第七章

贝叶斯决策理论

汪小我、张学工

清华大学自动化系

2020年春季学期

引言

统计模式识别：

用概率统计的观点和方法来解决模式识别问题

基本概念：

◆样本(sample) $\mathbf{x} \in R^d$

◆状态(state)　第一类：$\omega = \omega_1$，　第二类：$\omega = \omega_2$

◆先验概率 (*a priori* probablity or prior) $P(\omega_1)$，$P(\omega_2)$

◆样本分布密度(sample distribution density) $p(\mathbf{x})$

（总体概率密度）

◆ 类条件概率密度(class-conditional probablity density) :

$$p(\mathbf{x} \mid \omega_1), \quad p(\mathbf{x} \mid \omega_2)$$

◆后验概率(*a posteriori* probablity or posterior) :

$$P(\omega_1 \mid \mathbf{x}),\quad P(\omega_2 \mid \mathbf{x})$$

◆错误概率(probablity of error) :

$$P(e \mid \mathbf{x}) = \begin{cases} P(\omega_2 \mid \mathbf{x}) & \text{if } \mathbf{x} \text{ is assigned to } \omega_1 \\ P(\omega_1 \mid \mathbf{x}) & \text{if } \mathbf{x} \text{ is assigned to } \omega_2 \end{cases}$$

◆平均错误率(average probablity of error):

$$P(e) = \int P(e \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

◆正确率(proabality of correctness):

$$P(c) = 1 - P(e)$$

贝叶斯决策（统计决策理论）

是统计模式识别的基本方法和基础。

是“最优分类器”：使平均错误率最小

条件：

① 类别数一定， $\omega_i, i = 1, \ldots, c$

② 已知类先验概率和类条件概率密度不太容易满足

$$P(\omega_i),\ P(x \mid \omega_i),\ i = 1, \ldots, c$$

最小错误率贝叶斯决策

$$\min \quad \boldsymbol{P(e)=\int P(e\,|\,x)\,p(x)dx}$$

因为 $P(e|x)\geq 0$，$p(x)\geq 0$，所以上式等价于：$\min P(e|x)$ for all x .

而 $P(e\,|\,x)=\begin{cases} P(\omega_2\,|\,x) & \text{if assign } x\in\omega_1 \\ P(\omega_1\,|\,x) & \text{if assign } x\in\omega_2 \end{cases}$

$$\therefore \quad \boldsymbol{if} \quad \boldsymbol{P(\omega_1\,|\,x)} \gtrless \boldsymbol{P(\omega_2\,|\,x)}, \quad \textbf{assign} \quad \begin{matrix} \boldsymbol{x\in\omega_1} \\ \boldsymbol{x\in\omega_2} \end{matrix}$$

------ 最小错误率贝叶斯决策，简称**贝叶斯决策**

如何计算后验概率？

已知 $P(\omega_i)$， $p(x \mid \omega_i)$， $i = 1,2$

贝叶斯公式：（Bayes' Theorem）

$$P(\omega_i | x) = \frac{p(x \mid \omega_i) P(\omega_i)}{p(x)} = \frac{p(x \mid \omega_i) P(\omega_i)}{\sum_{j=1}^{2} p(x \mid \omega_j) P(\omega_j)},$$

分母上的$p(x)$与类别无关，对于决策没有影响，通常不用算

$$\therefore \quad \textbf{\textit{if}} \quad p(x|\omega_1)P(\omega_1) \underset{<}{>} p(x|\omega_2)P(\omega_2) \quad \textbf{assign} \quad \begin{matrix} x \in \omega_1 \\ x \in \omega_2 \end{matrix}$$

最小错误率贝叶斯决策规则的几种等价表达式：

(1) If $P(\omega_i \mid x) = \max\limits_{j=1,2} P(\omega_j \mid x)$, then $x \in \omega_i$

(2) If $p(x \mid \omega_i)P(\omega_i) = \max\limits_{j=1,2} p(x \mid \omega_j)P(\omega_j)$, then $x \in \omega_i$

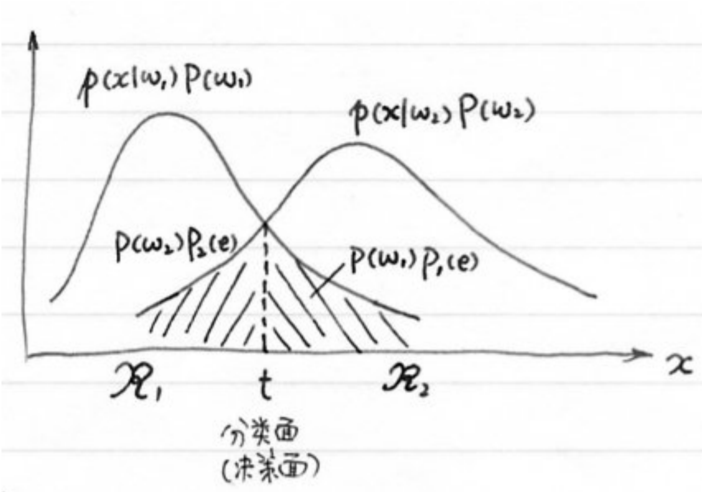
(3) If $l(x) = \dfrac{p(x \mid \omega_1)}{p(x \mid \omega_2)} \begin{matrix} > \\ < \end{matrix} \dfrac{P(\omega_2)}{P(\omega_1)}$, then $x \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$

(4) 定义 $h(x) = -\ln[l(x)] = -\ln p(x \mid \omega_1) + \ln p(x \mid \omega_2)$

If $h(x) \begin{matrix} \overset{<}{\cancel{>}} \\ \overset{>}{\cancel{<}} \end{matrix} \ln\left(\dfrac{P(\omega_1)}{P(\omega_2)}\right)$, then $x \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$

其中，$l(x)$：似然比， $\dfrac{P(\omega_1)}{P(\omega_2)}$：似然比阈值， $h(x)$：对数似然比

图示：



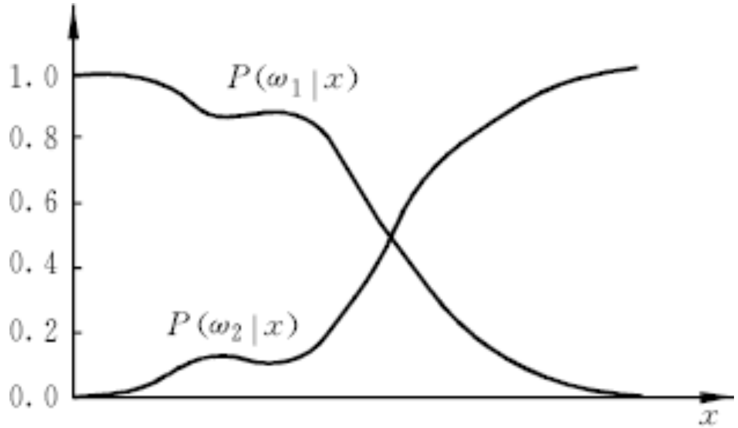


图2.2 后验概率

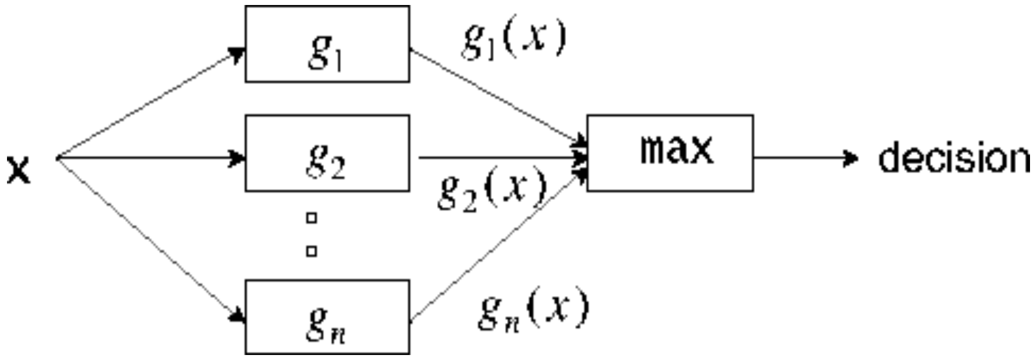
错误率：

$$
\begin{aligned}
P(e) &= P(\omega_2)P_2(e) + P(\omega_1)P_1(e) \\
&= P(\omega_2)\int_{\Re 1} p(x \mid \omega_2)dx + P(\omega_1)\int_{\Re 2} p(x \mid \omega_1)dx
\end{aligned}
$$

扩展到多类情况：

（1） If $P(\omega_i \mid x) = \max\limits_{j=1,\dots,c} P(\omega_j \mid x)$, then $x \in \omega_i$

（2） If $p(x \mid \omega_i) P(\omega_i) = \max\limits_{j=1,\dots,c} p(x \mid \omega_j) P(\omega_j)$, then $x \in \omega_i$



错误率： $P(e) = 1 - P(c) = 1 - \sum_{j=1}^{c} p(\omega_j) \int_{R_j} p(x \mid \omega_j)\, dx$

最小风险贝叶斯决策

- 最小错误率只考虑了错误
- 进一步可考虑不同错误所带来的损失（代价）

用决策论方法把问题表述如下：

（1）把样本x看作d维随机向量 $x=[x_1,x_2,\cdots,x_d]^T$

（2）状态空间Ω由c个可能的状态（c类）组成：$\Omega=\{\omega_1,\omega_2,\ldots,\omega_c\}$

（3）对随机向量x可能采取的决策组成了决策空间，它由k个决策组成：

$$A=\{\alpha_1,\alpha_2,\ldots,\alpha_k\}$$

（4）对于实际状态为ω_j的向量 x，采取决策α_i所带来的损失为

$$\lambda(\alpha_i,\omega_j),\ i=1,\ldots,k,\quad j=1,\ldots,c,$$

形成**损失函数**。

对于实际问题，损失函数通常以**决策表**形式给出。

$\alpha \backslash \omega$	ω_1	$\cdots$	ω_c
α_1	$\lambda(\alpha_1,\omega_1)$	$\cdots$	$\lambda(\alpha_1,\omega_c)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
α_k	$\lambda(\alpha_k,\omega_1)$	$\cdots$	$\lambda(\alpha_k,\omega_c)$

条件期望损失：对于特定的 $\mathbf{x}$ 采取决策α_i的期望损失：

$$R(\alpha_i \mid x)=E\left[\lambda(\alpha_i,\omega_j)\mid x\right]=\sum_{j=1}^{c}\lambda(\alpha_i,\omega_j)P(\omega_j \mid x),\ i=1,\cdots,k$$

期望风险：

对所有可能的 x 采取决策 $\alpha(x)$ 所造成的期望损失之和。

也称平均风险。 $$R(\alpha) = \int R(\alpha(x) \mid x) p(x) dx$$

最小风险决策： $\min R(\alpha)$ -- 期望风险最小化

对所有 x，使 $R(\alpha(x) \mid x)$ 最小，则可以使 $R(\alpha)$ 最小，因此有：

最小风险贝叶斯决策规则：

$$\text{if } R(\alpha_i \mid x) = \min_{j=1,\dots,k} R(\alpha_j \mid x),$$

$$\text{then } \alpha = \alpha_i$$

$$\text{if } R(\alpha_i \mid x) = \min_{j=1,\ldots,k} R(\alpha_j \mid x), \quad \text{then } \alpha = \alpha_i$$

计算：可采取以下步骤（对于给定的样本 x）：

（1）计算后验概率：$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j)P(\omega_j)}{\sum_{i=1}^{c} p(x \mid \omega_i)P(\omega_i)} \qquad j = 1, \ldots, c$$

（2）计算风险：$$R(\alpha_i \mid x) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j)P(\omega_j \mid x), \ i = 1, \ldots, k$$

（3）决策：$$\alpha = \arg\min_{i=1,\ldots,k} R(\alpha_i \mid x)$$

两类情况：（损失函数表λ_{11}，λ_{12}，λ_{21}，λ_{22}）

$$\lambda_{11}P(\omega_1 \mid x)+\lambda_{12}P(\omega_2 \mid x) \mathop{\gtrless}^{<}_{>} \lambda_{21}P(\omega_1 \mid x)+\lambda_{22}P(\omega_2 \mid x), \quad \text{then} \quad x \in \begin{cases}\omega_1 \\ \omega_2\end{cases}$$

(划为第一类的损失）　　　(划为第二类的损失）

显然，当$\lambda_{11}=\lambda_{22}=0$，$\lambda_{12}=\lambda_{21}=1$时，最小风险就是最小错误率。

例子：癌细胞识别　（课本例 2.1 和 2.2）

正常细胞：$P(\omega_1)=0.9$　　　癌细胞：$P(\omega_2)=0.1$

观察结果为x, 从类条件概率密度曲线上查得：

$$P(x|\omega_1)=0.2 \qquad P(x|\omega_2)=0.4$$

利用贝叶斯公式计算：　$P(x|\omega_1)\,P(\omega_1)= 0.2*0.9 =0.18$

$P(x|\omega_2)\,P(\omega_2) = 0.4*0.1 =0.04$

根据最小错误率准则，决策为第一类

决策表

	ω_1	ω_2
α_i	0	6
α_2	1	0

根据损失函数计算条件风险：

$$R(\alpha_1 \mid x) =1.092$$
$$R(\alpha_2 \mid x) =0.818$$

根据最小风险准则，决策为第二类

请认真学习体会同样数据情况下，不同的损失会导致不同的决策。

问题：Fisher 线性判别如何选分界面？

正态分布时的统计决策

为什么研究正态分布？

---- 简单，且较符合很多实际情况。

正态分布的知识回顾

单变量： $$p(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

$$\mu=E(x)=\int xp(x)dx\,,\qquad \sigma^2=\int(x-\mu)^2p(x)dx=E\left\{(x-\mu)^2\right\}$$

记作 $N(\mu,\sigma^2)$

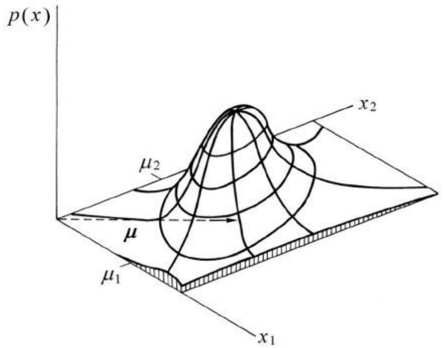
多变量：

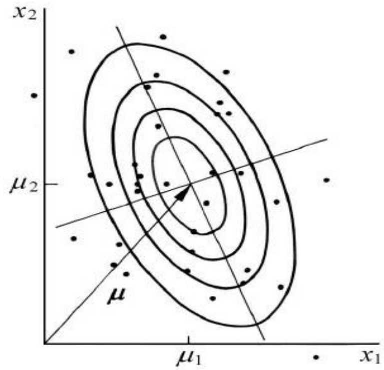
$$p(x)=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\},\quad x\in R^d$$

$\mu=E[x]$ 均值向量

$\Sigma=E[(x-\mu)(x-\mu)^T]$ 协方差矩阵（$d\times d$），对角线元素为方差

记作 $N(\mu,\Sigma)$





正态分布下的贝叶斯决策

$$p(x \mid \omega_i) = \frac{1}{(2\pi)^{2/d} \mid \sum_i \mid^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^T \sum_i^{-1} (x-\mu_i) \right]$$

考虑判别函数（对数后验概率）

$$g_i(x) = \ln[p(x \mid \omega_i) P(\omega_i)] = \ln p(x \mid \omega_i) + \ln P(\omega_i)$$

$$= -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln \mid \sum_i \mid - \frac{1}{2}(x-\mu_i)^T \sum_i^{-1} (x-\mu_i) + \ln P(\omega_i)$$

决策面方程 $g_i(x) = g_j(x)$

$$-\frac{1}{2}\left[(x-\mu_i)^T \textstyle\sum_i^{-1}(x-\mu_i) - (x-\mu_j)^T \sum_j^{-1}(x-\mu_j)\right] - \frac{1}{2}\ln\frac{\mid \sum_i \mid}{\mid \sum_j \mid} + \ln\frac{P(\omega_i)}{P(\omega_j)} \quad = \quad 0$$

下面研究一些特殊情况：

（一）$\Sigma_i = \sigma^2 I$，$i = 1, \ldots, c$ （各类协方差阵相等，且各特征独立，方差相等）

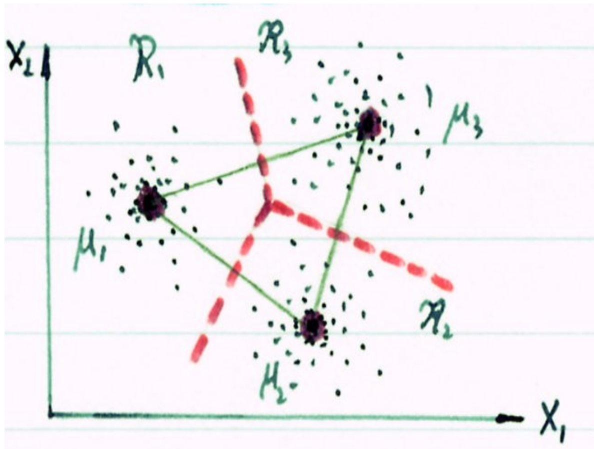
（1）如果$P(\omega_i)$，$i = 1, \ldots, c$ 相等，略去判别函数中与类别无关的项，得

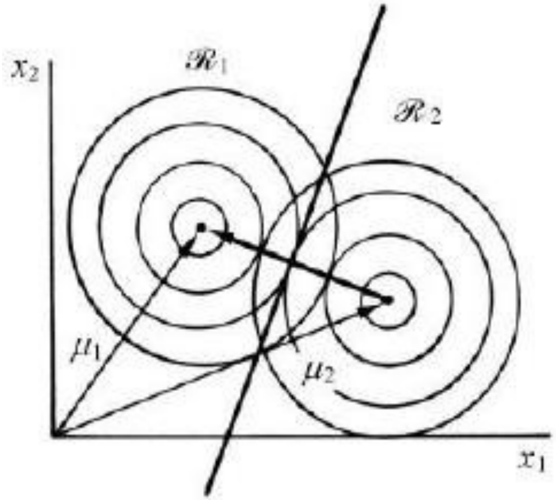
$$g_i(x) = -\frac{1}{2\sigma^2}(x-\mu_i)^T(x-\mu_i) = -\frac{1}{2\sigma^2}\left\|x-\mu_i\right\|^2$$

$$= -\frac{1}{2\sigma^2}\sum_{j=1}^{d}(x_j-\mu_{ij})^2$$

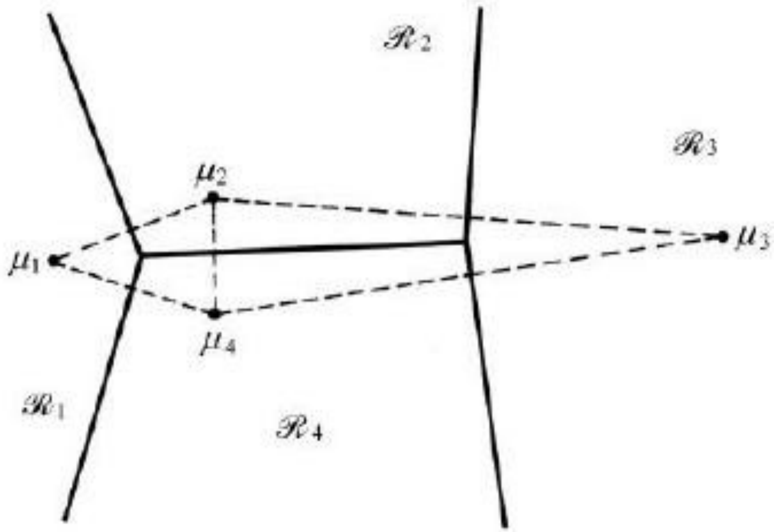
球状分布，各类先验概率相等，则分类只取决于样本到各类中心的距离。

—— 最小距离分类器，模板匹配




x_2
$\mathscr{R}_1$
$\mathscr{R}_2$
μ_1
μ_2
x_1

(a) 两类情况


$\mathscr{R}_2$
$\mathscr{R}_3$
μ_2
μ_1
μ_3
μ_4
$\mathscr{R}_1$
$\mathscr{R}_4$

(b) 多类情况

(2) 如果$P(\omega_i)$， $i=1,\ldots,c$不相等，得

$$g_i(x)=-\frac{1}{2\sigma^2}(x-\mu_i)^T(x-\mu_i)+\ln P(\omega_i)$$

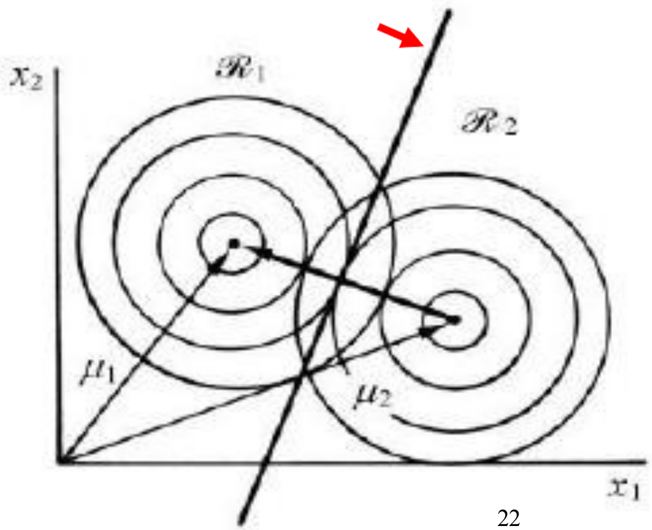
再略去与i无关的项 $x^T x$（g_i和 g_j里都有这一项），整理可得

$$g_i(x)=\boldsymbol{w}_i^T x+b_i$$ —— 线性判别函数

其中 $\boldsymbol{w}_i=\frac{1}{\sigma^2}\mu_i$

$$b_i=-\frac{1}{2\sigma^2}\mu_i^T\mu_i+\ln P(\omega_i)$$

决策面向先验概率小的方向偏移



（二）$\Sigma_i=\Sigma,\quad i=1,\ldots,c$ 各类协方差阵相等

$$g_i(x)=-\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i)+\ln P(\omega_i)$$

x 到 μ_i 的 Mahalanobis 距离（马氏距离）的平方

若 $P(\omega_i)$ 相等，则分类取决于样本到类中心的 Mahalanobis 距离。

Prasanta Chandra Mahalanobis
1893-1972
印度统计之父

一般情况下，可得

$$g_i(x)=-\frac{1}{2}(x^T\Sigma^{-1}x-\mu_i^T\Sigma^{-1}x-x^T\Sigma^{-1}\mu_i+\mu_i^T\Sigma^{-1}\mu_i)+\ln P(\omega_i)$$

略去 $x^T\Sigma^{-1}x$ 项，得

$$g_i(x)=w_i^Tx+b_i$$ ——线性判别函数

其中， $w_i = \Sigma^{-1}\mu_i \qquad b_i = -\frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i + \ln P(\omega_i)$

决策面方程： $g_i(x) = g_j(x)$

可写为 $w^T(x - x_0) = 0$

其中 $w = \Sigma^{-1}(\mu_i - \mu_j)$

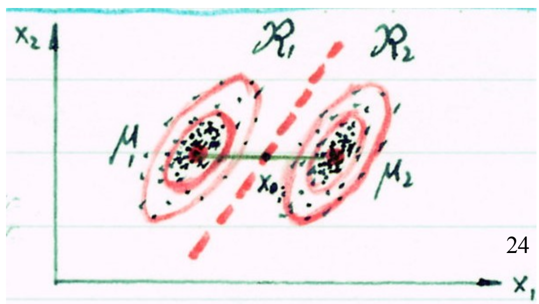
$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln(P(\omega_i)/P(\omega_j))}{(\mu_i - \mu_j)^T \Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)$$

μ_i 到 μ_j 的马氏距离平方。

当 $P(\omega_i) = P(\omega_j)$时，$x_0 = \frac{1}{2}\left(\mu_i + \mu_j\right)$

Mahalanobis 距离考虑了方差因素。

当 $\Sigma = I$ 时就是欧氏距离。



（三）一般情况，各类协方差不同

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

$$= x^T W_i x + w_i^T x + w_{i0}$$

其中 $W_i = -\frac{1}{2}\Sigma_i^{-1}$； $w_i = \Sigma_i^{-1}\mu_i$； $w_{i0} = -\frac{1}{2}\mu_i^T \Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$

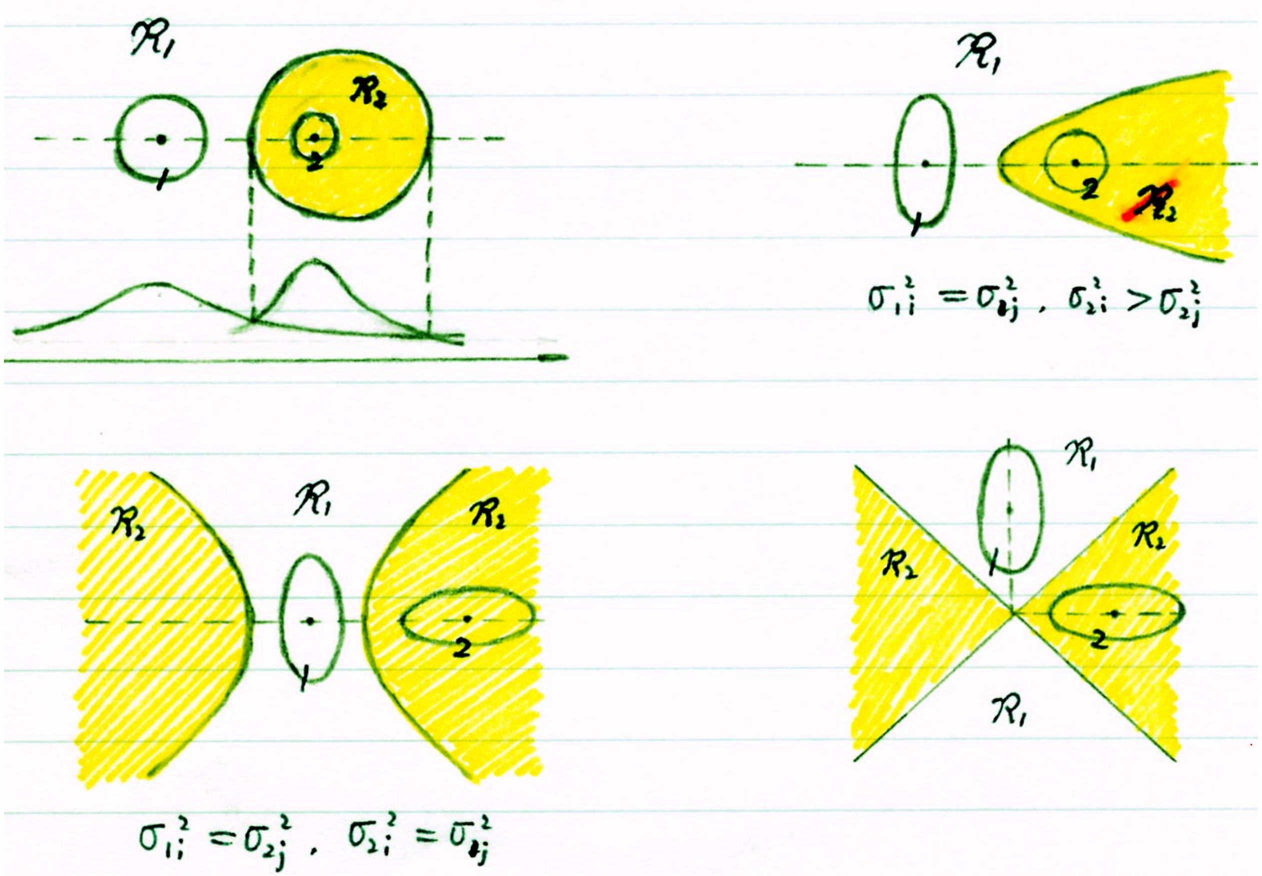
决策面 $g_i(x) = g_j(x)$,

$$x^T(W_i - W_j)x + (w_i - w_j)^T x + w_{i0} - w_{j0} = 0$$

为超二次曲面

举例：二维常见情况，

$x_{1,} x_2$相互独立， $P(\omega_i) = P(\omega_j)$



小结

- 贝叶斯决策理论是统计模式识别的重要理论基础
- 理论上讲，贝叶斯决策方法是最优的（在最小错误率或最小风险意义上）
- 应用中：需要首先得到先验概率和类条件概率密度
  - 方法一： 先估计概率密度，后求解决策规则
  - 方法二： 若已知或可假设概率密度为某种形式（比如正态分布），可先求出判决函数形式，再从样本估计其中的参数。
  - 方法三： 直接选择或假设某种判决函数形式，用样本确定其参数。