

法律声明

□ 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

□ 课程详情请咨询

■ 微信公众号：小象

■ 新浪微博：ChinaHadoop



第四章 判别分析

主讲教师 周仕君

第四章 判别分析

- 第一节 引言
- 第二节 距离判别分析
- 第三节 贝叶斯 (Bayes) 判别法
- 第四节 费歇 (Fisher) 判别法
- 第五节 R语言实现

第一节 引言

- 判别分析的基本思想是根据一批分类明确的样本在若干指标上的观察值，建立一个关于指标的判别函数和判别准则，然后根据这个判别函数和判别准则对新的样本进行分类，并且根据回代判别的准确率评估它的实用性。

□ 例如，以一批正常人和一批已经确诊的病人为样本，收集他们的各项检验指标，如化验指标、X线，心脑电图、超声波、CT等诊断指标，然后利用这批分类明确的样本在这些相同指标上的观察值，建立一个关于指标的判别函数和判别准则(区分病人和正常人的界限的方法)，使得按此准则来判断这批样本归属的正确率达到最高。

□ 把判别分析的思想用数学语言来表达，可以叙述如下：
设有 n 个样本，对每个样本测得 p 项指标（变量）的数据，
已知每个样本属于 k 个类别（或总体） $G_1, G_2, \dots, G_k$
中的某一类，且它们的分布函数分别为 $F_1(x), F_2(x), \dots, F_k(x)$ 。我们希望利用这些数据，找出一种
判别函数，使得这一函数具有某种最优性质，能把属于
不同类别的样本点尽可能地区别开来，并对测得同样 p
项指标（变量）数据的一个新样本，能判定这个样本归
属于哪一类。

-
- ❑ 判断分析按判别的总体数来区分，有两个总体判别分析和多总体判别分析；
 - ❑ 按区分不同总体所用的数学模型来分，有线性判别和非线性判别；
 - ❑ 常用的几种判别分析方法：距离判别法、Fisher判别法、Bayes判别法和逐步判别法。
 - ❑ 按判别时所处理的变量方法不同，有逐步判别和序贯判别等。

第二节 距离判别法

- 马氏距离的概念
- 距离判别的思想及方法
- 判别分析的实现

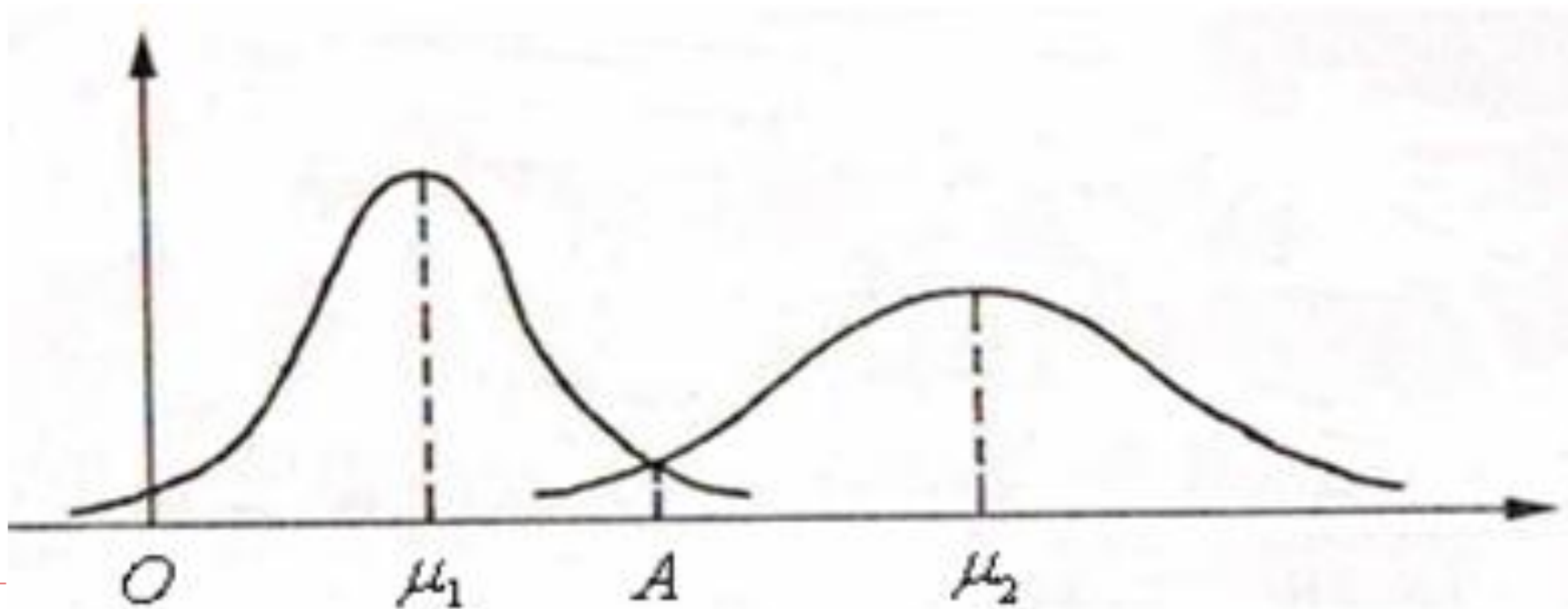
一、马氏距离的概念

□ 设 p 维欧氏空间 R^p 中的两点 $\mathbf{X} = (X_1, X_2, \dots, X_p)'$ 和 $\mathbf{Y} = (Y_1, Y_2, \dots, Y_p)'$ 通常我们所说的两点之间的距离，是指欧氏距离，即

$$d^2(\mathbf{X}, \mathbf{Y}) = (X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \dots + (X_p - Y_p)^2$$

在解决实际问题时，特别是针对多元数据的分析问题，欧氏距离就显示出了它的薄弱环节。

- 第一、设有两个正态总体, $X \sim N(\mu_1, \sigma^2)$ 和 $Y \sim N(\mu_2, 4\sigma^2)$, 现有一个样品位于如图4.1所示的A点, 距总体X的中心 2σ 远, 距总体Y的中心 3σ 远, 那么, A点处的样品到底离哪一个总体近呢?



第二、设有量度重量和长度的两个变量 X 与 Y ，以单位分别为 **kg** 和 **cm** 得到样本 $A(0,5)$ ， $B(10,0)$ ， $C(1,0)$ ， $D(0,10)$ 。
今按照欧氏距离计算，有

$$AB = \sqrt{10^2 + 5^2} = \sqrt{125} ;$$

$$CD = \sqrt{1^2 + 10^2} = \sqrt{101}$$

如果我们将长度单位变为 **mm**，那么，有

$$AB = \sqrt{10^2 + 50^2} = \sqrt{2600} ;$$

$$CD = \sqrt{1^2 + 100^2} = \sqrt{10001}$$

量纲的变化，将影响欧氏距离计算的结果。

-
- 为此，我们引入一种由印度著名统计学家马哈拉诺比斯（Mahalanobis, 1936）提出的“马氏距离”的概念。
 - 设 $\mathbf{X}$ 与 $\mathbf{Y}$ 是来自均值向量为 $\boldsymbol{\mu}$ ，协方差为 $\boldsymbol{\Sigma}$ 的总体 G 中的 p 维样品，则两个样品间的马氏距离为：

$$D^2(\mathbf{X}, \mathbf{Y}) = (\mathbf{X} - \mathbf{Y})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \mathbf{Y})$$

- 定义点到总体 G 的马氏距离为

$$D^2(\mathbf{X}, G) = (\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu})$$

- 这里应该注意到，当 $\boldsymbol{\Sigma} = \mathbf{I}$ （单位矩阵）时，即为欧氏距离的情形。

二、距离判别的思想及方法

1、两个总体的距离判别问题

 问题：设有协方差矩阵 Σ 相等的两个总体 G_1 和 G_2 ，其均值分别是 μ_1 和 μ_2 ，对于一个新的样品 X ，要判断它来自哪个总体。

□ 一般的想法是计算新样品 X 到两个总体的马氏距离 $D^2(X, G_1)$ 和 $D^2(X, G_2)$ ，并按照如下的判别规则进行判断。

$$\begin{cases} X \in G_1, & \text{如果 } D^2(X, G_1) \leq D^2(X, G_2) \\ X \in G_2, & \text{如果 } D^2(X, G_1) > D^2(X, G_2) \end{cases} \quad (4.4)$$

□ 这个判别规则的等价描述为：求新样品 X 到 G_1 的距离与到 G_2 的距离之差，如果其值为正， X 属于 G_2 ；否则 X 属于 G_1 。

□ 我们考虑

$$\begin{aligned} & D^2(\mathbf{X}, G_1) - D^2(\mathbf{X}, G_2) \\ &= (\mathbf{X} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}_1) - (\mathbf{X} - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}_2) \\ &= \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X} - 2\mathbf{X}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\mu}_1' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 - (\mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X} - 2\mathbf{X}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 + \boldsymbol{\mu}_2' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2) \\ &= 2\mathbf{X}' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \boldsymbol{\mu}_1' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 \\ &= 2\mathbf{X}' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &= -2 \left(\mathbf{X} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2} \right)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &= -2(\mathbf{X} - \bar{\boldsymbol{\mu}})' \boldsymbol{\alpha} = -2\boldsymbol{\alpha}'(\mathbf{X} - \bar{\boldsymbol{\mu}}) \end{aligned}$$

其中 $\bar{\mu} = \frac{1}{2}(\mu_1 + \mu_2)$ 是两个总体均值的平均值，
 $\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$ ，记

$$W(\mathbf{X}) = \alpha'(\mathbf{X} - \bar{\mu}) \quad (4.5)$$

则判别规则 (4.4) 式可表示为

$$\begin{cases} \mathbf{X} \in G_1, & \text{如果 } W(\mathbf{X}) \geq 0 \\ \mathbf{X} \in G_2, & \text{如果 } W(\mathbf{X}) < 0 \end{cases} \quad (4.6)$$

这里称 $W(\mathbf{X})$ 为两总体距离判别的判别函数，由于它是 $\mathbf{X}$ 的线性函数，故又称为线性判别函数， α 称为判别系数。

在实际应用中，总体的均值和协方差矩阵一般是未知的，可由样本均值和样本协方差矩阵分别进行估计。设 $\mathbf{x}_1^{(1)}, \dots, \mathbf{x}_{n_1}^{(1)}$ 来自总体 G_1 的样本， $\mathbf{x}_1^{(2)}, \dots, \mathbf{x}_{n_2}^{(2)}$ 是来自总体 G_2 的样本， μ_1 和 μ_2 的一个无偏估计分别为

$$\bar{\mathbf{X}}^{(1)} = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{X}_i^{(1)} \quad \text{和} \quad \bar{\mathbf{X}}^{(2)} = \frac{1}{n_2} \sum_{i=1}^{n_2} \mathbf{X}_i^{(2)}$$

Σ 的一个联合无偏估计为

$$\hat{\Sigma} = \frac{1}{n_1 + n_2 - 2} (\mathbf{S}_1 + \mathbf{S}_2)$$

这里

$$\mathbf{S}_\alpha = \sum_{i=1}^{n_\alpha} (\mathbf{X}_i^{(\alpha)} - \bar{\mathbf{X}}^{(\alpha)})(\mathbf{X}_i^{(\alpha)} - \bar{\mathbf{X}}^{(\alpha)})', \quad \alpha = 1, 2$$

此时，两总体距离判别的判别函数为 $\hat{W}(\mathbf{X}) = \hat{\alpha}'(\mathbf{X} - \bar{\mathbf{X}})$

其中 $\bar{\mathbf{X}} = \frac{1}{2}(\bar{\mathbf{X}}^{(1)} + \bar{\mathbf{X}}^{(2)})$, $\hat{\alpha} = \hat{\Sigma}^{-1}(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)})$ 。这样，判别规则为

$$\begin{cases} \mathbf{X} \in G_1, & \text{如果 } \hat{W}(\mathbf{X}) \geq 0 \\ \mathbf{X} \in G_2, & \text{如果 } \hat{W}(\mathbf{X}) < 0 \end{cases} \quad (4.7)$$

□ 这里我们应该注意到：

(1) 当 $p=1$ ， G_1 和 G_2 的分布分别为 $N(\mu_1, \sigma^2)$ 和 $N(\mu_2, \sigma^2)$ 时， μ_1, μ_2, σ^2 均为已知，且 $\mu_1 < \mu_2$ ，则判别系数为 $\alpha = \frac{\mu_1 - \mu_2}{\sigma^2} < 0$ ，判别函数为

$$W(x) = \alpha(x - \bar{\mu})$$

判别规则为

$$\begin{cases} x \in G_1, & \text{如果 } x \leq \bar{\mu} \\ x \in G_2, & \text{如果 } x > \bar{\mu} \end{cases}$$

(2) 当 $\mu_1 \neq \mu_2$, $\Sigma_1 \neq \Sigma_2$ 时, 我们采用 (4.4) 式作为判别规则的形式。选择判别函数为

$$\begin{aligned} W^*(\mathbf{X}) &= D^2(\mathbf{X}, G_1) - D^2(\mathbf{X}, G_2) \\ &= (\mathbf{X} - \mu_1)' \Sigma_1^{-1} (\mathbf{X} - \mu_1) - (\mathbf{X} - \mu_2)' \Sigma_2^{-1} (\mathbf{X} - \mu_2) \end{aligned}$$

它是 $\mathbf{X}$ 的二次函数, 相应的判别规则为

$$\begin{cases} \mathbf{X} \in G_1, & \text{如果 } W^*(\mathbf{X}) \leq 0 \\ \mathbf{X} \in G_2, & \text{如果 } W^*(\mathbf{X}) > 0 \end{cases}$$

□ 例1 今从1995年世界各国人文发展指数的排序中，选取发达国家和发展中国家各5个，观察两个重要指标权：出生时预期寿命(岁) X_1 ，成人识字率(%) X_2 。以此样本为训练样本建立判别模型，并对四个待判国家进行判别分析。

14 个国家的出生时预期寿命和成人识字率

类别	序号	国家名称	出生时预期寿命(岁)	成人识字率(%)
第一类 (发达国家)	1	美国	76.0	99.0
	2	日本	79.5	99.0
	3	瑞士	78.0	99.0
	4	阿根廷	72.1	95.9
	5	阿联酋	73.8	77.7
第二类 (发展中国家)	6	保加利亚	71.2	93.0
	7	古巴	75.3	94.9
	8	巴拉圭	70.0	91.2
	9	格鲁吉亚	72.8	99.0
	10	南非	62.9	80.6
待判样品	11	中国	68.5	79.3
	12	罗马尼亚	69.9	96.9
	13	希腊	77.6	93.8
	14	哥伦比亚	69.3	90.3

解:

(1) 计算各类均向量和协方差矩阵,及合并协方差矩阵。

均向量

协方差矩阵

第一类:

$$\begin{array}{l} x_1 \\ x_2 \end{array} \begin{bmatrix} 75.88 \\ 94.12 \end{bmatrix} \quad \begin{array}{l} S_1 \\ V_1 \end{array} = \begin{bmatrix} 9.0570 & 14.0055 \\ 14.0055 & 86.0570 \end{bmatrix}$$

第二类:

$$\begin{array}{l} x_1 \\ x_2 \end{array} \begin{bmatrix} 70.44 \\ 91.74 \end{bmatrix} \quad S_2 = \begin{bmatrix} 21.7030 & 29.4205 \\ 29.4205 & 47.1680 \end{bmatrix}$$

合并协方差矩阵:

$$\begin{array}{l} x_1 \\ x_2 \end{array} \quad \hat{\Sigma} = \begin{bmatrix} 15.3800 & 21.7130 \\ 21.7130 & 66.6125 \end{bmatrix}$$

合并协方差矩阵之逆:

$$\begin{array}{l} x_1 \\ x_2 \end{array} \quad \begin{array}{l} V_w^{-1} \\ \hat{\Sigma}^{-1} \end{array} = \begin{bmatrix} 0.120447 & -0.039261 \\ -0.039261 & 0.027810 \end{bmatrix}$$

$$d_1(2) = (x_1 - 70.44, x_2 - 91.74) \begin{pmatrix} 0.120447 & -0.039261 \\ -0.039261 & 0.027810 \end{pmatrix} \begin{pmatrix} x_1 - 70.44 \\ x_2 - 91.74 \end{pmatrix}$$

判别函数为：

$$W = (x_1 - 73.16, x_2 - 92.93) \begin{pmatrix} 0.120477 & -0.039261 \\ -0.039261 & 0.027810 \end{pmatrix} \begin{pmatrix} -5.44 \\ -2.38 \end{pmatrix}$$

展开后得： $W = 27.4034 - 0.5618x_1 + 0.1474x_2$

结果见表

因此,对任一样品,只要将其取值代入上式,依据判别准则,即可作出判别,而不必再计算马氏距离。

序号	类别	国家名称	W	判别各类
1.	第一类 (发达国家)	美国	-0.7008	1
2.		日本	-2.6671	1
3.		瑞士	-1.8244	1
4.		阿根廷	1.0332	2
5.		阿联酋	-2.6043	1
6.	第二类 (发展中 国家)	保加利亚	1.1114	2
7.		古巴	-0.9119	1
8.		巴拉圭	1.5203	2
9.		格鲁吉亚	1.0969	2
10.		南非	3.9466	2
11.	待判样品	中国	0.6090	2
12.		罗马尼亚	2.4166	2
13.		希腊	-2.3661	1
14.		哥伦比亚	1.7809	2

第一，10个样品中，有2个国家错判。一是将属于发达国家的阿根廷错判为发展中国家，二是将属于发展中国家的古巴错判为发达国家。错判(miscassified)率为20%。错判率是衡量判别效果的一个重要指标。这是对训练样本(已知类别)判别归类，目的是考察两类的判别效果。称为组内回代。

第二，对待判样品进行分类。4个国家中，希腊判为第一类，中国、罗马尼亚、哥伦比亚属于第二类。本步是对待判样品(未知类别)进行判别归类，目的是判定各样品的归属。

2、多个总体的距离判别问题

问题：设有 k 个总体 $G_1, G_2, \dots, G_k$ ，其均值和协方差矩阵分别是 $\mu_1, \mu_2, \dots, \mu_k$ 和 $\Sigma_1, \Sigma_2, \dots, \Sigma_k$ ，而且 $\Sigma_1 = \Sigma_2 = \dots = \Sigma_k = \Sigma$ 。对于一个新的样品 X ，要判断它来自哪个总体。

该问题与两个总体的距离判别问题的解决思想一样。计算新样品 X 到每一个总体的距离，即

$$\begin{aligned} D^2(X, G_\alpha) &= (X - \mu_\alpha)' \Sigma^{-1} (X - \mu_\alpha) \\ &= X' \Sigma^{-1} X - 2\mu_\alpha' \Sigma^{-1} X + \mu_\alpha' \Sigma^{-1} \mu_\alpha \\ &= X' \Sigma^{-1} X - 2(I_\alpha' X + C_\alpha) \end{aligned} \quad (4.8)$$

这里 $I_\alpha = \Sigma^{-1} \mu_\alpha$ ， $C_\alpha = -\frac{1}{2} \mu_\alpha' \Sigma^{-1} \mu_\alpha$ ， $\alpha = 1, 2, \dots, k$ 。

由（4.8）式，可以取线性判别函数为

$$W_{\alpha}(\mathbf{X}) = \mathbf{I}'_{\alpha} \mathbf{X} + C_{\alpha}, \quad \alpha = 1, 2, \dots, k$$

相应的判别规则为

$$\mathbf{X} \in G_i \quad \text{如果} \quad W_i(\mathbf{X}) = \max_{1 \leq \alpha \leq k} (\mathbf{I}'_{\alpha} \mathbf{X} + C_{\alpha}) \quad (4.9)$$

针对实际问题，当 $\mu_1, \mu_2, \dots, \mu_k$ 和 Σ 均未知时，可以通过相应的样本值来替代。

设 $\mathbf{X}_1^{(\alpha)}, \dots, \mathbf{X}_{n_\alpha}^{(\alpha)}$ 是来自总体 G_α 中的样本 ($\alpha = 1, 2, \dots, k$), 则 $\boldsymbol{\mu}_\alpha$ ($\alpha = 1, 2, \dots, k$) 和 $\boldsymbol{\Sigma}$ 可估计为

$$\bar{\mathbf{X}}^{(\alpha)} = \frac{1}{n_\alpha} \sum_{i=1}^{n_\alpha} \mathbf{X}_i^{(\alpha)}, \quad \alpha = 1, 2, \dots, k$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-k} \sum_{\alpha=1}^k \mathbf{S}_\alpha,$$

其中 $n = n_1 + n_2 + \dots + n_k$,

$$\mathbf{S}_\alpha = \sum_{i=1}^{n_\alpha} (\mathbf{X}_i^{(\alpha)} - \bar{\mathbf{X}}^{(\alpha)})(\mathbf{X}_i^{(\alpha)} - \bar{\mathbf{X}}^{(\alpha)})', \quad \alpha = 1, 2, \dots, k$$

同样，我们注意到，如果总体 $G_1, G_2, \dots, G_k$ 的协方差矩阵分别是 $\Sigma_1, \Sigma_2, \dots, \Sigma_k$ ，而且它们不全相等，则计算 $\mathbf{X}$ 到各总体的马氏距离，即

$$D^2(\mathbf{X}, G_\alpha) = (\mathbf{X} - \boldsymbol{\mu}_\alpha)' \boldsymbol{\Sigma}_\alpha^{-1} (\mathbf{X} - \boldsymbol{\mu}_\alpha) \quad \alpha = 1, 2, \dots, k$$

则判别规则为

$$\mathbf{X} \in G_i \quad \text{如果} \quad D^2(\mathbf{X}, G_i) = \min_{1 \leq \alpha \leq k} D^2(\mathbf{X}, G_\alpha) \quad (4.10)$$

当 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_k$ 和 $\Sigma_1, \Sigma_2, \dots, \Sigma_k$ 均未知时， $\boldsymbol{\mu}_\alpha$ ($\alpha = 1, 2, \dots, k$) 的估计同前， Σ_α ($\alpha = 1, 2, \dots, k$) 的估计为

$$\hat{\Sigma}_\alpha = \frac{1}{n_\alpha - 1} \mathbf{S}_\alpha, \quad \alpha = 1, 2, \dots, k$$

例：fisher于1936年发表的Iris数据已被广泛地作为判别分析的例子。他测量了某种植物的萼片长、宽和花瓣长、宽，以毫米为单位。样本取自三个品种，每个品种取50个样品。结果见右表。

某种植物三个品种的萼片长、宽和花萼长、宽测量值

编号	第一类				编号	第二类				编号	第三类			
	x1	x2	x3	x4		x1	x2	x3	x4		x1	x2	x3	x4
1.	51	35	14	3	51.	58	27	42	13	101.	63	25	50	19
2.	52	41	15	1	52.	70	32	47	14	102.	69	31	54	21
3.	50	34	16	2	53.	51	25	30	11	103.	76	30	66	21
4.	50	35	16	6	54.	64	29	43	13	104.	69	31	51	23
5.	46	32	14	2	55.	69	31	49	15	105.	67	31	56	24
6.	57	38	17	3	56.	49	24	33	10	106.	60	30	48	18
7.	52	35	15	2	57.	64	32	45	15	107.	64	28	56	22
8.	47	32	13	2	58.	58	27	41	10	108.	61	30	49	18
9.	54	39	17	4	59.	56	29	36	13	109.	72	30	58	16
10.	53	37	15	2	60.	58	26	40	12	110.	73	29	63	18
11.	48	34	18	2	61.	57	26	35	10	111.	68	32	59	23
12.	43	30	11	1	62.	62	28	43	13	112.	65	30	58	22
13.	55	42	14	2	63.	67	31	47	15	113.	69	32	57	23
14.	50	35	13	3	64.	60	22	40	10	114.	63	34	56	24
15.	45	23	13	3	65.	67	30	50	17	115.	58	27	51	19
16.	55	35	13	2	66.	58	27	39	12	116.	64	28	56	21
17.	50	36	14	2	67.	55	24	37	10	117.	57	25	50	20
18.	52	34	14	2	68.	50	20	35	10	118.	67	33	57	21
19.	50	33	14	2	69.	65	28	46	15	119.	63	29	56	18
20.	51	34	15	2	70.	52	27	39	14	120.	58	28	51	24
21.	46	36	10	2	71.	50	23	33	10	121.	74	28	61	19
22.	49	31	15	2	72.	55	26	44	12	122.	58	27	51	19
23.	51	35	14	2	73.	55	23	40	13	123.	63	33	60	25
24.	44	32	13	2	74.	67	31	44	14	124.	64	32	53	23
25.	51	33	17	5	75.	63	25	49	15	125.	67	25	58	18
26.	46	31	15	2	76.	61	28	40	13	126.	72	32	60	18
27.	54	34	15	4	77.	56	30	45	15	127.	64	27	53	19
28.	54	37	15	2	78.	56	30	41	13	128.	65	30	55	18
29.	46	34	14	3	79.	57	29	42	13	129.	59	30	51	18
30.	48	30	14	1	80.	59	32	48	18	130.	60	22	50	16
31.	51	37	15	4	81.	61	29	47	14	131.	68	30	55	21
32.	49	30	14	2	82.	61	28	47	12	132.	71	30	59	21
33.	57	44	15	4	83.	56	25	39	11	133.	65	30	52	20
34.	44	29	14	2	84.	60	27	51	16	134.	67	30	52	23
35.	47	32	16	2	85.	62	22	45	15	135.	77	30	61	23



第二类:

$$\begin{matrix} x1 \\ x2 \\ x3 \\ x4 \end{matrix} \begin{bmatrix} 59.36 \\ 27.70 \\ 42.60 \\ 13.26 \end{bmatrix} \begin{bmatrix} 26.64326531 & 8.51836735 & 18.28979592 & 5.57795918 \\ 8.51836735 & 9.84693878 & 8.26530612 & 4.12040816 \\ 18.28979592 & 8.26530612 & 22.08163265 & 7.31020408 \\ 5.57795918 & 4.12040816 & 7.31020408 & 3.91061224 \end{bmatrix}$$

第三类:

$$\begin{matrix} x1 \\ x2 \\ x3 \\ x4 \end{matrix} \begin{bmatrix} 65.88 \\ 29.74 \\ 55.52 \\ 20.26 \end{bmatrix} \begin{bmatrix} 40.43428571 & 9.37632653 & 30.32897959 & 4.90938776 \\ 9.37632653 & 10.40040816 & 7.13795918 & 4.76285714 \\ 30.32897959 & 7.13795918 & 30.45877551 & 4.88244898 \\ 4.90938776 & 4.76285714 & 4.88244898 & 7.54326531 \end{bmatrix}$$

合并协方差矩阵:

$$V_w = \begin{bmatrix} 26.50081633 & 9.27210884 & 16.75142857 & 3.84013605 \\ 9.27210884 & 11.53877551 & 5.52435374 & 3.27102041 \\ 16.75142857 & 5.52435374 & 18.51877551 & 4.26653061 \\ 3.84013605 & 3.27102041 & 4.26653061 & 4.18816327 \end{bmatrix}$$

组内回代。

资料的原分类与距离判别分类

原分类	判别分类			合计
	1	2	3	
1	50	0	0	50
2	0	48	2	50
3	0	1	49	50
合计	50	49	51	150

其中,与原分类不一致的样品有:80,84 和 145 号。

三、判别分析的实质

□ 我们知道，判别分析就是希望利用已经测得的变量数据，找出一种判别函数，使得这一函数具有某种最优性质，能把属于不同类别的样本点尽可能地区别开来。为了更清楚的认识判别分析的实质，以便能灵活的应用判别分析方法解决实际问题，我们有必要了解“划分”这样概念。

设 $R_1, R_2, \dots, R_k$ 是 p 维空间 R^p 的 k 个子集, 如果它们互不相交, 且它们的和集为 R^p , 则称 $R_1, R_2, \dots, R_k$ 为 R^p 的一个划分。

在两个总体的距离判别问题中, 利用 $W(\mathbf{X}) = \boldsymbol{\alpha}'(\mathbf{X} - \bar{\boldsymbol{\mu}})$ 可以得到空间 R^p 的一个划分

$$\begin{cases} R_1 = \{\mathbf{X} : W(\mathbf{X}) \geq 0\} \\ R_2 = \{\mathbf{X} : W(\mathbf{X}) < 0\} \end{cases} \quad (4.11)$$

新的样品 $\mathbf{X}$ 落入 R_1 推断 $\mathbf{X} \in G_1$, 落入 R_2 推断 $\mathbf{X} \in G_2$ 。

□ 这样我们将会发现，判别分析问题实质上就是在某种意义上，以最优的性质对 p 维空间 R^p 构造一个“划分”，这个“划分”就构成了一个判别规则。这一思想将在后面的各节中体现的更加清楚。

第三节 贝叶斯判别法

- Bayes判别的基本思想
- Bayes判别的基本方法

-
- 从上节看距离判别法虽然简单，便于使用。
但是该方法也有它明显的不足之处。
 - ✿ 第一，判别方法与总体各自出现的概率的大小无关；
 - ✿ 第二，判别方法与错判之后所造成的损失无关。
 - ✿ Bayes判别法就是为了解决这些问题而提出的一种判别方法。

- 在一般情况下，我们对所研究的总体不会一无所知，或多或少总有一些认识，这个认识称为先验的(prior)。
- 英国统计学家Savage,L,J.曾考虑这样一个例子，一位常饮奶茶的妇女声称她能辨别先倒进杯子里的是牛奶还是茶。对此做了十次试验，她都正确地说了出来。
- 假如被测试者是在猜测，每次成功的概率为0.5，那么10次都说对的概率为0.0009766。这是一个很小的概率，在一次试验中基本是不可能发生的。所以，“每次成功的概率为0.5”的假设应被拒绝，被测试者说正确的概率要比0.5大得多。这就不是猜测，而是她的经验帮了她的忙。
- 可见“经验”在推断中是不可忽视的。这里的经验就是一种先验信息。

- Bayes学派认为，一个事件的概率可以是人们根据经验对该事件发生的可能性所给出的个人信念。这样给出的概率常称为先验概率或事前概率。
 - 例如，一位脑外科大夫要对某病人动手术，他认为成功的概率是90%，这是他根据手术的难易程度和自己的手术经验对“手术成功”所提供的把握程度。
- 当总体有一定先验概率时，先验概率对判别有较大影响，例如某地历史上多年统计的结果，旱涝年先验概率很小，常年先验概率很大，当气象指标既接近常年又接近涝年时，应当判为常年，这就是贝叶斯方法。

-
- Bayes判别法的基本思想是在判别分析时考虑先验概率，并利用Bayes公式导出后验概率(posterior probability)，即各个样品属于每一类的概率。判别准则是按后验概率大小归类。

一、Bayes判别的基本思想

问题：设有 k 个总体 $G_1, G_2, \dots, G_k$ ，其各自的分布密度函数 $f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_k(\mathbf{x})$ 互不相同的。假设 k 个总体各自出现的概率分别为 $q_1, q_2, \dots, q_k$ （先验概率）， $q_i \geq 0$ ， $\sum_{i=1}^k q_i = 1$ 。

假设已知若将本来属于 G_i 总体的样品错判到总体 G_j 时造成的损失为 $C(j | i)$ ， $i, j = 1, 2, \dots, k$ 。

在这样的情形下，对于新的样品 $\mathbf{X}$ 判断其来自哪个总体。

下面我们对这一问题进行分析。首先应该清楚 $C(i|i) = 0$ 、 $C(j|i) \geq 0$ ，对于任意的 $i, j = 1, 2, \dots, k$ 成立。

设 k 个总体 $G_1, G_2, \dots, G_k$ 相应的 p 维样本空间为 $R_1, R_2, \dots, R_k$ ，即为一个划分，故我们可以简记一个判别规则为 $R = (R_1, R_2, \dots, R_k)$ 。

从描述平均损失的角度出发，如果原来属于总体 G_i 且分布密度为 $f_i(\mathbf{x})$ 的样品，正好取值落入了 R_j ，我们就将会错判为属于 G_j 。

故在规则 R 下，将属于 G_i 的样品错判为 G_j 的概率为

$$P(j | i, R) = \int_{R_j} f_i(\mathbf{x}) d\mathbf{x} \quad i, j = 1, 2, \dots, k \quad i \neq j$$

此概率值等于总体 G_i 的分布密度在区域 R_j 上的面积。此面积可以用该分布密度函数在该区域上的积分求出。

如果实属 G_i 的样品，错判到其它总体 $G_1, \dots, G_{i-1}, G_{i+1}, \dots, G_k$ 所造成的损失为 $C(1|i), \dots, C(i-1|i), C(i+1|i), \dots, C(k|i)$ ，则这种判别规则 R 对总体 G_i 而言，样品错判后所造成的平均损失为

$$r(i|R) = \sum_{j=1}^k [C(j|i)P(j|i, R)] \quad i = 1, 2, \dots, k$$

其中 $C(i|i) = 0$

由于 k 个总体 $G_1, G_2, \dots, G_k$ 出现的先验概率分别为 $q_1, q_2, \dots, q_k$ ，则用规则 R 来进行判别所造成的总平均损失为

$$\begin{aligned} g(R) &= \sum_{i=1}^k q_i r(i, R) \\ &= \sum_{i=1}^k q_i \sum_{j=1}^k C(j | i) P(j | i, R) \end{aligned} \quad (4.12)$$

这相当于假设 $q_1, q_2, \dots, q_k$ 是一个样品被随机地选择为属于 $G_1, G_2, \dots, G_k$ 的各个概率，也就是说，考虑 k 个总体的混合，并随机地从这个单一混合总体中抽出一个样品，抽出的该样品将来自 G_i 的先验概率为 q_i ，即假定这 k 个总体出现的先验概率为 $q_1, q_2, \dots, q_k$ ，则通过划分 $R_1, R_2, \dots, R_k$ 来判别的平均损失为式 (4.12)。

所谓 **Bayes** 判别法则，就是要选择 $R_1, R_2, \dots, R_k$ ，使得

$$\begin{aligned} g(R) &= \sum_{i=1}^k q_i r(i, R) \\ &= \sum_{i=1}^k q_i \sum_{j=1}^k C(j | i) P(j | i, R) \end{aligned}$$

表示的总平均损失 $g(R)$ 达到极小。

二、Bayes判别的基本方法

设每一个总体 G_i 的分布密度为 $f_i(\mathbf{x})$, $i = 1, 2, \dots, k$, 来自总体 G_i 的样品 $\mathbf{X}$ 被错判为来自总体 G_j ($i, j = 1, 2, \dots, k$) 时所造成的损失记为 $C(j | i)$, 并且 $C(i | i) = 0$ 。那么, 对于判别规则 $R = (R_1, R_2, \dots, R_k)$ 产生的误判概率记为 $P(j | i, R)$, 有

$$P(j | i, R) = \int_{R_j} f_i(\mathbf{x}) d\mathbf{x}$$

如果已知样品 X 来自总体 G_i 的先验概率为 q_i , $i = 1, 2, \dots, k$, 则在规则 R 下, 由 (4.12) 式知, 误判的总平均损失为

$$\begin{aligned} g(R) &= \sum_{i=1}^k q_i \sum_{j=1}^k C(j|i) P(j|i, R) \\ &= \sum_{i=1}^k q_i \sum_{j=1}^k C(j|i) \int_{R_j} f_i(\mathbf{x}) d\mathbf{x} \\ &= \sum_{j=1}^k \int_{R_j} \left(\sum_{i=1}^k q_i C(j|i) f_i(\mathbf{x}) \right) d\mathbf{x} \end{aligned} \quad (4.13)$$

令 $\sum_{i=1}^k q_i C(j|i) f_i(\mathbf{x}) = h_j(\mathbf{x})$, 那么 , (4.13) 式为

$$g(R) = \sum_{j=1}^k \int_{R_j} h_j(\mathbf{x}) d\mathbf{x}$$

如果空间 R^p 有另一种划分 $R^* = (R_1^*, R_2^*, \dots, R_k^*)$, 则它的总平均损失为

$$g(R^*) = \sum_{j=1}^k \int_{R_j^*} h_j(\mathbf{x}) d\mathbf{x}$$

那么, 在两种划分下的总平均损失之差为

$$g(R) - g(R^*) = \sum_{i=1}^k \sum_{j=1}^k \int_{R_i \cap R_j^*} [h_i(\mathbf{x}) - h_j(\mathbf{x})] d\mathbf{x} \quad (4.14)$$

由 R_i 的定义, 在 R_i 上 $h_i(\mathbf{x}) \leq h_j(\mathbf{x})$ 对一切 j 成立, 故(4.14)式小于或等于零, 这说明 $R_1, R_2, \dots, R_k$ 确能使总平均损失达到极小, 它是 Bayes 判别的解。

这样，我们以 Bayes 判别的思想得到的划分 $R = (R_1, R_2, \dots, R_k)$ 为

$$R_i = \{\mathbf{x} \mid h_i(\mathbf{x}) = \min_{1 \leq j \leq k} h_j(\mathbf{x})\} \quad i = 1, 2, \dots, k \quad (4.15)$$

具体说来，当抽取了一个未知总体的样本值 $\mathbf{X}$ ，要判断它属于哪个总体，只要前计算出 k 个按先验分布加权的误判平均损失

$$h_j(\mathbf{x}) = \sum_{i=1}^k q_i C(j \mid i) f_i(\mathbf{x}) \quad j = 1, 2, \dots, k \quad (4.16)$$

然后再比较这 k 个误判平均损失 $h_1(\mathbf{x}), h_2(\mathbf{x}), \dots, h_k(\mathbf{x})$ 的大小，选取其中最小的，则判定样品 $\mathbf{X}$ 来自该总体。

这里我们看一个特殊情形，当 $k = 2$ 时，由（4.16）式得

$$h_1(\mathbf{x}) = q_2 C(1|2) f_2(\mathbf{x}) \quad h_2(\mathbf{x}) = q_1 C(2|1) f_1(\mathbf{x})$$

从而

$$R_1 = \{\mathbf{x} \mid q_2 C(1|2) f_2(\mathbf{x}) \leq q_1 C(2|1) f_1(\mathbf{x})\}$$

$$R_2 = \{\mathbf{x} \mid q_2 C(1|2) f_2(\mathbf{x}) > q_1 C(2|1) f_1(\mathbf{x})\}$$

若令

$$V(\mathbf{x}) = \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}, \quad d = \frac{q_2 C(1|2)}{q_1 C(2|1)}$$

则判别规则可表示为

$$\begin{cases} \mathbf{x} \in G_1, & \text{当 } V(\mathbf{x}) \geq d \\ \mathbf{x} \in G_2, & \text{当 } V(\mathbf{x}) < d \end{cases} \quad (4.17)$$

如果在此, $f_1(\mathbf{x})$ 与 $f_2(\mathbf{x})$ 分别为 $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ 和 $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, 那么

$$\begin{aligned} V(\mathbf{x}) &= \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \\ &= \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) \right\} \\ &= \exp \left\{ [\mathbf{x} - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2]' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \right\} \\ &= \exp W(\mathbf{x}) \end{aligned}$$

其中 $W(\mathbf{x})$ 由 (4.5) 所定义。

于是, 判定样品 $\mathbf{X}$ 来自该总体时, 判别规则 (4.17) 成

$$\begin{cases} \mathbf{X} \in G_1, & \text{如果 } W(\mathbf{X}) \geq \ln d \\ \mathbf{X} \in G_2, & \text{如果 } W(\mathbf{X}) < \ln d \end{cases} \quad (4.18)$$

对比判别规则 (4.6)，唯一的差别仅在于阈值点，(4.6) 用 0 作为阈值点，而这里用 $\ln d$ 。当 $q_1 = q_2$ ， $C(1|2) = C(2|1)$ 时， $d = 1$ ， $\ln d = 0$ ，则 (4.6) 与 (4.18) 完全一致。

□ 例：在美国，有两家主要的专门研究证券优劣的机构。一是标准普尔公司，一是穆迪投资服务公司。他们根据发行证券的企业在信用方面可信赖程度的评价，把证券分别评定等级，通常把评为高级的证券(如标准普尔公司对股票评级为A+、A、A-和B+的股票)称为投资级证券，把评为低级的证券称为投机级证券，评级表每年重新评定一次，并予以公布，为证券投资者提供投资指导。证券评级是根据发行证券的企业的财务和信用状况评定的。

□ 假设评定股票所依据的指标主要有发行在外的股份数X1，季度盈利X2、市盈率X3、速功比率X4。并假设两类股票的这四个指标均服从四维正态分布，且有相同的协方差矩阵。每类股票的均值向量和共向协差阵列在下表中。

两类股票的各指标均值和共同协方差阵

指 标	均 值		协方差矩阵			
	投资级	投机级				
发行在外股份数 x_1	119.62	9.52	15724.313	16.585	167.831	26.887
季度盈利 x_2	0.3632	-0.0972	16.585	1.040	-1.970	0.227
市盈率 x_3	19.00000	19.09999	167.831	-1.970	172.393	-0.333
速动比率 x_4	1.5666	0.4846	26.887	0.227	-0.333	0.769

假设投资级股票和投机级股票出现的先验概率分别为 $2/3$ 和 $1/3$ ，试用贝叶斯判别准则建立判别函数。已知某种股票的四个指标值分别为 $x_1=43, x_2=0.17, x_3=19.06, x_4=1.20$ ，对该股票进行判别。

$$\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) = \begin{bmatrix} 0.00487409 \\ 0.09799348 \\ -0.00187398 \\ 1.20687200 \end{bmatrix}, \quad \bar{\mu} = \begin{bmatrix} 64.72 \\ 0.133 \\ 19.049995 \\ 1.0256 \end{bmatrix}$$

判别函数为：

$$\begin{aligned} \varphi(x) &= (x - \bar{\mu})' \Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \\ &= 0.0049(x_1 - 64.72) + 0.0980(x_2 - 0.133) \\ &\quad - 0.0019(x_3 - 19.05) + 1.2069(x_4 - 1.0256) \\ &= 0.0049x_1 + 0.0980x_2 - 0.0019x_3 \\ &\quad + 1.2069x_4 - 1.5358 \end{aligned}$$

判别函数值的临界值为：

$$\ln \frac{q_2}{q_1} = \ln \frac{1/3}{2/3} = \ln 0.5 = -0.6931$$

判别规则为：

$$\begin{cases} D_1 = \{x: \varphi(x) > -0.6931\}; \\ D_2 = \{x: \varphi(x) < -0.6931\}; \\ \text{待判}, \varphi(x) = -0.6931 \end{cases}$$

将需要进行判别的股票的指标值代入判别函数，得：

$$\begin{aligned} \varphi(x) &= 0.0049 \times 43 + 0.0980 \times 0.17 - 0.0019 \times 19.06 \\ &\quad + 1.2069 \times 1.20 - 1.5318 \\ &= 0.1076 \end{aligned}$$

因为 $\varphi(x) > -0.6931$ ，所以该股票应判为投资级股票。

注意：

- Bayes判别法优于距离判别法之处是它考虑了先验概率，然而，令人头疼的事也是先验概率。在实际应用中，要确定先验概率是很困难的事。
- 一般来说，在现场调查中，先验概率常取为各类型疾病的患病率，在医院的门诊诊断中，先验概率常取各种类型疾病的就诊比例；也有人将样本的各类样本含量之构成比作为先验概率的估计。
- 在无法确定时，可以取各类的先验概率相等，当然此时，Bayes判别失去了其优越性。

第四节 费歇（Fisher）判别法

- Fisher判别的基本思想
- Fisher判别函数的构造
- 线性判别函数的求法

-
- Fisher判别法是1936年提出来的，该方法的主要思想是通过将多维数据投影到某个方向上，投影的原则是将总体与总体之间尽可能的放开，然后再选择合适的判别规则，将新的样品进行分类判别。



一、Fisher判别的基本思想

从 k 个总体中抽取具有 p 个指标的样品观测数据，借助方差分析的思想构造一个线性判别函数

$$U(\mathbf{X}) = u_1 X_1 + u_2 X_2 + \cdots + u_p X_p = \mathbf{u}'\mathbf{X} \quad (4.19)$$

其中系数 $\mathbf{u} = (u_1, u_2, \cdots, u_p)'$ 确定的原则是使得总体之间区别最大，而使每个总体内部的离差最小。

有了线性判别函数后，对于一个新的样品，将它的 p 个指标值代入线性判别函数（4.19）式中求出 $U(\mathbf{X})$ 值，然后根据判别一定的规则，就可以判别新的样品属于哪个总体。

二、Fisher判别函数的构造

1、针对两个总体的情形

假设有两个总体 G_1, G_2 ，其均值分别为 μ_1 和 μ_2 ，协方差矩阵为 Σ_1 和 Σ_2 。当 $\mathbf{X} \in G_i$ 时，我们可以求出 $\mathbf{u}'\mathbf{X}$ 的均值和方差，即

$$E(\mathbf{u}'\mathbf{X}) = E(\mathbf{u}'\mathbf{X} | G_i) = \mathbf{u}'E(\mathbf{X} | G_i) = \mathbf{u}'\mu_i \triangleq \bar{\mu}_i, \quad i = 1, 2$$

$$D(\mathbf{u}'\mathbf{X}) = D(\mathbf{u}'\mathbf{X} | G_i) = \mathbf{u}'D(\mathbf{X} | G_i)\mathbf{u} = \mathbf{u}'\Sigma_i\mathbf{u} \triangleq \sigma_i^2$$

在求线性判别函数时，尽量使得总体之间差异大，也就是要求 $\mathbf{u}'\boldsymbol{\mu}_1 - \mathbf{u}'\boldsymbol{\mu}_2$ 尽可能的大，即 $\bar{\mu}_1 - \bar{\mu}_2$ 变大；同时要求每一个总体内的离差平方和最小，即 $\sigma_1^2 + \sigma_2^2$ ，则我们可以建立一个目标函数

$$\Phi(\mathbf{u}) = \frac{(\bar{\mu}_1 - \bar{\mu}_2)}{\sigma_1^2 + \sigma_2^2} \quad (4.20)$$

这样，我们就将问题转化为，寻找 $\mathbf{u}$ 使得目标函数 $\Phi(\mathbf{u})$ 达到最大。从而可以构造出所要求的线性判别函数。

2、针对多个总体的情形

假设有 k 个总体 $G_1, G_2, \dots, G_k$ ，其均值和协方差矩阵分别为 μ_i 和 Σ_i (> 0) ($i = 1, 2, \dots, k$)。同样，我们考虑线性判别函数 $\mathbf{u}'\mathbf{X}$ ，在 $\mathbf{X} \in G_i$ 的条件下，有

$$E(\mathbf{u}'\mathbf{X}) = E(\mathbf{u}'\mathbf{X} | G_i) = \mathbf{u}'E(\mathbf{X} | G_i) = \mathbf{u}'\mu_i \quad i = 1, 2, \dots, k$$

$$D(\mathbf{u}'\mathbf{X}) = D(\mathbf{u}'\mathbf{X} | G_i) = \mathbf{u}'D(\mathbf{X} | G_i)\mathbf{u} = \mathbf{u}'\Sigma_i\mathbf{u} \quad i = 1, 2, \dots, k$$

令

$$b = \sum_{i=1}^k (\mathbf{u}'\boldsymbol{\mu}_i - \mathbf{u}'\bar{\boldsymbol{\mu}})^2$$

$$e = \sum_{i=1}^k \mathbf{u}'\boldsymbol{\Sigma}_i\mathbf{u} = \mathbf{u}'\left(\sum_{i=1}^k \boldsymbol{\Sigma}_i\right)\mathbf{u} = \mathbf{u}'\mathbf{E}\mathbf{u}$$

$$\text{其中 } \bar{\boldsymbol{\mu}} = \frac{1}{k} \sum_{i=1}^k \boldsymbol{\mu}_i, \quad \mathbf{E} = \sum_{i=1}^k \boldsymbol{\Sigma}_i。$$

这里 b 相当于一元方差分析中的组间差， e 相当于组内差，应用方差分析的思想，选择 $\mathbf{u}$ 使得目标函数

$$\Phi(\mathbf{u}) = \frac{b}{e} \quad (4.21)$$

达到极大。

这里我们应该说明的是，如果我们得到线性判别函数 $\mathbf{u}'\mathbf{X}$ ，对于一个新的样品 $\mathbf{X}$ 可以这样构造一个判别规则，如果

$$\left| \mathbf{u}'\mathbf{X} - \mathbf{u}'\boldsymbol{\mu}_j \right| = \min_{1 \leq i \leq k} \left| \mathbf{u}'\mathbf{X} - \mathbf{u}'\boldsymbol{\mu}_i \right| \quad (4.22)$$

则判定 $\mathbf{X}$ 来自总体 G_j 。

三、线性判别函数的求法

针对多个总体的情形，我们讨论使目标函数（4.21）式达到极大的

求法。设 $\mathbf{X}$ 为 p 维空间的样品，那么 $\bar{\boldsymbol{\mu}} = \frac{1}{k} \sum_{i=1}^k \boldsymbol{\mu}_i = \frac{1}{k} \mathbf{M}' \mathbf{1}$

其中

$$\mathbf{M} = \begin{pmatrix} \mu_{11} & \mu_{21} & \cdots & \mu_{p1} \\ \mu_{12} & \mu_{22} & \cdots & \mu_{p2} \\ \cdots & \cdots & \cdots & \cdots \\ \mu_{1k} & \mu_{2k} & \cdots & \mu_{pk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}'_1 \\ \boldsymbol{\mu}'_2 \\ \cdots \\ \boldsymbol{\mu}'_k \end{pmatrix} \quad \mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix}$$

注意到 $\mathbf{M}' \mathbf{M} = (\boldsymbol{\mu}_1 \quad \boldsymbol{\mu}_2 \quad \cdots \quad \boldsymbol{\mu}_k) \begin{pmatrix} \boldsymbol{\mu}'_1 \\ \boldsymbol{\mu}'_2 \\ \cdots \\ \boldsymbol{\mu}'_k \end{pmatrix} = \sum_{i=1}^k \boldsymbol{\mu}_i \boldsymbol{\mu}'_i$

$$\begin{aligned}
\text{从而} \quad b &= \sum_{i=1}^k (\mathbf{u}'\boldsymbol{\mu}_i - \mathbf{u}'\bar{\boldsymbol{\mu}})^2 \\
&= \mathbf{u}' \sum_{i=1}^k (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})' \mathbf{u} \\
&= \mathbf{u}' \left[\sum_{i=1}^k \boldsymbol{\mu}_i \boldsymbol{\mu}_i' - k \bar{\boldsymbol{\mu}} \bar{\boldsymbol{\mu}}' \right] \mathbf{u} \\
&= \mathbf{u}' (\mathbf{M}'\mathbf{M} - \frac{1}{k} \mathbf{M}'\mathbf{1}\mathbf{1}'\mathbf{M}) \mathbf{u} \\
&= \mathbf{u}' \mathbf{M}' (\mathbf{I} - \frac{1}{k} \mathbf{J}) \mathbf{M} \mathbf{u} \\
&= \mathbf{u}' \mathbf{B} \mathbf{u}
\end{aligned}$$

这里, $\mathbf{B} = \mathbf{M}'(\mathbf{I} - \frac{1}{k} \mathbf{J})\mathbf{M}$, $\mathbf{I}_{p \times p}$ 为 $p \times p$ 的单位阵, $\mathbf{J} = \begin{pmatrix} 1 & \cdots & 1 \\ & \ddots & \\ 1 & \cdots & 1 \end{pmatrix}$ 。

即有 $\Phi(\mathbf{u}) = \frac{\mathbf{u}'\mathbf{B}\mathbf{u}}{\mathbf{u}'\mathbf{E}\mathbf{u}}$ (4.23) 求使得 (4.23) 式达到极大的 $\mathbf{u}$ 。

为了确保解的唯一性，不妨设 $\mathbf{u}'\mathbf{E}\mathbf{u} = 1$ ，这样问题转化为，在 $\mathbf{u}'\mathbf{E}\mathbf{u} = 1$ 的条件下，求 $\mathbf{u}$ 使得 $\mathbf{u}'\mathbf{B}\mathbf{u}$ 式达到极大。

考虑目标函数

$$\varphi(\mathbf{u}) = \mathbf{u}'\mathbf{B}\mathbf{u} - \lambda(\mathbf{u}'\mathbf{E}\mathbf{u} - 1) \quad (4.24)$$

$$\varphi(\mathbf{u}) = \mathbf{u}'\mathbf{B}\mathbf{u} - \lambda(\mathbf{u}'\mathbf{E}\mathbf{u} - 1) \quad (4.24)$$

对 (4.24) 式求导, 有

$$\begin{cases} \frac{\partial \varphi}{\partial \mathbf{u}} = 2(\mathbf{B} - \lambda\mathbf{E})\mathbf{u} = 0 \end{cases} \quad (4.25)$$

$$\begin{cases} \frac{\partial \varphi}{\partial \lambda} = \mathbf{u}'\mathbf{E}\mathbf{u} - 1 = 0 \end{cases} \quad (4.26)$$

对 (4.25) 式两边同乘 $\mathbf{u}'$, 有 $\mathbf{u}'\mathbf{B}\mathbf{u} = \lambda\mathbf{u}'\mathbf{E}\mathbf{u} = \lambda$

从而, $\mathbf{u}'\mathbf{B}\mathbf{u}$ 的极大值为 λ 。再用 $\mathbf{E}^{-1}$ 左乘 (4.25) 式, 有

$$(\mathbf{E}^{-1}\mathbf{B} - \lambda\mathbf{I})\mathbf{u} = 0 \quad (4.27)$$

由 (4.27) 式说明 λ 为 $\mathbf{E}^{-1}\mathbf{B}$ 特征值, $\mathbf{u}$ 为 $\mathbf{E}^{-1}\mathbf{B}$ 的特征向量。在此最大特征值所对应的特征向量 $\mathbf{u} = (u_1, u_2, \dots, u_p)'$ 为我们所求结果。

□ 这里值得注意的是，本书有几处利用极值原理求极值时，只给出了必要条件的数学推导，而有关充分条件的论证省略了，因为在实际问题中，往往根据问题本身的性质就能肯定有最大值（或最小值），如果所求的驻点只有一个，这时就不需要根据极值存在的充分条件判定它是极大还是极小而就能肯定这唯一的驻点就是所求的最大值（或最小值）。为了避免用较多的数学知识或数学上的推导，这里不追求数学上的完整性。

在解决实际问题时，当总体参数未知，需要通过样本来估计，我们仅对 $k=2$ 的情形加以说明。设样本分别为 $\mathbf{X}_1^{(1)}, \mathbf{X}_2^{(1)}, \dots, \mathbf{X}_{n_1}^{(1)}$ 和 $\mathbf{X}_1^{(2)}, \mathbf{X}_2^{(2)}, \dots, \mathbf{X}_{n_2}^{(2)}$ ，则

$$\bar{\mathbf{X}} = \frac{n_1 \bar{\mathbf{X}}^{(1)} + n_2 \bar{\mathbf{X}}^{(2)}}{n_1 + n_2}$$

$$\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}} = \frac{n_2}{n_1 + n_2} (\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)})$$

$$\bar{\mathbf{X}}^{(2)} - \bar{\mathbf{X}} = \frac{n_1}{n_1 + n_2} (\bar{\mathbf{X}}^{(2)} - \bar{\mathbf{X}}^{(1)})$$

那么

$$\begin{aligned} \hat{\mathbf{B}} &= n_1 (\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}})(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}})' + n_2 (\bar{\mathbf{X}}^{(2)} - \bar{\mathbf{X}})(\bar{\mathbf{X}}^{(2)} - \bar{\mathbf{X}})' \\ &= \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)})(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)})' \end{aligned}$$

当 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_k$ 和 $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \dots, \boldsymbol{\Sigma}_k$ 均未知时， $\boldsymbol{\mu}_\alpha$ ($\alpha = 1, 2, \dots, k$) 的估计同前， $\boldsymbol{\Sigma}_\alpha$ ($\alpha = 1, 2, \dots, k$) 的估计为

$$\hat{\boldsymbol{\Sigma}}_\alpha = \frac{1}{n_\alpha - 1} \mathbf{S}_\alpha, \quad \alpha = 1, 2, \dots, k$$

疑问

□ 问题答疑：<http://www.xxwenda.com/>

■ 可邀请老师或者其他人回答问题

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：小象
- 新浪微博：ChinaHadoop

