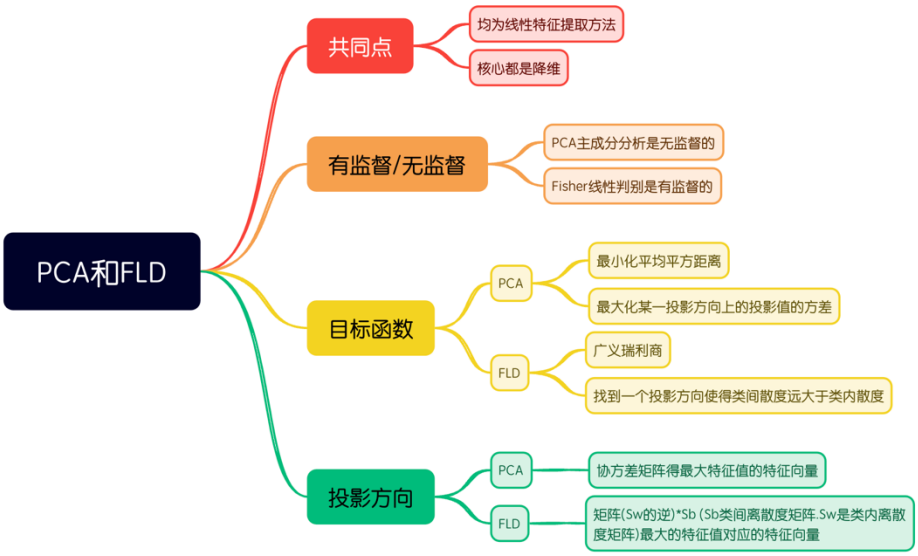
《模式识别》读书报告 第二部分 特征提取



第五章 PCA 主成分分析

主成分分析（Principal Component Analysis）技术是一种线性特征的提取方法。同时也是一种线性的降维技术。

1、二维情况：

数据集有 2 个自由度，我们需要两个值(x,y)来指定一个数据点。因此我们说该数据集的内在维度为 2，与其自然维度相同。

但是若 x 和 y 具有依赖关系，如下图所示，展示了 x 和 y 之间的线性相关性，y 是 x 的线性函数，我们说该图中的内在维度是 1，明显小于其自然维度。原始维度的线性组合是 PCA 的核心部分。

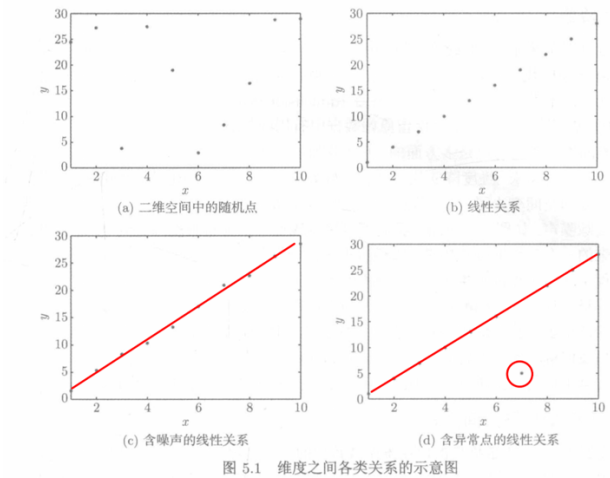
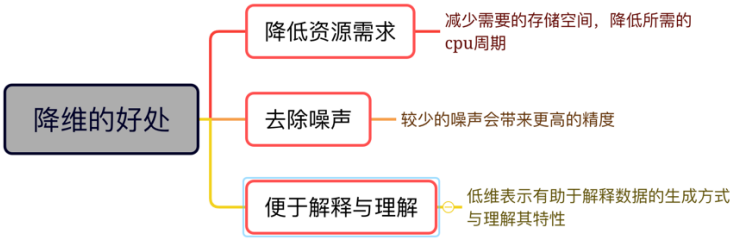


图 5.1 维度之间各类关系的示意图

2、降维

由原始的、相对高维的向量发现其低维表示的这一过程称为降维。尽管使用的维度变少了，但是我们希望降维的过程能够保留原始数据中有用的信息。



3、PCA 与子空间方法

子空间方法：把高维空间中松散分布的样本，通过线性或非线性变换压缩到一个低维的子空间中，在低维的子空间中使样本的分布更紧凑、更有利于分类，同时使计算复杂度减少。

PCA 只考虑线性关系，线性降维的问题可以被看作如何找到线性约束或发现低维的子空间，子空间方法在如何找到约束或子空间方面彼此不同。它们具有各异的评估指标或假设。

4、PCA 降维到零维子空间

想法-形式化-优化实践

想法：

假设我们有 x 的一组实例：

$$X = \{x_1, x_2, \dots, x_N\}$$

这构成了我们学习 PCA 参数所需的训练集，如果不含有噪声，并且存在一个零维的子空间来表示该集合，那么唯一的可能就是：

$$x_1 = x_2 = \cdots = x_N$$

存储x_1需要 D 维，则每个x_i需要的平均维数为$\frac{D}{N}$,而$\lim_{N\to\infty} \frac{D}{N} = 0$，因此认为这是个零维表示也是合理的。

噪声存在的情况下，那么就存在$1 \le i \le j \le N$，存在$x_i \ne x_j$。我们仍然需要找出一个向量 m 来表示 X 中的每个元素，关键在于如何确定最优性。

形式化：

假设噪声很小，我们想找到一个 m，该向量接近于 X 中的所有元素，接近即是“小距离”。

$$m^* = arg\min_m \frac{1}{N}\sum_{i=1}^{N}\|x_i - m\|^2$$

优化：

我们记

$$J = \frac{1}{N}\sum_{i=1}^{N}\|x_i - m\|^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)^T(x_i - m)$$

$$\frac{\partial J}{\partial m} = \frac{2}{N}\sum_{i=1}^{N}(m - x_i)$$

将此项置 0 可得以下最优性条件：

$$m = \frac{1}{N}\sum_{i=1}^{N}x_i$$

也就是说，最好的零维表示为所有训练样本的均值。

5、PCA 降维到一维子空间

形式化：

$$x = x_0 + aw$$

x_0与 w 由子空间确定，a 由元素 x 决定。

令$x_0 = \bar{x}$。

$$x_i \approx \bar{x} + a_i w$$

残差:

$$res = x_i - (\bar{x} + a_i w)$$

通常被认为是由噪声引起的，我们需要发现的参数是a_i,以及 w。我们记

$$a = (a_1, a_2, \dots, a_N)^T$$

并定义目标 J 来最小化平均平方距离:

$$J(w, a) = \frac{1}{N}\sum_{i=1}^{N} \|x_i - (\bar{x} + a_i w)\|^2$$

最优性条件与化简：

现在我们来计算偏导数并将其置为 0

$$\frac{\partial J}{\partial a_i} = \frac{2}{N}\left(a_i\|\boldsymbol{w}\|^2 - \boldsymbol{w}^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\right) = 0 \quad \forall i\,, \tag{5.9}$$

$$\frac{\partial J}{\partial \boldsymbol{w}} = \frac{2}{N}\sum_{i=1}^{N}\left(a_i^2\boldsymbol{w} - a_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\right) = \boldsymbol{0}\,. \tag{5.10}$$

公式 (5.9) 为我们提供了 a_i 的解, 为

$$a_i = \frac{\boldsymbol{w}^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}})}{\|\boldsymbol{w}\|^2} = \frac{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\boldsymbol{w}}{\|\boldsymbol{w}\|^2}\,. \tag{5.11}$$

注意到 $\boldsymbol{x}_i - \bar{\boldsymbol{x}}$ 在 $\boldsymbol{w}$ 上的投影是 $\frac{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\boldsymbol{w}}{\|\boldsymbol{w}\|^2}\boldsymbol{w}$, 我们可以得出结论: a_i 的最优值可以看作 $\boldsymbol{x}_i - \bar{\boldsymbol{x}}$ 投影到 $\boldsymbol{w}$ 上后带符号的长度. 当 $\boldsymbol{x}_i - \bar{\boldsymbol{x}}$ 与 $\boldsymbol{w}$ 之间的夹角大于 $90°$ 时, $a_i \leqslant 0$.

在前进到处理公式 (5.10) 之前, 对 $a_i\boldsymbol{w}$ 的进一步检查表明, 对于任意非零标量 $c \in \mathbb{R}$, 有

$$a_i\boldsymbol{w} = \frac{\boldsymbol{w}^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\boldsymbol{w}}{\|\boldsymbol{w}\|^2} = \frac{(c\boldsymbol{w})^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(c\boldsymbol{w})}{\|c\boldsymbol{w}\|^2}\,. \tag{5.12}$$

令||w||=1,这一假设会极大简化问题，而不改变问题的解。我们有

$$a_i = \boldsymbol{w}^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\boldsymbol{w}\,.$$

$$J(\boldsymbol{w}, \boldsymbol{a}) = \frac{1}{N}\sum_{i=1}^{N}\left[\|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|^2 - a_i^2\right]\,.$$

因此，我们知道最佳的参数通过最大化$\frac{1}{N}\sum_{i=1}^{N} a_i^2$得到。

6、与特征分解的联系

$$Cov(x)w = \frac{\sum_{i=1}^{N} a_i^2}{N}w$$

其中$Cov(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^T(x_i - \bar{x})$,为 x 的协方差矩阵。因此我们可以得到最优解 w 必定是 x 的协方差矩阵的特征向量，并且特征值为$\frac{\sum_{i=1}^{N} a_i^2}{N}$。因此，为了最大化$\frac{1}{N}\sum_{i=1}^{N} a_i^2$，一维降维需要我们选取对应最大特征值的特征向量。

7、PCA 投影到更多维度

谱分解表明:

$$Cov(x) = \sum_{i=1}^{D} \lambda_i \xi_i \xi_i^T$$

其中 Cov(x)有 D 个特征向量ξ_i,对应的特征值是λ_i.满足$\lambda_1 > \lambda_2 > \lambda_3 > , \ldots , > \lambda_D$。
$\xi_i^T \xi_j = 0$, $||\xi_i|| = 1$。若构造一个矩阵 E，其第 i 列由ξ_i组成。

$$EE^T = E^T E = I .$$

那么, 我们就可以证明

$$\begin{aligned} \boldsymbol{x} &= \bar{\boldsymbol{x}} + (\boldsymbol{x} - \bar{\boldsymbol{x}}) \\ &= \bar{\boldsymbol{x}} + EE^T(\boldsymbol{x} - \bar{\boldsymbol{x}}) \\ &= \bar{\boldsymbol{x}} + (\boldsymbol{\xi}_1^T(\boldsymbol{x} - \bar{\boldsymbol{x}}))\boldsymbol{\xi}_1 + (\boldsymbol{\xi}_2^T(\boldsymbol{x} - \bar{\boldsymbol{x}}))\boldsymbol{\xi}_2 + \cdots + (\boldsymbol{\xi}_D^T(\boldsymbol{x} - \bar{\boldsymbol{x}}))\boldsymbol{\xi}_D \end{aligned}$$

由此可以得到ξ_i是第 i 个投影方向，系数为:$\xi_i^T(x - \bar{x})$。

8、完整的 PCA 算法

算法 5.1 PCA 算法

1: **输入**: 一个 D 维训练集 $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ 和一个新的 (更低的) 维度 d $(d < D)$.
2: 计算均值

$$\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i .$$

3: 计算协方差矩阵

$$\mathrm{Cov}(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T .$$

4: 找到 $\mathrm{Cov}(\boldsymbol{x})$ 的谱分解, 得到特征向量 $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_D$ 及其对应的特征值 $\lambda_1, \lambda_2, \ldots, \lambda_D$. 请注意, 特征值是按序排列的, 使得 $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_D \geqslant 0$.
5: 对于任一 $\boldsymbol{x} \in \mathbb{R}^D$, 其新的低维表示是

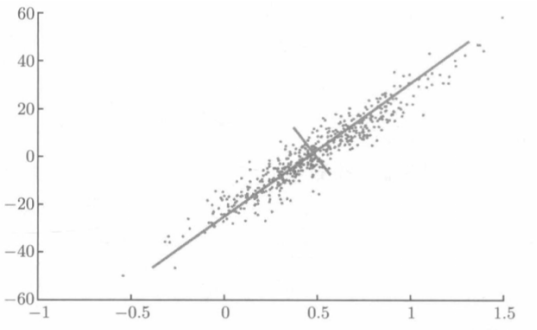
$$\boldsymbol{y} = \left(\boldsymbol{\xi}_1^T(\boldsymbol{x} - \bar{\boldsymbol{x}}), \boldsymbol{\xi}_2^T(\boldsymbol{x} - \bar{\boldsymbol{x}}), \ldots, \boldsymbol{\xi}_d^T(\boldsymbol{x} - \bar{\boldsymbol{x}})\right)^T \in \mathbb{R}^d , \tag{5.30}$$

原始的 $\boldsymbol{x}$ 可通过下式近似得到

$$\boldsymbol{x} \approx \bar{\boldsymbol{x}} + (\boldsymbol{\xi}_1^T(\boldsymbol{x} - \bar{\boldsymbol{x}}))\boldsymbol{\xi}_1 + (\boldsymbol{\xi}_2^T(\boldsymbol{x} - \bar{\boldsymbol{x}}))\boldsymbol{\xi}_2 + \cdots + (\boldsymbol{\xi}_d^T(\boldsymbol{x} - \bar{\boldsymbol{x}}))\boldsymbol{\xi}_d . \tag{5.31}$$

9、从最大化方差出发

PCA 正是在最大化投影到某一方向上后投影值的方差。



两条新线表示协方差矩阵的两个特征向量，长线投影值的方差最大。换句话说 PCA 也可以通过最大化某一投影方向上投影值的方差推导得到，并且该形式化方法的最优解与最小化平均平方距离所获得的最优解相同。

10、高斯数据的 PCA

假设$x \sim N(\mu, \Sigma)$,在一般情况下我们不知道方差和均值的值，我们通过极大似然估计得到这些项为$\bar{x}$和$Cov(x)$。

令Λ是 Cov(x)的特征值构成的对角矩阵。根据正态分布特性，很容易得到新的 PCA 表示的 y 也是服从正态分布的，如果使用了全部的投影方向，则参数估计为

$$y \sim N(0, \Lambda)$$

也就是说 PCA 先对$\bar{x}$执行平移，再进行旋转，使得正态分布的轴能与坐标轴平行。

如果只使用了前 d 个特征向量，那么我们可以定义一个 d*d 的矩阵$\Lambda_d = diag(\lambda_1, \lambda_2, \ldots, \lambda_d)$，并且$y_d \sim N(0, \Lambda_d)$。

11、非高斯数据的 PCA

如果数据不是高斯的，那么 y 在 PCA 之后的均值为 0，协方差矩阵为Λ，因此我们知道 y 的各个维度是不相关的。但是因为 x 是非高斯的，y 就不是多元正态分布，故它们不一定独立。

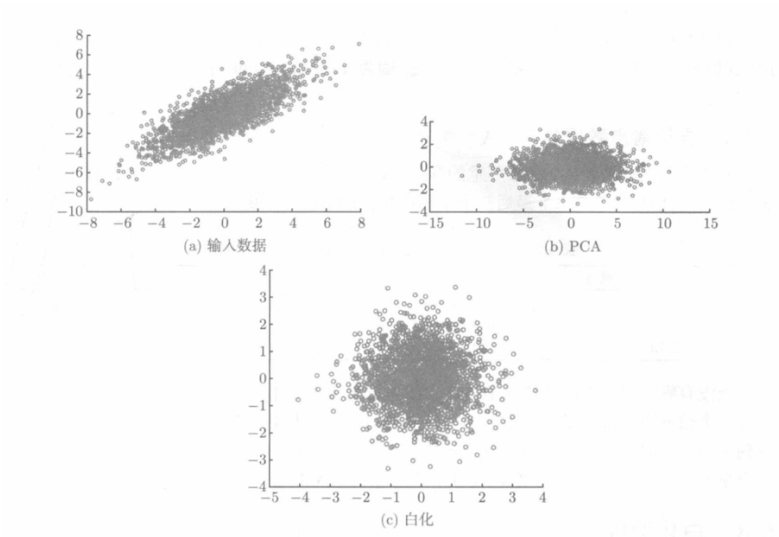


图 5.3 将 PCA 与白化变换应用到高斯数据上. 图 5.3a 是二维输入数据. 在 PCA 之后, 数据被旋转了, 使得数据的两个主轴平行于图 5.3b 中的坐标轴 (即正态分布变成了椭圆体形状的分布). 在白化变换之后, 数据在图 5.3c 中两个主轴上具有相同的长度 (即正态分布变成了球形的分布). 请注意, 不同子图

12、白化变换

白化变换是 PCA 的简单变形，白化变换可以保证 y 中的各个维度具有大致相同的数值范围,即$\mathbb{E}(y_1^2) = \mathbb{E}(y_2^2) = \mathbb{E}(y_3^2) = \cdots \mathbb{E}(y_d^2)$

白化变换通过下式获得新的低维表示

$$y = \left(E_d \Lambda_d^{-\frac{1}{2}}\right)^T (x - \bar{x})$$

在白化变换中我们必须去除任何特征值为 0 的投影方向。

第六章 Fisher 线性判别(FLD)

Fisher 线性判别与主成分分析的最大区别：FLD 是有监督学习，PCA 是无监督学习。如下图所示，将任意的 x 投影到标记为 FLD 的投影方向上，就很容易区分这两个类别，在投影后，正类样本将聚集到一个很小的范围的值域中，不同类别的投影范围几乎不重叠。也就是说一个恰当的线性降维有助于求解二分类问题。

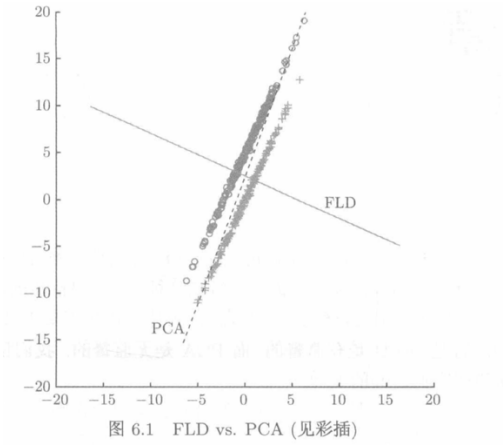
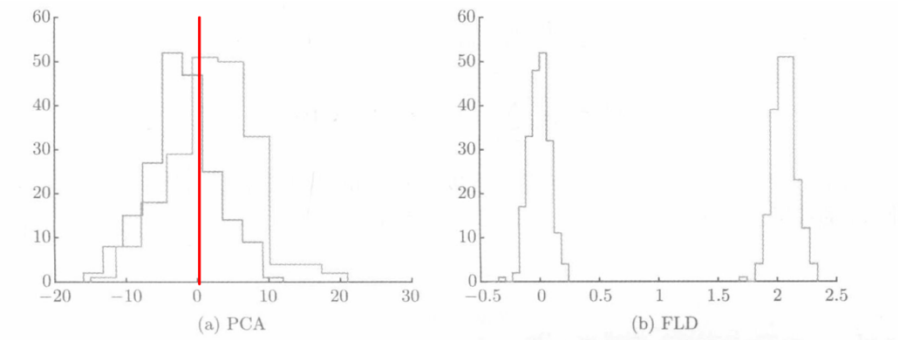


图 6.1 FLD vs. PCA (见彩插)

1、用于二分类的 FLD

给定任意投影方向$w(||w|| = 1)$,我们可以计算投影值为$w^T x$，并且要求正类投影值彼此隔得很远。

以上图为例，我们可以作出上面两个类别 PCA 和 FLD 的直方图:



可见 FLD 的图，两个类的直方图都很尖，这意味着两类样本的投影值在一个很小的范围内。这种情况下我们就可以用每个类的均值来代表该类。且两类的分离程度可以用两个均值来度量。相反 PCA 的直方图就有着比较差的分离能力。

对于二分类问题的 FLD 而言，我们希望最大化以下两个值的比率:

1、 两个均值之间的距离

2、 两个标准差

2、数学语言

正类样本：$X_1 = \{x_i | 1 \le i \le N, y_i = 1\}$,$N_1 = |X_1|$是该子集的大小，也是 X 中正类样本的数量。同理负类样本为：$X_2 = \{x_i | 1 \le i \le N, y_i = 2\}$,$N_2 = |X_2|$。

均值为：$m_1 = \frac{1}{N_1}\sum_{x \in X_1} x$，$m_2 = \frac{1}{N_2}\sum_{x \in X_2} x$

协方差矩阵分别为：$C_1 = \frac{1}{N_1}\sum_{x \in X_1}(x - m_1)(x - m_1)^T$，$C_2 = \frac{1}{N_2}\sum_{x \in X_2}(x - m_2)(x - m_2)^T$

投影均值为：$M_1 = m_1^T w$，$M_2 = m_2^T w$。

正类样本样本投影值方差为：$\frac{1}{N_1}\sum_{x \in X_1}(x^T w - m_1^T w)^2 = w^T C_1 w$

所以正类、负类样本的标准差分别为：$\sigma_1 = \sqrt{w^T C_1 w}$，$\sigma_2 = \sqrt{w^T C_2 w}$

因此我们想要最大化的目标函数可以写为：$\frac{|m_1 - m_2|}{\sigma_1 + \sigma_2}$

3、散度矩阵

最大化式子$\frac{|m_1 - m_2|}{\sigma_1 + \sigma_2}$可能比较复杂，我们可以最大化如下的目标函数：

$$\frac{(m_1 - m_2)^2}{{\sigma_1}^2 + {\sigma_2}^2}$$

也就是

$$\frac{w^T(m_1 - m_2)(m_1 - m_2)^T w}{w^T(C_1 + C_2)w}$$

$C_1 + C_2$那一项利用协方差矩阵来衡量它们有多分散。我们也可以使用散度矩阵来衡量一组点的分散程度，对于一个若干点的集合$z_1, z_2, \ldots, z_n$，其散度矩阵为：

$$S = \sum_{i=1}^{n}(z_i - \bar{z})(z_i - \bar{z})^T$$

传统意义上的 FLD 使用散度矩阵构造目标函数：

$$\frac{w^T(m_1 - m_2)(m_1 - m_2)^T w}{w^T(S_1 + S_2)w}$$

4、类间散度矩阵、类内散度矩阵

类间散度矩阵：

$$S_B = (m_1 - m_2)(m_1 - m_2)^T$$

类间散度矩阵用于衡量两个不同类之间的离散程度。

类内散度矩阵：

$$S_W = S_1 + S_2$$

类内散度矩阵用于衡量原始输入集合每个类别内部的离散程度。

现在可以形式化的定义 FLD 的目标函数为（广义瑞利商）：

$$J = \frac{w^T S_B w}{w^T S_W w}$$

即该目标函数的优化目标是找到一个投影方向使得类间散度远大于类内散度。

上述公式被称为广义瑞利商，找到 J 关于 w 的导数并置 0，可以得到最优性的必要条件为：

$$S_B w = \frac{w^T S_B w}{w^T S_W w} S_W w$$

$\frac{w^T S_B w}{w^T S_W w}$是标量值，所以 w 应该是$S_B$和$S_W$的广义特征向量，$\frac{w^T S_B w}{w^T S_W w}$是其广义特征值。

$$S_B w = \frac{w^T S_B w}{w^T S_W w} S_W w$$

$$= \frac{w^T S_w w}{w^T S_B w} (m_1 - m_2)^T w (m_1 - m_2)$$

$$= c(m_1 - m_2)$$

其中$c = \frac{w^T S_w w}{w^T S_B w} (m_1 - m_2)^T w$，因此该最优性条件给出了最优投影方向为：

$$S_w^{-1}(m_1 - m_2)$$

5、二分类问题 FLD 算法

算法 6.1 二分类问题的 FLD 算法

1: **输入**: 一个 D 维的二分类训练集 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$.

2: 按照本章的公式计算 $\boldsymbol{m}_1$、$\boldsymbol{m}_2$ 和 S_W.

3: 计算

$$\boldsymbol{w} \leftarrow S_W^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)\,.$$

4: 规范化:

$$\boldsymbol{w} \leftarrow \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\,.$$

若S_W不可逆，则采用伪逆来代替S_w的逆矩阵。

6、伪逆

由于 S_W 是两个协方差矩阵的和, 它是半正定的. 因此, 我们可以发现其谱分解为

$$S_W = E\Lambda E^T\,,$$

其中对角矩阵 Λ 包含 S_W 的特征值, 正交矩阵 E 的列包含 S_W 的特征向量. 那么 S_W 的 Moore-Penrose 伪逆是

$$S_W^+ = E\Lambda^+ E^T\,. \tag{6.35}$$

请注意, 当 S_W 可逆时有 $S_W^+ = S_W^{-1}$.

7、多分类的 FLD

类别 K>2 时：

类内散度，即 K 个子集的散度矩阵之和：

$$S_W = \sum_{k=1}^{K} S_k$$

类间散度：

$$S_B = \sum_{k=1}^{K} N_k (m_k - m)(m_k - m)^T$$

总散度矩阵：

$$S_T = \sum_{i=1}^{N}(x_i - m)(x_i - m)^T$$

其中$m = \frac{1}{N}\sum_{i=1}^{N} x_i$，是所有训练点的均值。

$$S_T = S_B + S_W$$

8、求解

多分类问题中，S_B不再是一个秩为 1 的矩阵，算法 6.1 不再适用，但是我们仍然可以通过求解广义特征值问题来找到最佳投影方向：

$$S_B w = \lambda S_W w$$

S_W可逆时，广义特征值问题等价于求解下述特征值问题：

$${S_W}^{-1} S_B w = \lambda w$$

对于一个 K 分类问题，我们最多可以提取 K-1 个有意义的特征（即投影值）。