

有朋友问：为什么感觉对机器学习和神经网络，似乎总是隔着一层纱那样，不清晰，虽然能跑代码，但总是不放心，不知道自己做的是否有根据。这是因为：你和机器学习之间，缺少了一个数学。

费雪的线性判别分析

作者：老齐

Github: <https://github.com/qiwsir/>

1. 缘起

对于线性判别分析的介绍，在网上或者有关书籍中，均能找到不少资料。但是，诸多内容中，未能给予线性判别分析完整地讲解，本文不揣冒昧，斗胆对其中的一部分内容，即费雪的线性判别分析进行整理，权当自己的学习笔记，在这里发布出来，供朋友们参考。

2. 费雪的线性判别分析

英国统计学家费雪（Ronald Fisher）提出的专门为含有两个类别样本的有监督的降维方法，称为“费雪的线性判别分析（Fisher Linear Discriminant Analysis）”。

费雪的线性判别分析基本思想是（如图1所示）：

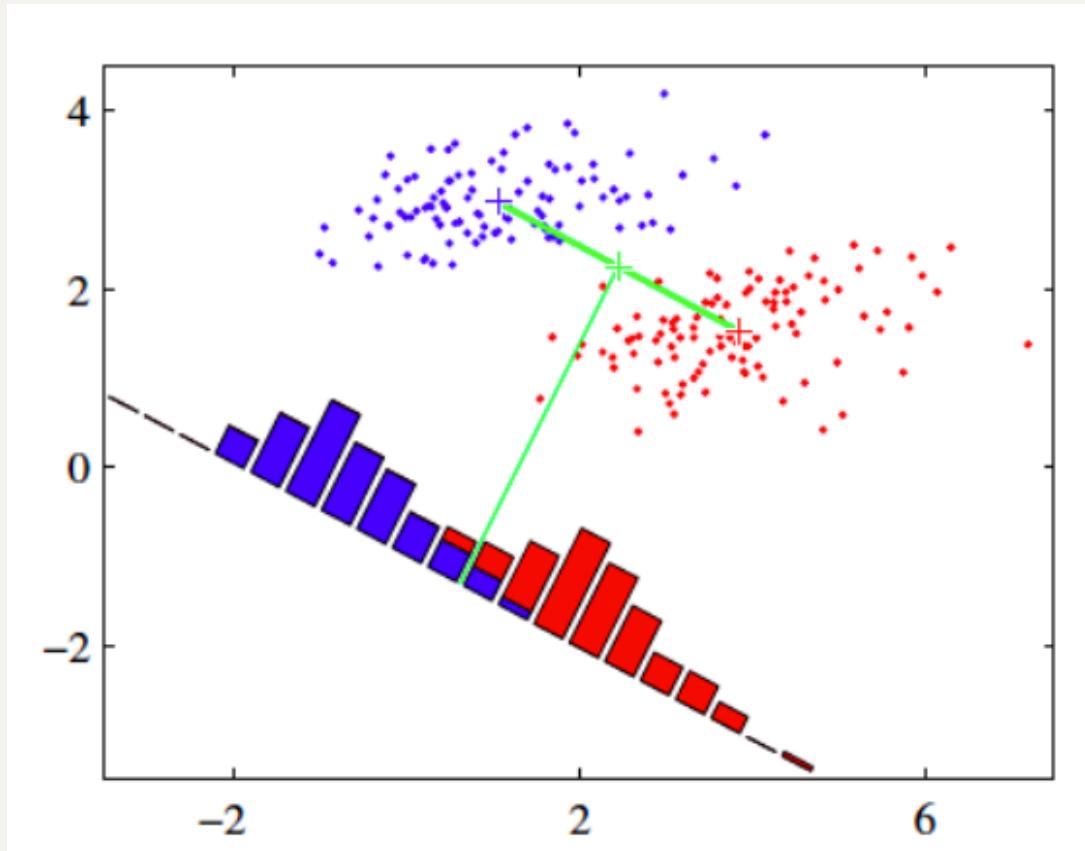


图 1

获得数据在某直线（超平面）上的投影，并同时要求：

- 类内散度最小
- 类间散度最大

多数资料中介绍线性判别分析的时候，都是按照上述费雪所提出的线性判别分析思想讲解的。

本文也首先介绍上述基本思想，而后引申出其他相关问题。

注意：以下内容是以数学推导为主，对此方面如读者有感不足，请参阅：[《机器学习数学基础》](#)

2.1 二分类的样本数据

设数据样本 $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \in \mathbb{R}^d$ ，样本大小为 n ，特征数（维数）是 d 。

假设此样本分为两类，类别 C_1 ，样本数为 n_1 ；类别 C_2 ，样本数量为 n_2 。并且 $n = n_1 + n_2$ 。

2.2 问题：投影可能重叠

设有一条直线 L，用单位向量 \mathbf{w} ($\|\mathbf{w}\|^2 = \mathbf{w}^\top \mathbf{w} = 1$) 表示此直线的方向。

若将样本中的任意一个向量 \mathbf{x} 向此直线投影，得到投影量 $y\mathbf{w}$ ，其中 y 表示投影的大小（长度）。

由于 $\mathbf{x} - y\mathbf{w}$ 与单位向量 \mathbf{w} 正交，即： $\mathbf{w}^\top (\mathbf{x} - y\mathbf{w}) = 0$ ，所以：

$$y = \frac{\mathbf{w}^\top \mathbf{x}}{\mathbf{w}^\top \mathbf{w}} = \mathbf{w}^\top \mathbf{x} \quad (1)$$

故样本中每个样本 $\mathbf{x}_1, \dots, \mathbf{x}_n$ 在直线 L 上的投影大小 y_1, \dots, y_n 为：

$$y_i = \mathbf{w}^\top \mathbf{x}_i, \quad (1 \leq i \leq n) \quad (2)$$

将 $\mathbf{x}_1, \dots, \mathbf{x}_n$ 所在的空间称为 x 空间，投影 y_1, \dots, y_n 所在的空间称为 y 空间。

根据 (2) 式结果，可以得到 y 空间的每个类别样本的投影大小的平均数：

$$\hat{m}_j = \frac{1}{n_j} \sum_{i \in C_j} y_i = \frac{1}{n_j} \sum_{i \in C_j} \mathbf{w}^\top \mathbf{x}_i = \mathbf{w}^\top \left(\frac{1}{n_j} \sum_{i \in C_j} \mathbf{x}_i \right), \quad (j = 1, 2) \quad (3)$$

令 $\mathbf{m}_j = \frac{1}{n_j} \sum_{i \in C_j} \mathbf{x}_i$ ($j = 1, 2$)，代表 x 空间的某个类别的所有样本的平均数向量（样本平均），故（3）式可以继续表示为：

$$\hat{m}_j = \mathbf{w}^T \mathbf{m}_j, \quad (j = 1, 2) \quad (4)$$

由于可以用均值表示数据的集中趋势^[2]，那么向量 \mathbf{m}_j ，就可以作为对应类别的中心的度量。则（4）式即表示将 x 空间的每个类别的样本平均（即该类别的中心，或称“类别中心”），投影到直线 L，得到了 y 空间上每个类别样本投影的平均数（即 \hat{m}_j ，表示该类别样本投影的中心，或称“类别投影中心”）。

于是得到两个类别投影中心的距离：

$$|\hat{m}_2 - \hat{m}_1| = |\mathbf{w}^T \mathbf{m}_2 - \mathbf{w}^T \mathbf{m}_1| = |\mathbf{w}^T (\mathbf{m}_2 - \mathbf{m}_1)| \quad (5)$$

(5) 式说明，类别投影中心的距离 ($|\hat{m}_2 - \hat{m}_1|$) 等于类别中心的距离（即 $|\mathbf{m}_2 - \mathbf{m}_1|$ ）的投影 $|\mathbf{w}^T (\mathbf{m}_2 - \mathbf{m}_1)|$ 。

数据点与数据点之间的距离，表征了数据的分散程度，统计学中使用方差衡量数据的分散程度（样本方差： $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ ）^[2]。因此，可以使用 $(\hat{m}_2 - \hat{m}_1)^2$ 度量不同类别投影中心的分散程度，并将其命名为类间散度（Between-class scatter），即：

$$(\hat{m}_2 - \hat{m}_1)^2 = (\mathbf{w}^T (\mathbf{m}_2 - \mathbf{m}_1))^2 \quad (6)$$

散度与方差有相同的地方，都表示了数据相对平均值的分散程度。图 2 左边表示了散度较大，右边较小。

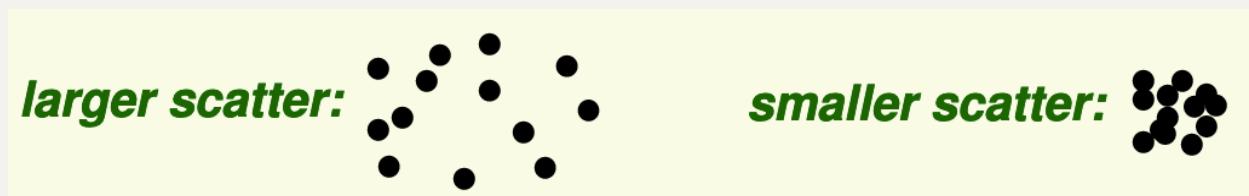


图 2

根据前述费雪的线性判别基本思想，要找到一条适合的直线，用作样本数据投影，并且要能够满足类间散度最大，即找到适合的 \mathbf{w} ，使得 $(\mathbf{w}^T (\mathbf{m}_2 - \mathbf{m}_1))^2$ 最大化。

下面使用拉格朗日乘数法^[3]解决这个问题，但是，做一下转化，将“最大化”转化为“最小化”，在最大化的表达式前面添加一个负号。

$$\begin{aligned} & \text{minimize} && -(\mathbf{w}^T (\mathbf{m}_2 - \mathbf{m}_1))^2 \\ & \text{subject to} && \mathbf{w}^T \mathbf{w} = 1 \end{aligned} \quad (7)$$

定义拉格朗日函数：

$$\begin{aligned} L(\lambda, \mathbf{w}) &= -(\mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_1))^2 + \lambda(\mathbf{w}^T \mathbf{w} - 1) \\ &= -\mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w} + \lambda(\mathbf{w}^T \mathbf{w} - 1) \end{aligned} \quad (8)$$

其中 λ 是拉格朗日乘数。为了计算极值，必须要计算 $\frac{\partial L}{\partial \mathbf{w}}$:

$$\frac{\partial L}{\partial \mathbf{w}} = -2(\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w} + 2\lambda \mathbf{w} \quad (9)$$

要实现 (7) 式中第一个式子的最小化，须令 $\frac{\partial L}{\partial \mathbf{w}} = 0$ ，则：

$$\mathbf{w} = \frac{1}{\lambda}(\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w} \quad (10)$$

因为 $\frac{1}{\lambda}(\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w}$ 是数量（标量），所以：

$$\mathbf{w} \propto (\mathbf{m}_2 - \mathbf{m}_1) \quad (11)$$

这说明直线 L 的方向与两个类别中心的距离矢量方向平行。

但是，如果按照 (11) 式的方式确定直线 L 方向，样本的投影有可能出现图 1 所示的重叠现象。

从降维的角度看，假设是将高维数据降到一维，图 1 演示的降维效果并不会有利于后续分类过程。

对此，费雪提出，应该兼顾类别之间和同一类别之内的样本投影的方差：

- 不同类别之间的样本投影的方差越大越好（如以上说明）
- 同一类之内的样本投影的方差越小越好

这样，在直线（或超平面）上的投影，不同的类别投影中心的距离就尽可能大；同一类别之内的样本的投影尽可能聚集在一起。

2.3 费雪准则

前面已经确定，可以用类别的样本数据的平均数表示每个类别的样本数据的中心（即类别中心）：

$$\mathbf{m}_j = \frac{1}{n_j} \sum_{i \in C_j} \mathbf{x}_i, \quad (j = 1, 2) \quad (12)$$

以下将 \mathbf{m}_j 表述为每个类别分别在 x 空间的平均数向量。

在直线 L 上，类别同样中心用的样本投影的平均数表示：

$$\hat{m}_j = \frac{1}{n_j} \sum_{i \in C_j} y_i = \frac{1}{n_j} \sum_{i \in C_j} \mathbf{w}^T \mathbf{x}_i = \mathbf{w}^T \left(\frac{1}{n_j} \sum_{i \in C_j} \mathbf{x}_i \right) = \mathbf{w}^T \mathbf{m}_j, \quad (j = 1, 2) \quad (13)$$

以下将 \hat{m}_j 表述为在 y 空间的平均数。

仿照方差的定义形式，定义 y 空间衡量类别内数据投影相对本类别投影中心的分散程度的量： y 空间类内散度 (Within-class scatter)：

$$\hat{s}_j^2 = \sum_{i \in C_k} (y_i - \hat{m}_j)^2, \quad (j = 1, 2) \quad (14)$$

前述 (6) 式定义了 y 空间的类间散度：

$$(\hat{m}_2 - \hat{m}_1)^2 = (\mathbf{w}^T (\mathbf{m}_2 - \mathbf{m}_1))^2 \quad (15)$$

根据费雪的思想，既要实现 y 空间的类间散度最大化，同时又要实现 y 空间的类内散度最小化。也就是实现下述函数最大化：

$$J(\mathbf{w}) = \frac{(\hat{m}_2 - \hat{m}_1)^2}{\hat{s}_1^2 + \hat{s}_2^2} \quad (16)$$

2.4 散度矩阵

以下分别写出 x 空间的类内、类间散度的矩阵表示形式，分别称为散度矩阵：

- x 空间的每个类的类内散度矩阵：

$$\mathbf{S}_j = \sum_{i \in C_j} (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T, \quad (j = 1, 2) \quad (17)$$

- x 空间整体的类内散度矩阵：

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2 \quad (18)$$

- x 空间的类间散度矩阵：

$$\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T \quad (19)$$

其中 $\mathbf{m}_j, (j = 1, 2)$ 见 (12) 式。

- y 空间的类内散度：

根据 (14) 式和 (2)、(13) 式， y 空间的类内散度 \hat{s}_j^2 等于：

$$\begin{aligned}
\hat{s}_j^2 &= \sum_{i \in C_k} (y_i - \hat{m}_j)^2, \quad (j = 1, 2) \\
&= \sum_{i \in C_j} (\mathbf{w}^T \mathbf{x}_i - \mathbf{w}^T \mathbf{m}_j)^2 \quad (\text{将 (2) (13) 式代入}) \\
&= \sum_{i \in C_k} \mathbf{w}^T (\mathbf{x}_i - \mathbf{m}_j) (\mathbf{x}_i - \mathbf{m}_j)^T \mathbf{w} \\
&= \mathbf{w}^T \left(\sum_{i \in C_k} (\mathbf{x}_i - \mathbf{m}_j) (\mathbf{x}_i - \mathbf{m}_j)^T \right) \mathbf{w} \quad (\text{将 (17) 式代入, 得到下步结果}) \\
&= \mathbf{w}^T \mathbf{S}_j \mathbf{w}
\end{aligned} \tag{20}$$

故:

$$\hat{s}_1^2 + \hat{s}_2^2 = \mathbf{w}^T \mathbf{S}_1 \mathbf{w} + \mathbf{w}^T \mathbf{S}_2 \mathbf{w} = \mathbf{w}^T \mathbf{S}_W \mathbf{w} \tag{21}$$

- y 空间的类间散度:

y 空间的类间散度 $(\hat{m}_2 - \hat{m}_1)^2$ 等于:

$$\begin{aligned}
(\hat{m}_2 - \hat{m}_1)^2 &= (\mathbf{w}^T \mathbf{m}_2 - \mathbf{w}^T \mathbf{m}_1)^2 \quad (\text{根据 (13) 式}) \\
&= \mathbf{w}^T (\mathbf{m}_2 - \mathbf{m}_1) (\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w} \\
&= \mathbf{w}^T \mathbf{S}_B \mathbf{w} \quad (\text{将 (19) 式代入后})
\end{aligned} \tag{22}$$

于是, (16) 式的费雪准则, 可以用 (21) 式和 (22) 式的结果表示为:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \tag{23}$$

由 (17) 和 (18) 式可知, \mathbf{S}_W 是半正定矩阵, 如果样本大小 n 大于维数 d (这种情况比较常见), 则 \mathbf{S}_W 一般为正定, 且可逆。

由 (19) 式可知, \mathbf{S}_B 是半正定矩阵, 若 $\mathbf{m}_2 \neq \mathbf{m}_1$, 则 $\text{rank} \mathbf{S}_B = 1$ 。

2.5 最优化问题求解

对 (23) 式的最大化求解, 可以有多种方法:

- 法1: 直接计算 $\frac{\partial J}{\partial \mathbf{w}} = 0$ 求解
- 法2: 用线性代数方法: 因为 (23) 式也称为广义瑞利商, 故可以根据冠以特征值求解
- 法3: 拉格朗日乘数法: 参考资料 [1] 中使用的这个方法, 但推导过程不详细。

法1:

最直接的思路：

$$\begin{aligned}\frac{\partial J}{\partial \mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}} \left(\frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \right) \\ &= (\mathbf{w}^T \mathbf{S}_W \mathbf{w}) \frac{\partial(\mathbf{w}^T \mathbf{S}_B \mathbf{w})}{\partial \mathbf{w}} - (\mathbf{w}^T \mathbf{S}_B \mathbf{w}) \frac{\partial(\mathbf{w}^T \mathbf{S}_W \mathbf{w})}{\partial \mathbf{w}} = 0\end{aligned}\quad (24)$$

所以，得到（以下使用了矩阵导数，请见参考资料 [2]）：

$$(\mathbf{w}^T \mathbf{S}_W \mathbf{w}) 2 \mathbf{S}_B \mathbf{w} - (\mathbf{w}^T \mathbf{S}_B \mathbf{w}) 2 \mathbf{S}_W \mathbf{w} = 0 \quad (25)$$

上式两侧同时除以 $\mathbf{w}^T \mathbf{S}_W \mathbf{w}$ ，得：

$$\begin{aligned}\left(\frac{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \right) \mathbf{S}_B \mathbf{w} - \left(\frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \right) \mathbf{S}_W \mathbf{w} &= 0 \\ \mathbf{S}_B \mathbf{w} - J \mathbf{S}_W \mathbf{w} &= 0 \quad (\text{将 (23) 式代入}) \\ \mathbf{S}_B \mathbf{w} &= J \mathbf{S}_W \mathbf{w}\end{aligned}\quad (26)$$

对 (26) 式最后结果等号两边同时左乘 \mathbf{S}_W^{-1} （其中 J 是数量（标量）函数，见 (23) 式），得：

$$\begin{aligned}\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w} &= J \mathbf{S}_W^{-1} \mathbf{S}_W \mathbf{w} \\ J \mathbf{w} &= \mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w} \\ J \mathbf{w} &= \mathbf{S}_W^{-1} (\mathbf{m}_2 - \mathbf{m}_1) (\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w}\end{aligned}\quad (27)$$

又因为上式中的 $(\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w}$ 是数量（标量），故令： $c = (\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w}$ ，则 (27) 式进一步写成：

$$\begin{aligned}J \mathbf{w} &= c \mathbf{S}_W^{-1} (\mathbf{m}_2 - \mathbf{m}_1) \\ \mathbf{w} &= \frac{c}{J} \mathbf{S}_W^{-1} (\mathbf{m}_2 - \mathbf{m}_1)\end{aligned}\quad (28)$$

由此，可知：直线 L 的方向 \mathbf{w} 满足：

$$\mathbf{w} \propto \mathbf{S}_W^{-1} (\mathbf{m}_2 - \mathbf{m}_1) \quad (29)$$

此时能够实现费雪准则的要求，即 $J(\mathbf{w})$ 最大化。

这样，就找到了最佳投影的直线（或超平面），从而实现了最佳降维的目的。

法2^[6]：

由 (18) 和 (19) 式可知， $\mathbf{S}_B^T = \mathbf{S}_B$ 、 $\mathbf{S}_W^T = \mathbf{S}_W$ ，即 \mathbf{S}_B 和 \mathbf{S}_W 都是实对称矩阵，亦即厄米矩阵，故 (23) 式可以看做广义瑞利商^[4]。

从而对 $J(\mathbf{w})$ 的最大化问题，等价于广义特征值问题^[4]，如下述方程：

$$\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w} \quad (30)$$

因为 \mathbf{S}_W 可逆, 故:

$$\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w} = \lambda \mathbf{w} \quad (31)$$

又因为 \mathbf{S}_W^{-1} 和 \mathbf{S}_B 都是半正定矩阵, 所以 $\mathbf{S}_W^{-1} \mathbf{S}_B$ 的特征值 λ 是非负数, 则特征向量 \mathbf{w} 是实特征向量。

于是, 可以对 (30) 式等号两边同时左乘 \mathbf{w}^T , 得:

$$\mathbf{w}^T \mathbf{S}_B \mathbf{w} = \lambda \mathbf{w}^T \mathbf{S}_W \mathbf{w} \quad (32)$$

将 (32) 式代入到 (23) 式, 可得:

$$J(\mathbf{w}) = \lambda \quad (33)$$

由此可知, 只要找出最大的特征值, 即可得到 $J(\mathbf{w})$ 的最大值。

又因为 $\text{rank}(\mathbf{S}_W^{-1} \mathbf{S}_B) = \text{rank} \mathbf{S}_B = 1$, 则说明 (31) 式中的 $\mathbf{S}_W^{-1} \mathbf{S}_B$ 只有一个大于零的特征值。

将 (19) 式中的 \mathbf{S}_B 代入到 (31) 式中, 得:

$$\mathbf{S}_W^{-1} (\mathbf{m}_2 - \mathbf{m}_1) (\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w} = \lambda \mathbf{w} \quad (34)$$

上式中的 $(\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w}$ 是数量 (标量), 故令: $c = (\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w}$, 则上式进一步写成:

$$\mathbf{w} = \frac{c}{\lambda} \mathbf{S}_W^{-1} (\mathbf{m}_2 - \mathbf{m}_1) \quad (35)$$

从而得到与 (29) 式同样的结论和解释。

法3:

对于 (23) 式, 因为分子分母都有 \mathbf{w} , 故 $J(\mathbf{w})$ 与 \mathbf{w} 的长度无关, 可以设 $\mathbf{w}^T \mathbf{S}_W \mathbf{w} = 1$, 则对 (23) 式的最大化问题, 即转化为^[1]:

$$\begin{aligned} & \min && -\mathbf{w}^T \mathbf{S}_B \mathbf{w} \\ & \text{subject to} && \mathbf{w}^T \mathbf{S}_W \mathbf{w} = 1 \end{aligned} \quad (36)$$

根据拉格朗日乘数法^[5]:

$$L(\mathbf{w}, \lambda) = -\mathbf{w}^T \mathbf{S}_B \mathbf{w} + \lambda (\mathbf{w}^T \mathbf{S}_W \mathbf{w} - 1) \quad (37)$$

计算 $\frac{\partial L}{\partial \mathbf{w}}$:

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{w}} &= -\frac{\partial(\mathbf{w}^T \mathbf{S}_B \mathbf{w})}{\partial \mathbf{w}} + \lambda \frac{\partial(\mathbf{w}^T \mathbf{S}_W \mathbf{w} - 1)}{\partial \mathbf{w}} \\
&= -(\mathbf{S}_B + \mathbf{S}_B^T) \mathbf{w} + \lambda (\mathbf{S}_W + \mathbf{S}_W^T) \mathbf{w} \\
&= -2\mathbf{S}_B \mathbf{w} + 2\lambda \mathbf{S}_W \mathbf{w} \quad (\because \mathbf{S}_B^T = \mathbf{S}_B, \mathbf{S}_W^T = \mathbf{S}_W)
\end{aligned} \tag{38}$$

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

$$\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w} \tag{39}$$

因为 \mathbf{S}_W 可逆, 所以有: $\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w} = \lambda \mathbf{w}$, 再将 (19) 式的 \mathbf{S}_B 代入此时, 得到:

$$\mathbf{S}_W^{-1} (\mathbf{m}_2 - \mathbf{m}_1) (\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w} = \lambda \mathbf{w} \tag{40}$$

这与 (34) 式雷同, 只不过 (40) 中的 λ 是拉格朗日乘数, 但形式一样, 故可得:

$$\mathbf{w} \propto \mathbf{S}_W^{-1} (\mathbf{m}_2 - \mathbf{m}_1) \tag{41}$$

与 (29) 同样的结果。

2.6 小结

在以上讨论中, 使用三种方法找到了 \mathbf{w} 的方向, 至此, 事实上实现的是数据的降维 (注意区别于 PCA), 并没有实现对数据属于哪一个类别的“判别分析”——这一点非常重要, 有的资料就此为止, 未继续阐明如何判断所属类别。

当找到了 \mathbf{w} 方向之后, 假设有数据 \mathbf{x} , 要判断属于哪一个类别, 必须还要有阈值 w_0 , 即:

- 当 $\mathbf{w}^T \mathbf{x} \leq -w_0$ 时, \mathbf{x} 属于 C_1 类
- 否则, 属于 C_2 类

而 w_0 是多少? 这在费雪的线性判别分析中并未提及。所以需要进一步探讨。

下面要探讨 w_0 是多少, 并进而实现对数据所述类别的判别分析。

3. 计算判别阈值

如果要判别某个样本属于哪一类, 必须计算出阈值 w_0 , 求解方法有两种:

1. 贝叶斯方法。此方法在另外一篇《线性判别分析》中详解
2. 最小二乘法。此处演示此方法的求解过程

3.1 最小二乘法^[6]

关于最小二乘法的详细讲解, 请阅读参考资料 [2] 的有关章节, 在其中对最小二乘法通过多个角度给予了理论和应用的介绍。

将两个类别的线性边界写作:

$$g(\mathbf{x}_i) = \mathbf{w}^T \mathbf{x}_i + w_0 \quad (42)$$

相应的最小平方误差函数:

$$E = \frac{1}{2} \sum_{i=1}^n (g(\mathbf{x}_i) - r_i)^2 = \frac{1}{2} \sum_{i=1}^n (\mathbf{w}^T \mathbf{x}_i + w_0 - r_i)^2 \quad (43)$$

其中, r_i 是样本数据的类别标签, 即样本类别的真实值。若 $i \in C_1$, 则 r_i 为正例, 否则为负例, 不妨分别假设两个类别的标签分别是:

$$r_{i,i \in C_1} = \frac{n}{n_1}, r_{i,i \in C_2} = -\frac{n}{n_2} \quad (43-2)$$

将 (43) 式分别对 w_0 和 \mathbf{w} 求导:

$$\frac{\partial E}{\partial w_0} = \sum_{i=1}^n (\mathbf{w}^T \mathbf{x}_i + w_0 - r_i) \quad (44)$$

$$\frac{\partial E}{\partial \mathbf{w}} = \sum_{i=1}^n (\mathbf{w}^T \mathbf{x}_i + w_0 - r_i) \mathbf{x}_i \quad (45)$$

令 (44) 式为零, 即:

$$\begin{aligned} \frac{\partial E}{\partial w_0} &= \sum_{i=1}^n (\mathbf{w}^T \mathbf{x}_i + w_0 - r_i) = 0 \\ \sum_{i=1}^n (\mathbf{w}^T \mathbf{x}_i) + \sum_{i=1}^n w_0 - \sum_{i=1}^n r_i &= 0 \\ \sum_{i=1}^n (\mathbf{w}^T \mathbf{x}_i) + nw_0 - \sum_{i=1}^n r_i &= 0 \end{aligned} \quad (46)$$

所以:

$$\begin{aligned} w_0 &= -\frac{1}{n} \sum_{i=1}^n (\mathbf{w}^T \mathbf{x}_i) + \frac{1}{n} \sum_{i=1}^n r_i \\ &= -\mathbf{w}^T \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \right) + \frac{1}{n} \sum_{i=1}^n r_i \end{aligned} \quad (47)$$

其中：

- $\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$ 是样本平均值（向量），记作 \mathbf{m} ；
- 前述 (43-2) 式 $r_{i,i \in C_1} = \frac{n}{n_1}, r_{i,i \in C_2} = -\frac{n}{n_2}$ ，则 $\frac{1}{n} \sum_{i=1}^n r_i = \frac{1}{n} (n_1 \frac{n}{n_1} - n_2 \frac{n}{n_2}) = 0$ 。

所以，(47) 最终得：

$$\mathbf{w}_0 = -\mathbf{w}^\top \mathbf{m} \quad (48)$$

而对于 \mathbf{w} ，在前述最优化求解中，已经得到： $\mathbf{w} \propto \mathbf{S}_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$ ，（如 (41) 式），又因为 \mathbf{w} 的表示的是方向（单位向量），或者说直线的长度不影响边界，故可直接令：

$$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1) \quad (49)$$

于是，可以用：

$$\mathbf{w}^\top \mathbf{x} \leq \mathbf{w}^\top \mathbf{m} \quad (50)$$

作为判别标准。

3.2 检验

若 $\mathbf{x} = \mathbf{m}_1$ ，很显然，此样本属于 C_1 类，利用 (50) 式对此结论进行检验：

$$\begin{aligned} \mathbf{w}^\top \mathbf{x} - \mathbf{w}^\top \mathbf{m} &= \mathbf{w}^\top (\mathbf{x} - \mathbf{m}) = \mathbf{w}^\top (\mathbf{m}_1 - \mathbf{m}) \\ &= (\mathbf{S}_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1))^\top (\mathbf{m}_1 - \mathbf{m}) \quad (\text{将 (49) 式代入}) \\ &= (\mathbf{m}_2 - \mathbf{m}_1)^\top (\mathbf{S}_W^{-1})^\top (\mathbf{m}_1 - \mathbf{m}) \end{aligned} \quad (51)$$

- 由 (18) 式知： $\mathbf{S}_W^\top = \mathbf{S}_W \implies (\mathbf{S}_W^{-1})^\top = \mathbf{S}_W^{-1}$
- $\mathbf{m} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \frac{1}{n} (n_1 \mathbf{m}_1 + n_2 \mathbf{m}_2)$

于是，(51) 式继续计算如下：

$$\begin{aligned} \mathbf{w}^\top \mathbf{x} - \mathbf{w}^\top \mathbf{m} &= (\mathbf{m}_2 - \mathbf{m}_1)^\top \mathbf{S}_W^{-1} (\mathbf{m}_1 - \frac{1}{n} (n_1 \mathbf{m}_1 + n_2 \mathbf{m}_2)) \\ &= (\mathbf{m}_2 - \mathbf{m}_1)^\top \mathbf{S}_W^{-1} (\frac{n_2}{n} \mathbf{m}_1 - \frac{n_2}{n} \mathbf{m}_2) \\ &= -\frac{n_2}{n} (\mathbf{m}_2 - \mathbf{m}_1)^\top \mathbf{S}_W^{-1} (\mathbf{m}_2 - \mathbf{m}_1) < 0 \end{aligned} \quad (52)$$

其中的 \mathbf{S}_W^{-1} (半) 正定。

检验成功。

3.3 用最小二乘法计算 \mathbf{w}

用最小二乘法能够计算出 \mathbf{w}_0 ，也可以计算 \mathbf{w} 。前面已经计算过了 \mathbf{w} 的方向，这里再用最小二乘法计算，作为对此方法的深入理解。

在 (45) 式中，得到了 $\frac{\partial E}{\partial \mathbf{w}}$ ，令它等于零，则可以计算 \mathbf{w} ，但是步骤比较繁琐，以下仅供参考。

由 (17) 式可得：

$$\begin{aligned}
 \mathbf{S}_j &= \sum_{i \in C_j} (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T, \quad (j = 1, 2) \\
 &= \sum_{i \in C_j} \mathbf{x}\mathbf{x}_i^T - \mathbf{m}_j \sum_{i \in C_j} \mathbf{x}_i^T - \sum_{i \in C_j} \mathbf{x}_i \mathbf{m}_j^T + n_j \mathbf{m}_j \mathbf{m}_j^T \\
 &= \sum_{i \in C_j} \mathbf{x}\mathbf{x}_i^T - \mathbf{m}_j(n_j \mathbf{m}_j) - (n_j \mathbf{m}_j) \mathbf{m}_j^T + n_j \mathbf{m}_j \mathbf{m}_j^T \\
 &= \sum_{i \in C_j} \mathbf{x}\mathbf{x}_i^T - \mathbf{m}_j(n_j \mathbf{m}_j)
 \end{aligned} \tag{53}$$

所以：

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2 = \sum_{i=1}^n \mathbf{x}\mathbf{x}_i^T - n_1 \mathbf{m}_1 \mathbf{m}_1^T - n_2 \mathbf{m}_2 \mathbf{m}_2^T \tag{54}$$

令 (45) 式的 $\frac{\partial E}{\partial \mathbf{w}} = 0$ ，即：

$$\begin{aligned}
 \sum_{i=1}^n (\mathbf{w}^T \mathbf{x}_i + w_0 - r_i) \mathbf{x}_i &= 0 \\
 \sum_{i=1}^n (\mathbf{w}^T \mathbf{x}_i - \mathbf{w}^T \mathbf{m} - r_i) \mathbf{x}_i &= 0 \quad (\text{将 (48) 式代入}) \\
 \sum_{i=1}^n (\mathbf{w}^T \mathbf{x}_i - \mathbf{w}^T \mathbf{m}) \mathbf{x}_i &= \sum_{i=1}^n r_i \mathbf{x}_i \\
 \sum_{i=1}^n \mathbf{x}_i (\mathbf{x}_i^T - \mathbf{m}^T) \mathbf{w} &= \sum_{i=1}^n r_i \mathbf{x}_i
 \end{aligned} \tag{55}$$

下面对上式等号左右两边分别计算：

$$\begin{aligned}
\text{左边} &= \sum_{i=1}^n \mathbf{x}_i (\mathbf{x}_i^T - \mathbf{m}^T) \mathbf{w} \\
&= \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T - \sum_{i=1}^n \mathbf{x}_i \mathbf{m}^T \right) \mathbf{w} \\
&= \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T - n \mathbf{m} \mathbf{m}^T \right) \mathbf{w} \\
&= \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T - \frac{1}{n} (n_1 \mathbf{m}_1 + n_2 \mathbf{m}_2) (n_1 \mathbf{m}_1 + n_2 \mathbf{m}_2)^T \right) \mathbf{w} \\
&= \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T - n_1 \mathbf{m}_1 \mathbf{m}_1^T - n_2 \mathbf{m}_2 \mathbf{m}_2^T + \frac{n_1 n_2}{n} (\mathbf{m}_2 - \mathbf{m}_1) (\mathbf{m}_2 - \mathbf{m}_1)^T \right) \mathbf{w} \\
&= \left(\mathbf{S}_W + \frac{n_1 n_2}{n} \mathbf{S}_B \right) \mathbf{w} \quad (\text{代入 (54) 式和 (19) 式})
\end{aligned}$$

$$\begin{aligned}
\text{右边} &= \sum_{i=1}^n r_i \mathbf{x}_i \\
&= \frac{n}{n_1} \sum_{i \in C_1} \mathbf{x}_i - \frac{n}{n_2} \sum_{i \in C_2} \mathbf{x}_i \\
&= n(\mathbf{m}_1 - \mathbf{m}_2)
\end{aligned}$$

$$\begin{aligned}
\therefore \quad &\left(\mathbf{S}_W + \frac{n_1 n_2}{n} \mathbf{S}_B \right) \mathbf{w} = n(\mathbf{m}_1 - \mathbf{m}_2) \\
&\mathbf{S}_W \mathbf{w} + \frac{n_1 n_2}{n} \mathbf{S}_B \mathbf{w} = n(\mathbf{m}_1 - \mathbf{m}_2) \\
&\mathbf{S}_W \mathbf{w} = -\frac{n_1 n_2}{n} (\mathbf{m}_2 - \mathbf{m}_1) (\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w} + n(\mathbf{m}_1 - \mathbf{m}_2) \\
&= \left(-\frac{n_1 n_2}{n} (\mathbf{m}_2 - \mathbf{m}_1) (\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w} - n \right) (\mathbf{m}_2 - \mathbf{m}_1)
\end{aligned}$$

$$\therefore \quad \mathbf{w} = \mathbf{S}_W^{-1} \left(-\frac{n_1 n_2}{n} (\mathbf{m}_2 - \mathbf{m}_1) (\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w} - n \right) (\mathbf{m}_2 - \mathbf{m}_1)$$

因为 $\left(-\frac{n_1 n_2}{n} (\mathbf{m}_2 - \mathbf{m}_1) (\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w} - n \right)$ 是数量（标量），所以：

$$\mathbf{w} \propto \mathbf{S}_W^{-1} (\mathbf{m}_2 - \mathbf{m}_1) \quad (57)$$

4. 多类别的判别分析^[6]

前面讨论的问题是基于 2.1 节中数据假设，即二分类问题，如果将上述两个类别下的类内散度和类间散度矩阵推广到多类别，就可以实现多类别的判别分析。

4.1 多类别的类内散度矩阵

(18) 式定义了 x 空间两个类别的类内散度矩阵，将其定义方式可以直接推广到多类别的类内散度矩阵：

$$\mathbf{S}_W = \sum_{j=1}^k \mathbf{S}_j \quad (58)$$

其中：

- $\mathbf{S}_j = \sum_{i \in C_j} (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T, \quad j = 1, 2, \dots, k,$
- 且 $\mathbf{m}_j = \frac{1}{n_j} \sum_{i \in C_j} \mathbf{x}_i \quad j = 1, \dots, k$, 共计有 k 个类别。
- $n = n_1 + \dots + n_k$ 表示总样本数等于每个类别样本数的和。

4.2 多类别的类间散度矩阵

多类别的类间散度矩阵，不能由 (19) 式直接推广。

令 \mathbf{m} 表示 x 空间的全体样本的平均数（向量），即：

$$\mathbf{m} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \frac{1}{n} \sum_{j=1}^k \sum_{i \in C_j} \mathbf{x}_i = \frac{1}{n} \sum_{j=1}^k (n_j \mathbf{m}_j) \quad (59)$$

对于所有的样本，仿照样本（有偏）方差的定义，可以定义针对 x 空间的所有样本的 **Total scatter matrix**（参考资料 [1] 中称为“全局散度矩阵”。愚以为，由于此概念是针对当前数据集所有样本而言——依然是抽样所得，如果用“全局”一词，容易引起与“总体”的错误联系，其真正含义是：本数据集所有样本散度矩阵）：

$$\mathbf{S}_T = \sum_{i=1}^n (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T \quad (60)$$

将 (60) 式进一步写成：

$$\begin{aligned} \mathbf{S}_T &= \sum_{i=j}^k \sum_{i \in C_j} (\mathbf{x}_i - \mathbf{m}_j + \mathbf{m}_j - \mathbf{m})(\mathbf{x}_i - \mathbf{m}_j + \mathbf{m}_j - \mathbf{m})^T \\ &= \sum_{j=1}^k \sum_{i \in C_j} (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T + \sum_{j=1}^k \sum_{i \in C_j} (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T \\ &\quad + \sum_{j=1}^k \sum_{i \in C_j} (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{m}_j - \mathbf{m})^T + \sum_{j=1}^k (\mathbf{m}_j - \mathbf{m}) \sum_{i \in C_j} (\mathbf{x}_i - \mathbf{m}_j)^T \end{aligned} \quad (61)$$

因为：

- 由 (58) 式可得: $\sum_{j=1}^k \sum_{i \in C_j} (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T = \mathbf{S}_W$
- $\sum_{j=1}^k \sum_{i \in C_j} (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T = \sum_{j=1}^k n_j (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T$
- 因为 $\sum_{i \in C_j} (\mathbf{x}_i - \mathbf{m}_j) = 0$ (每个样本与平均数的差求和, 结果为 0), 故得:
 $\sum_{j=1}^k \sum_{i \in C_j} (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{m}_j - \mathbf{m})^T = \sum_{j=1}^k (\mathbf{m}_j - \mathbf{m}) \sum_{i \in C_j} (\mathbf{x}_i - \mathbf{m}_j)^T = 0$

所以, (61) 式为:

$$\mathbf{S}_T = \mathbf{S}_W + \sum_{j=1}^k n_j (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T \quad (62)$$

令:

$$\mathbf{S}_B = \sum_{j=1}^k n_j (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T \quad (63)$$

即为多类别的类间散度矩阵。

对于 (63) 式, 如果 $k = 2$, 即为二类别下的类间散度矩阵:

$$\begin{aligned} \mathbf{S}_{B_2} &= n_1(\mathbf{m}_1 - \mathbf{m})(\mathbf{m}_1 - \mathbf{m})^T + n_2(\mathbf{m}_2 - \mathbf{m})(\mathbf{m}_2 - \mathbf{m})^T \\ \because \mathbf{m}_1 - \mathbf{m} &= \mathbf{m}_1 - \frac{1}{n}(n_1 \mathbf{m}_1 + n_2 \mathbf{m}_2) = \frac{n_2}{n}(\mathbf{m}_1 - \mathbf{m}_2) \\ \mathbf{m}_2 - \mathbf{m} &= \mathbf{m}_2 - \frac{1}{n}(n_1 \mathbf{m}_1 + n_2 \mathbf{m}_2) = \frac{n_1}{n}(\mathbf{m}_2 - \mathbf{m}_1) \\ \therefore \mathbf{S}_{B_2} &= \frac{n_1 n_2}{n}(\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T \end{aligned} \quad (64)$$

(64) 式最终得到的二类别下的类间散度矩阵和 (19) 式相比, 多了系数 $\frac{n_1 n_2}{n}$, 因为它是常数, 不会影响对 $J(\mathbf{w})$ 最大化求解。

4.3 多类别样本下的费雪准则

假设由 q 个单位向量 $\mathbf{w}_1, \dots, \mathbf{w}_q$ 作为 x 空间样本数据 $\mathbf{x} \in \mathbb{R}^d$ 投影的超平面 (直线) 方向, 得到:

$$y_l = \mathbf{w}_l^T \mathbf{x}, \quad (l = 1, \dots, q) \quad (65)$$

写成矩阵形式:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_q \end{bmatrix} = \begin{bmatrix} \mathbf{w}_1^T \mathbf{x} \\ \vdots \\ \mathbf{w}_q^T \mathbf{x} \end{bmatrix} = [\mathbf{w}_1 \quad \cdots \quad \mathbf{w}_q]^T \mathbf{x} = \mathbf{W}^T \mathbf{x} \quad (66)$$

其中 $\mathbf{W} = [\mathbf{w}_1 \quad \cdots \quad \mathbf{w}_q]$ 是 $d \times q$ 矩阵。

由此，得到了 x 空间的样本 $\mathbf{x}_1, \dots, \mathbf{x}_n$ 投影到 $\mathbf{w}_l, (1 \leq l \leq q)$ 上的投影，即得到 y 空间的数据 $\mathbf{y}_i, (i = 1, \dots, n)$ ：

$$\mathbf{y}_i = \mathbf{W}^T \mathbf{x}_i, \quad (i = 1, \dots, n) \quad (67)$$

仿照 x 空间计算平均值（向量）的方法，计算 y 空间投影数据的平均值：

$$\hat{\mathbf{m}}_j = \frac{1}{n_j} \sum_{i \in C_j} \mathbf{y}_i = \frac{1}{n_j} \sum_{i \in C_j} \mathbf{W}^T \mathbf{x}_i = \mathbf{W}^T \mathbf{m}_j, \quad (j = 1, \dots, k) \quad (68)$$

$$\hat{\mathbf{m}} = \frac{1}{n} \sum_{j=1}^k n_j \hat{\mathbf{m}}_j = \frac{1}{n} \sum_{j=1}^k n_j \mathbf{W}^T \mathbf{m}_j = \mathbf{W}^T \mathbf{m} \quad (69)$$

从而定义 y 空间的类内散度矩阵和类间散度矩阵

$$\hat{\mathbf{S}}_W = \sum_{j=1}^k \sum_{i \in C_j} (\mathbf{y}_i - \hat{\mathbf{m}}_j)(\mathbf{y}_i - \hat{\mathbf{m}}_j)^T \quad (70)$$

$$\hat{\mathbf{S}}_B = \sum_{j=1}^k n_j (\hat{\mathbf{m}}_j - \hat{\mathbf{m}})(\hat{\mathbf{m}}_j - \hat{\mathbf{m}})^T \quad (71)$$

将 (67) (68) 代入到 (70)，得到：

$$\begin{aligned} \hat{\mathbf{S}}_W &= \sum_{j=1}^k \sum_{i \in C_k} (\mathbf{W}^T \mathbf{x}_i - \mathbf{W}^T \mathbf{m}_j)(\mathbf{W}^T \mathbf{x}_i - \mathbf{W}^T \mathbf{m}_j)^T \\ &= \sum_{j=1}^k \sum_{i \in C_k} \mathbf{W}^T (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)^T \mathbf{W} \\ &= \mathbf{W}^T \mathbf{S}_W \mathbf{W} \quad (\text{根据 (58) 式}) \end{aligned} \quad (72)$$

将 (68) (69) 带入到 (71)，得到：

$$\begin{aligned} \hat{\mathbf{S}}_B &= \sum_{j=1}^k n_j (\mathbf{W}^T \mathbf{m}_j - \mathbf{W}^T \mathbf{m})(\mathbf{W}^T \mathbf{m}_j - \mathbf{W}^T \mathbf{m})^T \\ &= \sum_{j=1}^k n_j \mathbf{W}^T (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T \mathbf{W} \\ &= \mathbf{W}^T \mathbf{S}_B \mathbf{W} \quad (\text{根据 (63) 式}) \end{aligned} \quad (73)$$

由此，多类别下的费雪准则，其目标函数有两种表示方式：

- 第一种：用矩阵的迹表示

$$J_1(\mathbf{W}) = \text{trace} \left(\hat{\mathbf{S}}_W^{-1} \hat{\mathbf{S}}_B \right) = \text{trace} \left((\mathbf{W}^T \mathbf{S}_W \mathbf{W})^{-1} (\mathbf{W}^T \mathbf{S}_B \mathbf{W}) \right) \quad (74)$$

- 第二种：用行列式表示

$$J_2(\mathbf{W}) = \frac{|\hat{\mathbf{S}}_B|}{|\hat{\mathbf{S}}_W|} = \frac{|\mathbf{W}^T \mathbf{S}_B \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_W \mathbf{W}|} \quad (75)$$

不论以上哪种形式，最终均可得到如下最优化条件：

$$\mathbf{S}_B \mathbf{w}_l = \lambda_l \mathbf{S}_W \mathbf{w}_l, \quad (l = 1, \dots, q) \quad (76)$$

由上式得： $\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w}_l = \lambda_l \mathbf{w}_l$ 。参考（30）式之后的推导， λ_l 为广义特征值， \mathbf{w}_l 是广义特征向量。以（74）式为例，可得如下结论（详细推导过程，请见参考资料[7]中的推导过程）

$$J_1(\mathbf{W}) = \lambda_1 + \dots + \lambda_q \quad (77)$$

因特征值非负，故 $q = \text{rank}(\hat{\mathbf{S}}_W^{-1} \hat{\mathbf{S}}_B)$ 。

又因为 $\hat{\mathbf{m}}$ 是 $\hat{\mathbf{m}}_1, \dots, \hat{\mathbf{m}}_k$ 的线性组合，故：

$$\text{rank}(\hat{\mathbf{S}}_W^{-1} \hat{\mathbf{S}}_B) = \text{rank}(\hat{\mathbf{S}}_B) = \dim \text{span}\{\hat{\mathbf{m}}_1 - \hat{\mathbf{m}}, \dots, \hat{\mathbf{m}}_k - \hat{\mathbf{m}}\} \leq k - 1 \quad (78)$$

故 $q \leq k - 1$ 。

给定包含 $k \geq 2$ 个类别的样本，多类别判别分析所能产生有效线性特征总数最多是 $k - 1$ ，即降维的最大特征数。

参考资料

- [1]. 周志华. 机器学习. 北京：清华大学出版社
- [2]. 齐伟. 机器学习数学基础. 北京：电子工业出版社
- [3]. 拉格朗日乘数法
- [4]. 广义特征值与极小极大原理。以下为此参考文献部分内容摘抄：

1、定义：设 \mathbf{A} 、 \mathbf{B} 为 n 阶方阵，若存在数 λ ，使得方程 $\mathbf{Ax} = \lambda \mathbf{Bx}$ 存在非零解，则称 λ 为 \mathbf{A} 相对于 \mathbf{B} 的广义特征值， \mathbf{x} 为 \mathbf{A} 相对于 \mathbf{B} 的属于广义特征值 λ 的特征向量。

- 当 $\mathbf{B} = \mathbf{I}$ (单位矩阵) 时, 广义特征值问题退化为标准特征值问题。
- 特征向量是非零的
- 广义特征值的求解

$$(\mathbf{A} - \lambda \mathbf{B})\mathbf{x} = \mathbf{0} \text{ 或者 } (\lambda \mathbf{B} - \mathbf{A})\mathbf{x} = \mathbf{0}$$

特征方程: $\det(\mathbf{A} - \lambda \mathbf{B}) = 0$

求得 λ 后代回原方程 $\mathbf{Ax} = \lambda \mathbf{Bx}$ 可求出 \mathbf{x}

2、等价表述

\mathbf{B} 正定, 且可逆, 即 \mathbf{B}^{-1} 存在, 则: $\mathbf{B}^{-1} \mathbf{Ax} = \lambda \mathbf{x}$, 广义特征值问题化为了标准特征值问题。

3、广义瑞丽商

若 \mathbf{A} 、 \mathbf{B} 为 n 阶厄米矩阵 (Hermitian matrix, 或译为“艾米尔特矩阵”、“厄米特矩阵”等), 且 \mathbf{B} 正定, 则:

$$R(\mathbf{x}) = \frac{\mathbf{x}^H \mathbf{Ax}}{\mathbf{x}^H \mathbf{Bx}}, (\mathbf{x} \neq 0) \text{ 为 } \mathbf{A} \text{ 相对于 } \mathbf{B} \text{ 的瑞丽商。}$$

[5]. 谢文睿, 秦州. 机器学习公式详解. 北京: 人民邮电出版社

[6]. 线代启示录: 费雪的判别分析与线性判别分析