_{\omega_1} U = \Lambda_1 = \text{diag}\{ \lambda_1^{(1)}, \lambda_2^{(1)}, \dots, \lambda_n^{(1)} \}$ \leftarrow $y^{(1)}$ 的各分量不相关

(2) 再对 U 作变换：

$$y^{(1)} = \Lambda_1^{-\frac{1}{2}} U^T x^{(1)} \quad \leftarrow$$
 白化变换矩阵

求新变换后的协方差：

$$B = U \Lambda_1^{-\frac{1}{2}} \quad \leftarrow$$
 白化变换

$E[(y^{(1)} - \bar{y}^{(1)})^2] = E[(\Lambda_1^{-\frac{1}{2}} U^T x^{(1)} - \bar{y}^{(1)})^2] = E[(U^T x^{(1)} - \Lambda_1^{\frac{1}{2}} \bar{y}^{(1)})^2]$

$$= \Lambda_1^{-\frac{1}{2}} E[(U^T x^{(1)} - \bar{y}^{(1)})^2] = \Lambda_1^{-\frac{1}{2}} (U^T x^{(1)})^T \Lambda_1^{-\frac{1}{2}} = I$$

即： $B^T S_{\omega_1} B = \Lambda_1^{-\frac{1}{2}} U^T S_{\omega_1} U \Lambda_1^{-\frac{1}{2}} = \Lambda_1^{-\frac{1}{2}} \Lambda_1 \Lambda_1^{-\frac{1}{2}} = I$

这样 $y^{(1)}$ 的协方差矩阵为单位阵，即各分量也不相关。

- 用 U 消除各分量的相关性
- 用 $\Lambda_1^{-\frac{1}{2}}$ 进行白化

这一步是坐标尺度变换，相当于是椭圆整形成圆

(3) 再转换： $\tilde{S}_{\omega_1} = B^T S_{\omega_1} B$ \leftarrow 3.1 步

设 V 为 \tilde{S}_{ω_1} 的特征矢量矩阵，令 $W^T = V^T B = V^T \Lambda_1^{-\frac{1}{2}} U^T$

W 作为最终的特征矢量，变量：最显着的特征为向量 z_1

$$z = W^T x = V^T \Lambda_1^{-\frac{1}{2}} U^T x \quad x \in \omega_1 \text{ 或 } x \in \omega_2$$

第2类方差的特征矢量 z^2 \leftarrow 第1类方差的特征矢量 z_1

上式的各分量是不相等的， $z^2 = x^T U (\Lambda_1^{-\frac{1}{2}} V)^T V^T \Lambda_1^{-\frac{1}{2}} U^T x = x^T U \Lambda_1^{-\frac{1}{2}} \Lambda_1^{-\frac{1}{2}} U^T x$

相关的分量互不相关，于是根据这种准则选择 $z = x^T U \Lambda_1^{-\frac{1}{2}} U^T x = x^T \Lambda_1^{-\frac{1}{2}} x$

分量以降低维数。

上式为特征矢量的正交化。

$B = U \Lambda_1^{-\frac{1}{2}} \quad \leftarrow$ 3.1 步

$\tilde{S}_{\omega_1} = B^T S_{\omega_1} B$

Step 1: $U^T S_{\omega_1} U = \Lambda_1$

Step 2: 白化变换

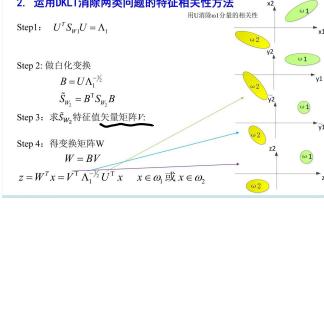
$$B = U \Lambda_1^{-\frac{1}{2}}$$

Step 3: \tilde{S}_{ω_1} 特征值欠量矩阵

Step 4: 得变换矩阵 W

$$W = B^T = V^T \Lambda_1^{-\frac{1}{2}} U^T x \quad x \in \omega_1 \text{ 或 } x \in \omega_2$$

$z = W^T x = V^T \Lambda_1^{-\frac{1}{2}} U^T x \quad x \in \omega_1 \text{ 或 } x \in \omega_2$



3. 基于总的类内离差矩阵 S_w 的特征提取

先按 S_w 提供的信息产生相应的 K-L 坐标系，把原问题各分量相关性消除，得到在新坐标系中的类内离差矩阵 S_w ，再对均值向量在新坐标系中的分离程度 S_w 做出判断

构造特征函数 $J(y) = \frac{u^T S_w u}{u^T S_w u + u^T S_b u}$

其中： $S_w = \frac{u^T S_w u}{u^T S_w u + u^T S_b u}$

该判据表征变换后的特征 $y_i = u^T x$ 的分类性能，表示在新坐标轴上的后 S_w 是类间离差矩阵，而 y_i 表示在新坐标系中该坐标轴包含更多的类内离差矩阵。

$J(y_i) > J(y_j) \geq \dots \geq J(y_k)$ ，取前 d 个较大的 $J(i)$ 值对应的特征值 u_i ($i=1, 2, \dots, d$)，组成变换矩阵 W ，有

$$Y = W^T X$$

4. 依据 S_w, S_b 作 DKL 以降低特征维数的最优压缩方法

设 Λ 和 Ω 是对角矩阵， Λ 的特征对角阵和特征矢量矩阵， Ω 的特征对角阵和特征矢量矩阵

$$V^T \Lambda^{-\frac{1}{2}} U^T S_w U \Lambda^{-\frac{1}{2}} V = \hat{\Lambda}$$

存在正交矩阵 V 使得 $V^T \Lambda^{-\frac{1}{2}} U^T S_w U \Lambda^{-\frac{1}{2}} V = \hat{\Lambda}$

其中 $\hat{\Lambda}$ 是白化变换后的总的类间离差矩阵 Λ' ， Λ' 的特征值对角阵

由于 S_w 的秩不大于 1 ，对于 C 类问题，由于总的类间离差矩阵 S_w 的秩不大于 $c-1$ ，故最多有 $c-1$ 个非零特征值，所以 S_w 最多只有 $c-1$ 个非零特征值，设非零特征值共有 d 个，用这 d 个非零特征值对应的特征向量 v_i ($i=1, 2, \dots, d$) 作变换矩阵， $V^T \Lambda^{-\frac{1}{2}} U^T S_w U$