应用统计II 第6讲 判别分析 (Discriminant analysis)

Instructor: 郝壮

haozhuang@buaa.edu.cn
School of Economics and Management
Beihang University

May 28, 2022

判别分析 (Discriminant analysis)

- 教材:第十二章
- 判别分析(Discriminant analysis)在机器学习中应用广泛. 用于对个体进行分类, 从而基于分类进行决策.

一. 判别分析概述

判别分析的工作目的:

- 已知某客观事物按照某种标准(如p个指标)可分为k个总体$G_1, G_2, \ldots, G_k$
- 根据已掌握的各个总体的样本信息, 总结事物分类的规律.
- 建立合理有效的判别规则.
- 对于未知类别的个体, 根据它的p个指标判断其属于哪个总体(类别).

例如:

- 根据病人的诸项检验指标, 进行疾病诊断.
- 根据已有的气象资料来进行气象预报.
- 根据心理测试问题, 判断受试者的基本心理特征, 如PHQ-9测分后对depression进行判别.
- 垃圾邮件

判别分析案例: 根据个人信用资料, 做违约风险评估

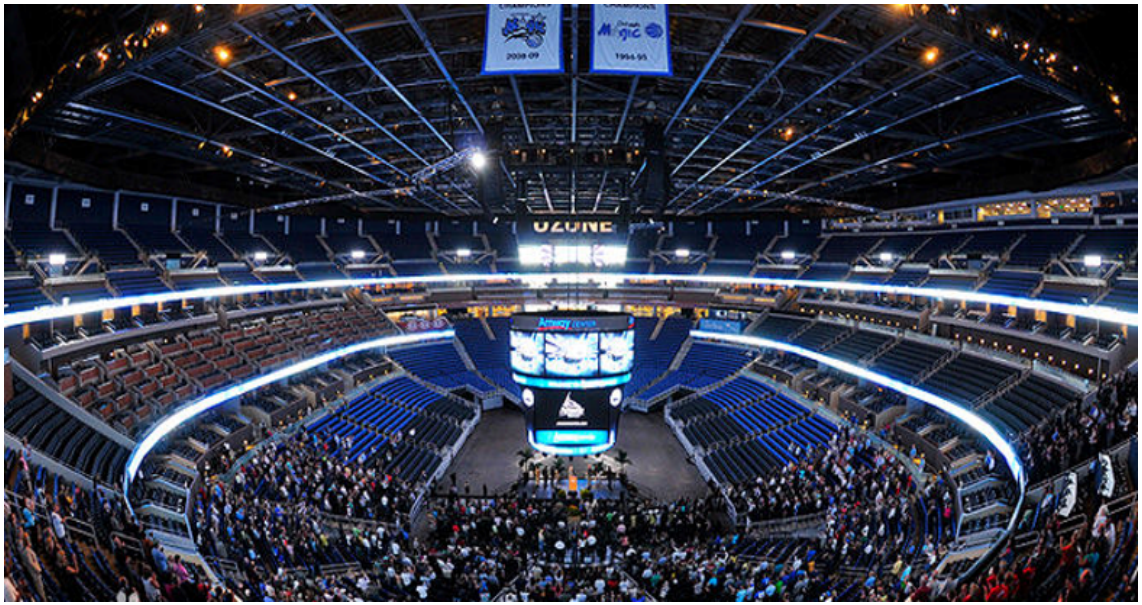
数据:

- 个人基本资料 (性别, 年龄, 学历, 婚姻)
- 实物资产(房, 车), 收入 (工资, 股票, 配偶收入)
- 社交资产 (微信好友个数, 电话本好友数, QQ好友数, 等等)
- 贷款模式 (贷款总额, 还款年限, 月付)
- 违约记录

目的:

- (1)识别: 人群类型 + 贷款模式 $\Rightarrow$ 违约风险
- (2)向客户建议贷款总额和贷款模式 (还款年限, 月付)

Dota2国际邀请赛(2014年)



判别分析: 战斗行为变量

项目	观测标准
单杀次数	观测整盘游戏中每一次英雄死亡的原因，如果其伤害完全来自敌方的一个英雄，则视为单杀，单杀次数+1。
团杀次数	观测整盘游戏中每一次英雄死亡的原因，如果其伤害来自敌方的二个及以上英雄，则视为团杀，团杀次数+1。
发起战斗次数	观测整盘游戏中每次战斗由哪一方战队的英雄先发出攻击，则对应的战队发起战斗次数+1。
开雾次数	观测整盘游戏中双方战队开雾次数。
视野眼个数	观测整盘游戏中双方战队插的视野眼个数。在小地图观察，颜色代表阵营，实心点代表视野眼。
反隐眼个数	观测整盘游戏中双方战队插的反隐眼个数。在小地图观察，颜色代表阵营，虚心点代表视野眼。
控符次数	每两分钟观察一次上下路河道，查看符文由哪方战队控制，则对应战队控符次数+1。
正补	最后一刀杀死敌方小兵，则所在战队正补次数+1。
反补	最后一刀杀死己方小兵，则所在战队反补次数+1。

原始数据

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	单杀次数	团杀次数	单杀次数	发起战斗沙	开雾次数	视野眼个数	反隐眼个数	控符次数	正补	反补	国籍（1为	输赢（1为赢）	
2	1	8	1	3	1	5	5	4	243	60	1	1	
3	0	20	0	9	6	15	9	5	797	28	1	0	
4	1	5	1	2	4	7	3	5	245	29	1	1	
5	2	23	2	9	3	12	8	9	764	30	1	1	
6	0	25	0	9	1	9	19	8	775	37	1	0	
7	6	24	6	11	1	10	4	10	744	34	1	1	
8	3	21	3	8	4	16	22	10	922	31	1	1	
9	1	24	1	12	5	12	1	8	692	28	1	1	
10	2	9	2	12	5	16	9	18	1366	59	1	0	
11	0	14	0	12	5	12	14	9	680	29	1	0	
12	7	20	7	14	6	13	13	11	1979	54	1	1	
13	1	8	1	5	1	4	6	4	173	39	1	1	
14	2	11	2	10	3	6	3	6	612	36	1	0	
15	6	22	6	9	2	7	9	10	508	34	1	1	
16	4	15	4	12	4	6	11	5	703	31	1	0	
17	5	8	5	7	1	6	5	5	396	38	1	1	
18	6	10	6	6	2	5	8	10	460	26	1	1	
19	6	22	6	14	4	7	7	15	788	35	1	1	
20	7	18	7	11	3	9	9	13	875	60	1	1	
21	4	13	4	21	4	13	7	12	1065	57	1	0	
22	2	17	2	13	9	16	19	5	1828	62	1	0	
23	1	21	1	10	3	7	2	9	352	28	1	1	
24	2	9	2	7	4	11	14	10	545	53	1	0	
25	0	35	0	14	6	15	4	20	1508	76	1	1	

例题

有三种鸢尾花的花瓣, 花萼的长宽数据. 共收集了三种鸢尾花, 每种50个观测量, 共150个观测量的数据.
鸢尾花品种Spno:

- 1 刚毛鸢尾花 Setoaq
- 2 变色鸢尾花 Versicolor
- 3 佛吉尼亚鸢尾花 Virginica

鸢尾花特征:

- Slen: 花萼长 sepal length
- Swid: 花萼宽 sepal width
- Plen: 花瓣长 petal length
- Plen: 花瓣长 petal length

根据4个特征判断是哪类鸢尾花

No	Slen	Swid	Plan	Pwid	Spno
1	50	33	14	2	1
2	67	31	56	24	3
3	89	31	51	23	3
...	...	...	...	...	...
10	70	32	47	14	2
...	...	...	...	...	...
150	...	...	...	...	...

判别分析的输入数据为:

K个总体, p个指标:

$$
\begin{array}{c}
 \\
G_1 \\
\cdot \\
\cdot \\
\cdot \\
\cdot \\
G_K
\end{array}
\begin{bmatrix}
x_{11}^{(1)} & \cdots & x_{1p}^{(1)} \\
\cdots & \cdots & \cdots \\
x_{n_1 1}^{(1)} & \cdots & x_{n_1 p}^{(1)} \\
\cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots \\
x_{11}^{(k)} & \cdots & x_{1p}^{(k)} \\
\cdots & \cdots & \cdots \\
x_{n_k 1}^{(k)} & \cdots & x_{n_k p}^{(k)}
\end{bmatrix}
$$

有监督模型和无监督模型

思考: 判别和聚类最大的不同在哪里?

有监督模型和无监督模型

思考: 判别和聚类最大的不同在哪里?

判别分析是一个有监督模型: 判别是可以观察到目标变量(被解释变量), 机器学习的语言中被称为标签(label).

拓展知识: 机器学习中将模型分为有监督模型和无监督模型两大类, 区别在于有无目标变量.

- 有监督模型
  - 连续型目标变量: 回归
  - 离散型目标变量: 判别
- 无监督模型
  - 目标变量未知 (无目标): 聚类, PCA

本课主要介绍的两种判别法

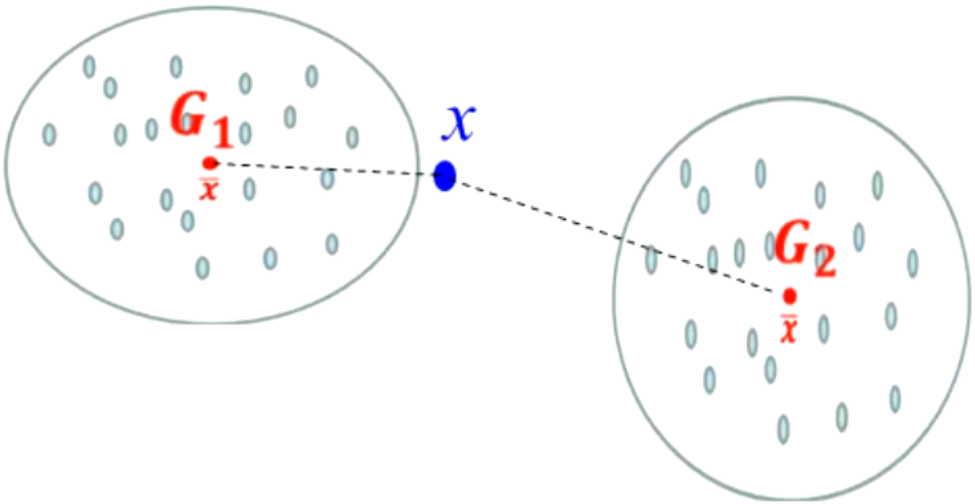
- 距离判别法
- Fisher 判别法 (线性判别法)

二. 距离判别法

原理: “离哪个类的距离最近, 就属于哪一类"

1. 判别规则

$$\begin{cases} x \in G_1, & \text{if } d(x, G_1) < d(x, G_2) \\ x \in G_2, & \text{if } d(x, G_2) < d(x, G_1) \\ \text{待判}, & \text{if } d(x, G_1) = d(x, G_2) \end{cases}$$



计算距离: 马哈拉诺比斯(Mahalanobis)距离

判别中用到的统计距离: Mahalanobis距离/马氏距离/马哈拉诺比斯距离

$$d^2(\boldsymbol{x}, G_i) = (\boldsymbol{x} - \boldsymbol{\mu}_i)' \Sigma_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)$$

其中

- μ_i: 第i类的均值 ($p \times 1$维)
- Σ_i^{-1}: 第i类方差协方差矩阵的逆 ($p \times p$ 维)

即标准化后距离 (可以消除量纲).

(在实际应用中, $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_k$ 和 $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \cdots, \boldsymbol{\Sigma}_k$ 一般都是未知的, 它们的值可由相应的样本估计值代替.)

判别分析: 两个总体情况

1. 构造判别函数(discriminant function)

$$
\begin{aligned}
W(x) &= d^2(x, G_2) - d^2(x, G_1) \\
&= (x-\mu_2)' \sum_2^{-1} (x-\mu_2) - (x-\mu_1)' \sum_1^{-1} (x-\mu_1)
\end{aligned}
$$

2. 根据 $W(x)$ 建立判别规则:

$$
\begin{cases}
x \in G_1, & W(x) > 0 \\
x \in G_2, & W(x) < 0 \\
x \text{ 待判}, & W(x) = 0
\end{cases}
$$

特殊情况: 若判别变量只有一个 ($p = 1$)

若仅有一个判别变量且可假设个总体方差相等

(a) $\sigma_1 = \sigma_2 = \sigma$ (σ^2 的估计量为: s^2). 可以推导出,

$$W(x) = \frac{(x - \bar{\mu})}{s^2}\left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)$$

其中,

$$\bar{\mu} = \frac{\bar{x}^{(1)} + \bar{x}^{(2)}}{2}$$

判别规则非常简单, 只需考察$\bar{\mu}$. 当 $\bar{x}^{(1)} > \bar{x}^{(2)}$ 时 :

$$\begin{cases} x \in G_1, & x > \bar{\mu} \\ x \in G_2, & x < \bar{\mu} \\ x \text{ 待判}, & x = \bar{\mu} \end{cases}$$

特殊情况: 若判别变量只有一个 ($p = 1$)

若仅有一个判别变量但不能假设个总体方差相等

(b) $\sigma_1 \neq \sigma_2$.

σ_1 的估计量为: s_1

σ_2 的估计量为: s_2

可以推导出, 判别阈值为:

$$\mu^* = \frac{s_1 \bar{x}^{(2)} + s_2 \bar{x}^{(1)}}{s_1 + s_2}$$

判别规则($\mu_1 > \mu_2$)

$$\begin{cases} x \in G_1, & x > \mu^* \\ x \in G_2, & x < \mu^* \\ x \text{ 待判}, & x = \mu^* \end{cases}$$

* 样本标准差越小, 数据越集中, 阈值更像该方向倾斜

多个总体的距离判别法 (p维空间)

设有 k 个总体 $\mathrm{G}_1, \mathrm{G}_2, \ldots\ldots \mathrm{G}_\mathrm{k}$, 他们的均值协方差矩阵分别为

$$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_k$$
$$\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \cdots, \boldsymbol{\Sigma}_k$$

现对于任意新的样本点$\forall x \in R^p$, 要判别 x 属于哪个总体G_j?

1. 计算 $d^2(x, G_j)$
2. 选取 $G_i : d^2(x, G_i) = \min_{j=1,2,\ldots,k} \left\{ d^2(x, G_j) \right\}$

(在实际应用中, $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_k$ 和 $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \cdots, \boldsymbol{\Sigma}_k$ 一般都是未知的, 它们的值可由相应的样本估计值代替.)

- 特别地, 当 $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k = \Sigma$ 时:

$$\begin{aligned} d^2(x, G_j) &= (x - \mu_j)^T \Sigma^{-1} (x - \mu_j) \\ &= x^T \Sigma^{-1} x - 2\left[x^T \Sigma^{-1} \mu_j - \tfrac{1}{2} \mu_j^T \Sigma^{-1} \mu_j\right] \end{aligned}$$

- 将中括号中线性函数定义为 $f(x) = x^T \Sigma^{-1} \mu_j - \frac{1}{2} \mu_j^T \Sigma^{-1} \mu_j$ (**分类函数,classification functions**), 则

$$\begin{aligned} d^2(x, G_1) &= x^T \Sigma^{-1} x - 2 f_1(x) \\ d^2(x, G_2) &= x^T \Sigma^{-1} x - 2 f_2(x) \\ &\vdots \\ d^2(x, G_K) &= x^T \Sigma^{-1} x - 2 f_K(x) \end{aligned}$$

- 所以我们有如下等价表达:

$$\begin{aligned} x \in G_i &\Leftrightarrow d^2(x, G_i) = \min_{j=1,2,\ldots k} \left\{d^2(x, G_j)\right\} \\ &\Leftrightarrow \max_{j=1,2,\ldots,k} \left\{f_j(x)\right\} \end{aligned}$$

多总体判别分析例题: 鸢尾花

有三种鸢尾花的花瓣, 花萼的长宽数据. 共收集了三种鸢尾花, 每种50个观测, 共150个观测的数据.
鸢尾花品种Spno:

1. 刚毛鸢尾花 Setoaq
2. 变色鸢尾花 Versicolor
3. 佛吉尼亚鸢尾花 Virginica

鸢尾花特征:

- Slen: 花萼长 sepal length
- Swid: 花萼宽 sepal width
- Plen: 花瓣长 petal length
- Plen: 花瓣长 petal length

根据4个特征判断是哪类鸢尾花

No	Slen	Swid	Plan	Pwid	Spno
1	50	33	14	2	1
2	67	31	56	24	3
3	89	31	51	23	3
...	...	...	...	...	...
10	70	32	47	14	2
...	...	...	...	...	...
150	...	...	...	...	...

(假设可以认为各类鸢尾花总体方差协方差矩阵相同)
经计算, 得到三个线性**分类函数** $f_1(x), f_2(x), f_3(x)$:

1. 刚毛鸢尾花 $f_1(x) = 2.309*Slen+2.380*Swid-1.584Plen-1.792Pwid-85.892$
2. 变色鸢尾花 $f_2(x) = 1.553 * Slen + 0.709 * Swid + 0.551 * Plan + 0.629 * Pwid - 73.000$
3. 佛吉尼亚鸢尾花 $f_3(x) = 1.238 * Slen + 0.364 * Swid + 1.301 * Plan + 2.107 * Pwid - 104.738$

如果有一个新的样本点: $Slen = 50, Swid = 33, Plan = 14, Pwid = 2$ 则 $f_1 = 82.3172, f_2 = 37.0909, f_3 = -8.39790$

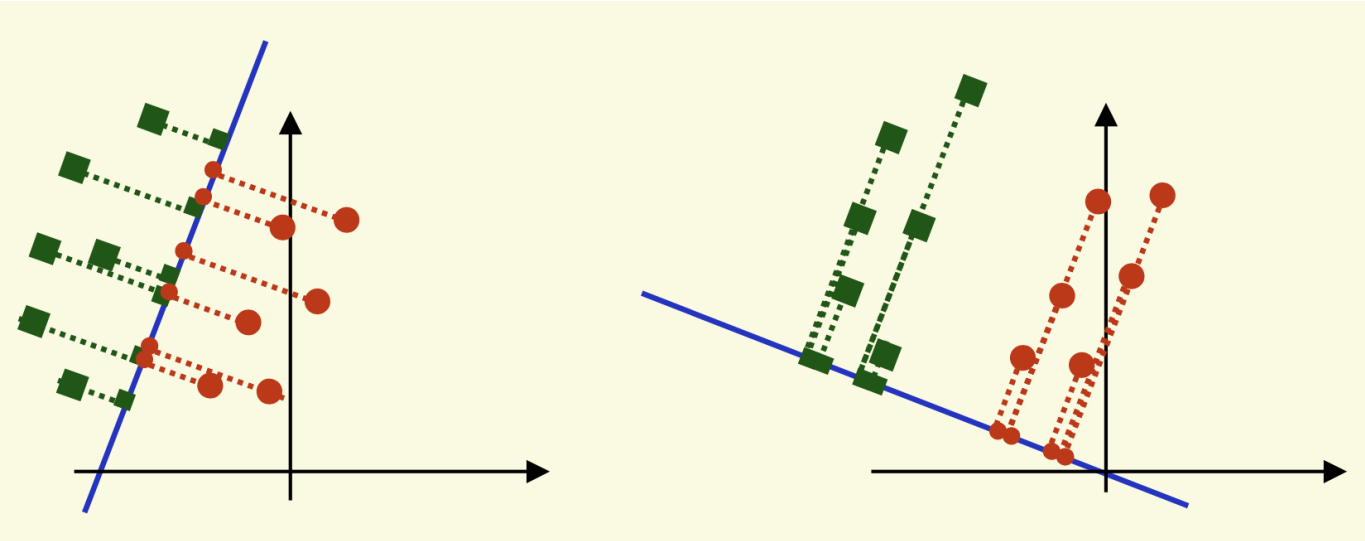
所以:判断该样本点是属于刚毛鸢尾花.

三. Fisher 判别法

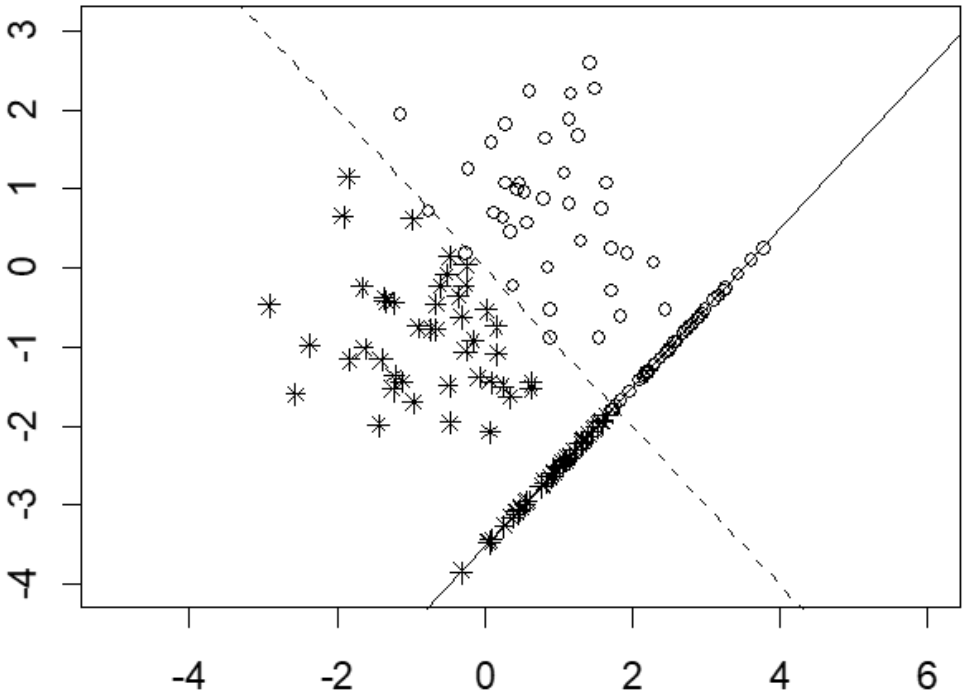
更常用的判别法是Fisher 判别法也称 Fisher's linear discriminant analysis (LDA)

1. 基本思想方法

将 K 组的P 维数据投影在某一个方向, 对投影点来说能使得组与组之间尽可能地分开.



(注意与PCA的不同. PCA: 通过平移 + 旋转省去数据变异不大方向的信息)



假定只有两种类型的训练样本(“o"和“*"). 按照原来的变量(横坐标和纵坐标), 很难将这两种点分开.
寻找一个方向进行投影, 这些点在该直线上的投影形成一维空间点的集合; 可以很容易分开; 而如果向其他方向投影, 判别效果不会更好.

2. 方法推导

分别从k个总体分别取得 p维的样本观测值

$$G_1 : \boldsymbol{x}_1^{(1)} \ldots\ldots\ldots \boldsymbol{x}_{n_1}^{(1)} \quad \bar{\boldsymbol{x}}^{(1)} = \frac{1}{n_1}\sum_{i=1}^{n_1} \boldsymbol{x}_i^{(1)} \text{第1类平均值}$$

$$\ldots \quad \ldots \quad \ldots$$

$$G_k : \boldsymbol{x}_1^{(k)} \ldots\ldots\ldots \boldsymbol{x}_{n_k}^{(k)} \quad \bar{\boldsymbol{x}}^{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k} \boldsymbol{x}_i^{(k)} \text{第k类平均值}$$

其中,

$$\mathbf{x}_i^{(\alpha)} \in R^p$$

表示第α类的观察第i个观察. 这个观察有p维, 表示p个变量.
定义总平均值:

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{\alpha=1}^{k}\sum_{i=1}^{n_\alpha} \boldsymbol{x}_i^{(\alpha)}$$

其中, $n = n_1 + n_2 + \ldots + n_k$

- 令: $\boldsymbol{a} \in R^p$, 则 $\mu(X) = a^T\boldsymbol{x}$ 为$\boldsymbol{x}$在a方向的投影.
  上述投影数据为:

$$
\begin{aligned}
G_1 &: \boldsymbol{a}^T\boldsymbol{x}_1^{(1)} \ldots\ldots\ldots \boldsymbol{a}^T\boldsymbol{x}_{n_1}^{(1)} \\
&\ldots \quad \ldots \quad \ldots \\
G_k &: \boldsymbol{a}^T\boldsymbol{x}_1^{(k)} \ldots\ldots\ldots \boldsymbol{a}^T\boldsymbol{x}_{n_k}^{(k)}
\end{aligned}
$$

- 计算各类投影平均值 (投影平均值等于平均值的投影).

$$
\begin{aligned}
\boldsymbol{a}^T\bar{\boldsymbol{x}}^{(1)} &= \frac{1}{n_1}\sum_{i=1}^{n_1}\boldsymbol{a}^T\boldsymbol{x}_i^{(1)} \\
&\ldots \quad \ldots \quad \ldots \\
\boldsymbol{a}^T\bar{\boldsymbol{x}}^{(k)} &= \frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{a}^T\boldsymbol{x}_i^{(k)}
\end{aligned}
$$

- 另外, 总投影均值 $\boldsymbol{a}^T\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{\alpha=1}^{k}\sum_{i=1}^{n_\alpha}\boldsymbol{a}^T\boldsymbol{x}_i^{(\alpha)}$.

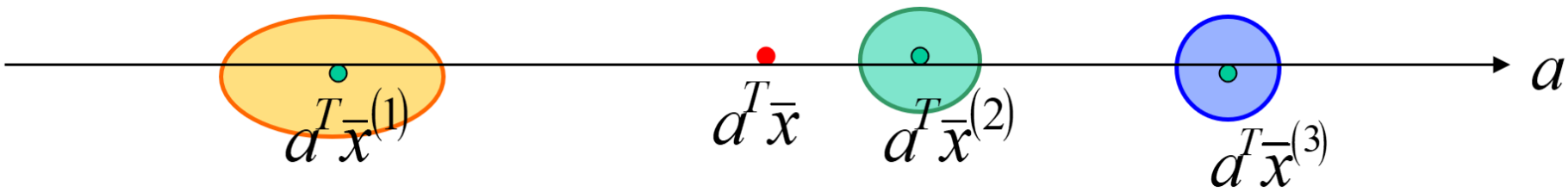
注意: p维数据投影以后的数据都是标量 (scalar)

最好的分类结果是: 使得类与类之间的分辨度尽可能大, 而类内的点尽可能聚合.

我们的问题转化为: 想寻找$a \in \boldsymbol{R}^p$, 使投影数据的组间离差平方和SSG大, 组内离差平方和SSE小.

$$\text{组间离差平方和 }: SSG = \sum_{\alpha=1}^{k} n_\alpha \left(\boldsymbol{a}^T \bar{\boldsymbol{x}}^{(\alpha)} - \boldsymbol{a}^T \bar{\boldsymbol{x}}\right)^2$$

$$\text{组内离差平方和 }: SSE = \sum_{\alpha=1}^{k} \sum_{i=1}^{n_\alpha} \left(\boldsymbol{a}^T \boldsymbol{x}_i^{(\alpha)} - \boldsymbol{a}^T \bar{\boldsymbol{x}}^{(\alpha)}\right)^2$$



组内离差

$$
\begin{aligned}
SSE &= \sum_{\alpha=1}^{k}\sum_{i=1}^{n_\alpha}\left(\boldsymbol{a}^T\boldsymbol{x}_i^{(\alpha)} - \boldsymbol{a}^T\bar{\boldsymbol{x}}^{(\alpha)}\right)^2 \\
&= \sum_{\alpha=1}^{k}\sum_{i=1}^{n_\alpha}\left[\boldsymbol{a}^T\left(\boldsymbol{x}_i^{(\alpha)} - \bar{\boldsymbol{x}}^{(\alpha)}\right)\right]^2 \\
&= \sum_{\alpha=1}^{k}\sum_{i=1}^{n_\alpha}\boldsymbol{a}^T\left(\boldsymbol{x}_i^{(\alpha)} - \bar{\boldsymbol{x}}^{(\alpha)}\right)\left(\boldsymbol{x}_i^{(\alpha)} - \bar{\boldsymbol{x}}^{(\alpha)}\right)^T\boldsymbol{a} \\
&= \boldsymbol{a}^T\left[\sum_{\alpha=1}^{k}\sum_{i=1}^{n_\alpha}\left(\boldsymbol{x}_i^{(\alpha)} - \bar{\boldsymbol{x}}^{(\alpha)}\right)\left(\boldsymbol{x}_i^{(\alpha)} - \bar{\boldsymbol{x}}^{(\alpha)}\right)^T\right]\boldsymbol{a} \\
&= \boldsymbol{a}^T E\boldsymbol{a}, \quad E\text{代表中括号所有项}
\end{aligned}
$$

组间离差

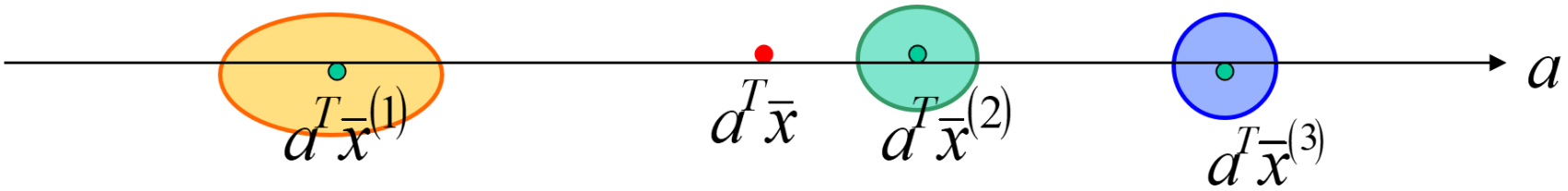
$$
\begin{aligned}
SSG &= \sum_{\alpha=1}^{k} n_\alpha \left(\boldsymbol{a}^T \bar{\boldsymbol{x}}^{(\alpha)} - \boldsymbol{a}^T \bar{\boldsymbol{x}}\right)^2 \\
&= \sum_{\alpha=1}^{k} n_\alpha \left[\boldsymbol{a}^T \left(\bar{\boldsymbol{x}}^{(\alpha)} - \bar{\boldsymbol{x}}\right)\right]^2 \\
&= \sum_{\alpha=1}^{k} n_\alpha \boldsymbol{a}^T \left(\bar{\boldsymbol{x}}^{(\alpha)} - \bar{\boldsymbol{x}}\right) \left(\bar{\boldsymbol{x}}^{(\alpha)} - \bar{\boldsymbol{x}}\right)^T \boldsymbol{a} \\
&= \boldsymbol{a}^T \left[\sum_{\alpha=1}^{k} n_\alpha \left(\bar{\boldsymbol{x}}^{(\alpha)} - \bar{\boldsymbol{x}}\right) \left(\bar{\boldsymbol{x}}^{(\alpha)} - \bar{\boldsymbol{x}}\right)^T\right] \boldsymbol{a} \\
&= \boldsymbol{a}^T B \boldsymbol{a}, \quad B\text{代表中括号所有项}
\end{aligned}
$$

希望组与组之间尽可能分开:

$$SSG \rightarrow \text{大}$$
$$SSE \rightarrow \text{小}$$

相当于求$\boldsymbol{a}$使得比值最大

$$\Delta(\boldsymbol{a}) = \frac{\boldsymbol{a}^T B \boldsymbol{a}}{\boldsymbol{a}^T E \boldsymbol{a}} \rightarrow \max$$



上述最大化问题的解可由如下方法给出 (证明超纲, 见拓展内容):

对于矩阵

$$E^{-1}B$$

设 $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_r > 0$ 是矩阵$E^{-1}B$的全部非零特征根 $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_r$ 是对应的特征向量.
则有:

$$\boldsymbol{a} = \boldsymbol{l}_1$$

(判别效率最高的方向)

$$\lambda_1$$

为该方向的最大判别能力.

Fisher判别方法: 求$E^{-1}B$的最大特征根λ_1和所对应的特征向量$\boldsymbol{l}_1$.

拓展内容：Fisher and Kernel Fisher Discriminant Analysis: Tutorial

https://arxiv.org/abs/1906.09436

Fisher判别函数 (discriminant function)

- 对于一个新的p维样本点$\boldsymbol{x}$, 计算其在$\boldsymbol{l}_1$方向的投影:
  判别函数, discriminant function:

$$\mu(\boldsymbol{x}) = \boldsymbol{l}_1^T \boldsymbol{x}$$

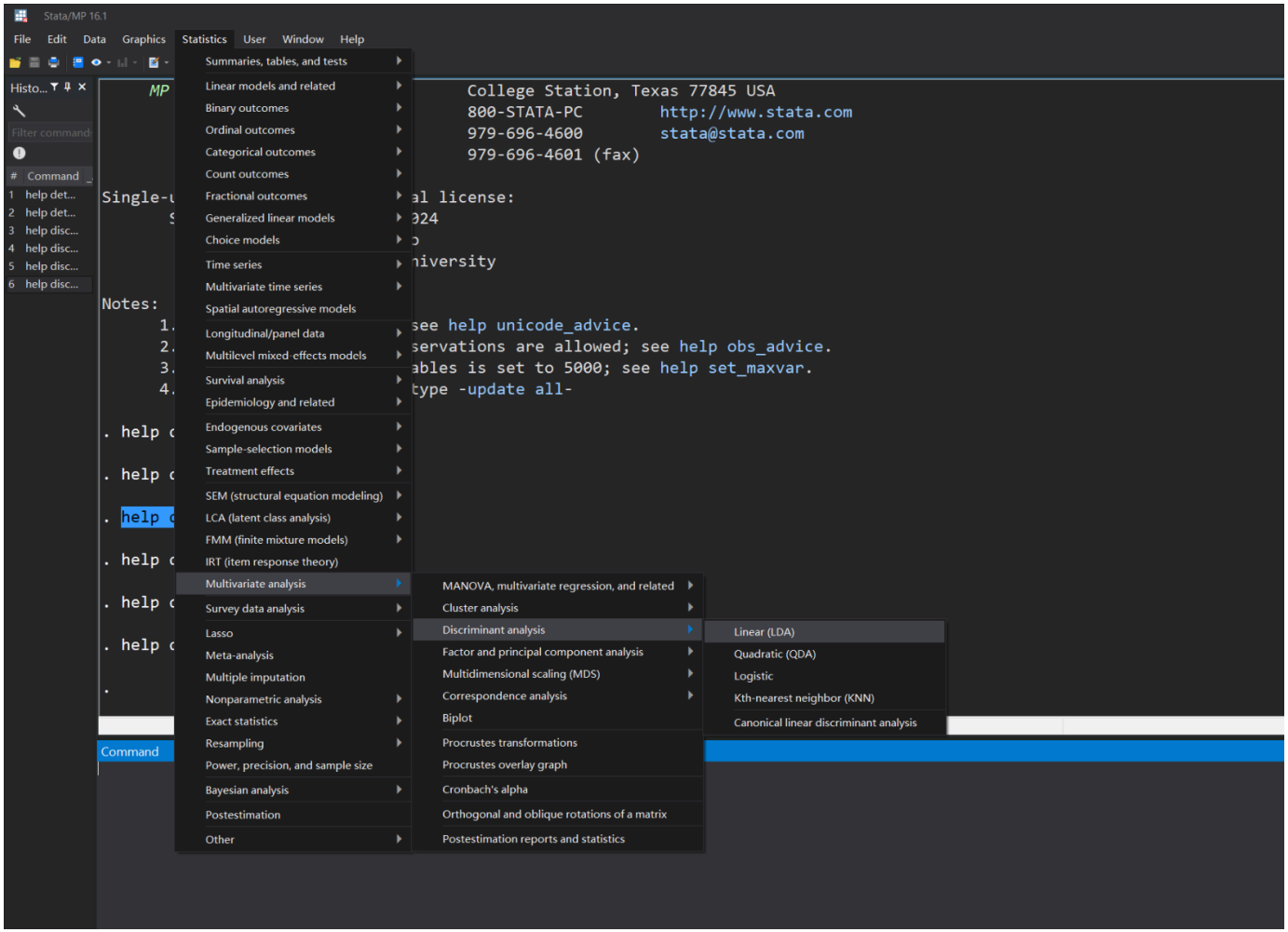
  将p维$\boldsymbol{x}$降到1维$\mu(\boldsymbol{x})$.
- 接下来, 用距离判别法判别新的观察x属于哪一类.
- **判别效率:**

$$Q_m = \frac{\lambda_1}{\sum_{h=1}^{r} \lambda_h}$$

- 若使用一维判别函数判别效率太低, 可采用m个线性判别函数$\boldsymbol{l}_1, \ldots, \boldsymbol{l}_m$.
- **判别方法:** 先将所有样本点在$\boldsymbol{l}_1, \ldots, \boldsymbol{l}_m$方向上投影 ($p$维降到$m$维), 然后按 $p > 1$的情形使用距离判别法. (在此$\boldsymbol{l}_1, \ldots, \boldsymbol{l}_m$不要求正交, 这是LDA和PCA的区别)
- **判别效率:**

$$Q_m = \frac{\sum_{j=1}^{m} \lambda_1}{\sum_{j=1}^{r} \lambda_j}$$

使用Stata软件做判别分析



help discrim

help discrim lda postestimation

Fisher判别例: Fisher (1936)鸢尾花案例

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7: 179–188. (renamed Annals of Human Genetics in 1954)

鸢尾花数据

- The data consist of four features measured on 50 observations from each of three iris species.
- The four features are the length and width of the sepal and petal.
- The three species are Iris setosa, Iris versicolor, and Iris virginica.

Fisher判别例: Fisher (1936)鸢尾花案例

```
. use https://www.stata-press.com/data/r16/iris
(Iris data)

. discrim lda seplen sepwid petlen petwid, group(iris)

Linear discriminant analysis
Resubstitution classification summary
```

```
    +---------+
    | Key     |
    |---------|
    | Number  |
    | Percent |
    +---------+
```

True iris	Classified setosa	versicolor	virginica	Total
setosa	50	0	0	50
	100.00	0.00	0.00	100.00
versicolor	0	48	2	50
	0.00	96.00	4.00	100.00
virginica	0	1	49	50
	0.00	2.00	98.00	100.00
Total	50	49	51	150
	33.33	32.67	34.00	100.00
Priors	0.3333	0.3333	0.3333	

Fisher判别例: Fisher (1936)

对哪些观察的判别是错误的?

```
. estat list, misclassified

+-------------------------------------------------------------+
|      | Classification        | Probabilities                |
|      |                       |                              |
| Obs.|      True    Class.    | setosa  versicolor  virginica |
|-----+----------------------+------------------------------|
|  71 | versicol  virginic * | 0.0000      0.2532     0.7468 |
|  84 | versicol  virginic * | 0.0000      0.1434     0.8566 |
| 134 | virginic  versicol * | 0.0000      0.7294     0.2706 |
+-------------------------------------------------------------+
* indicates misclassified observations
```

典型判别函数(canonical discriminant functions)

以下命令同时给出针对未标准化数据和标准化后数据的判别方程参数

```
. estat loadings, unstandardized standardized

Canonical discriminant function coefficients
                | function1  function2
    ------------+----------------------
          seplen | -.8293776  -.0241021
          sepwid | -1.534473  -2.164521
          petlen |  2.201212   .9319212
          petwid |   2.81046  -2.839188
           _cons | -2.105106   6.661473

Standardized canonical discriminant function coefficients
                | function1  function2
    ------------+----------------------
          seplen | -.4269548  -.0124075
          sepwid | -.5212417  -.7352613
          petlen |  .9472572   .4010378
          petwid |  .5751608  -.5810399
```

*Fisher线性判别降维后线性函数, 用于计算投影坐标, 注意与下面讲到的分类函数的区别

Fisher判别例: Fisher (1936)

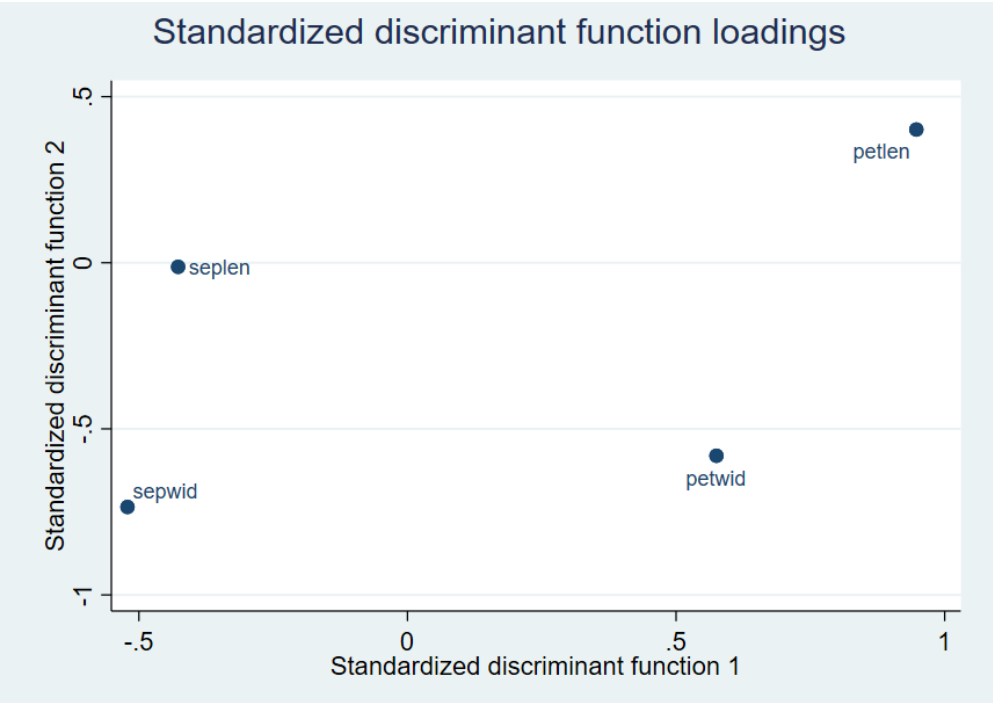
思考: 判别方程中哪个变量影响比较大, 哪个变量相对影响较小?

Fisher判别例: Fisher (1936)

思考: 判别方程中哪个变量影响比较大, 哪个变量相对影响较小?

1. petlen在第一判别函数中影响最大 (根据标准化后判别系数决定).
2. 也可用载荷图更可视化地判定重要性

`loadingplot \\` 标准化后数据图



判别函数的重要程度

```
. estat canontest

Canonical linear discriminant analysis

     |                                    | Like-
     | Canon.   Eigen-     Variance       | lihood
 Fcn | Corr.    value   Prop.   Cumul.    | Ratio      F      df1    df2  Prob>F
 ----+-----------------------------------+-----------------------------------
   1 | 0.9848  32.1919  0.9912  0.9912    | 0.0234  199.15     8    288  0.0000 e
   2 | 0.4712  .285391  0.0088  1.0000    | 0.7780  13.794     3    145  0.0000 e
 ---------------------------------------------------------------------------
 Ho: this and smaller canon. corr. are zero;                    e = exact F
```

投影的重要性是和特征值的贡献率有关. 该表说明第一个函数的贡献率已经是99%了, 而第二个只有1%.
检验各个判别函数有无统计学意义上的显著性.

判别方程(canonical discriminant functions)

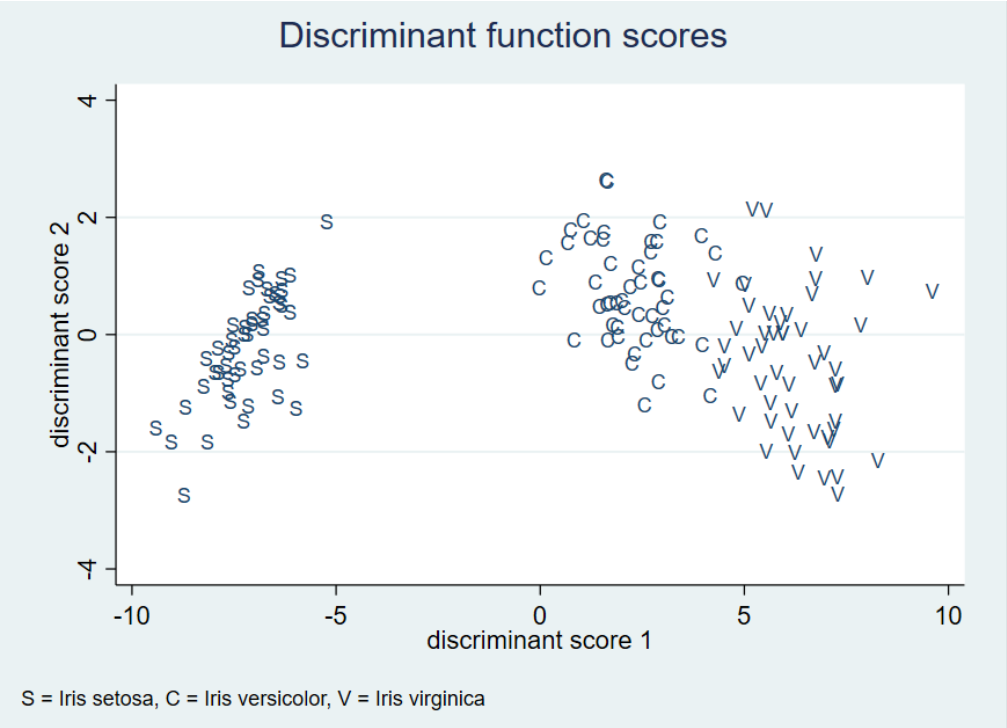
对于新的样本点, 在两个判别维度上的投影坐标怎么计算?

1. 可先对数据进行标准化, 再用标准化数据判别方程参数乘以标准化后样本点各维度值
2. 也可以用针对未标准化数据的判别方程直接乘以未标准化数据. 在本例中,

$$F_1 = -2.105 - .829x_1 - 1.534x_2 + 2.201x_3 - 2.810x_4$$
$$F_2 = 6.661 - .024x_1 - 2.164x_2 + .932x_3 - 2.839x_4$$

投影到2维平面的可视化表达

```
. label language oneletter
. scoreplot, msymbol(i)
> note("S = Iris setosa, C = Iris versicolor, V =Iris virginica")
```



1. Setosa很好地和另两类区分, 但Versicolor 和 Virginica 有重叠.
2. 用第一维判别函数就可以进行较好的判别, 第二维作用不大(纵轴上差异不大).

组质心处的函数

3组原始鸢尾花种类质心在第一和第二典型判别函数上的函数值.

```
.   //输出各组分类的质心值
.   estat grmeans, canonical

Group means on canonical variables

             iris | function1   function2
------------------+----------------------
           setosa |   -7.6076    -.215133
       versicolor |   1.825049    .7278996
        virginica |    5.78255   -.5127666
```

训练样本之外, 如果有新观察, 如何判定其属于哪类?

方法一： 利用分类函数 (假定Σ相同时的距离判别法)

分类函数(classification functions)系数

```
. estat classfunction
Classification functions
                 | iris
                 |     setosa  versico~r  virginica
-----------------+---------------------------------
          seplen |   23.54417   15.69821   12.44585
          sepwid |   23.58787    7.07251    3.68528
          petlen |  -16.43064   5.211451   12.76654
          petwid |  -17.39841   6.434229   21.07911
           _cons |  -85.20986    -71.754  -103.2697
-----------------+---------------------------------
          Priors |   .3333333   .3333333   .3333333
```

对于新观察, 把每个观察带入三个分类函数, 取值最大的分类即为该观察的判别分类.

$$x \in G_i \Leftrightarrow d^2(x, G_i) = \min_{j=1,2\ldots,k} \left\{ d^2(x, G_j) \right\}$$
$$\Leftrightarrow \max_{j=1,2,\ldots,k} \left\{ f_j(x)_i \right\}$$

训练样本之外, 如果有新观察, 如何判定其属于哪类?

方法二： 利用Fisher典型判别函数+距离判别

问题： 假设有新的观察, 其各指标数据为 seplen = 5; sepwid = 3; petlen = 5; petwid = 6. 假设三种鸢尾花在第一判别函数上的方差没有显著差别, 用第一典型判别函数, 判别此观察属于哪类鸢尾花.

答：

训练样本之外, 如果有新观察, 如何判定其属于哪类?

方法二： 利用Fisher典型判别函数+距离判别

问题： 假设有新的观察, 其各指标数据为 seplen = 5; sepwid = 3; petlen = 5; petwid = 6. 假设三种鸢尾花在第一判别函数上的方差没有显著差别, 用第一典型判别函数, 判别此观察属于哪类鸢尾花.

答：

第一步： 计算该观察在第一判别函数上的得分

$$F_1 = -2.105 - .829x_1 - 1.534x_2 + 2.201x_3 - 2.810x_4 =$$
$$-2.105 - .829 * 5 - 1.534 * 3 + 2.201 * 5 - 2.810 * 6 = -16.707$$

训练样本之外, 如果有新观察, 如何判定其属于哪类?

第二步： 由于总体方差相同, 根据组质心在第一典型判别函数上的函数值

iris	function1	function2
setosa	-7.6076	-.215133
versicolor	1.825049	.7278996
virginica	5.78255	-.5127666

可计算出判别临界值:

$$\mu_{12}^{*} = \frac{Y_1 + Y_2}{2} = \frac{5.78255 + 1.825049}{2} = 3.8037995$$

$$\mu_{23}^{*} = \frac{Y_2 + Y_3}{2} = \frac{1.825049 - 7.6076}{2} = -2.8912755.$$

训练样本之外, 如果有新观察, 如何判定其属于哪类?

即判别规则为

$$\begin{cases} x \in virginica, F_1(x) > \mu_{12}^*; \\ x \in versicolor, \mu_{23}^* < F_1(x) < \mu_{12}^*; \\ x \in setosa, F_1(x) < \mu_{23}^* \\ \text{待判}, F_1(x) = \mu_{12}^* \text{ 或 } \mu_{23}^*. \end{cases}$$

第三步： 判别由于 $-16.707 < \mu_{23}$ 可见, 待分析的样本点属于setosa.

训练样本之外, 如果有新观察, 如何判定其属于哪类?

也可利用统计软件自动给出(在线性判别分析中, Stata默认用多维距离判别法给出判别)

stata postestimation command: predict

```
// add 1 random observations to the original sample
set obs 151
. replace seplen = 5 in 151
(1 real change made)
. replace sepwid = 3 in 151
(1 real change made)
. replace petlen = 5 in 151
(1 real change made)
. replace petwid = 6 in 151
(1 real change made)

. predict g, classification

. list  in 151
     +-----------------------------------------------+
     | iris   seplen   sepwid   petlen   petwid   g |
     |-----------------------------------------------|
151. |    .      5.0      3.0      5.0      6.0   3 |
     +-----------------------------------------------+
```

Fisher判别例: Fisher (1936)

以上大多数主要结果也可以直接使用典型线性判别分析命令整体给出

```
. candisc seplen sepwid petlen petwid, group(iris)

Canonical linear discriminant analysis
```

其他相关命令

help discrim lda
help discrim lda postestimation

课堂练习

上机操作观察: 如果只用2个变量进行判别, 计算结果中图形的分辨能力下降, 判别的正确率低于前面分析

拓展知识: 1. 贝叶斯判别模型

- 设有两个总体, 它们的先验概率分别为p_1, p_2, 各总体的密度函数分别为 $f_1(x)$, $f_2(x)$
- 在观测到一个样本 x 的情况下, 可用贝叶斯公式, 计算它来自第 k 个总体的后验概率为:

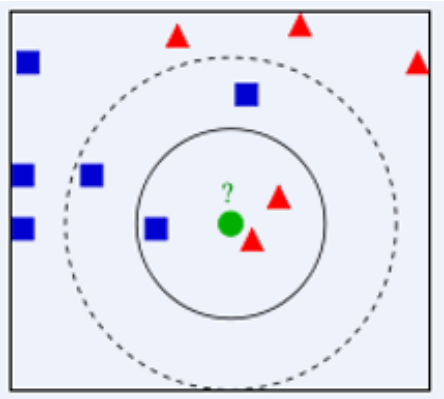
$$\mathrm{P}\left(G_k \mid x\right)=\frac{p_k f_k(x)}{p_1 f_1(x)+p_2 f_2(x)}, k=1,2$$

- 判别准则: 对于待判样本 x, 如果在所有的$P(G_k|x)$中, $P(G_h|x)$是最大的, 则判定x 属于第h 总体
- 可用样本频率作为各总体的先验概率; 并假设 $f_1(x)$, $f_2(x)$ 服从某特定分布

(马氏距离判别和贝叶斯判别在先验概率相同时是等价的)

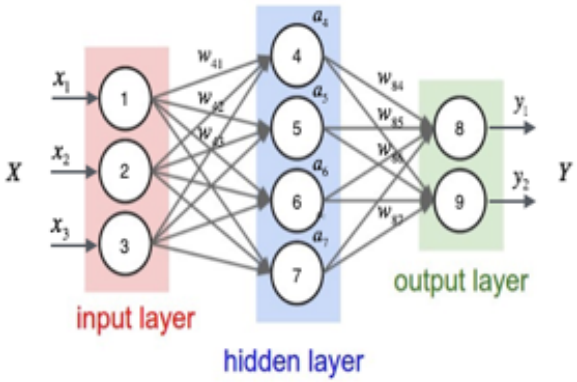
拓展知识: 2. K近邻 (k-nearest neighbors , KNN)

- 核心思想: 物以类聚, 人以群分
- 对于待分类的样本点 x, 从训练集中找出与之空间距离最近的k个点, 取这k个点的众数, 作为该样本点的类别
- 距离定义: 根据问题 (明氏距离 Minkowski Distance, 马氏距离, cos距离)
- 参数 k 的选取: 采用交叉验证法, 将k由小变大逐渐递增, 制作k与分类误差的曲线图, 选取合适的k (一般说来: $k < \sqrt{n}$)



拓展知识: 3. 机器学习中的诸多有监督的分类方法都是判别分析

- 神经网络
- 支持向量机
- K邻近
- 决策树
- 随机森林
- Boosting



应用统计学II作业5暨上机实验12

某大型航空公司人力资源部调研了3种工作的雇员性格, 这三类工作分别是1)客服 customer service personnel, 2)飞机维修 mechanics and 3)调度员 dispatchers.

人力资源部领导想要知道是否特定工种和对于特定性格的雇员更匹配. 每个雇员回答了一系列心理测验问题衡量了三种主要的性格特征 1) 喜好户外活动 (interest in outdoor activity), 2)喜欢社交(sociability), 3) 性格保守(conservativeness). (分值越大特征越明显)

运用Fisher判别法, 对该数据 (AirlineHR.xls)进行分析, 并对结果进行评述总结.