多元统计

陈崇双

西南交通大学数学学院统计系

ccsmars@swjtu.edu.cn

2018-2019学年

1. 判别分析
   - 案例
   - 基本思想
   - 距离判别法
   - 贝叶斯判别法
   - Fisher判别法

第一节：案例

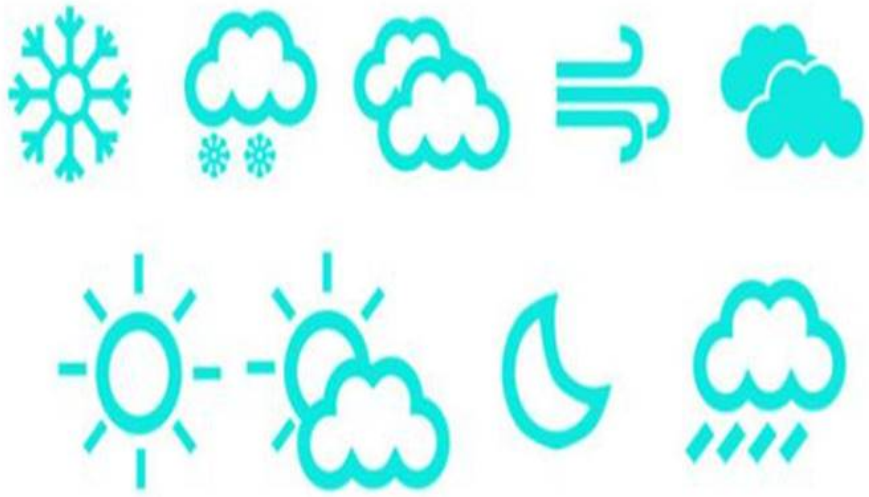
主要内容：四个判别分析案例

例1

病症判断。依据：身体的信息(血糖、血脂、血压、心电等)。

例2

天气预报。依据：气象的信息(气温、气压、湿度、云图、风等)。

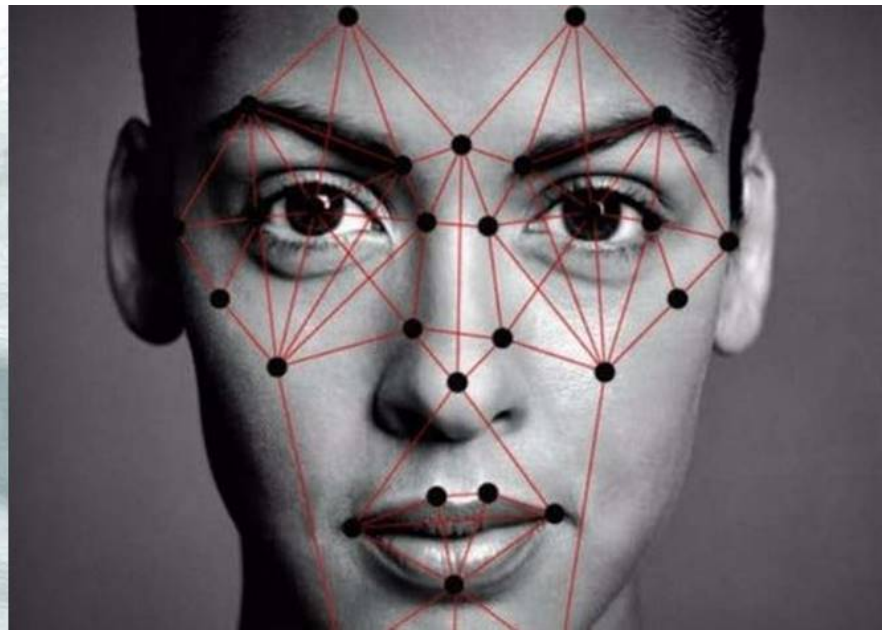


例3

股市涨跌。依据：股票的信息(成交量、振幅、成交额、换手率、总市值、市净率、流通市值、市盈率等)。

例4

人脸识别。依据：人脸的信息(轮廓、肤色、眼鼻嘴等器官特征)。



案例小结

1. 如何根据连续的自变量取值，预测离散的因变量的值；
2. 传统的回归分析难以解决此类问题，而判别分析是合适的统计分析方法。

其他应用案例

地质灾害判别；

油、气、矿、岩勘探定位；

工程、设施设备的状态检测（失效或正常）；

动植物（界、门、纲、目、科、属、种）判断；

垃圾邮件识别；

银行风险评估，等

截止至2018.06


1.输入检索控制条件：
期刊年期：从 不限 年 到 不限 年
指定期：
请输入
更新时间：
不限
来源期刊：
输入期刊名称，ISSN,CN均可
模糊
来源类别：
EI来源期刊
支持基金：
输入基金名称
模糊
作者
输入作者姓名
精确
作者单位：
输入作者单位，全称、简称、曾用名均可
模糊
2.输入内容检索条件：
主题
判别分析
词频
并且包含
输入检索词
词频
精确
或者
关键词
判别分析
词频
并且包含
词频
精确
仅限优先出版论文
中英文扩展检索
在结果中检索
检索文献
定制本次检索式
3.您可以按如下文献分组排序方式选择文献：(分组只对前4万条记录分组,排序只在800万条记录以内有效)
文献分组浏览：学科类别 期刊名称 研究资助基金 研究层次 文献作者 作者单位 中文关键词 发表年度 不分组
"紫色"刊名为"中国知网"独家出版刊物
文献排序浏览：发表时间 相关度 被引频次 下载频次
列表显示
每页记录数：10 20 50
找到 491 条结果 共25页
1 2 3 4 5 6 7 8 9 后页
全选 清除 存盘 定制

计算机软件及计算机应用(92)	化学(71)
轻工业手工业(63)	自动化技术(54)
矿业工程(34)	地质学(28)
电信技术(24)	建筑科学与工程(22)
园艺(19)	安全科学与灾害防治(18)
植物保护(17)	数学(16)
工业通用技术及设备(14)	物理学(14)
电力工业(13	石油天然气工业(13)
中药学(12)	农作物(11)
公路与水路运输(10)	无线电电子学(9)
机械工业(9)	

第二节：判别分析基本思想

主要内容：判别分析的统计模型、基本要求、基本假设；

判别分析的统计模型

有m个p维总体$G_1, G_2, \cdots, G_m$, 分别服从一定的分布$F_1(x)$, $F_2(x), \cdots, F_m(x)$。现在有一个新的样品$\boldsymbol{x} = (x_1, x_2, \cdots, x_p)$, 它可能来自这$m$个总体中的某一个。要依据该样品的$p$项指标，判别它最可能来自哪一个总体。

判别分析的基本要求

1. 分组类型(总体)至少两组；
2. 每组样本数量至少一个；
3. 解释变量必须可测量，才能够计算样本平均值和方差等，从而估计总体的特征。

判别分析的假设条件

1. 每一个自变量不能是其他判别变量的线性组合，即不存在多重共线性问题。
2. 各组变量的协方差矩阵都相等。判别分析最简单和最常用的形式是采用线性判别函数，它们是判别变量的简单线性组合。
3. 判别变量服从多元正态分布，即每个变量的条件分布也为正态分布。

常用的判别分析方法

- 距离判别法
- Bayes判别法
- Fisher判别法

第三节：距离判别法

主要内容：两总体情形，多总体情形

两总体情形

设有两个p维总体G_1和G_2, 待判别的样本$\boldsymbol{x} = (x_1, x_2, \cdots, x_p)^{\mathsf{T}}$，是来自$G_1$或是来自$G_2$。

两总体情形

设有两个p维总体G_1和G_2, 待判别的样本$\boldsymbol{x} = (x_1, x_2, \cdots, x_p)^{\mathsf{T}}$，是来自$G_1$或是来自$G_2$。

理想地，如果总体G_1与G_2的取值范围互不重叠。那么只需视样品$\boldsymbol{x}$落入哪个的取值范围，就可作出准确无误的判别。

两总体情形

设有两个p维总体G_1和G_2, 待判别的样本$\boldsymbol{x} = (x_1, x_2, \cdots, x_p)^{\mathsf{T}}$，是来自$G_1$或是来自$G_2$。

理想地，如果总体G_1与G_2的取值范围互不重叠。那么只需视样品$\boldsymbol{x}$落入哪个的取值范围，就可作出准确无误的判别。

现实中，两总体的取值范围大多有重叠的部分，当新的样品正好落在重叠部分时，关于$\boldsymbol{x}$ 归属的判别难以做到绝对准确。

两总体情形

设有两个p维总体G_1和G_2, 待判别的样本$\boldsymbol{x} = (x_1, x_2, \cdots, x_p)^{\mathsf{T}}$，是来自$G_1$或是来自$G_2$。

理想地，如果总体G_1与G_2的取值范围互不重叠。那么只需视样品$\boldsymbol{x}$落入哪个的取值范围，就可作出准确无误的判别。

现实中，两总体的取值范围大多有重叠的部分，当新的样品正好落在重叠部分时，关于$\boldsymbol{x}$ 归属的判别难以做到绝对准确。

希望建立客观的判别准则，使得判别尽可能准确。

两总体情形

若能客观合理地定义$\boldsymbol{x}$ 到G_1和G_2的距离，分别记为$d(\boldsymbol{x}, G_1)$和$d(\boldsymbol{x}, G_2)$, 则可以按如下规则进行判别:

两总体情形

若能客观合理地定义$\boldsymbol{x}$ 到G_1和G_2的距离，分别记为$d(\boldsymbol{x}, G_1)$和$d(\boldsymbol{x}, G_2)$, 则可以按如下规则进行判别:

$$
\begin{aligned}
&\text{如果} d(\boldsymbol{x}, G_1) < d(\boldsymbol{x}, G_2), \quad \text{则判} \boldsymbol{x} \in G_1 \\
&\text{如果} d(\boldsymbol{x}, G_1) > d(\boldsymbol{x}, G_2), \quad \text{则判} \boldsymbol{x} \in G_2 \\
&\text{如果} d(\boldsymbol{x}, G_1) = d(\boldsymbol{x}, G_2), \quad \text{则待判}
\end{aligned}
$$

两总体情形

G_1, G_2是p维随机变量总体，而不是具体的点，解决办法

两总体情形

G_1, G_2是p维随机变量总体，而不是具体的点，解决办法

1. 在G_1, G_2中各找一个最具代表性的点，用样品到代表点的距离定义为样品到相应总体的距离。

两总体情形

G_1, G_2是p维随机变量总体，而不是具体的点，解决办法

1. 在G_1, G_2中各找一个最具代表性的点，用样品到代表点的距离定义为样品到相应总体的距离。通常将各总体的均值向量作为代表点，因为它是总体取值以概率加权后的“中心”点。

两总体情形

G_1, G_2是p维随机变量总体，而不是具体的点，解决办法

1. 在G_1, G_2中各找一个最具代表性的点，用样品到代表点的距离定义为样品到相应总体的距离。通常将各总体的均值向量作为代表点，因为它是总体取值以概率加权后的“中心”点。
2. 马氏距离，

$$d(\boldsymbol{x}, G_i) = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)}$$

其中$\boldsymbol{\mu}_i$和$\boldsymbol{\Sigma}_i$分别为总体G_i的均值向量和协差阵。

两总体情形

定义样品$\boldsymbol{x}$到总体G_i的距离为$d(\boldsymbol{x}, G_i), i = 1, 2$，并令判别函数

$$W(\boldsymbol{x}) = \frac{1}{2}[d^2(\boldsymbol{x}, G_2) - d^2(\boldsymbol{x}, G_1)]$$

两总体情形

定义样品$\boldsymbol{x}$到总体G_i的距离为$d(\boldsymbol{x}, G_i), i = 1, 2$，并令判别函数

$$W(\boldsymbol{x}) = \frac{1}{2}[d^2(\boldsymbol{x}, G_2) - d^2(\boldsymbol{x}, G_1)]$$

并按下述规则进行判别

$$\text{如果} W(\boldsymbol{x}) > 0, \text{则判} \boldsymbol{x} \in G_1$$

$$\text{如果} W(\boldsymbol{x}) < 0, \text{则判} \boldsymbol{x} \in G_2$$

$$\text{如果} W(\boldsymbol{x}) = 0, \text{则待判}$$

这种方法称为距离判别法。

两总体情形

$W(\boldsymbol{x})$与0的差距越大，样品到两总体的距离差异越明显，作出正确判断的把握越大; $W(\boldsymbol{x})$越接近于0，越容易造成误判。

两总体情形

$W(\boldsymbol{x})$与0的差距越大，样品到两总体的距离差异越明显，作出正确判断的把握越大; $W(\boldsymbol{x})$越接近于0，越容易造成误判。

距离判别法的本质即为去重叠化，两总体G_1和G_2的取值范围本有重叠，通过判别函数$W(\boldsymbol{x})$将其划分成两个不相重叠的部分,

$$R_1 = \{x \in \mathbb{R}^p | W(\boldsymbol{x}) > 0\}, R_2 = \{x \in \mathbb{R}^p | W(\boldsymbol{x}) < 0\}$$

新样品落在哪部分，就判它来自那个对应的总体。

两总体情形

若两总体的协方差阵相等，即$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, 采用马氏距离，则有

$$
\begin{aligned}
W(\boldsymbol{x}) =& d^2(\boldsymbol{x}, G_2) - d^2(\boldsymbol{x}, G_1) \\
=& (\boldsymbol{x} - \boldsymbol{\mu}_2)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2) - (\boldsymbol{x} - \boldsymbol{\mu}_1)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1) \\
=& \boldsymbol{x}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - 2\boldsymbol{x}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 + \boldsymbol{\mu}_2^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 \\
& \qquad - [\boldsymbol{x}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - 2\boldsymbol{x}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1] \\
=& 2\boldsymbol{x}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \boldsymbol{\mu}_2^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 \\
=& 2\boldsymbol{x}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + (\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \\
=& 2 \left(\boldsymbol{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2} \right)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)
\end{aligned}
$$

两总体情形

令$\overline{\boldsymbol{\mu}} = \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}, \boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$，则有

$$W(\boldsymbol{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \overline{\boldsymbol{\mu}}) = \boldsymbol{\alpha}^{\mathsf{T}}(\boldsymbol{x} - \overline{\boldsymbol{\mu}})$$

两总体情形

令$\overline{\boldsymbol{\mu}} = \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}, \boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$，则有

$$W(\boldsymbol{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \overline{\boldsymbol{\mu}}) = \boldsymbol{\alpha}^{\mathsf{T}}(\boldsymbol{x} - \overline{\boldsymbol{\mu}})$$

由于$W(\boldsymbol{x})$是$\boldsymbol{x}$线性函数，又称为线性判别函数, $\boldsymbol{\alpha}$称为判别系数(类似于回归系数)。

Linear discriminant analysis (LDA), normal discriminant analysis (NDA), or discriminant function analysis is a generalization of Fisher's linear discriminant, a method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification.

LDA is closely related to analysis of variance (ANOVA) and regression analysis, which also attempt to express one dependent variable as a linear combination of other features or measurements. However, ANOVA uses categorical independent variables and a continuous dependent variable, whereas discriminant analysis has continuous independent variables and a categorical dependent variable (i.e. the class label). Logistic regression and probit regression are more similar to LDA than ANOVA is, as they also explain a categorical variable by the values of continuous independent variables. These other methods are preferable in applications where it is not reasonable to assume that the independent variables are normally distributed, which is a fundamental assumption of the LDA method.

LDA is also closely related to principal component analysis (PCA) and factor analysis in that they both look for linear combinations of variables which best explain the data. LDA explicitly attempts to model the difference between the classes of data. PCA on the other hand does not take into account any difference in class, and factor analysis builds the feature combinations based on differences rather than similarities. Discriminant analysis is also different from factor analysis in that it is not an interdependence technique: a distinction between independent variables and dependent variables (also called criterion variables) must be made.

LDA works when the measurements made on independent variables for each observation are continuous quantities. When dealing with categorical independent variables, the equivalent technique is discriminant correspondence analysis.

`https://en.wikipedia.org/wiki/Linear_discriminant_analysis`

两总体情形

实际中，$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$常常都未知。为此，从两个总体中分分别抽取样本，估计它们的均值向量$\boldsymbol{\mu}_i$和协差阵$\boldsymbol{\Sigma}_i$。

两总体情形

实际中，$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$常常都未知。为此，从两个总体中分分别抽取样本，估计它们的均值向量$\boldsymbol{\mu}_i$和协差阵$\boldsymbol{\Sigma}_i$。

具体估计方法如下：设从总体G抽取了n个样本(p维)，样本观察值为$\boldsymbol{x}_{(i)} = (x_{i1}, x_{i2}, \cdots, x_{ip})^{\mathsf{T}}$，$i = 1, 2, \cdots, n$。其中$x_{ij}$表示第$i$个样本中第$j$个指标的观察值。

两总体情形

实际中，$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$常常都未知。为此，从两个总体中分分别抽取样本，估计它们的均值向量$\boldsymbol{\mu}_i$和协差阵$\boldsymbol{\Sigma}_i$。

具体估计方法如下：设从总体G抽取了n个样本(p维)，样本观察值为$\boldsymbol{x}_{(i)} = (x_{i1}, x_{i2}, \cdots, x_{ip})^{\mathsf{T}}$，$i = 1, 2, \cdots, n$。其中$x_{ij}$表示第$i$个样本中第$j$个指标的观察值。

用样本均值向量估计总体均值向量$\boldsymbol{\mu}$，即

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{x}_{(k)}$$

两总体情形

实际中，$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$常常都未知。为此，从两个总体中分分别抽取样本，估计它们的均值向量$\boldsymbol{\mu}_i$和协差阵$\boldsymbol{\Sigma}_i$。

具体估计方法如下：设从总体G抽取了n个样本(p维)，样本观察值为$\boldsymbol{x}_{(i)} = (x_{i1}, x_{i2}, \cdots, x_{ip})^{\mathsf{T}}$，$i = 1, 2, \cdots, n$。其中$x_{ij}$表示第$i$个样本中第$j$个指标的观察值。

用样本均值向量估计总体均值向量$\boldsymbol{\mu}$，即

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{x}_{(k)}$$

用样本协差阵估计总体协差阵$\boldsymbol{\Sigma}$，即

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{k=1}^{n}(\boldsymbol{x}_{(k)} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{(k)} - \bar{\boldsymbol{x}})^{\mathsf{T}}$$

两总体情形

对$\boldsymbol{\mu}$的估计，相当于对总体均值向量的每一个分量都用对相应分量观测值的平均值来估计。

$$\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{x}_{(k)} = \frac{1}{n}\left[\begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1p} \end{pmatrix} + \begin{pmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2p} \end{pmatrix} + \cdots + \begin{pmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{np} \end{pmatrix}\right]$$

两总体情形

对$\boldsymbol{\mu}$的估计，相当于对总体均值向量的每一个分量都用对相应分量观测值的平均值来估计。

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{x}_{(k)} &= \frac{1}{n}\left[\begin{pmatrix}x_{11}\\x_{12}\\\vdots\\x_{1p}\end{pmatrix}+\begin{pmatrix}x_{21}\\x_{22}\\\vdots\\x_{2p}\end{pmatrix}+\cdots+\begin{pmatrix}x_{n1}\\x_{n2}\\\vdots\\x_{np}\end{pmatrix}\right]\\ &= \frac{1}{n}\begin{pmatrix}x_{11}+x_{21}+\cdots+x_{n1}\\x_{12}+x_{22}+\cdots+x_{n2}\\\vdots\\x_{1p}+x_{2p}+\cdots+x_{np}\end{pmatrix}\end{aligned}$$

两总体情形

对$\boldsymbol{\mu}$的估计，相当于对总体均值向量的每一个分量都用对相应分量观测值的平均值来估计。

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{x}_{(k)} &= \frac{1}{n}\left[\begin{pmatrix}x_{11}\\x_{12}\\\vdots\\x_{1p}\end{pmatrix}+\begin{pmatrix}x_{21}\\x_{22}\\\vdots\\x_{2p}\end{pmatrix}+\cdots+\begin{pmatrix}x_{n1}\\x_{n2}\\\vdots\\x_{np}\end{pmatrix}\right]\\ &= \frac{1}{n}\begin{pmatrix}x_{11}+x_{21}+\cdots+x_{n1}\\x_{12}+x_{22}+\cdots+x_{n2}\\\vdots\\x_{1p}+x_{2p}+\cdots+x_{np}\end{pmatrix}=\begin{pmatrix}\bar{x}_1\\\bar{x}_2\\\vdots\\\bar{x}_p\end{pmatrix}\end{aligned}$$

两总体情形

对$\boldsymbol{\mu}$的估计，相当于对总体均值向量的每一个分量都用对相应分量观测值的平均值来估计。

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{x}_{(k)} &= \frac{1}{n}\left[\begin{pmatrix}x_{11}\\x_{12}\\\vdots\\x_{1p}\end{pmatrix}+\begin{pmatrix}x_{21}\\x_{22}\\\vdots\\x_{2p}\end{pmatrix}+\cdots+\begin{pmatrix}x_{n1}\\x_{n2}\\\vdots\\x_{np}\end{pmatrix}\right]\\ &= \frac{1}{n}\begin{pmatrix}x_{11}+x_{21}+\cdots+x_{n1}\\x_{12}+x_{22}+\cdots+x_{n2}\\\vdots\\x_{1p}+x_{2p}+\cdots+x_{np}\end{pmatrix}=\begin{pmatrix}\bar{x}_1\\\bar{x}_2\\\vdots\\\bar{x}_p\end{pmatrix}\end{aligned}$$

类与类之间的距离

$\hat{\mathbf{\Sigma}}$中第i行j列元素为

$$\hat{\Sigma}_{i,j} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

类与类之间的距离

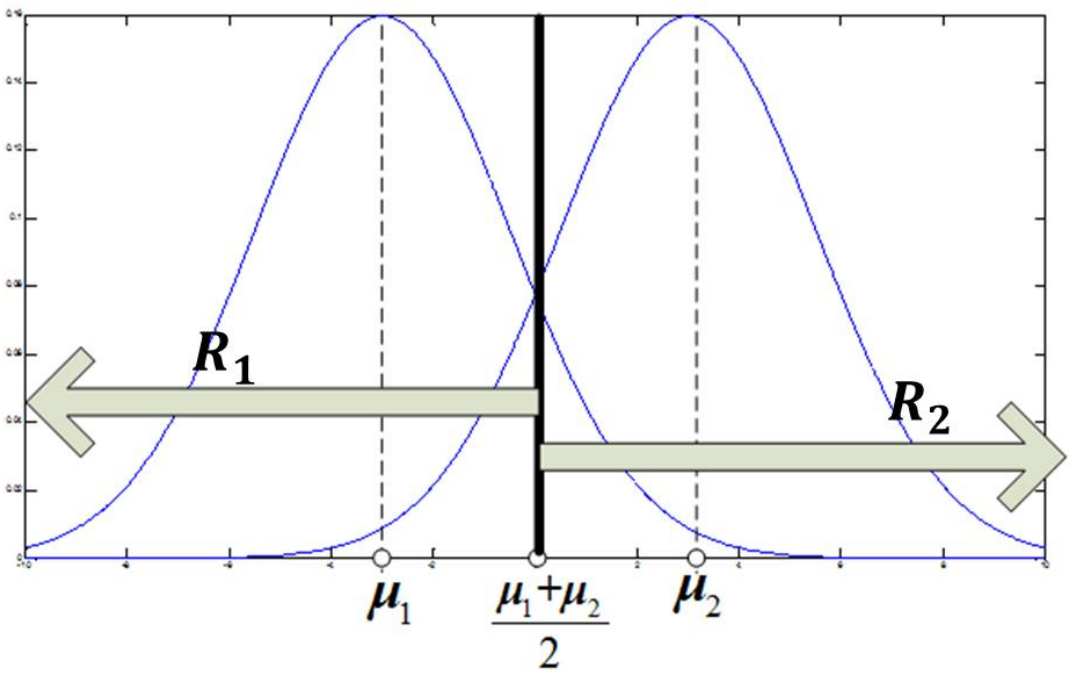
$\hat{\mathbf{\Sigma}}$中第i行j列元素为

$$\hat{\Sigma}_{i,j} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

故对$\mathbf{\Sigma}$的估计，相当于对总体各分量两两间的协方差分别用其样本协方差估计。

例5 (误判问题)

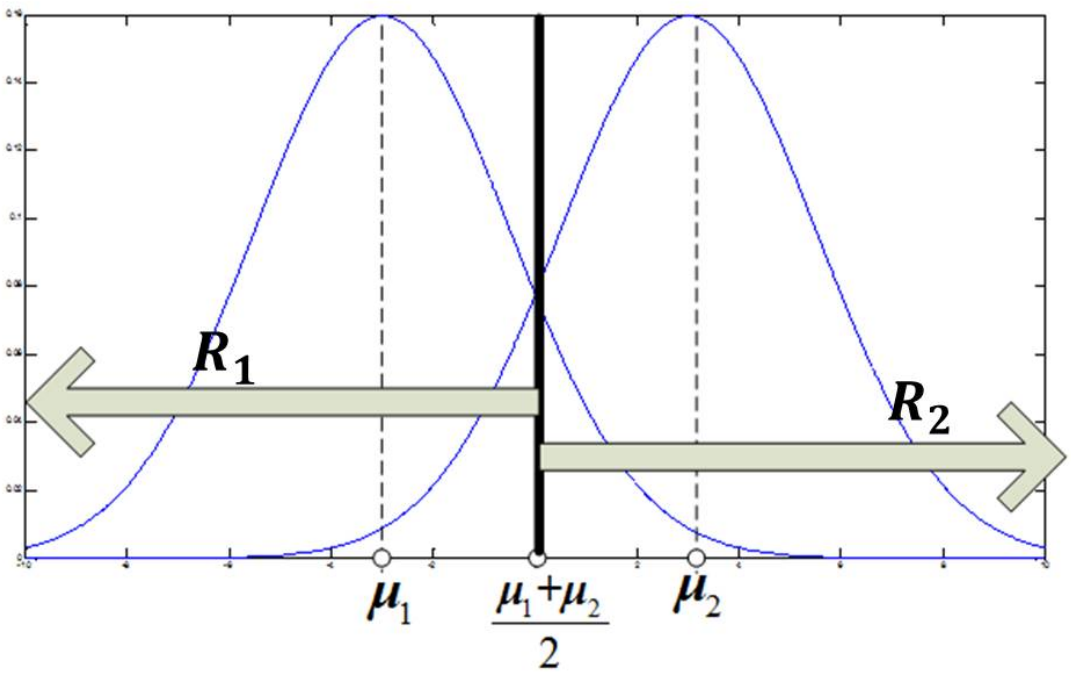
以两个一维正态分布$(\Sigma_1 = \Sigma_2 = \sigma^2)$进行说明。



判别函数$W(\boldsymbol{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \overline{\boldsymbol{\mu}})$，其中$\overline{\boldsymbol{\mu}} = \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}$，则有

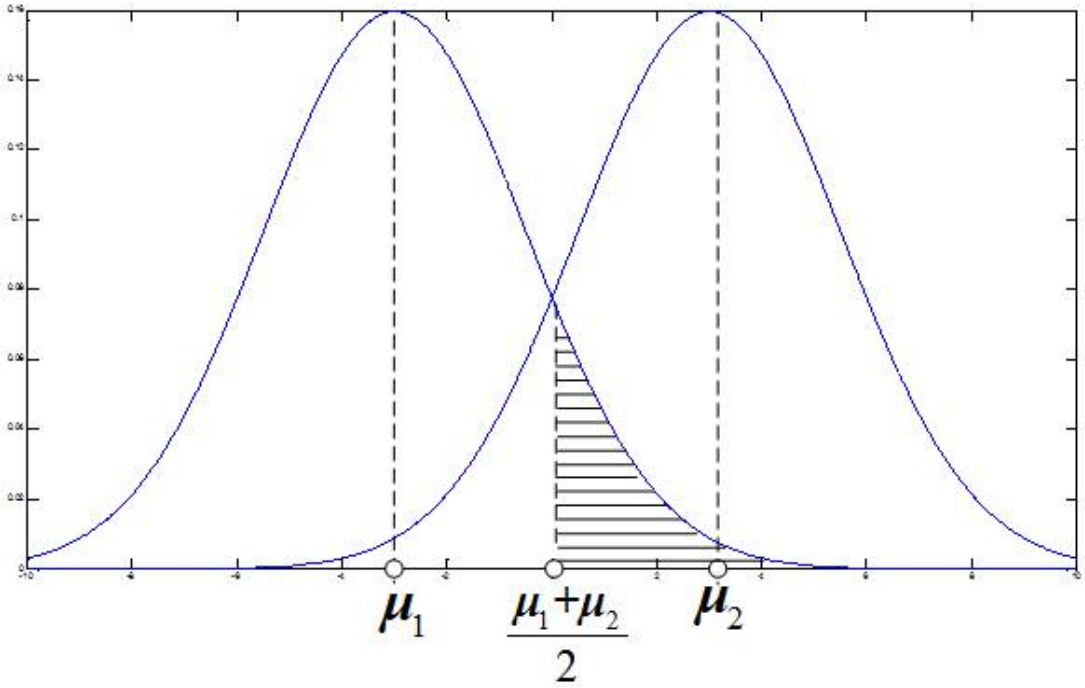
例5 (误判问题)

以两个一维正态分布($\Sigma_1 = \Sigma_2 = \sigma^2$)进行说明。

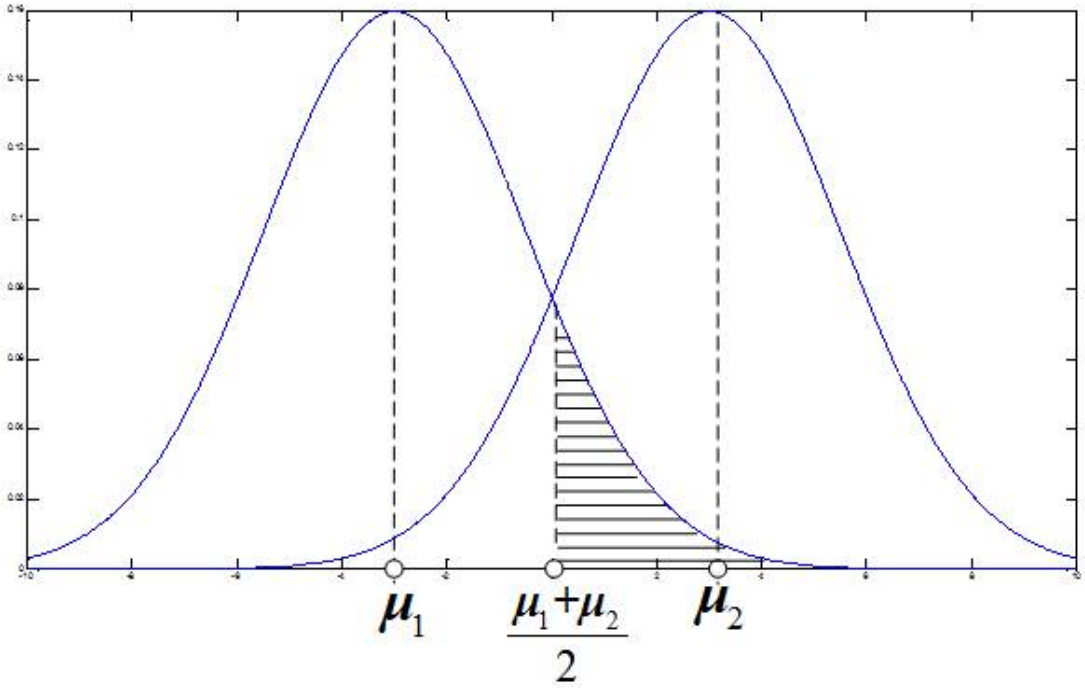


判别函数$W(\boldsymbol{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \overline{\boldsymbol{\mu}})$，其中$\overline{\boldsymbol{\mu}} = \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}$，则有

$$R_1 = \{x \le \overline{\boldsymbol{\mu}}\}, R_2 = \{x \ge \overline{\boldsymbol{\mu}}\}$$

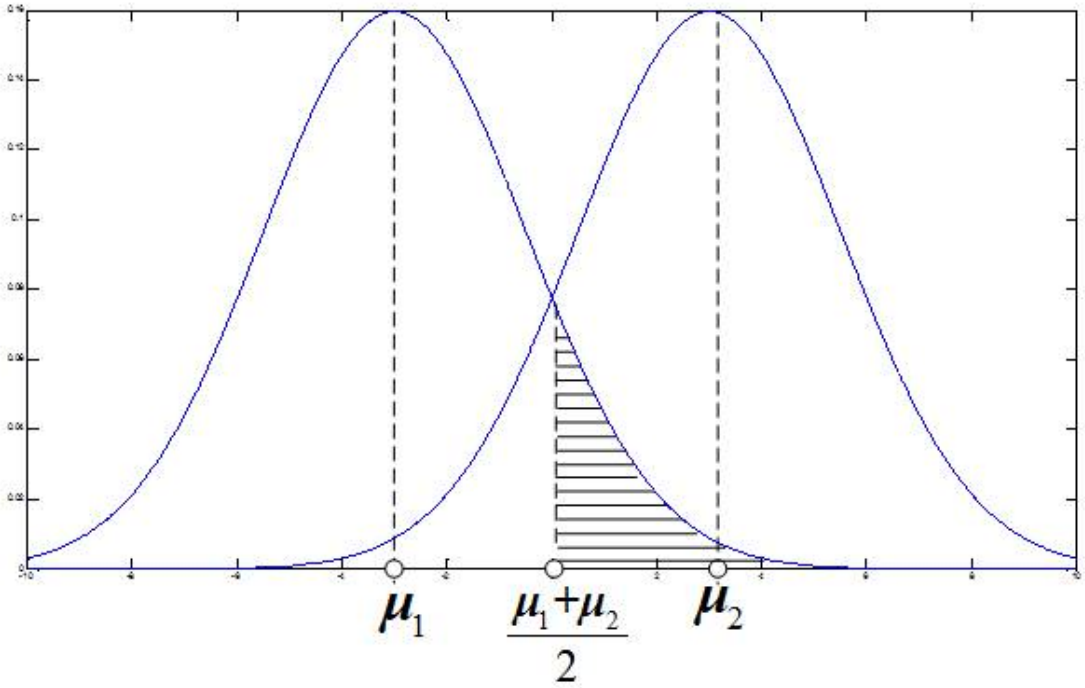


当$x \in (\frac{\mu_1+\mu_2}{2}, \mu_2)$, 总有$d(x, G_1) > d(x, G_2)$, 故误判在所难免。



当$x \in (\frac{\mu_1+\mu_2}{2}, \mu_2)$, 总有$d(x, G_1) > d(x, G_2)$, 故误判在所难免。记将来自$G_i$的样本判断来自$G_j$的概率为$p(j|i)$, 则

$$p(2|1) = p(1|2) = 1 - \Phi\left(\frac{\mu_2 - \mu_1}{2\sigma}\right)$$



当$x \in (\frac{\mu_1+\mu_2}{2}, \mu_2)$, 总有$d(x, G_1) > d(x, G_2)$, 故误判在所难免。 记将来自$G_i$的样本判断来自$G_j$的概率为$p(j|i)$, 则

$$p(2|1) = p(1|2) = 1 - \Phi\left(\frac{\mu_2 - \mu_1}{2\sigma}\right)$$

同时也表明， 当$\mu_2 - \mu_1$越大时， 错判概率越小。

多总体情形

多总体的判别问题，与两总体的判别问题基本思想是一致的。

多总体情形

多总体的判别问题，与两总体的判别问题基本思想是一致的。

设有m个p维总体$G_1, G_2, \cdots, G_m$, 它们的均值向量分别为$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_m$，协差阵分别为$\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \cdots, \boldsymbol{\Sigma}_m$。对一个新的样品, $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)^{\mathsf{T}}$, 需要判别其归属。

多总体情形

多总体的判别问题，与两总体的判别问题基本思想是一致的。

设有m个p维总体$G_1, G_2, \cdots, G_m$，它们的均值向量分别为$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_m$，协差阵分别为$\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \cdots, \boldsymbol{\Sigma}_m$。对一个新的样品，$\boldsymbol{x} = (x_1, x_2, \cdots, x_p)^{\mathsf{T}}$，需要判别其归属。

从数学形式上看，就是基于距离对样本空间进行划分。令

$$R_i = \{\boldsymbol{x} | D^2(\boldsymbol{x}, G_i) = \min_{1 \le j \le m} D^2(\boldsymbol{x}, G_j)\}$$

即区域R_i内的样本点，到G_i的距离比到其它各总体$G_1, G_2, \cdots, G_{i-1}, G_{i+1}, \cdots, G_m$的距离都近。

多总体情形

多总体的判别问题，与两总体的判别问题基本思想是一致的。

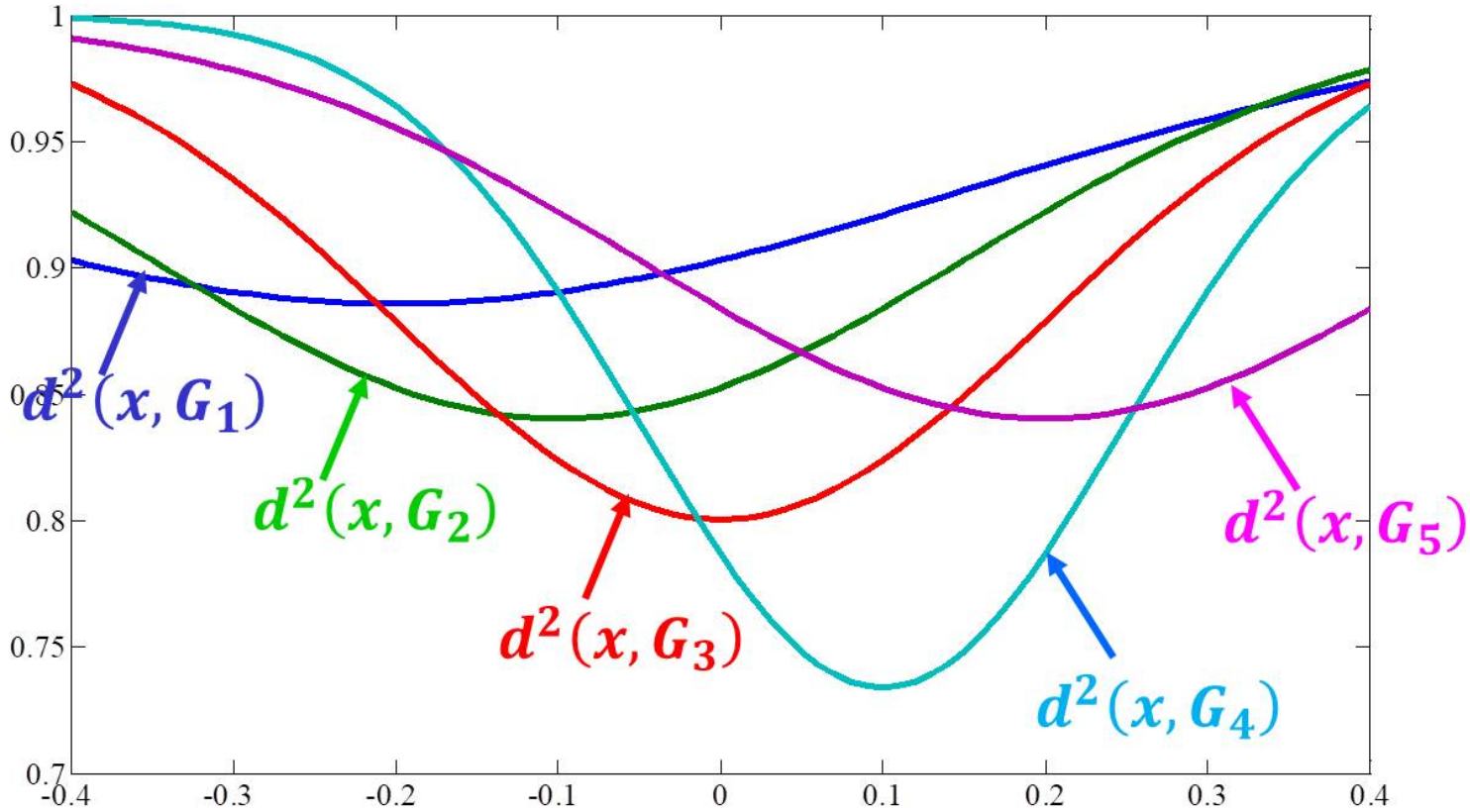
设有m个p维总体$G_1, G_2, \cdots, G_m$，它们的均值向量分别为$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_m$，协差阵分别为$\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \cdots, \boldsymbol{\Sigma}_m$。对一个新的样品，$\boldsymbol{x} = (x_1, x_2, \cdots, x_p)^{\mathsf{T}}$，需要判别其归属。

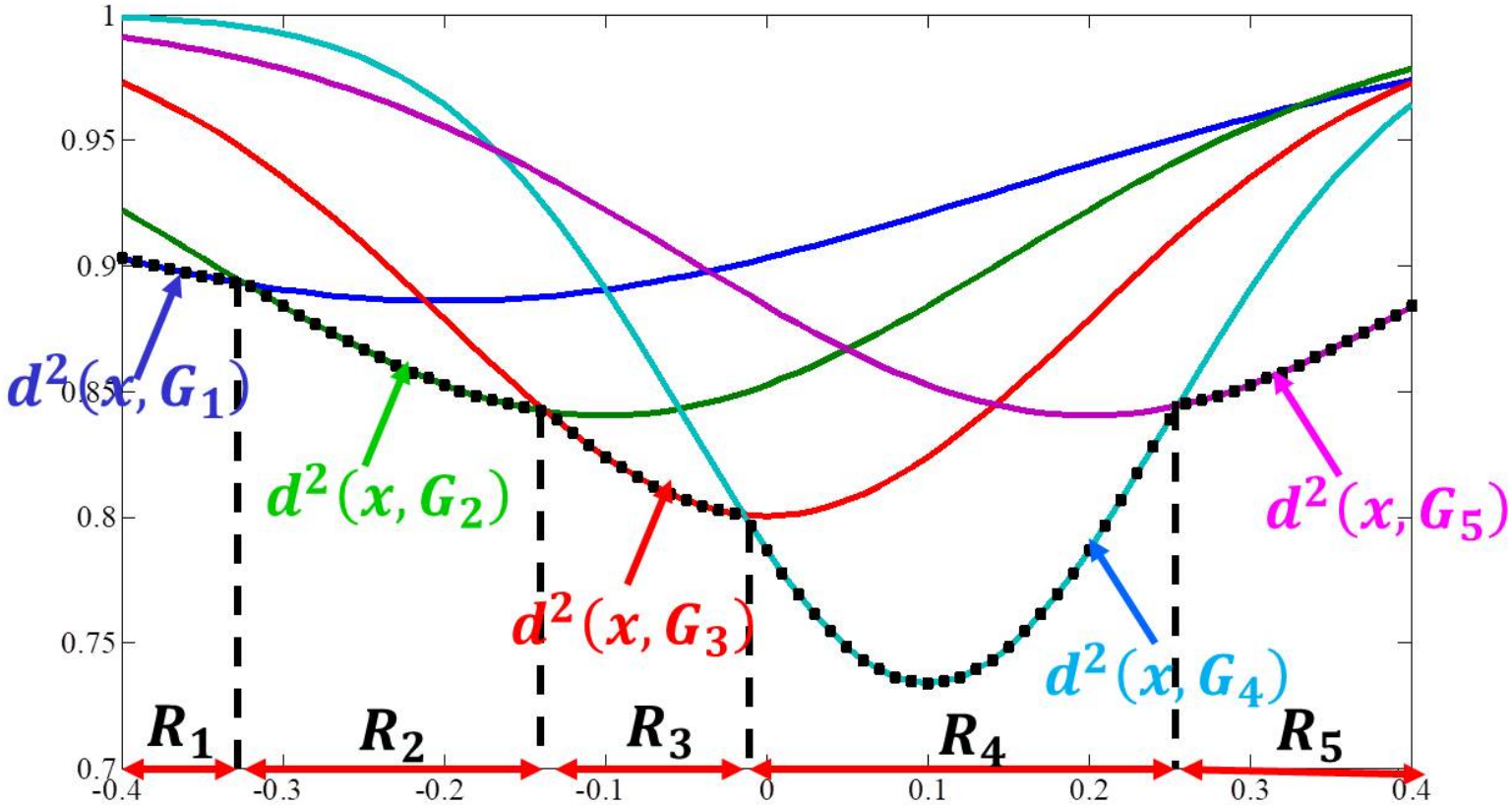
从数学形式上看，就是基于距离对样本空间进行划分。令

$$R_i = \{\boldsymbol{x} | D^2(\boldsymbol{x}, G_i) = \min_{1 \le j \le m} D^2(\boldsymbol{x}, G_j)\}$$

即区域R_i内的样本点，到G_i的距离比到其它各总体$G_1, G_2, \cdots, G_{i-1}, G_{i+1}, \cdots, G_m$的距离都近。

从而有判别规则：若$\boldsymbol{x} \in R_i$，则判$\boldsymbol{x} \in G_i$。


1
0.95
0.9
0.85
0.8
0.75
0.7
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
$d^2(x, G_1)$
$d^2(x, G_2)$
$d^2(x, G_3)$
$d^2(x, G_4)$
$d^2(x, G_5)$



样本空间划分: $R = R_1 \cup R_2 \cup R_3 \cup R_4 \cup R_5$

第四节：贝叶斯判别法

主要内容：贝叶斯判别法的基本思想，两总体情形，与距离判别法的联系，多总体情形

距离判别法特点

1. 优点：思路清晰，计算简单；
2. 缺点：
   - 将各总体置于同等的地位，用这种不偏不倚的态度处理问题固然公允，但当总体本身要求区别对待的时候，就会显露出不足。
   - 判别规则没有考虑: 错判的类型、错判的可能性大小以及错判导致的损失。

距离判别法特点

1. 优点：思路清晰，计算简单；
2. 缺点：
   - 将各总体置于同等的地位，用这种不偏不倚的态度处理问题固然公允，但当总体本身要求区别对待的时候，就会显露出不足。
   - 判别规则没有考虑: 错判的类型、错判的可能性大小以及错判导致的损失。

当各总体出现的可能性大小存在差异时，将这种差异用概率来度量，得到一个离散型分布。该分布往往是根据历史数据或经验来确定，称之为各总体的先验分布。

贝叶斯判别法

设有m个总体$G_1, G_2, \cdots, G_m$, 其概率密度分别为$f_1(x), f_2(x), \cdots, f_m(x)$。假定$m$个总体各自出现的概率分别为$q_1, q_2, \cdots, q_m$。将属于$G_i$的样品错判到$G_j$的损失记为$c(j|i)$。显然有$c(i|i) = 0, c(j|i) \geq 0$。

贝叶斯判别法

设有m个总体$G_1, G_2, \cdots, G_m$, 其概率密度分别为$f_1(x), f_2(x), \cdots, f_m(x)$。假定$m$个总体各自出现的概率分别为$q_1, q_2, \cdots, q_m$。将属于$G_i$的样品错判到$G_j$的损失记为$c(j|i)$。显然有$c(i|i) = 0, c(j|i) \geq 0$。

判别规则为：若得到样本空间的一个划分$R = (R_1, R_2, \cdots, R_m)$。则当样品$x \in R_i$, 则判$x \in G_i$。

贝叶斯判别法

设有m个总体$G_1, G_2, \cdots, G_m$, 其概率密度分别为$f_1(x), f_2(x), \cdots, f_m(x)$。假定$m$个总体各自出现的概率分别为$q_1, q_2, \cdots, q_m$。将属于$G_i$的样品错判到$G_j$的损失记为$c(j|i)$。显然有$c(i|i) = 0, c(j|i) \geq 0$。

判别规则为：若得到样本空间的一个划分$R = (R_1, R_2, \cdots, R_m)$。则当样品$x \in R_i$, 则判$x \in G_i$。

错判损失有多少之别，发生可能性有大小之分。显然考虑平均损失是合理的，如何度量呢？

贝叶斯判别法

属于G_i的样品若取值落在R_j中，那么按规则会把它错判为G_j的样品，这种错判的概率为

$$P(j|i,R)=\int_{R_j} f_i(x)\mathrm{d}x \quad i,j=1,2,\cdots,m,i\neq j$$

贝叶斯判别法

属于G_i的样品若取值落在R_j中，那么按规则会把它错判为G_j的样品，这种错判的概率为

$$P(j|i,R)=\int_{R_j} f_i(x)\mathrm{d}x \quad i,j=1,2,\cdots,m,i\neq j$$

属于G_i的样品，可能有多种错判$j=1,2,\cdots,n,j\neq i$，且错判损失$c(j|i)$多少不尽相同，

将G_i的样品错判到	G_1	$\cdots$	G_{i-1}	G_{i+1}	$\cdots$
错判造成的损失	$c(1\|i)$	$\cdots$	$c(i-1\|i)$	$c(i+1\|i)$	$\cdots$
发生错判的概率	$P(1\|i,R)$	$\cdots$	$P(i-1\|i,R)$	$P(i+1\|i,R)$	$\cdots$

贝叶斯判别法

属于G_i的样品若取值落在R_j中，那么按规则会把它错判为G_j的样品，这种错判的概率为

$$P(j|i,R)=\int_{R_j} f_i(x)\mathrm{d}x \quad i,j=1,2,\cdots,m,i\neq j$$

属于G_i的样品，可能有多种错判$j=1,2,\cdots,n,j\neq i$，且错判损失$c(j|i)$多少不尽相同，

将G_i的样品错判到	G_1	$\cdots$	G_{i-1}	G_{i+1}	$\cdots$
错判造成的损失	$c(1\|i)$	$\cdots$	$c(i-1\|i)$	$c(i+1\|i)$	$\cdots$
发生错判的概率	$P(1\|i,R)$	$\cdots$	$P(i-1\|i,R)$	$P(i+1\|i,R)$	$\cdots$

属于G_i的样品，平均错判损失为

$$r(i,R)=\sum_{j=1}^{m}[c(j|i)P(j|i,R)]$$

贝叶斯判别法

待判样本来自每个总体都有可能，进而考虑总平均损失

$$
\begin{aligned}
g(R) &= \sum_{i=1}^{m} q_i r(i,R) \\
&= \sum_{i=1}^{m} q_i \left\{ \sum_{j=1}^{m} \left[c(j|i) \int_{R_j} f_i(x)\mathrm{d}x \right] \right\}
\end{aligned}
$$

贝叶斯判别法

待判样本来自每个总体都有可能，进而考虑总平均损失

$$
\begin{aligned}
g(R) &= \sum_{i=1}^{m} q_i r(i,R) \\
&= \sum_{i=1}^{m} q_i \left\{ \sum_{j=1}^{m} \left[c(j|i) \int_{R_j} f_i(x)\mathrm{d}x \right] \right\}
\end{aligned}
$$

在总体分布密度$f_i(x)$, 先验概率q_i, 损失$c(j|i)$都给定的情况下，总平均损失$g(R)$只依赖于判别规则(样本空间划分)$R = (R_1, R_2, \cdots, R_m)$。

贝叶斯判别法

待判样本来自每个总体都有可能，进而考虑总平均损失

$$
\begin{aligned}
g(R) &= \sum_{i=1}^{m} q_i r(i,R) \\
&= \sum_{i=1}^{m} q_i \left\{ \sum_{j=1}^{m} \left[c(j|i) \int_{R_j} f_i(x)\mathrm{d}x \right] \right\}
\end{aligned}
$$

在总体分布密度$f_i(x)$, 先验概率q_i, 损失$c(j|i)$都给定的情况下，总平均损失$g(R)$只依赖于判别规则(样本空间划分)$R = (R_1, R_2, \cdots, R_m)$。贝叶斯判别法就是适当地选取某个划分$R$, 使得总平均损失$g(R)$达到最小。

两总体情形

此时，$m=2, R=R_1\cup R_2, q_1+q_2=1$, 总平均损失

$$
\begin{aligned}
g(R) &= q_1c(2|1)\int_{R_2} f_1(x)\mathrm{d}x + q_2c(1|2)\int_{R_1} f_2(x)\mathrm{d}x \\
&= \int_{R_2} q_1c(2|1)f_1(x)\mathrm{d}x - \int_{R_2} q_2c(1|2)f_2(x)\mathrm{d}x \\
&+ \int_{R_2} q_2c(1|2)f_2(x)\mathrm{d}x + \int_{R_1} q_2c(1|2)f_2(x)\mathrm{d}x \\
&= \int_{R_2} [q_1c(2|1)f_1(x) - q_2c(1|2)f_2(x)]\mathrm{d}x + \int_{R} q_2c(1|2)f_2(x)\mathrm{d}x \\
&= \int_{R_2} [q_1c(2|1)f_1(x) - q_2c(1|2)f_2(x)]\mathrm{d}x + q_2c(1|2)
\end{aligned}
$$

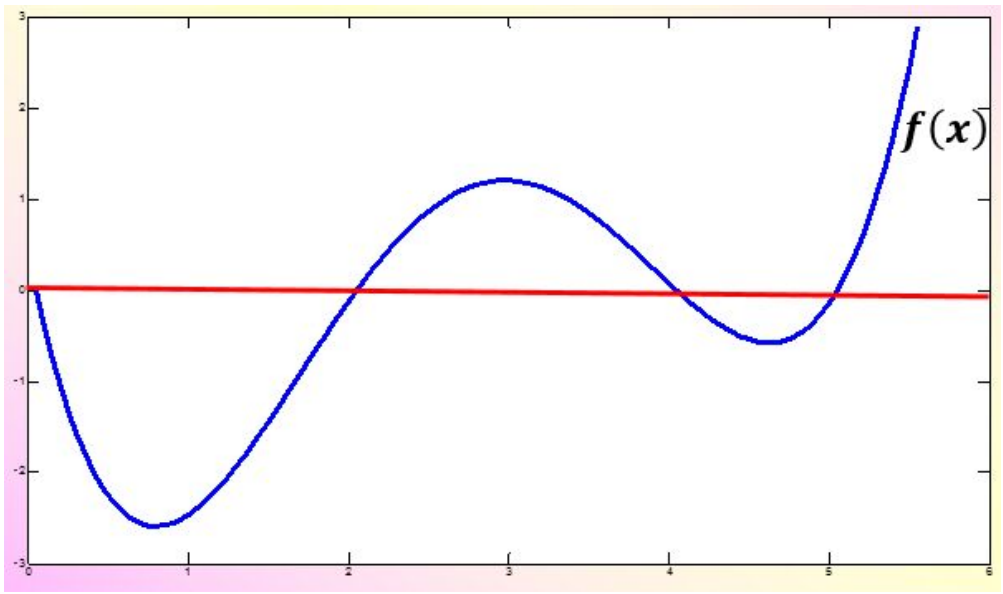
两总体情形

此时，$m=2, R=R_1\cup R_2, q_1+q_2=1$, 总平均损失

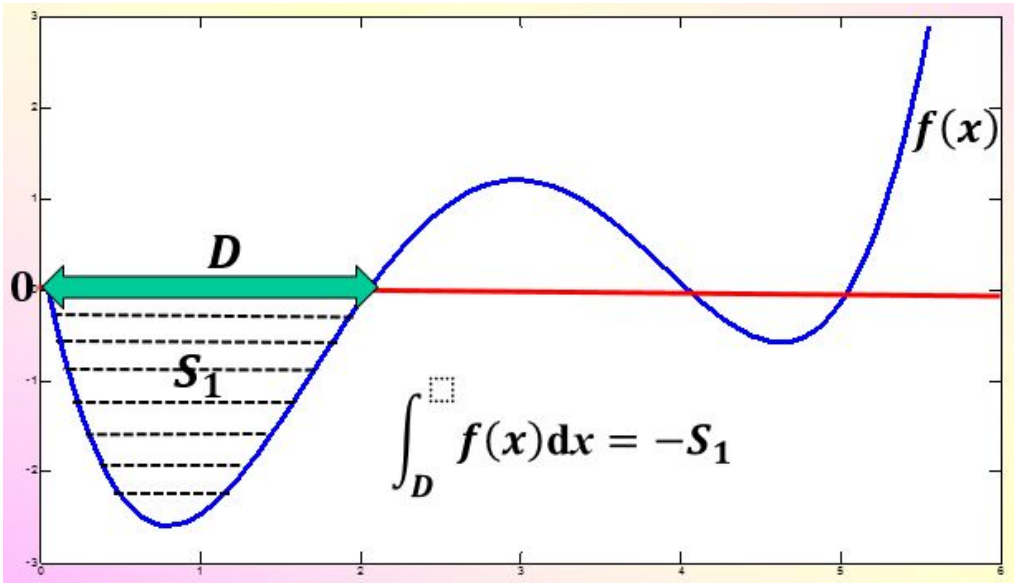
$$
\begin{aligned}
g(R) &= q_1c(2|1)\int_{R_2} f_1(x)\mathrm{d}x + q_2c(1|2)\int_{R_1} f_2(x)\mathrm{d}x \\
&= \int_{R_2} q_1c(2|1)f_1(x)\mathrm{d}x - \int_{R_2} q_2c(1|2)f_2(x)\mathrm{d}x \\
&+ \int_{R_2} q_2c(1|2)f_2(x)\mathrm{d}x + \int_{R_1} q_2c(1|2)f_2(x)\mathrm{d}x \\
&= \int_{R_2} [q_1c(2|1)f_1(x) - q_2c(1|2)f_2(x)]\mathrm{d}x + \int_{R} q_2c(1|2)f_2(x)\mathrm{d}x \\
&= \int_{R_2} [q_1c(2|1)f_1(x) - q_2c(1|2)f_2(x)]\mathrm{d}x + q_2c(1|2)
\end{aligned}
$$

要使$g(R)$最小，只需选择R_2使得第一项达到最小。

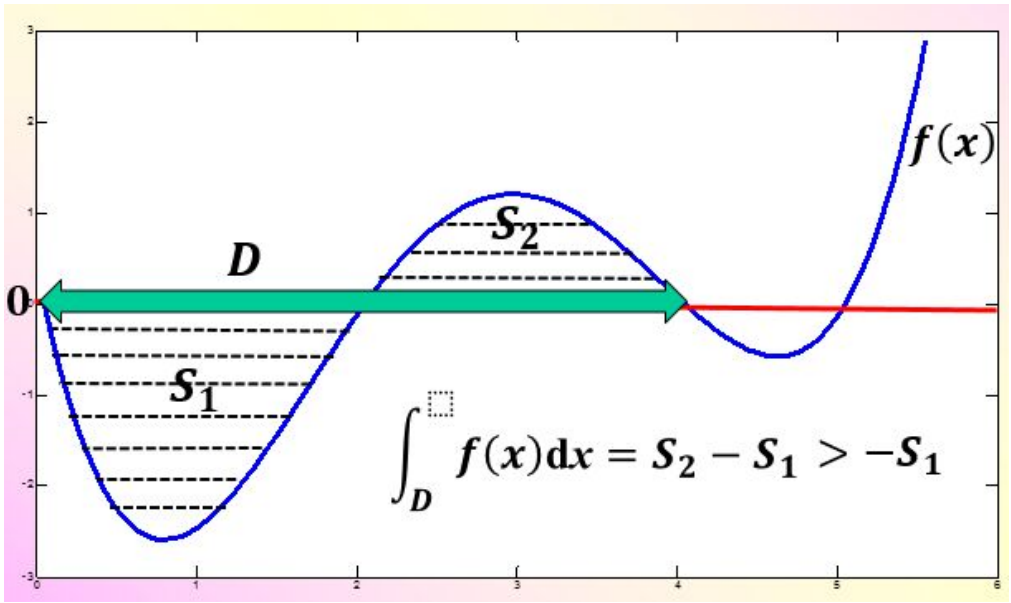
该问题即为：$\min \int_D f(x)\mathrm{d}x$, 如何选取积分区域$D$呢？



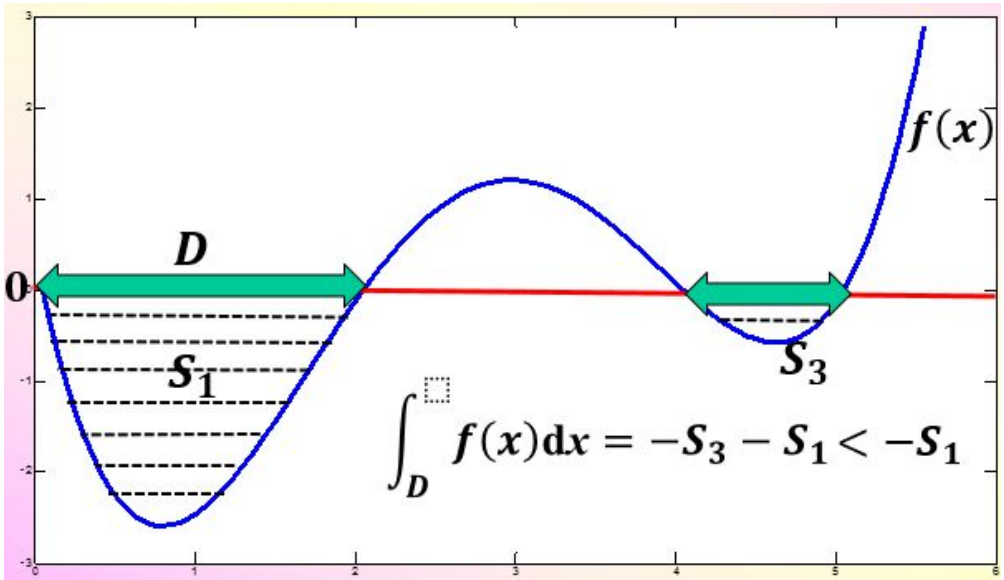
该问题即为：$\min \int_D f(x)\mathrm{d}x$, 如何选取积分区域$D$呢？



该问题即为：$\min \int_D f(x)\mathrm{d}x$, 如何选取积分区域$D$呢？



该问题即为：$\min \int_D f(x)\mathrm{d}x$, 如何选取积分区域$D$呢？



$$D = \{x | f(x) \leq 0\}$$

$$g(R) = \int_{R_2} [q_1 c(2|1) f_1(x) - q_2 c(1|2) f_2(x)] \mathrm{d}x + q_2 c(1|2)$$

因而要使$g(R)$最小，只需选择

$$R_2 = \{x | q_1 c(2|1) f_1(x) - q_2 c(1|2) f_2(x) < 0\}$$
$$R_1 = \{x | q_1 c(2|1) f_1(x) - q_2 c(1|2) f_2(x) \geq 0\}$$

即

$$R_1 = \left\{ x \left| \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2) q_2}{c(2|1) q_1} \right. \right\}, R_2 = \left\{ x \left| \frac{f_1(x)}{f_2(x)} < \frac{c(1|2) q_2}{c(2|1) q_1} \right. \right\}$$

$$g(R) = \int_{R_2} [q_1 c(2|1) f_1(x) - q_2 c(1|2) f_2(x)] \mathrm{d}x + q_2 c(1|2)$$

因而要使$g(R)$最小，只需选择

$$R_2 = \{x | q_1 c(2|1) f_1(x) - q_2 c(1|2) f_2(x) < 0\}$$
$$R_1 = \{x | q_1 c(2|1) f_1(x) - q_2 c(1|2) f_2(x) \geq 0\}$$

即

$$R_1 = \left\{ x \left| \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2) q_2}{c(2|1) q_1} \right. \right\}, R_2 = \left\{ x \left| \frac{f_1(x)}{f_2(x)} < \frac{c(1|2) q_2}{c(2|1) q_1} \right. \right\}$$

两总体情形的贝叶斯判别规则:

- 若待判样品点$x \in R_1$, 则判x来自G_1;
- 若待判样品点$x \in R_2$, 则判x来自G_2。

多总体情形

总平均损失

$$g(R) = \sum_{i=1}^{m} q_i \left[\sum_{j=1}^{m} c(j|i) \int_{R_j} f_i(x)\mathrm{d}x \right]$$

$$= \sum_{j=1}^{m} \int_{R_j} \left[\sum_{i=1}^{m} q_i c(j|i) f_i(x) \right] \mathrm{d}x \triangleq \sum_{j=1}^{m} \int_{R_j} h_j(x)\mathrm{d}x$$

其中$h_j(x) = \sum_{i=1}^{m} q_i c(j|i) f_i(x)$。

多总体情形

总平均损失

$$g(R)=\sum_{i=1}^{m}q_i\left[\sum_{j=1}^{m}c(j|i)\int_{R_j}f_i(x)\mathrm{d}x\right]$$
$$=\sum_{j=1}^{m}\int_{R_j}\left[\sum_{i=1}^{m}q_ic(j|i)f_i(x)\right]\mathrm{d}x\triangleq\sum_{j=1}^{m}\int_{R_j}h_j(x)\mathrm{d}x$$

其中$h_j(x)=\sum_{i=1}^{m}q_ic(j|i)f_i(x)$。$g(R)$达到最小，等价于在$R_j$上，$h_j(x)$是$h_1(x),h_2(x),\cdots,h_m(x)$中的最小，即取$R_j=\{x\,|h_j(x)=\min_{1\le i\le m}\{h_i(x)\}\}$。

多总体情形

总平均损失

$$g(R)=\sum_{i=1}^{m}q_i\left[\sum_{j=1}^{m}c(j|i)\int_{R_j}f_i(x)\mathrm{d}x\right]$$
$$=\sum_{j=1}^{m}\int_{R_j}\left[\sum_{i=1}^{m}q_ic(j|i)f_i(x)\right]\mathrm{d}x\triangleq\sum_{j=1}^{m}\int_{R_j}h_j(x)\mathrm{d}x$$

其中$h_j(x)=\sum_{i=1}^{m}q_ic(j|i)f_i(x)$。$g(R)$达到最小，等价于在$R_j$上，$h_j(x)$是$h_1(x)$, $h_2(x),\cdots,h_m(x)$中的最小，即取$R_j=\{x\,|h_j(x)=\min_{1\le i\le m}\{h_i(x)\}\}$。

如需对一个样品x进行判别，只需计算各函数$h_j(x)$在该样品x处的值$h_1(x),h_2(x),\cdots,h_m(x)$, 然后比较大小。

多总体情形

总平均损失

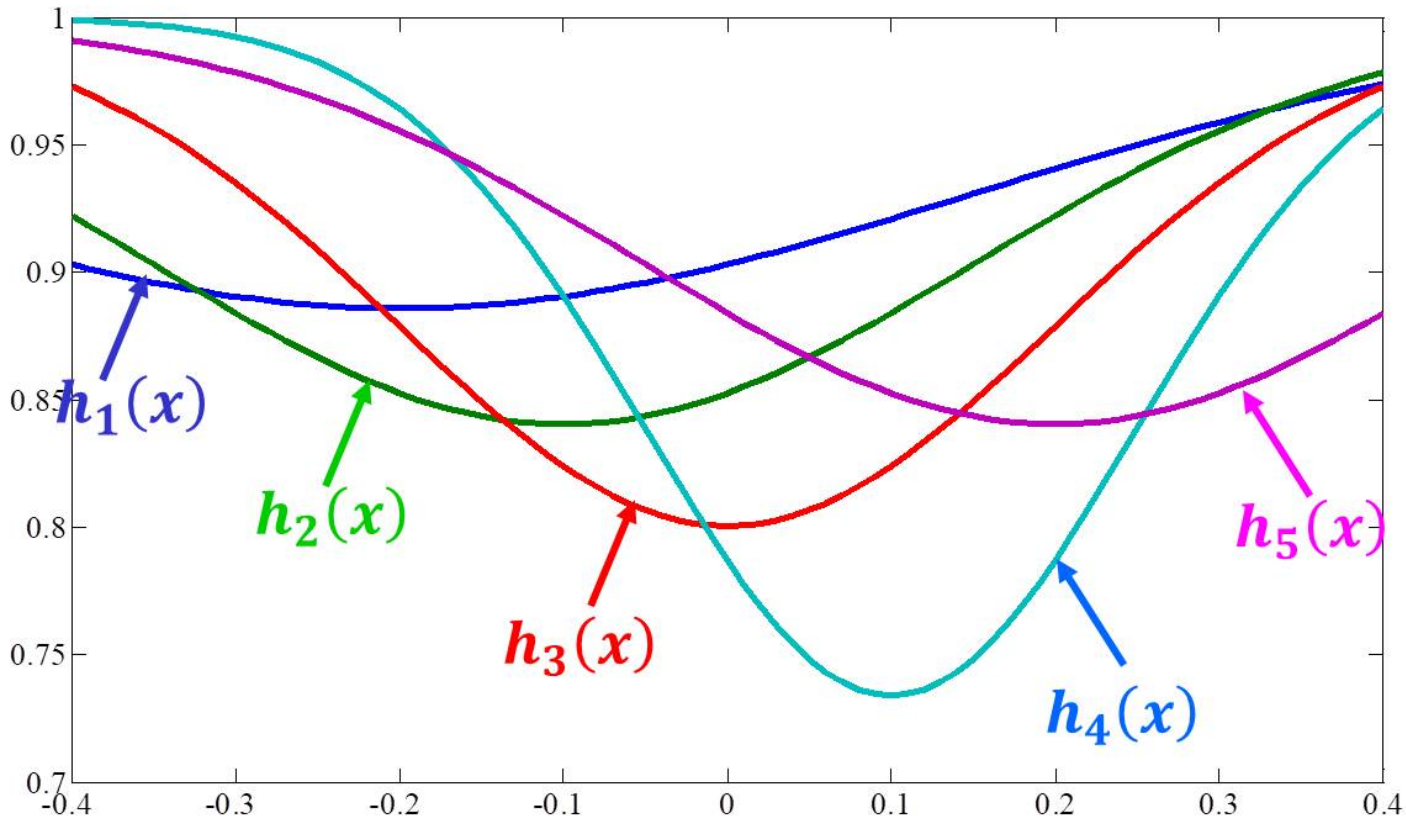
$$g(R) = \sum_{i=1}^{m} q_i \left[\sum_{j=1}^{m} c(j|i) \int_{R_j} f_i(x)\mathrm{d}x \right]$$

$$= \sum_{j=1}^{m} \int_{R_j} \left[\sum_{i=1}^{m} q_i c(j|i) f_i(x) \right] \mathrm{d}x \triangleq \sum_{j=1}^{m} \int_{R_j} h_j(x)\mathrm{d}x$$

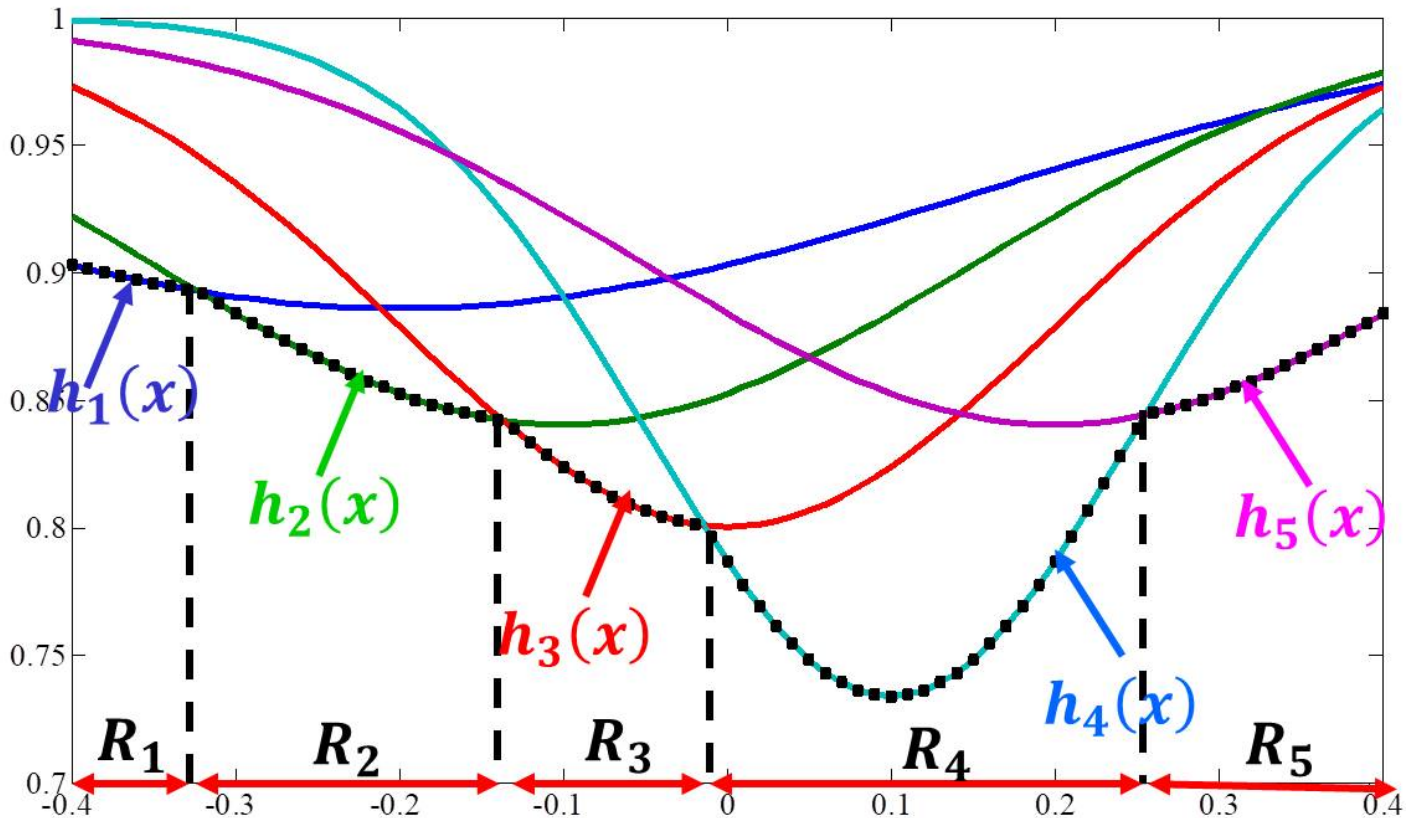
其中$h_j(x) = \sum_{i=1}^{m} q_i c(j|i) f_i(x)$。$g(R)$达到最小，等价于在$R_j$上，$h_j(x)$是$h_1(x), h_2(x), \cdots, h_m(x)$中的最小，即取$R_j = \{x \,|h_j(x) = \min_{1\le i\le m}\{h_i(x)\}\}$。

如需对一个样品x进行判别，只需计算各函数$h_j(x)$在该样品x处的值$h_1(x), h_2(x), \cdots, h_m(x)$, 然后比较大小。若$h_k(x)$ 最小，则判$x \in G_k$。

多总体情形



多总体情形



样本空间划分: $R = R_1 \cup R_2 \cup R_3 \cup R_4 \cup R_5$

两种判别法的对比

(1)距离判别法，只需已知总体的均值向量和协差阵，勿需分布细节。

(2)贝叶斯判别法，需要总体分布的密度函数，处理更加细致。考虑总体的先验分布（样本出现的频率进行估计）和错判损失（根据错判的类型和后果进行评定），判别规则使总平均损失达到最小。

两种判别法的联系

若两个总体$G_1 \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $G_2 \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$，即协差阵相同。

两种判别法的联系

若两个总体$G_1 \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $G_2 \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$，即协差阵相同。

(1)按距离判别法，判别函数为

$$W(\boldsymbol{x}) = \frac{1}{2}[(\boldsymbol{x}-\boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) - (\boldsymbol{x}-\boldsymbol{\mu}_1)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)]$$

两种判别法的联系

若两个总体$G_1 \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $G_2 \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$，即协差阵相同。

(1)按距离判别法，判别函数为

$$W(\boldsymbol{x}) = \frac{1}{2}[(\boldsymbol{x}-\boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) - (\boldsymbol{x}-\boldsymbol{\mu}_1)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)]$$

判别规则为:

$$\text{如果} W(\boldsymbol{x}) > 0, \text{则判} \boldsymbol{x} \in G_1$$

$$\text{如果} W(\boldsymbol{x}) < 0, \text{则判} \boldsymbol{x} \in G_2$$

$$\text{如果} W(\boldsymbol{x}) = 0, \text{则待判}$$

两种判别法的联系

若两个总体$G_1 \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $G_2 \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$，即协差阵相同。

(2)按贝叶斯判别法

$$\begin{aligned}\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} &= \exp\left\{\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right\} \\ &= \exp\{W(\boldsymbol{x})\}\end{aligned}$$

若记$K = \dfrac{c(1|2)q_2}{c(2|1)q_1}$, 则有$R_1 = \{\boldsymbol{x}|W(\boldsymbol{x}) \geq \ln K\}, R_2 = \{\boldsymbol{x}|W(\boldsymbol{x}) < \ln K\}$。

两种判别法的联系

若两个总体$G_1 \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $G_2 \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$，即协差阵相同。

(2)按贝叶斯判别法

$$\begin{aligned}\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} &= \exp\left\{\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right\} \\ &= \exp\{W(\boldsymbol{x})\}\end{aligned}$$

若记$K = \dfrac{c(1|2)q_2}{c(2|1)q_1}$, 则有$R_1 = \{\boldsymbol{x}|W(\boldsymbol{x}) \geq \ln K\}, R_2 = \{\boldsymbol{x}|W(\boldsymbol{x}) < \ln K\}$。

因而，协差阵相同时的两正态总体的判别问题，距离判别法与贝叶斯判别法的判别函数形式相同，只是临界值不同，前者为0，后者为$\ln K$。

两种判别法的联系

若两个总体$G_1 \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $G_2 \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$，即协差阵相同。

(2)按贝叶斯判别法

$$\begin{aligned}\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} &= \exp\left\{\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right\} \\ &= \exp\{W(\boldsymbol{x})\}\end{aligned}$$

若记$K = \dfrac{c(1|2)q_2}{c(2|1)q_1}$, 则有$R_1 = \{\boldsymbol{x}|W(\boldsymbol{x}) \geq \ln K\}, R_2 = \{\boldsymbol{x}|W(\boldsymbol{x}) < \ln K\}$。

因而，协差阵相同时的两正态总体的判别问题，距离判别法与贝叶斯判别法的判别函数形式相同，只是临界值不同，前者为0，后者为$\ln K$。进一步，若$K = 1$时, 两种判别法一致。

第五节：Fisher判别法

主要内容：基本思想

基本思想

判别分析问题：将一个p维样品$\boldsymbol{x}$判别来自m个p维总体$G_1, G_2, \cdots, G_m$的某一个。Fisher判别的主要思想，首先将高维问题通过投影变换，降为低维来处理；然后再借助其他判别法（如距离法）进行判别。

基本思想

判别分析问题：将一个p维样品$\boldsymbol{x}$判别来自m个p维总体$G_1, G_2, \cdots, G_m$的某一个。Fisher判别的主要思想，首先将高维问题通过投影变换，降为低维来处理；然后再借助其他判别法（如距离法）进行判别。

适当选择c个轴（即c个方向，一般地$c < p$），将所有的样品点都投影到这些轴上，每个轴上的投影值都将是一个数（即新坐标系下的坐标），多个轴的投影结果构成一个向量。

基本思想

判别分析问题：将一个p维样品$\boldsymbol{x}$判别来自m个p维总体$G_1, G_2, \cdots, G_m$的某一个。Fisher判别的主要思想，首先将高维问题通过投影变换，降为低维来处理；然后再借助其他判别法（如距离法）进行判别。

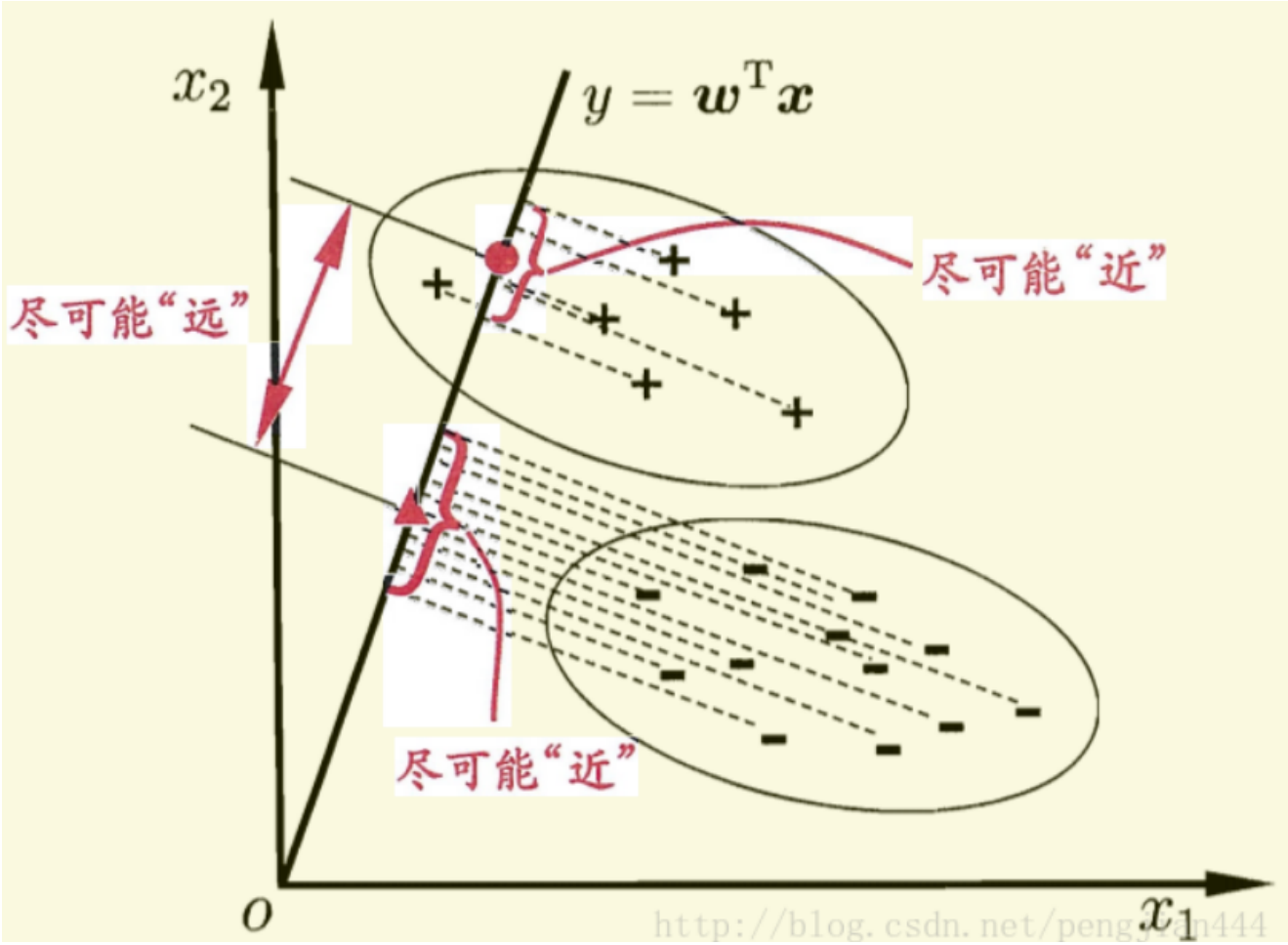
适当选择c个轴（即c个方向，一般地$c < p$），将所有的样品点都投影到这些轴上，每个轴上的投影值都将是一个数（即新坐标系下的坐标），多个轴的投影结果构成一个向量。对于这些投影向量（c维数据），希望同一类的类内离差尽可能小，而不同类的类间离差尽可能大。

基本思想

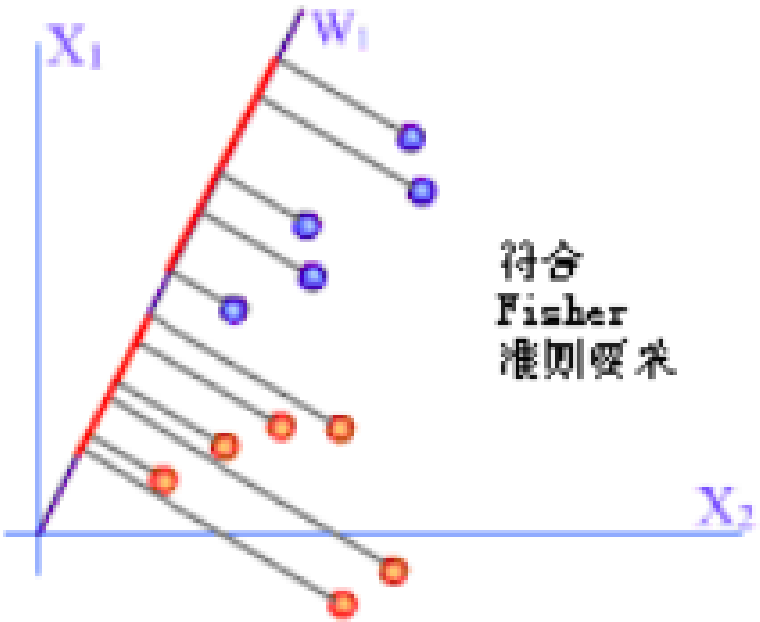
判别分析问题：将一个p维样品$\boldsymbol{x}$判别来自m个p维总体$G_1, G_2, \cdots, G_m$的某一个。Fisher判别的主要思想，首先将高维问题通过投影变换，降为低维来处理；然后再借助其他判别法（如距离法）进行判别。

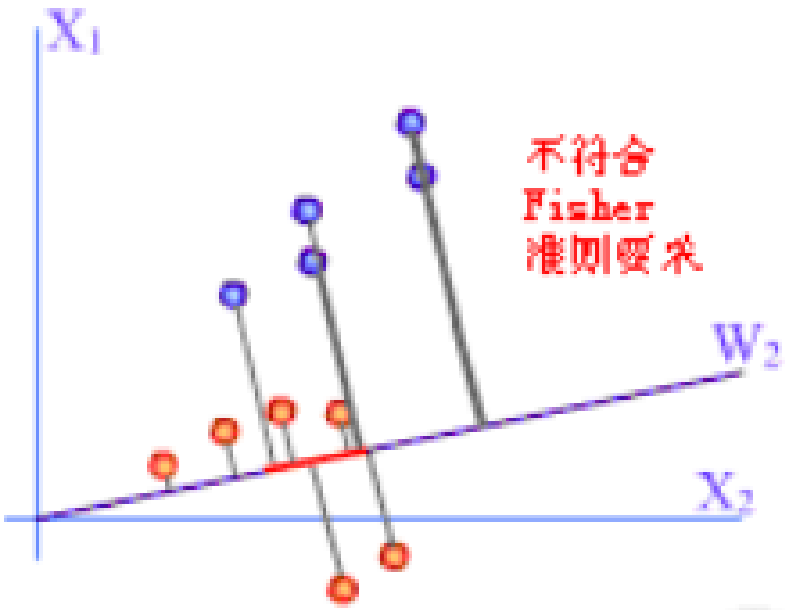
适当选择c个轴（即c个方向，一般地$c < p$），将所有的样品点都投影到这些轴上，每个轴上的投影值都将是一个数（即新坐标系下的坐标），多个轴的投影结果构成一个向量。对于这些投影向量（c维数据），希望同一类的类内离差尽可能小，而不同类的类间离差尽可能大。

后文关于Fisher判别方法的叙述，都假定$c = 1$。


x_2
$y = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$
尽可能“远”
尽可能“近”
尽可能“近”
O
x_1
http://blog.csdn.net/pengjian444

投影方向





两总体情形

2个p维总体的均值向量分别为$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$，协方差阵分别为$\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$。从这两个总体分别取得$n_1, n_2$个样品：

$$G_1 : \boldsymbol{x}_1^{(1)}, \boldsymbol{x}_2^{(1)}, \cdots, \boldsymbol{x}_{n_1}^{(1)}$$
$$G_2 : \boldsymbol{x}_1^{(2)}, \boldsymbol{x}_2^{(2)}, \cdots, \boldsymbol{x}_{n_2}^{(2)}$$

两总体情形

2个p维总体的均值向量分别为$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$，协方差阵分别为$\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$。从这两个总体分别取得$n_1, n_2$个样品：

$$G_1 : \boldsymbol{x}_1^{(1)}, \boldsymbol{x}_2^{(1)}, \cdots, \boldsymbol{x}_{n_1}^{(1)}$$
$$G_2 : \boldsymbol{x}_1^{(2)}, \boldsymbol{x}_2^{(2)}, \cdots, \boldsymbol{x}_{n_2}^{(2)}$$

令$\boldsymbol{w} \in R^p$为某投影直线的方向，则$\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}$为$\boldsymbol{x}$在该直线上的投影，即投影后的数据为：

$$G_1 : \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_1^{(1)}, \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_2^{(1)}, \cdots, \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_{n_1}^{(1)}$$
$$G_2 : \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_1^{(2)}, \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_2^{(2)}, \cdots, \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_{n_2}^{(2)}$$

两总体情形

两个总体的均值向量的投影（一个数）分别为$\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\mu}_1, \boldsymbol{w}^{\mathsf{T}}\boldsymbol{\mu}_2$。

两个总体的协方差阵的投影（一个数）分别为$\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\Sigma}_1\boldsymbol{w}, \boldsymbol{w}^{\mathsf{T}}\boldsymbol{\Sigma}_2\boldsymbol{w}$。

两总体情形

两个总体的均值向量的投影（一个数）分别为$\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\mu}_1, \boldsymbol{w}^{\mathsf{T}}\boldsymbol{\mu}_2$。
两个总体的协方差阵的投影（一个数）分别为$\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\Sigma}_1\boldsymbol{w}, \boldsymbol{w}^{\mathsf{T}}\boldsymbol{\Sigma}_2\boldsymbol{w}$。
投影直线选取的原则，

1. 来自同一总体的样本的投影点尽可能接近，可以让这些投影点的方差$\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\Sigma}_1\boldsymbol{w} + \boldsymbol{w}^{\mathsf{T}}\boldsymbol{\Sigma}_2\boldsymbol{w}$尽可能小；
2. 来自不同总体的样本的投影点尽可能远离，可以让这些投影点的中心之间的距离$||\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\mu}_1 - \boldsymbol{w}^{\mathsf{T}}\boldsymbol{\mu}_2||_2^2$尽可能大；

两总体情形

根据方差分析的知识，若投影后的2组数据存在显著性差异，则

$$\frac{||\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\mu}_1 - \boldsymbol{w}^{\mathsf{T}}\boldsymbol{\mu}_2||_2^2}{\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\Sigma}_1\boldsymbol{w} + \boldsymbol{w}^{\mathsf{T}}\boldsymbol{\Sigma}_2\boldsymbol{w}} = \frac{\boldsymbol{w}^{\mathsf{T}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{w}}{\boldsymbol{w}^{\mathsf{T}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\boldsymbol{w}}$$

应充分地大。

两总体情形

根据方差分析的知识，若投影后的2组数据存在显著性差异，则

$$\frac{||\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\mu}_1 - \boldsymbol{w}^{\mathsf{T}}\boldsymbol{\mu}_2||_2^2}{\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\Sigma}_1\boldsymbol{w} + \boldsymbol{w}^{\mathsf{T}}\boldsymbol{\Sigma}_2\boldsymbol{w}} = \frac{\boldsymbol{w}^{\mathsf{T}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{w}}{\boldsymbol{w}^{\mathsf{T}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\boldsymbol{w}}$$

应充分地大。基于此考虑最大化问题

$$max_{\boldsymbol{w}} \quad \frac{\boldsymbol{w}^{\mathsf{T}}\boldsymbol{S}_b\boldsymbol{w}/(2-1)}{\boldsymbol{w}^{\mathsf{T}}\boldsymbol{S}_w\boldsymbol{w}/(n_1+n_2-2)}$$

或者

$$max_{\boldsymbol{w}} \quad \Delta(\boldsymbol{w}) = \frac{\boldsymbol{w}^{\mathsf{T}}\boldsymbol{S}_b\boldsymbol{w}}{\boldsymbol{w}^{\mathsf{T}}\boldsymbol{S}_w\boldsymbol{w}}$$

两总体情形

其中$\boldsymbol{S}_b = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathsf{T}}$称之为类间散度矩阵（between-class scatter matrix），$\boldsymbol{S}_w = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$称之为类内散度矩阵（within-class scatter matrix）。

两总体情形

其中$\boldsymbol{S}_b = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathsf{T}}$称之为类间散度矩阵（between-class scatter matrix），$\boldsymbol{S}_w = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$称之为类内散度矩阵（within-class scatter matrix）。

注1：当总体的均值向量和协方差阵未知时，可以基于样本进行估计$\bar{\boldsymbol{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{x}_j^{(i)}$, $\hat{\boldsymbol{\Sigma}}_i = \sum_{j=1}^{n_i}(\boldsymbol{x}_j^{(i)} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_j^{(i)} - \bar{\boldsymbol{x}}_i)^{\mathsf{T}}$, $i = 1, 2$。

两总体情形

其中$\boldsymbol{S}_b = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathsf{T}}$称之为类间散度矩阵（between-class scatter matrix），$\boldsymbol{S}_w = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$称之为类内散度矩阵（within-class scatter matrix）。

注1：当总体的均值向量和协方差阵未知时，可以基于样本进行估计$\bar{\boldsymbol{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{x}_j^{(i)}$, $\hat{\boldsymbol{\Sigma}}_i = \sum_{j=1}^{n_i}(\boldsymbol{x}_j^{(i)} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_j^{(i)} - \bar{\boldsymbol{x}}_i)^{\mathsf{T}}$, $i = 1, 2$。

注2：使得$\Delta(\boldsymbol{w})$达到最大的$\boldsymbol{w}$并不唯一。提示：$\Delta(c\boldsymbol{w}) = \Delta(\boldsymbol{w})$。

两总体情形

其中$\boldsymbol{S}_b = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathsf{T}}$称之为类间散度矩阵（between-class scatter matrix），$\boldsymbol{S}_w = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$称之为类内散度矩阵（within-class scatter matrix）。

注1：当总体的均值向量和协方差阵未知时，可以基于样本进行估计$\bar{\boldsymbol{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{x}_j^{(i)}$, $\hat{\boldsymbol{\Sigma}}_i = \sum_{j=1}^{n_i}(\boldsymbol{x}_j^{(i)} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_j^{(i)} - \bar{\boldsymbol{x}}_i)^{\mathsf{T}}$, $i = 1, 2$。

注2：使得$\Delta(\boldsymbol{w})$达到最大的$\boldsymbol{w}$并不唯一。提示：$\Delta(c\boldsymbol{w}) = \Delta(\boldsymbol{w})$。

注3：$\Delta(\boldsymbol{w})$可衡量判别函数$\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}$的效果，称为判别效率。

两总体情形

注4：判别效率最大化问题

$$max_{\boldsymbol{w}} \quad \frac{\boldsymbol{w}^{\mathsf{T}}\boldsymbol{S}_b\boldsymbol{w}}{\boldsymbol{w}^{\mathsf{T}}\boldsymbol{S}_w\boldsymbol{w}}$$

等价于一个带约束的极值问题

$$\begin{aligned} max_{\boldsymbol{w}} \quad & \boldsymbol{w}^{\mathsf{T}}\boldsymbol{S}_b\boldsymbol{w} \\ s.t. \quad & \boldsymbol{w}^{\mathsf{T}}\boldsymbol{S}_w\boldsymbol{w} = c \end{aligned}$$

其中c为一个正常数。

两总体情形

引入lagrange乘子λ，化为无约束的极值问题

$$max \quad F(\boldsymbol{w}, \lambda) = \boldsymbol{w}^{\mathsf{T}} \boldsymbol{S}_b \boldsymbol{w} - \lambda(\boldsymbol{w}^{\mathsf{T}} \boldsymbol{S}_w \boldsymbol{w} - c)$$

两总体情形

引入lagrange乘子λ，化为无约束的极值问题

$$max \quad F(\boldsymbol{w},\lambda)=\boldsymbol{w}^{\mathsf{T}}\boldsymbol{S}_b\boldsymbol{w}-\lambda(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{S}_w\boldsymbol{w}-c)$$

极值点存在的必要条件：

$$\frac{\partial F(\boldsymbol{w},\lambda)}{\partial \boldsymbol{w}}=2\boldsymbol{S}_b\boldsymbol{w}-2\lambda\boldsymbol{S}_w\boldsymbol{w}=0$$

即有

$$(\boldsymbol{S}_b-\lambda\boldsymbol{S}_w)\boldsymbol{w}=0$$

两总体情形

引入lagrange乘子λ，化为无约束的极值问题

$$max \quad F(\boldsymbol{w}, \lambda) = \boldsymbol{w}^{\mathsf{T}} \boldsymbol{S}_b \boldsymbol{w} - \lambda(\boldsymbol{w}^{\mathsf{T}} \boldsymbol{S}_w \boldsymbol{w} - c)$$

极值点存在的必要条件：

$$\frac{\partial F(\boldsymbol{w}, \lambda)}{\partial \boldsymbol{w}} = 2\boldsymbol{S}_b \boldsymbol{w} - 2\lambda \boldsymbol{S}_w \boldsymbol{w} = 0$$

即有

$$(\boldsymbol{S}_b - \lambda \boldsymbol{S}_w)\boldsymbol{w} = 0$$

由矩阵知识，可知$\Delta(\cdot)$的极大值为$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$的最大特征根，$\boldsymbol{w}$取为相应的特征向量。

两总体情形

注5：由于仅关心向量$\boldsymbol{w}$的方向而不关心长度，故计算可以简化。

$$\boldsymbol{S}_b\boldsymbol{w} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\underline{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{w}} = \alpha(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

两总体情形

注5：由于仅关心向量$\boldsymbol{w}$的方向而不关心长度，故计算可以简化。

$$\boldsymbol{S}_b\boldsymbol{w} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\underline{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{w}} = \alpha(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

从而有

$$\lambda\boldsymbol{S}_w\boldsymbol{w} = \alpha(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

可求得$\boldsymbol{w} = \boldsymbol{S}_w^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$。考虑到数值解的稳定性，实际中一般都借助奇异值分解$\boldsymbol{S}_w = U\Lambda V^{\mathsf{T}}$，其中$\Lambda$为对角矩阵，$U, V$为正交矩阵。从而有$\boldsymbol{S}_w^{-1} = V\Lambda^{-1}U^{\mathsf{T}}$。

两总体情形

注6：结果解释。

$\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$为连接两类均值的向量，从两类样本被分割最远的效果来看，与$\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$平行的投影直线将两类分得最开。

两总体情形

注6：结果解释。

$\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$为连接两类均值的向量，从两类样本被分割最远的效果来看，与$\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$平行的投影直线将两类分得最开。

但还要兼顾到类内密集程度较高，则需根据两类样本的分布离散程度($\boldsymbol{S}_w$)对投影方向作调整，这就体现在对向量$\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$作线性变换($\boldsymbol{S}_w^{-1}$)。

多总体情形

推广到多个情形。假定存在m个总体，第i个总体的样本数为n_i

$$
\begin{aligned}
G_1 &: \boldsymbol{x}_1^{(1)}, \boldsymbol{x}_2^{(1)}, \cdots, \boldsymbol{x}_{n_1}^{(1)} \\
G_2 &: \boldsymbol{x}_1^{(2)}, \boldsymbol{x}_2^{(2)}, \cdots, \boldsymbol{x}_{n_2}^{(2)} \\
&\vdots \\
G_m &: \boldsymbol{x}_1^{(m)}, \boldsymbol{x}_2^{(m)}, \cdots, \boldsymbol{x}_{n_m}^{(m)}
\end{aligned}
$$

其中，$\boldsymbol{x}_j^{(i)}$表示从第i个总体中抽取的第j个样本，$i = 1, 2, \cdots, m$，$j = 1, 2, \cdots, n_i$。则第i个总体的样本均值为$\bar{\boldsymbol{x}}^{(i)} = \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{x}_j^{(i)}$，全部样本数$n = \sum_{i=1}^{m} n_i$，全部样本的均值为$\bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \boldsymbol{x}_j^{(i)}$。

多总体情形

实现多分类的线性判别，可转化为最大化问题

$$max_{\boldsymbol{w}} \quad \frac{tr\left(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{S}_b\boldsymbol{w}\right)}{tr\left(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{S}_w\boldsymbol{w}\right)}$$

其中，类间散度矩阵$\boldsymbol{S}_b = \sum_{i=1}^{m} n_i(\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}})(\overline{\boldsymbol{x}}^{(i)} - \overline{\boldsymbol{x}})^{\mathsf{T}}$，类内散度矩阵$\boldsymbol{S}_w = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\boldsymbol{x}_j^{(i)} - \overline{\boldsymbol{x}}^{(i)})(\boldsymbol{x}_j^{(i)} - \overline{\boldsymbol{x}}^{(i)})^{\mathsf{T}}$。

多总体情形

实现多分类的线性判别，可转化为最大化问题

$$max_{\boldsymbol{w}} \quad \frac{tr(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{S}_b\boldsymbol{w})}{tr(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{S}_w\boldsymbol{w})}$$

其中，类间散度矩阵$\boldsymbol{S}_b = \sum_{i=1}^{m} n_i(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}^{(i)} - \bar{\boldsymbol{x}})^{\mathsf{T}}$，类内散度矩阵$\boldsymbol{S}_w = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\boldsymbol{x}_j^{(i)} - \bar{\boldsymbol{x}}^{(i)})(\boldsymbol{x}_j^{(i)} - \bar{\boldsymbol{x}}^{(i)})^{\mathsf{T}}$。

等价于求解：

$$(\boldsymbol{S}_b - \lambda\boldsymbol{S}_w)\boldsymbol{w} = 0$$

投影矩阵$\boldsymbol{w}$即为$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$的$c$个最大非零特征值对应的特征向量组成的矩阵。

案例

Fisher于1936年发表的鸢尾花（Iris）数据。

三种鸢尾花：Setosa、Versicolor、Virginica各取容量为50的样本，测量其花萼长、花萼宽、花瓣长、花瓣宽，单位为mm。如excel附件。

I. versicolor photo by D. Langlois [Wikimedia Commons]

I. virginica photo by Frank Mayfield [Wikimedia Commons]

案例

第1步：数据标准化。首先计算全部样本（共150个）的均值和标准差。然后每个样本减去整体均值后再除以整体标准差。

	花萼长	花萼宽	花瓣长	花瓣宽
均值	5.84	3.06	3.76	1.20
标准差	0.83	0.44	1.77	0.76

案例

第1步：数据标准化。首先计算全部样本（共150个）的均值和标准差。然后每个样本减去整体均值后再除以整体标准差。

	花萼长	花萼宽	花瓣长	花瓣宽
均值	5.84	3.06	3.76	1.20
标准差	0.83	0.44	1.77	0.76

第2步：计算标准化后每个类的均值。

	花萼长	花萼宽	花瓣长	花瓣宽
Setosa	-1.01	0.85	-1.30	-1.25
Versicolor	0.11	-0.66	0.28	0.17
Virginica	0.90	-0.19	1.02	1.08

案例

第3步：计算类内散度矩阵S_w和类间散度矩阵S_b。

$$\begin{pmatrix} 56.81 & 37.76 & 16.85 & 8.94 \\ 37.76 & 89.28 & 10.55 & 14.47 \\ 16.85 & 10.55 & 8.74 & 4.66 \\ 8.94 & 14.47 & 4.66 & 10.60 \end{pmatrix}, \begin{pmatrix} 92.19 & -55.28 & 113.05 & 112.93 \\ -55.28 & 59.72 & -74.39 & -69.03 \\ 113.05 & -74.39 & 140.26 & 138.81 \\ 112.93 & -69.03 & 138.81 & 138.40 \end{pmatrix}$$

案例

第3步：计算类内散度矩阵S_w和类间散度矩阵S_b。

$$\begin{pmatrix} 56.81 & 37.76 & 16.85 & 8.94 \\ 37.76 & 89.28 & 10.55 & 14.47 \\ 16.85 & 10.55 & 8.74 & 4.66 \\ 8.94 & 14.47 & 4.66 & 10.60 \end{pmatrix}, \begin{pmatrix} 92.19 & -55.28 & 113.05 & 112.93 \\ -55.28 & 59.72 & -74.39 & -69.03 \\ 113.05 & -74.39 & 140.26 & 138.81 \\ 112.93 & -69.03 & 138.81 & 138.40 \end{pmatrix}$$

第4步：计算矩阵$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$的特征值和特征向量。

$$\begin{pmatrix} 32.19 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.29 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 \end{pmatrix}, \begin{pmatrix} -0.15 & 0.01 & 0.62 & 0.15 \\ -0.15 & 0.33 & 0.02 & -0.17 \\ 0.86 & -0.57 & 0.24 & -0.78 \\ 0.47 & 0.75 & -0.74 & 0.58 \end{pmatrix}$$

若取前两个特征值，即投影矩阵为特征向量矩阵的前两列。

案例

第5步：计算每个类均值向量的投影。

	投影方向1	投影方向2
Setosa	-1.68	0.07
Versicolor	0.40	-0.25
Virginica	1.27	0.18

案例

第5步：计算每个类均值向量的投影。

	投影方向1	投影方向2
Setosa	-1.68	0.07
Versicolor	0.40	-0.25
Virginica	1.27	0.18

第6步：计算每个样本的投影，然后分别计算这些投影到每个类均值投影的欧式距离，根据距离判别法进行判断。识别结果如下：

	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	48	2
Virginica	0	3	47

识别正确率96.67%。

案例

识别错误的样本如下：

花萼长	花萼宽	花瓣长	花瓣宽	原来的类别	判别类别
5.9	3.2	4.8	1.8	Versicolor	Virginica
6	2.7	5.1	1.6	Versicolor	Virginica
6	2.2	5	1.5	Virginica	Versicolor
6.3	2.8	5.1	1.5	Virginica	Versicolor
6.1	2.6	5.6	1.4	Virginica	Versicolor

对照每个类的均值（标准化前）。

类别	花萼长	花萼宽	花瓣长	花瓣宽
Setosa	5.01	3.43	1.46	0.25
Versicolor	5.94	2.77	4.26	1.33
Virginica	6.59	2.97	5.55	2.03

案例

若只取一个投影方向，识别结果如下：

	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	48	2
Virginica	0	0	50

识别正确率98.67%。

总结

多分类的线性判别问题将p维样本投影到c维空间，其中$c < p$。通过投影，减少了样本的维数，且投影过程中使用了类别信息。因而，线性判别也被视为是一种有监督（supervised）的降维技术。

本章作业

（1）必做题。讨论距离判别法与Bayes判别法的联系。

（2）选做题。汇报Fisher判别法。要求：限5人组队，其中1人汇报；课件或黑板手写，限时5-10min；严禁直接抄袭现有材料（如网上、教材等）。建议：原理+步骤+案例。

（3）必做题。收集一个判别问题的数据（最好含有类别信息），分别采用距离判别法，Bayes判别法，Fisher判别法进行计算，对比三者的识别正确率。要求：限5人组队；需提供数据出处，原始数据；若采用spss等不编程，需注明操作步骤并解释计算结果，严禁直接粘贴软件的图表；若采用matlab,R等编程，需提供代码；提交纸质版，封面为组内成员的姓名和学号。