



第三章 判别分析

§ 3.1 判别分析的基本思想

§ 3.2 距离判别

§ 3.3 Fisher判别

§ 3.4 Bayes判别

第三章 判别分析

判别分析的基本思想

- 在我们的日常生活和工作实践中，常常会遇到判别分析问题，即根据已知类别的资料确定一种判别方法，判定一个新的样品**归属**哪一类。
- 例如，某医院有患有肺炎、肝炎、冠心病、糖尿病等的病人的资料，记录了患者若干项指标数据。现在想利用现有的这些资料找出一种方法，以便根据新的病人的指标数据判定其患有哪种病。

第三章 判别分析

判别分析的基本思想

- 根据已掌握的每个类别的若干样本的数据信息，建立判别公式和判别准则。
- 当遇到新的样本点时，根据总结出来的判别公式和判别准则，即能判别该样本点所属的类别。

第三章 判别分析

判别分析的几种方法

- 两个总体判别分析和多总体判别分析
- 常用的判别分析方法
 - 距离判别法(马氏距离)
 - Fisher判别法
 - Bayes判别法

第三章 判别分析

距离判别法

- 两个总体的距离判别问题：设两个总体 G_1 和 G_2 ，对于一个新的样品 X ，要判断它来自哪个总体。
- 方法：按就近原则归类。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 。
- 根据上述准则可以推导出一个判别函数 W ，把待判样品的值代入判别函数，根据计算结果是否大于0得出判别结论。

第三章 判别分析



距离判别法

欧氏距离

马氏距离

第三章 判别分析

距离判别法

欧氏距离

简单的距离概念——欧氏距离, 或称直线距离.

如几何平面上的点 $p=(x_1, x_2)$ 到原点 $0=(0, 0)$ 的欧氏距离, 依勾股定理有

$$d(0, p) = (x_1^2 + x_2^2)^{1/2}$$

每个坐标对欧氏距离的贡献是同等的。

缺点:

- a. 当坐标轴表示测量值时, 它们往往带有大小不等的随机波动.
- b. 当各个分量为不同性质的量时, “距离”的大小与指标的单位有关。

第三章 判别分析

距离判别法

横轴 x_1 代表重量（以kg为单位），纵轴 x_2 代表长度（以cm为单位）。有四个点A、B、C、D见图1，它们的坐标如图1所示

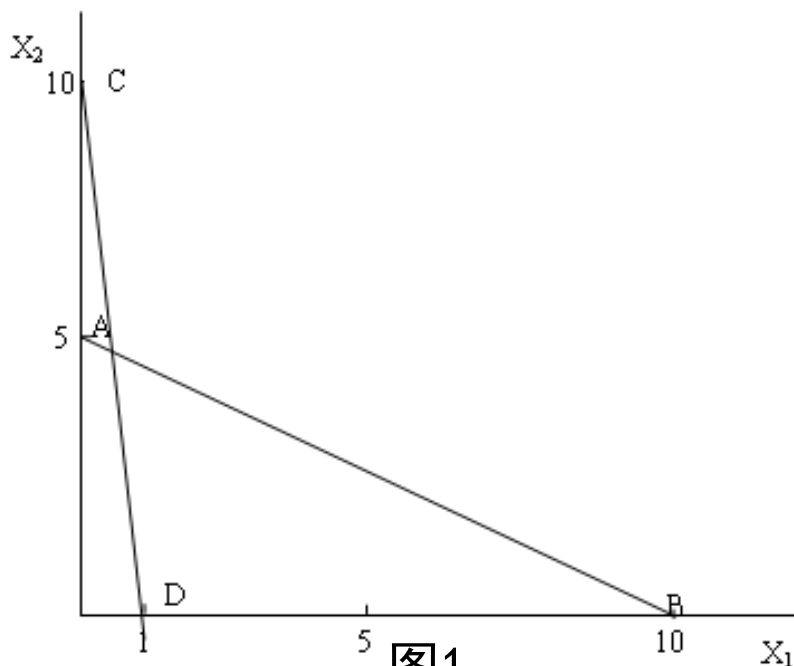


图1

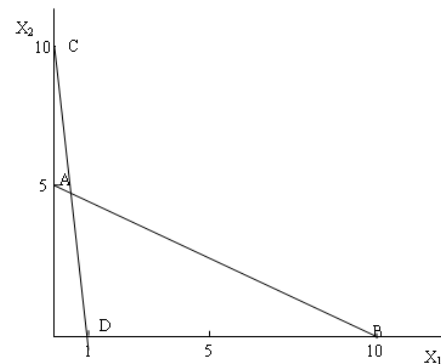
第三章 判别分析

距离判别法

这时

$$AB = \sqrt{5^2 + 10^2} = \sqrt{125}$$
$$CD = \sqrt{10^2 + 1^2} = \sqrt{101}$$

显然 AB 比 CD 要长。



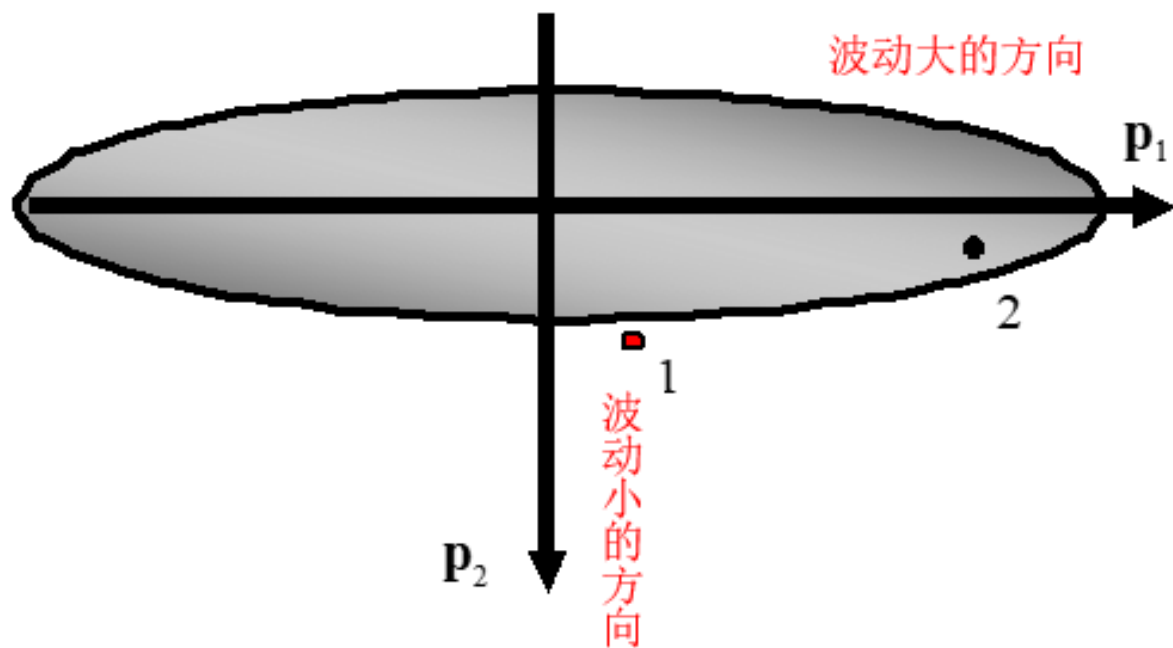
现在，如果 x_2 用 mm 作单位， x_1 单位保持不变，此时A坐标为 $(0, 50)$ ，C坐标为 $(0, 100)$ ，则

$$AB = \sqrt{50^2 + 10^2} = \sqrt{2600}$$
$$CD = \sqrt{100^2 + 1^2} = \sqrt{10001}$$

结果 CD 反而比 AB 长！

第三章 判别分析

距离判别法



哪个点距离总体近？

第三章 判别分析

距离判别法

有必要建立一种距离：

- a) 能够体现各个变量在变差大小上的不同
- b) 能够体现各个变量存在着的相关性
- c) 距离与各变量所用的单位无关。

因此：

选择的距离要依赖于样本方差和协方差。

采用“统计距离”这个术语以区别通常习惯用的欧氏距离。

最常用的一种统计距离是印度统计学家马哈拉诺比斯

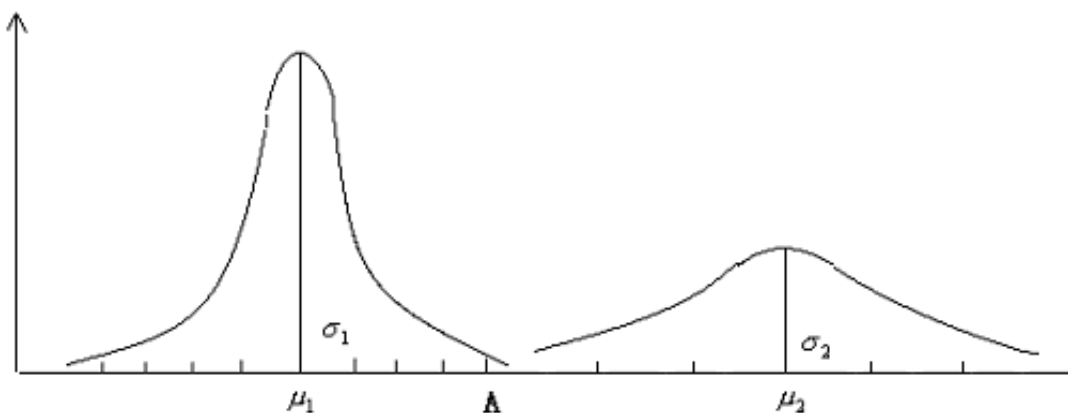
(*Mahalanobis*) 于1936年引入的距离，称为“**马氏距离**”。

第三章 判别分析

距离判别法

欧氏距离与马氏距离在概率上的差异：

设有两个一维正态总体 $G_1 : (\mu_1, \sigma_1^2)$ 和 $G_2 : (\mu_2, \sigma_2^2)$ 。若有一个样品，其值在 A 处， A 点距离哪个总体近些呢？



$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

图2

第三章 判别分析

距离判别法

马氏距离

设 \mathbf{X} 、 \mathbf{Y} 从均值向量为 $\boldsymbol{\mu}$ ，协方差阵为 $\boldsymbol{\Sigma}$ 的总体 G 中抽取的两个样品。

定义 \mathbf{X} 、 \mathbf{Y} 两点之间的马氏距离为

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

定义 \mathbf{X} 与总体 G 的马氏距离为

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

第三章 判别分析



距离判别法

马氏距离

为什么聚类用欧氏距离？

距离判别法的问题？

第三章 判别分析

Fisher判别法

- 借助方差分析的思想构造一个线性判别函数：

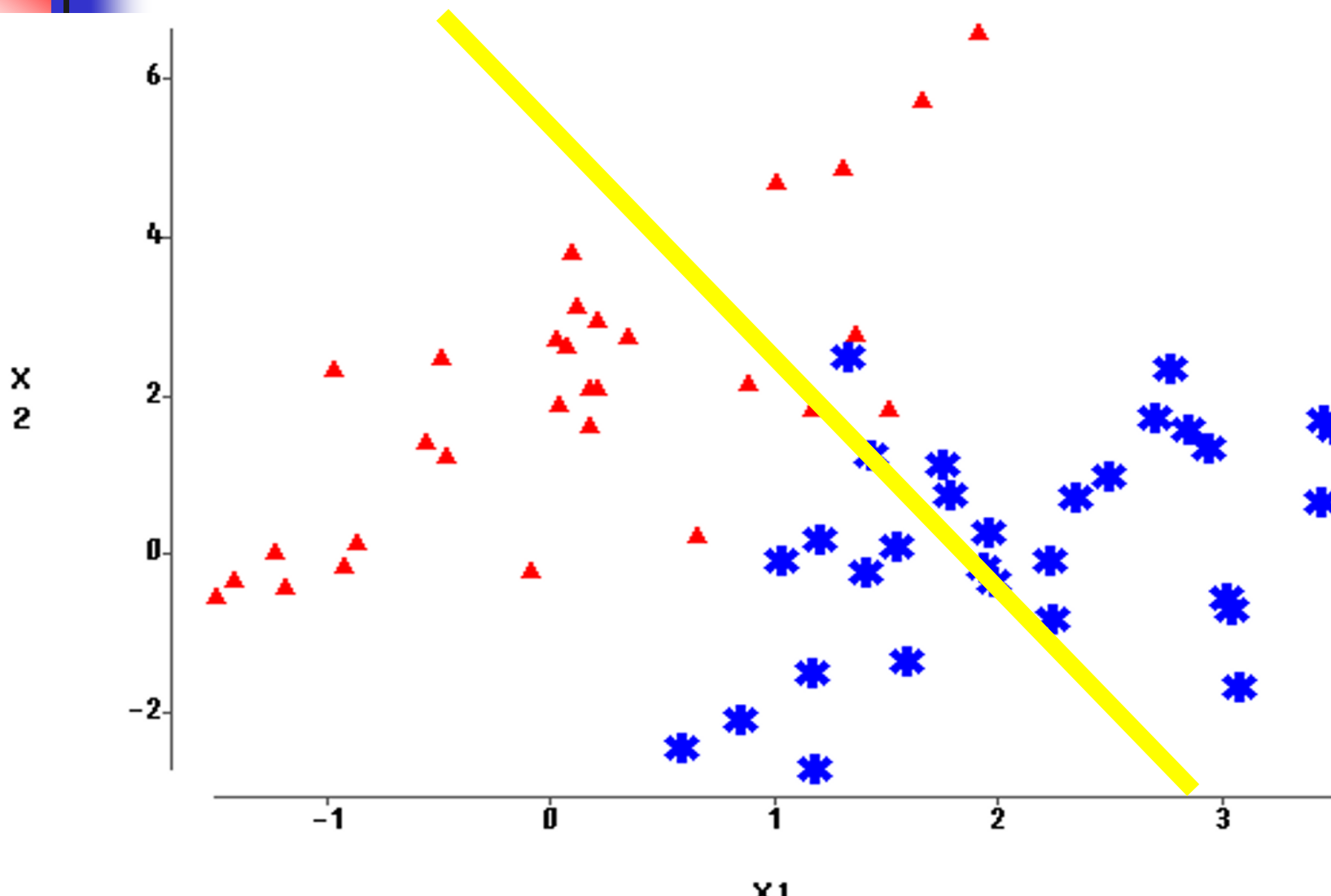
$$W = b_0 + b_1 X_1 + \cdots + b_p X_p$$

- 确定判别函数系数时要求使得总体之间区别最大，而使每个总体内部的离差平方和最小。
- 从几何的角度看，判别函数就是p维向量X在某种方向上的投影。使得变换后的数据同类别的点“尽可能聚在一起”，不同类别的点“尽可能分离”，以此达到分类的目的。

第三章 判别分析

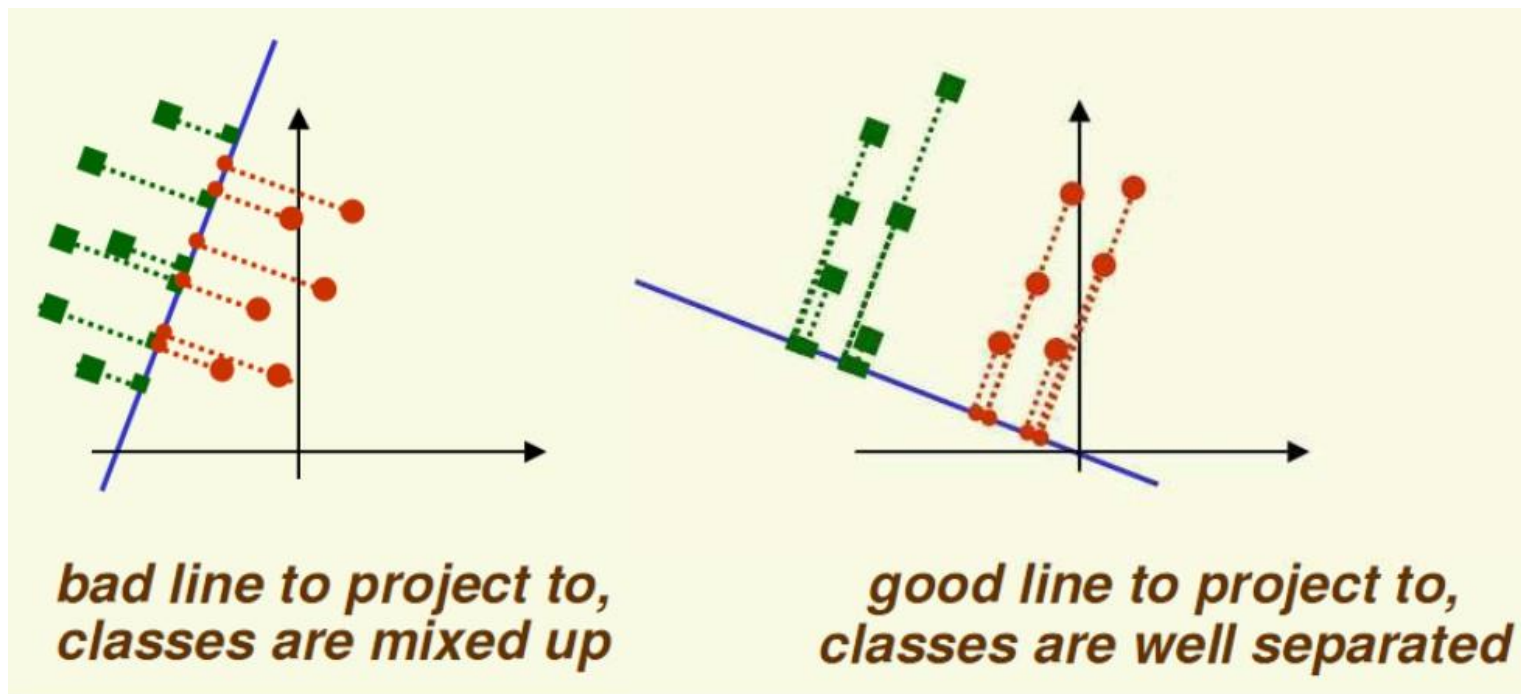
寻找方向 a ,使两组数据投影后在一维直线上尽可能区分开

两类Fisher判别示意图



第三章 判别分析

两类Fisher判别示意图



$$y_i = w^T x_i$$

使得d维向量 x_i 变成了一个标量

满足类内紧凑，类间分离的原则

第三章 判别分析



Fisher判别

怎么评估类内波动和类间波动？

类间的波动——方差，类内的波动——方差

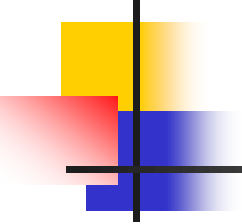
在特征空间中的方差——
在原始测量空间中则为协方差？



协方差阵的估计

$$\hat{\Sigma}_p = \frac{1}{n} L = \frac{1}{n} \sum_{i=1}^n (X_{(i)} - \bar{X})(X_{(i)} - \bar{X})'$$

$$= \frac{1}{n} \begin{bmatrix} \sum_{i=1}^n (X_{i1} - \bar{X}_1)^2 & \cdots & \sum_{i=1}^n (X_{i1} - \bar{X}_1)(X_{ip} - \bar{X}_p) \\ \vdots & \sum_{i=1}^n (X_{i2} - \bar{X}_2)^2 \cdots \sum_{i=1}^n (X_{i2} - \bar{X}_2)(X_{ip} - \bar{X}_p) & \vdots \\ \cdots & \cdots & \sum_{i=1}^n (X_{ip} - \bar{X}_p)^2 \end{bmatrix}$$



其中L是**离差阵**，它是每一个样品（向量）与样本均值（向量）的离差积形成的 n 个 $p \times p$ 阶对称阵的和。同一元相似， $\hat{\Sigma}_p$ 不是 Σ 的无偏估计，为了得到无偏估计我们常用样本协差阵 $\hat{\Sigma} = \frac{1}{n-1} L$ 作为总体协差阵的估计。



B是类内离差和，**E**是类间离差

$$\begin{aligned} \min(\mathbf{w}^T \mathbf{B} \mathbf{w}) \\ \text{s.t. } \mathbf{w}^T \mathbf{E} \mathbf{w} = \mathbf{c} \end{aligned}$$

$$\mathbf{E}^{-1} \mathbf{B} \mathbf{w} = \lambda \mathbf{w}$$

特征根分解问题

第三章 判别分析

Fisher判别法计算步骤

(1) 先将原始数据写成矩阵形式。组**A**的数据矩阵:

$$W^0 = \begin{bmatrix} x_{11}^0 & x_{12}^0 & \cdots & x_{1p}^0 \\ x_{21}^0 & x_{22}^0 & \cdots & x_{2p}^0 \\ \vdots & \vdots & \vdots & \vdots \\ x_{s1}^0 & x_{s2}^0 & \cdots & x_{sp}^0 \end{bmatrix}$$

组**B**的数据矩阵

$$W^1 = \begin{bmatrix} x_{11}^1 & x_{12}^1 & \cdots & x_{1p}^1 \\ x_{21}^1 & x_{22}^1 & \cdots & x_{2p}^1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{t1}^1 & x_{t2}^1 & \cdots & x_{tp}^1 \end{bmatrix}$$

矩阵 W^0 和 W^1 的列平均数分别为 $(\bar{x}_1^0, \bar{x}_2^0, \cdots \bar{x}_p^0)$ 和 $(\bar{x}_1^1, \bar{x}_2^1, \cdots \bar{x}_p^1)$

第三章 判别分析

Fisher判别法计算步骤

(2) 算出各组数据的代表, 即平均值

$$\bar{x}_j^0 = \frac{1}{s} \sum_{i=1}^s x_{ij}^0 \quad j = 1, 2 \cdots p$$

$$\bar{x}_j^1 = \frac{1}{t} \sum_{i=1}^t x_{ij}^1 \quad j = 1, 2 \cdots p$$

$$\bar{x}_j = \frac{1}{s+t} \sum_{i=1}^{s+t} x_{ij} \quad j = 1, 2 \cdots p$$

(3) 作新的矩阵 **A**, **B** 及两组的离差矩阵 S_1 S_2 与组间离差

$$A = \begin{bmatrix} x_{11}^0 - \bar{x}_1^0 & x_{12}^0 - \bar{x}_2^0 & \cdots & x_{1p}^0 - \bar{x}_p^0 \\ x_{21}^0 - \bar{x}_1^0 & x_{22}^0 - \bar{x}_2^0 & \cdots & x_{2p}^0 - \bar{x}_p^0 \\ \vdots & \vdots & \vdots & \vdots \\ x_{s1}^0 - \bar{x}_1^0 & x_{s2}^0 - \bar{x}_2^0 & \cdots & x_{sp}^0 - \bar{x}_p^0 \end{bmatrix}$$

$$B = \begin{bmatrix} x_{11}^1 - \bar{x}_1^1 & x_{12}^1 - \bar{x}_2^1 & \cdots & x_{1p}^1 - \bar{x}_p^1 \\ x_{21}^1 - \bar{x}_1^1 & x_{22}^1 - \bar{x}_2^1 & \cdots & x_{2p}^1 - \bar{x}_p^1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{t1}^1 - \bar{x}_1^1 & x_{t2}^1 - \bar{x}_2^1 & \cdots & x_{tp}^1 - \bar{x}_p^1 \end{bmatrix}$$

$$C = \begin{bmatrix} \bar{x}_{11}^0 - \bar{x}_1 & \cdots & \bar{x}_{1p}^0 - \bar{x}_p \\ \bar{x}_{11}^1 - \bar{x}_1 & \cdots & \bar{x}_{1p}^1 - \bar{x}_p \end{bmatrix}$$

$$S_1 = A^T A, S_2 = B^T B, S = S_1 + S_2$$

$$W = C^T C$$

第三章 判别分析

Fisher判别法计算步骤

(4) 可以证明, 最优判别函数系数 c_1, c_2, \dots, c_p 为 $S^{-1}W$ 特征根分解

(5) 写出判别函数

$$y = c_1 x_1 + c_2 x_2 + \dots + c_p x_p$$

(6) 算出组**A**, 组**B**的代表的判别值

$$\bar{y}_A = c_1 \bar{x}_1^0 + c_2 \bar{x}_2^0 + \dots + c_p \bar{x}_p^0$$

$$\bar{y}_B = c_1 \bar{x}_1^1 + c_2 \bar{x}_2^1 + \dots + c_p \bar{x}_p^1$$

$$y_0 = \frac{s\bar{y}_A + t\bar{y}_B}{s + t}$$

第三章 判别分析

Fisher判别法计算步骤

(7) 作判别。 有一判别的对象若其数据为 $(x_{01}, x_{02}, \dots, x_{0p})$ 则其判别值为 $y = c_1 x_{01} + c_2 x_{02} + \dots + c_p x_{0p}$

1) 当 $\bar{y}_A > y_0$ 时, 若 $y > y_0$ 则判别该对象属于组**A**, 若 $y < y_0$ 判别该对象属于组**B**。

2) 当 $\bar{y}_B > y_0$ 时, 若 $y > y_0$ 判别该对象属于组**B**, 则若 $y < y_0$ 则判别该对象属于组**A**。

第三章 判别分析

Fisher判别法应用举例

例 设某外贸公司生产一种产品，为正式上市之前，将样品寄往**12个国家**的进口代理商，并附意见调查表，要求**对该产品进行评估**。评估的内容有**式样，包装，耐久性**三个方面。评估的结果采用**10分制**计分，**评估后并被要求说明是否愿意购买**，调查结果列入表**1**中，表中的分数，高者表示代理商认为其特性良好，否则即较差。

今有第**13**个国家的进口代理商对该产品的评分分别是：式样**9**分，包装**5**分，耐久性**4**分，要预测该国是否愿意购买该产品。

第三章 判别分析

Fisher判别法应用举例

	编号	式样X ₁	包装X ₂	耐久性X ₃		编号	式样X ₁	包装X ₂	耐久性X ₃
购买者	1	9	8	7	非购买者	8	8	4	4
	2	7	6	6		9	3	6	6
	3	10	7	8		10	6	3	3
	4	8	4	5		11	6	4	5
	5	9	9	3		12	8	2	2
	6	8	6	7					
	7	7	5	6					

第三章 判别分析

Fisher判别法应用举例

我们用**Fisher**判别函数解答上述问题：

(1) 计算两组的平均值

购买者平均得分为 \bar{x}_A (**8.29, 6.43, 6.00**) ;

非购买者平均得分为 \bar{x}_B (**6.20, 3.80, 4.00**) .

(2) 计算两组资料的离差矩阵

$$A = \begin{pmatrix} 0.71 & 1.57 & 1 \\ -1.29 & -0.43 & 0 \\ 1.71 & 0.57 & 2 \\ -0.29 & -2.43 & -1 \\ 0.71 & 2.57 & -3 \\ -0.29 & -0.43 & 1 \\ -1.29 & -1.43 & 0 \end{pmatrix}$$

$$B = \begin{pmatrix} 1.8 & 0.2 & 0 \\ -3.2 & 2.2 & 2 \\ -0.2 & -0.8 & -1 \\ -0.2 & 0.2 & 1 \\ 1.8 & -1.8 & -2 \end{pmatrix}$$

第三章 判别分析

Fisher判别法应用举例

两组的离差矩阵分别为

$$S_1 = A'A = \begin{pmatrix} 7.43 & 7.14 & 2 \\ 7.14 & 17.71 & -3 \\ 2 & -3 & 16 \end{pmatrix} \quad S_2 = B'B = \begin{pmatrix} 16.8 & -9.8 & -10 \\ -9.8 & 8.8 & 9 \\ -10 & 9 & 10 \end{pmatrix}$$

$$S = S_1 + S_2 = \begin{pmatrix} 24.23 & -2.66 & -8 \\ -2.66 & 26.51 & 6 \\ -8 & 6 & 26 \end{pmatrix} \quad C = \begin{bmatrix} \bar{x}_{A1} - \bar{x}_1 & \cdots & \bar{x}_{Ap} - \bar{x}_p \\ \bar{x}_{B1} - \bar{x}_1 & \cdots & \bar{x}_{Bp} - \bar{x}_p \end{bmatrix}$$

组内离差

$W = C^T C$ 组间离差

(3) 特征根分解得判别系数

$$c_1 = 0.128$$

$$c_2 = 0.090$$

$$c_3 = 0.095$$

第三章 判别分析

Fisher判别法应用举例

(4) 根据计算结果，得判别函数为

$$y = 0.128x_1 + 0.090x_2 + 0.095x_3$$

(5) 求出判别临界值

购买组的平均值对应的判别值

$$\bar{y}_A = 0.128 \times 8.29 + 0.090 \times 6.43 + 0.095 \times 6.00 = 2.210$$

非购买组的平均值对应的判别值为

$$\bar{y}_B = 0.128 \times 6.20 + 0.090 \times 3.80 + 0.095 \times 4.00 = 1.516$$

从而临界值为

$$y_0 = \frac{2.210 \times 7 + 1.516 \times 5}{7 + 5} = 1.921$$

第三章 判别分析

Fisher判别法应用举例

(6) 作判别预测

按判别规则，由于 $\bar{y}_A > y_0$ ，故凡判别值大于 y_0 (**1.912**) 者，即判别其属于购买组。今第**13**个国家的判别值为

$$y = 0.128 \times 9 + 0.090 \times 5 + 0.095 \times 4 = 1.982$$

因**1.982**>**1.912**，故预测该国属于购买者范围。

第三章 判别分析

Fisher判别法

- 如果有多个类别，Fisher判别可能需要两个或者更多的判别函数才能完成分类。
- 一般来说判别函数的个数等于类别的个数减一。
- 得到判别函数后，计算待判样品的判别函数值，根据判别函数的值计算待判样品到各类的重心的距离，从而完成分类。

第三章 判别分析

Bayes判别法

贝叶斯（Bayes）统计的思想是：假定对研究的对象已有一定的认识，常用先验概率分布来描述这种认识，然后我们取得一个样本，用样本来修正已有的认识（先验概率分布），得到后验概率分布，各种统计推断都通过后验概率分布来进行。将贝叶斯思想用于判别分析，就得到贝叶斯判别。

- 朴素贝叶斯分类（Naïve Bayesian Model）是基于贝叶斯条件概率定理的概率分类器。
- 最大特点：该模型假设特征之间相互独立、彼此不相关。这就是它“朴素”之处。这也是很多人对它最担心之处。
- 人们往往先入为主地认为，其根本性假设都不对，那么效果一定好不到哪里去。但事实是，它在很多应用中表现很好。

第三章 判别分析

Bayes判别法

条件概率

$$P(B|A) = \frac{P(AB)}{P(A)}$$

乘法定理

$$P(AB) = P(A)P(B|A)$$

全概率公式

$$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \cdots + P(B_n)P(A|B_n)$$

贝叶斯公式

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^n P(B_j)P(A|B_j)}, \quad i = 1, 2, \dots, n.$$

第三章 判别分析

Bayes判别法

称 $P(B_i)$ 为**先验概率**，它是由以往的经验得到的，它是事件 **A** 的原因

称 $P(B_i|A)$ $i=0, 1, 2, 3, 4$ 为**后验概率**，它是

得到了信息 — **A** 发生，再对导致 A 发生的原因发生的可能性大小重新加以修正

全概率公式的主要用途在于它可以将一个复杂事件的概率计算问题，分解为若干个简单事件的概率计算问题，最后应用概率的可加性求出最终结果。

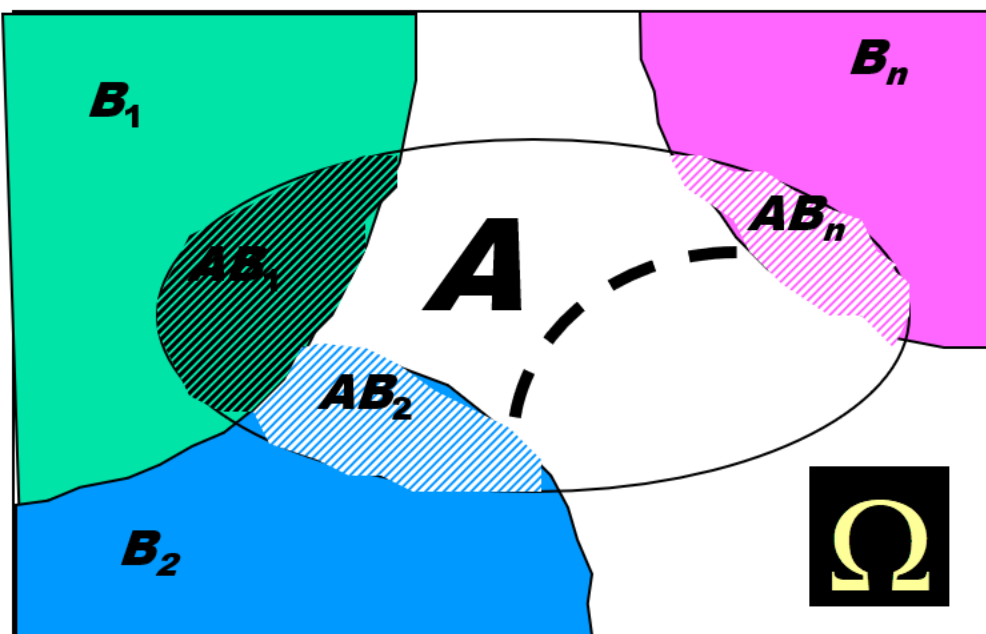
化整为零 各个突破

设 Ω 为试验 E 的样本空间, A 为 E 的事件,

B_1, B_2, \dots, B_n 为 Ω 的一个划分,

且 $P(B_i) > 0 (i = 1, 2, \dots, n)$, 则

$$P(A) = \sum_{i=1}^n P(AB_i) = \sum_{i=1}^n P(B_i) \cdot P(A|B_i)$$



设 Ω 为试验 E 的样本空间, A 为 E 的事件,

B_1, B_2, \dots, B_n 为 Ω 的一个划分, 且 $P(A) > 0$,

$P(B_i) > 0 (i = 1, 2, \dots, n)$, 则

$$P(B_k|A) = \frac{P(AB_k)}{P(A)} = \frac{P(B_k)P(A|B_k)}{\sum_{i=1}^n P(B_i)P(A|B_i)}$$

第三章 判别分析

Bayes判别法实例

例1: 办公室新来了一个雇员小王，小王是好人还是坏人大家都在猜测。按人们主观意识，一个人是好人或坏人的概率均为0.5。坏人总是要做坏事，好人总是做好事，偶尔也会做一件坏事，一般好人做好事的概率为0.9，坏人做好事的概率为0.2，一天，小王做了一件好事，小王是好人的概率有多大，你现在把小王判为何种人。

第三章 判别分析

Bayes判别法实例

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^n P(B_j)P(A|B_j)}, i=1,2,\dots,n.$$

$$\begin{aligned} & P(\text{好人} / \text{做好事}) \\ &= \frac{P(\text{好人})P(\text{做好事} / \text{好人})}{P(\text{好人})P(\text{做好事} / \text{好人}) + P(\text{坏人})P(\text{做好事} / \text{坏人})} \\ &= \frac{0.5 \times 0.9}{0.5 \times 0.9 + 0.5 \times 0.2} = 0.82 \end{aligned}$$

$$\begin{aligned} & P(\text{坏人} / \text{做好事}) \\ &= \frac{P(\text{坏人})P(\text{做好事} / \text{坏人})}{P(\text{好人})P(\text{做好事} / \text{好人}) + P(\text{坏人})P(\text{做好事} / \text{坏人})} \\ &= \frac{0.5 \times 0.2}{0.5 \times 0.9 + 0.5 \times 0.2} = 0.18 \end{aligned}$$

0.82 > 0.18, 判别结果: 小王为好人

第三章 判别分析

Bayes判别法实例

假定用血清甲胎蛋白法诊断肝癌，以 **C** 表示“被检验者患有肝癌”这一事件，以 **A** 表示“判断被检验者患有肝癌”这一事件。假设这一检验法相应的概率为

$$P(A|C) = 0.95, \quad P(\bar{A}|\bar{C}) = 0.90.$$

又设在人群中 $P(C) = 0.0004$

现在若有一人被此检验法诊断为患有肝癌，求此人真正患有肝癌的概率 $P(C|A)$ 。

解 因为 $P(A|C) = 0.95$,

$$P(A|\bar{C}) = 1 - P(\bar{A}|\bar{C}) = 0.1,$$

$$P(C) = 0.0004, \quad P(\bar{C}) = 0.9996$$



第三章 判别分析

Bayes判别法实例

由贝叶斯公式得所求概率为

$$\begin{aligned} P(C|A) &= \frac{P(C)P(A|C)}{P(C)P(A|C) + P(\bar{C})P(A|\bar{C})} \\ &= \frac{0.0004 \times 0.95}{0.0004 \times 0.95 + 0.9996 \times 0.1} \\ &= 0.0038. \end{aligned}$$



即平均**10000**个具有阳性反应的人中大约只有**38**人患有癌症。

第三章 判别分析

Bayes判别法应用

- 经典应用——**垃圾邮件过滤**：
把邮件自动标记为垃圾邮件或正常邮件。
- 朴素贝叶斯模型会通过邮件中的诸多垃圾邮件**标志物**来判断邮件是否是垃圾邮件。这些标志物可能是词汇（例如是否有货币符号），也可能是其他特征（例如是否群发）。



第三章 判别分析

Bayes判别法应用

- 一封邮件需要多高的概率才值得贴上垃圾邮件的标签？这取决于三个信息：
 - $\Pr(\text{垃圾邮件标志物} | \text{垃圾邮件})$ 。垃圾邮件中包含这个标志物（关键字等）的概率，即这个标志物是否经常出现在垃圾邮件中。如果这个标志物在垃圾邮件中出现并不频繁，那么它显然不是个好的标志物。
 - $\Pr(\text{垃圾邮件})$ 。一封垃圾邮件出现的基本概率，即先验概率。如果垃圾邮件经常出现，那么显然我们正在考察的这封邮件也更有可能是垃圾邮件。
 - $\Pr(\text{垃圾邮件标志物})$ 。即标志物出现的概率。如果标志物在很多邮件、甚至所有邮件中都出现，那么它就不是个好的标志物。

第三章 判别分析

Bayes判别法应用

根据这三个信息，可以得到**后验概率**：即在出现垃圾邮件标志物的前提下，邮件为垃圾邮件的可能性：

$$\Pr(\text{垃圾邮件}|\text{标志物}) = \frac{\Pr(\text{标志物}|\text{垃圾邮件})\Pr(\text{垃圾邮件})}{\Pr(\text{标志物})}$$

第三章 判别分析



为什么朴素Bayes的独立性假设是可行的？

- 只要正确类的后验概率比其他类要高就可以得到正确的分类。所以即使概率估计不精确，一定程度上也不影响正确做出分类。
- 在数据预处理环节，通常会进行变量选择，把对于高度相关的变量只保留其中一个，剩下的变量之间就接近于相互独立了。

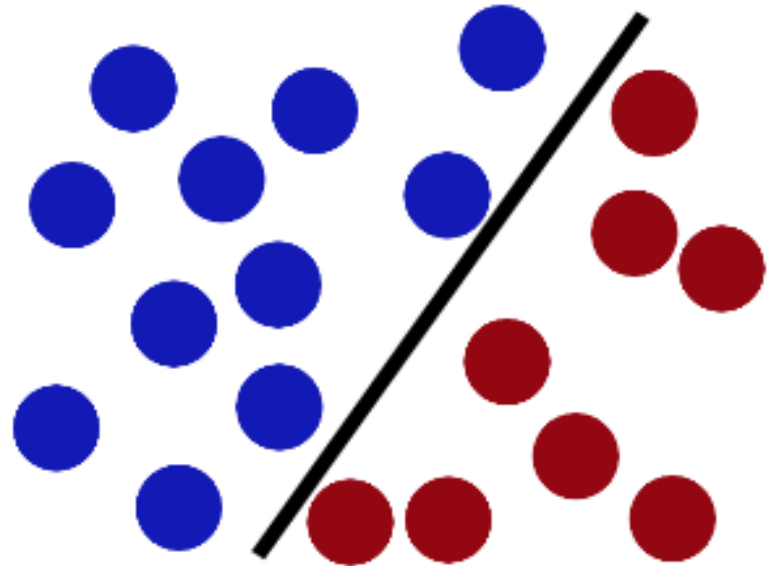
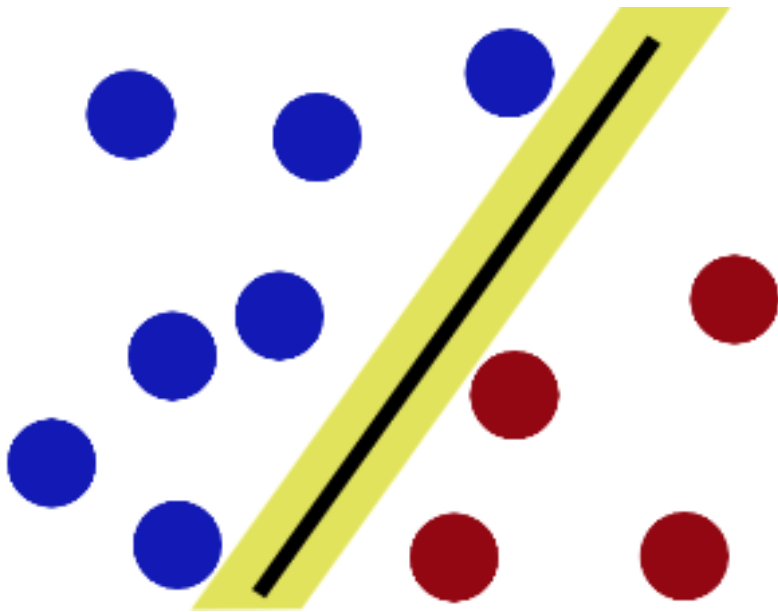
知识点拓展-SVM

- 从拯救公主的故事讲起





知识点拓展-SVM





知识点拓展-SVM



知识点拓展-SVM

如果一个线性函数能够将样本分开，称这些数据样本是线性可分的。那么什么是线性函数呢？其实很简单，在二维空间中就是一条直线，在三维空间中就是一个平面，以此类推，如果不考虑空间维数，这样的线性函数统称为超平面。

我们看一个简单的二维空间的例子，+代表正类，-代表负类，样本是线性可分的，但是很显然不只有这一条直线可以将样本分开，而是有无数条，我们所说的线性可分支持向量机就对应着能将数据正确划分并且间隔最大的直线。

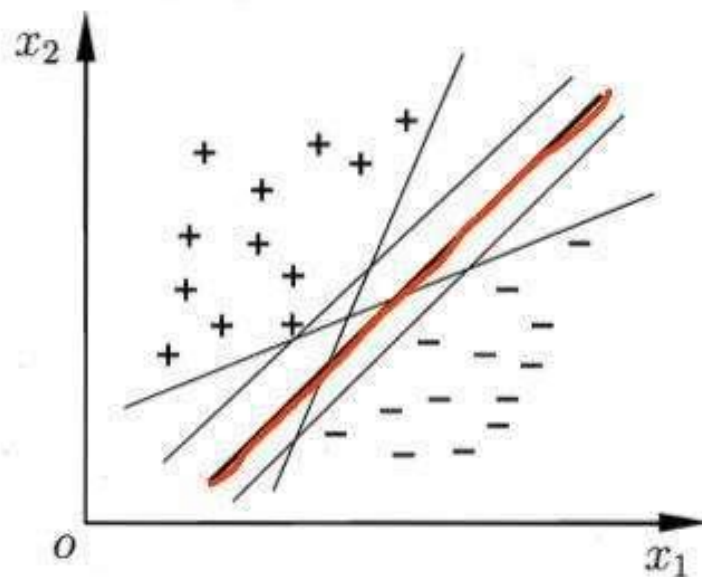
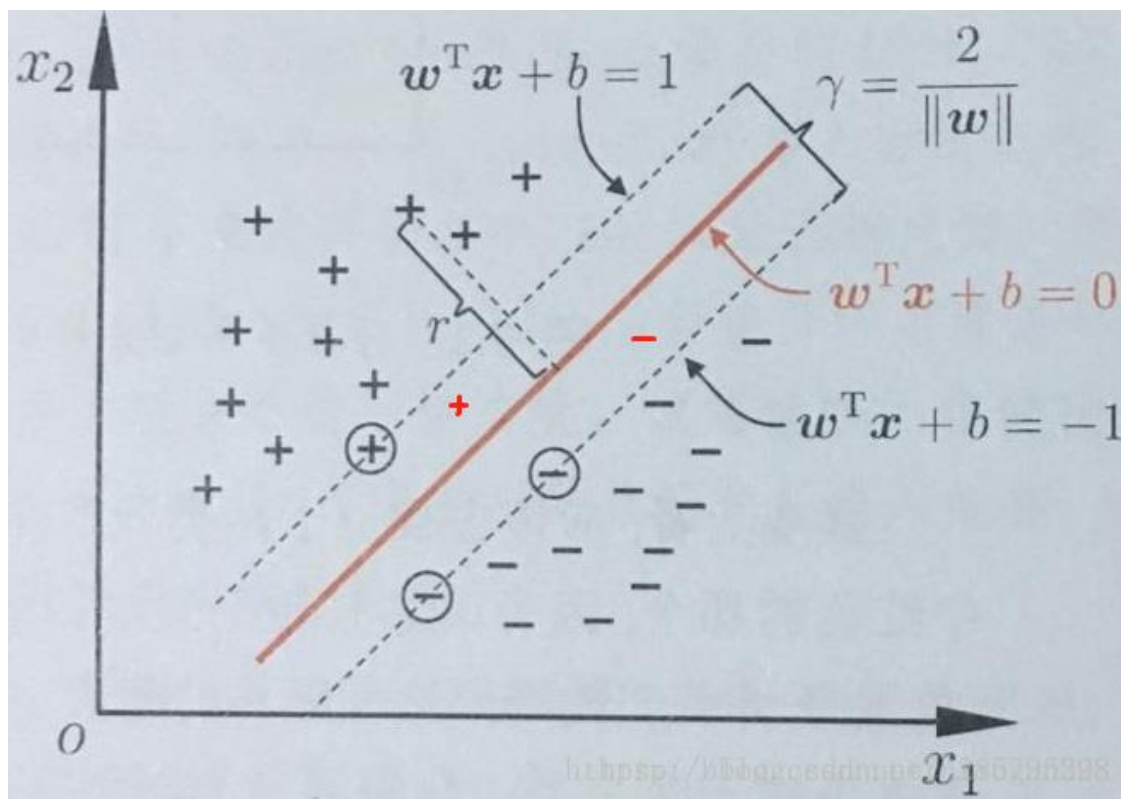


图 6.1 存在多个划分超平面将两类训练样本分开

https://blog.csdn.net/qq_35992440

SVM基本思想



$$\begin{cases} w^T x_i + b \geq +1, y_i = +1; \\ w^T x_i + b \leq -1, y_i = -1 \end{cases}$$

$$\begin{aligned} & \max_{w, b} \frac{2}{\|w\|} \\ & s.t. y_i(w^T x_i + b) \geq 1, i = 1, 2, \dots, m \end{aligned}$$

这等价于

$$\begin{aligned} & \min_{w, b} \frac{1}{2} \|w\|^2 \\ & s.t. y_i(w^T x_i + b) \geq 1, i = 1, 2, \dots, m \end{aligned}$$



知识点拓展-SVM

当训练样本线性可分时，通过**硬间隔**最大化，学习一个线性可分支持向量机；

当训练样本近似线性可分时，通过**软间隔**最大化，学习一个线性支持向量机；

当训练样本线性不可分时，通过**核技巧和软间隔**最大化，学习一个非线性支持向量机；



分类算法应用场景实例二十则

<https://blog.csdn.net/liulingyuan6/article/details/53637129>

整理了20个天池、DataCastle、DataFountain等中出现的，可使用分类算法处理的问题场景实例。

国家电网客户用电异常行为分析

希望基于国家电网公司提供的关于用户用电量、电能表停走、电流失流、计量们打开灯计量异常情况、窃电行为等相关数据，以及经过现场电工人员现场确认的窃电用户清单，利用大数据分析算法与技术，发现窃电用户的行为特征，形成窃电用户行为画像，准确识别窃电用户，提高窃电监测效率，降低窃电损失。



分类方法

END



世界上最高危的职业：韩国总统

1948-1960 李承晚，流放海外。
1960-1962 尹普善，被判监禁。
1963-1979 朴正熙，被当众暗杀。
1979-1980 崔圭夏，判刑监禁。
1980-1988 全斗焕，判无期。
1988-1993 卢泰愚，判刑监禁。
1993-1998 金泳三，驱逐出境。
1998-2003 金大中，判刑监禁。
2003-2008 卢武铉，被查处时自杀。
2008-2013 李明博。已逮捕，待审
2013- 2017 朴槿惠，弹劾下台判刑
2017-今 文在寅 ?

如不好好工作，就让你
去韩国做总统！！

内涵段子



李明博（第17届）：唯一“例外” 安享晚年



正态分布：我是正太，不是变态

特征工程：锦上添花

PCA：浓缩的都是精品

线性回归：

OLS——我是耿直的直男

RR ——我比较圆滑

PCR ——人无远虑必有近忧

PLS ——两手都要抓，两手都要硬

聚类：物以类聚，人以群分

K近邻——近朱者赤近墨者黑

判别：你的孤独，是因为你找不到归属感

FDA——快刀斩乱麻

朴素贝叶斯——打雷啦，收衣服啊