第三章　贝叶斯决策理论

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：

类别数一定， $\omega_1, i=1,\cdots,c$

已知类先验概率和类条件概率密度 $P(\omega_i)$, $P(x\,|\,\omega_i), i=1,\cdots,c$

最小错误率贝叶斯决策

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

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

$\therefore$ **if** $\boldsymbol{P(\omega_1\,|\,x)}\ \begin{matrix} > \\ < \end{matrix}\ \boldsymbol{P(\omega_2\,|\,x)}$, **assign** $\begin{matrix} \boldsymbol{x\in\omega_1} \\ \boldsymbol{x\in\omega_2} \end{matrix}$

------ 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）

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

$$i=1,2$$

If $\boldsymbol{P}(\boldsymbol{\omega_1} \mid \boldsymbol{x}) \underset{<}{>} \boldsymbol{P}(\boldsymbol{\omega_2} \mid \boldsymbol{x})$, then assign $\begin{matrix} \boldsymbol{x} \in \boldsymbol{\omega}_1 \\ \boldsymbol{x} \in \boldsymbol{\omega}_2 \end{matrix}$

最小错误率贝叶斯决策规则的几种等价表达形式：

（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 $x \in \omega_i$

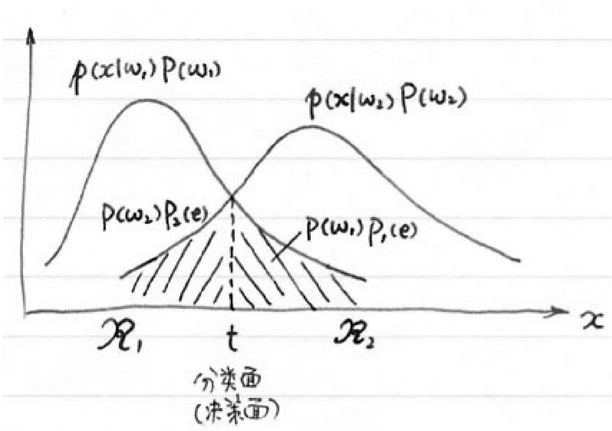
（3） If $l(x) = \dfrac{p(x \mid \omega_1)}{p(x \mid \omega_2)} \mathop{\lessgtr}\limits^{>}_{<} \dfrac{P(\omega_2)}{P(\omega_1)}$, then $x \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$

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

If $h(x) \mathop{\lessgtr}\limits^{<}_{>} \ln\left(\dfrac{P(\omega_1)}{P(\omega_2)}\right)$, 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：



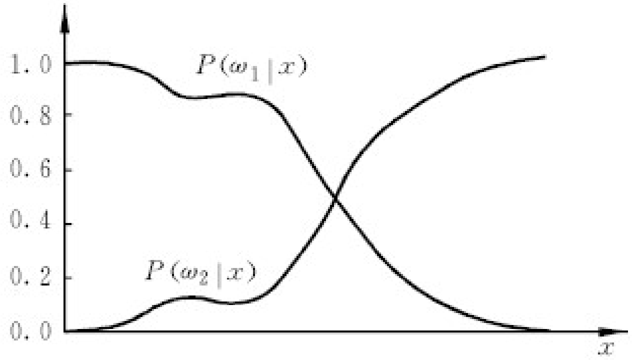


图 2.2 后验概率

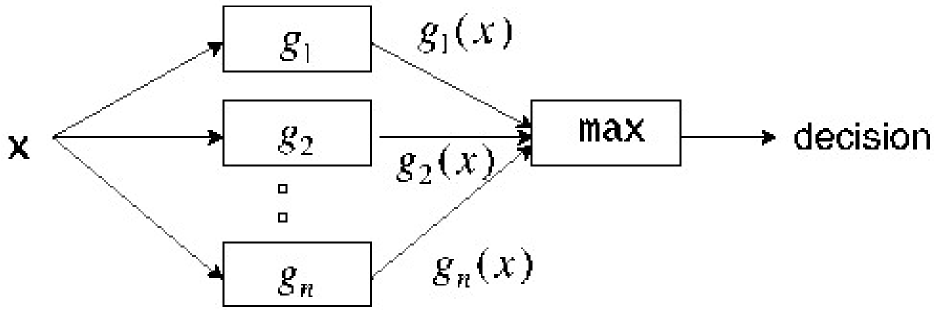
错误率：

$$P(e) = P(\omega_2)P_2(e) + P(\omega_1)P_1(e)$$
$$= P(\omega_2)\int_{\Re_1} p(x \mid \omega_2)dx + P(\omega_1)\int_{\Re_2} p(x \mid \omega_1)dx$$

多类情况：

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

（2） If $p(x \mid \omega_i)P(\omega_i) = \max\limits_{j=1,\cdots,c} p(x \mid \omega_j)P(\omega_j)$， 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：$\Omega=\{\omega_1,\omega_2,\cdots,\omega_c\}$

（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：

$$\mathcal{A}=\{\alpha_1,\alpha_2,\cdots,\alpha_k\}$$

（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

$$\lambda(\alpha_1,\omega_j),\ i=1,\cdots,k,\ \ j=1,\cdots,c\ ,$$

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。

λ \ ω / α	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。

$$R(\alpha) = \int R(\alpha(x) \mid x) p(x) dx$$

也称平均风险。

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：

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

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

计算：可采取以下步骤（对于给定的样本 x ）：

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

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

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

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

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

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

请认真学习课本例 2.1 和 2.2，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？

朴素贝叶斯分类器（Naïve Bayes）

多维特征情况下的

似然函数：$p(x_1, x_2, \cdots, x_d \mid \omega_i)$　　　后验概率：$p(\omega_i \mid x_1, x_2, \cdots, x_d)$

对于多维的特征变量情况下的联合概率，利用链式法则

$$p(x_1, x_2, \cdots, x_d, \omega_i) = p(x_1 \mid x_2, \cdots, x_d, \omega_i) p(x_2 \mid x_3, \cdots, x_d, \omega_i) \cdots p(x_d \mid \omega_i) p(\omega_i)$$

朴素贝叶斯分类器假设特征之间条件于类别独立,上式化简为：

$$p(x_1, x_2, \cdots, x_d, \omega_i) = p(x_1 \mid \omega_i) p(x_2 \mid \omega_i) \cdots p(x_d \mid \omega_i) p(\omega_i)$$

决策函数：$\hat{\omega} = \arg\max_{i} p(\omega_i) \prod_{k=1}^{d} p(x_k \mid \omega_i)$

例子：https://en.wikipedia.org/wiki/Naive_Bayes_classifier

1. 垃圾邮件的分类

2. 男女生的分类

Gender	Height (feet)	weight (lbs)	Foot size (inches)
male	6	180	12
male	5.92	190	11
male	5.58	170	12
male	5.92	165	10
female	5	100	6
female	5.5	150	8
female	5.42	130	7
female	5.75	150	9

测试样本：

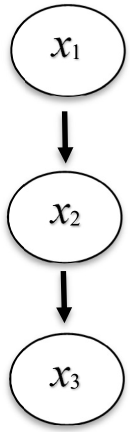
???	6	130	8

*贝叶斯网络(Bayesian network)

- 概率图模型（Graphical Model），用图来表示随机变量之间的关联
- 贝叶斯网络是一种有向无环图（DAG：Directed acyclic graph）
- 节点表示随机变量，边表示依赖关系，箭头表示“因果”
- 利用条件独立大大简化计算量和推理（inference）过程

$$p(x_1x_2x_3) = p(x_3|x_1x_2)p(x_2|x_1)p(x_1)$$
$$= p(x_3|x_2)p(x_2|x_1)p(x_1)$$

$$x_3 \perp x_1|x_2$$



例子：普通感冒与流感

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 $N(\mu,\sigma^2)$

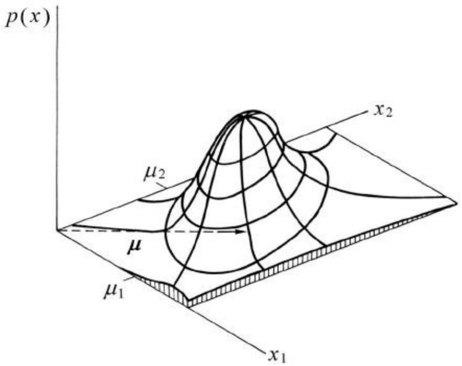
多变量：

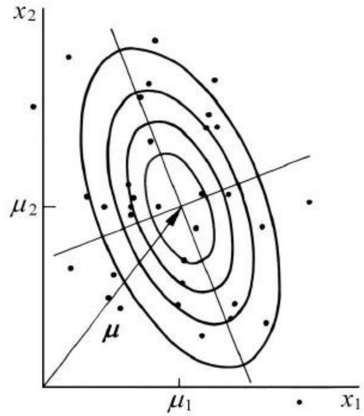
$$p(x)=\frac{1}{(2\pi)^{d/2}\,|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\},\quad x\in R^d$$

$\mu = E[x]$ 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

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

考虑判别函数

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

$$= -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\sum\nolimits_i\right| - \frac{1}{2}(x-\mu_i)^T \sum\nolimits_i^{-1}(x-\mu_i) + \ln P(\omega_i)$$

决策面方程　$g_i(x) = g_j(x)$

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

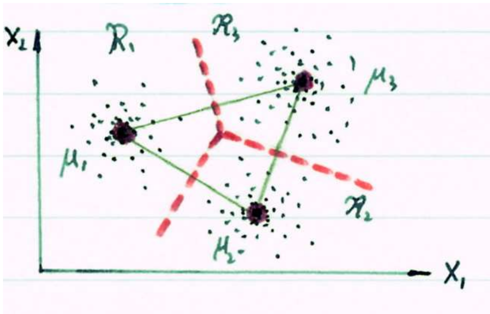
下面研究一些特殊情况：

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

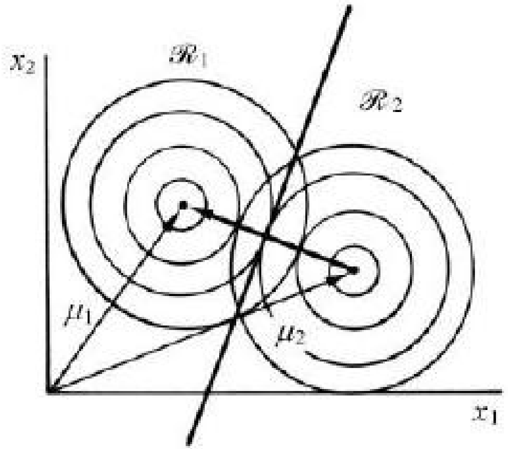
- 如果 $P(\omega_i)$， $i = 1, \cdots, c$ 相等，略去判别函数中与类别无关的项，得

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

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

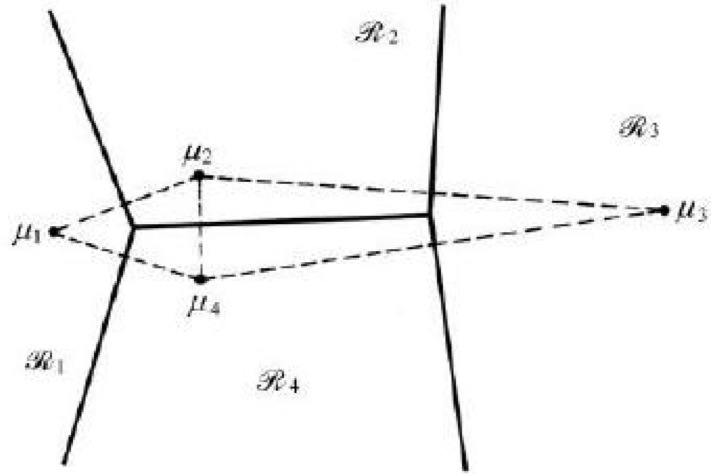


球状分布，各类先验概率相等，则分类只取决于样本到各类中心的距离。 —— **最小距离分类器，模板匹配**



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

- 如果 $P(\omega_i)$，$i=1,\cdots,c$ 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

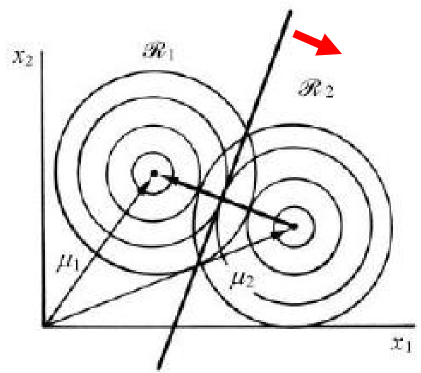
$$g_i(x) = \boldsymbol{w}_i^T x + b_i$$ —— 线性判别函数

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）$\Sigma_i = \Sigma$，$i = 1, \cdots, c$　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，记 γ^2

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

一般，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

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

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

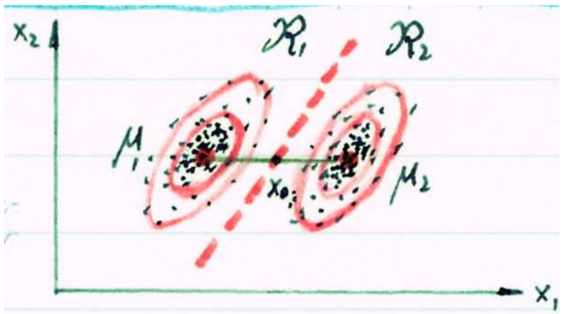
决策面方程： $g_i(x) = g_j(x)$

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，$x_0 = \frac{1}{2}(\mu_i + \mu_j)$

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

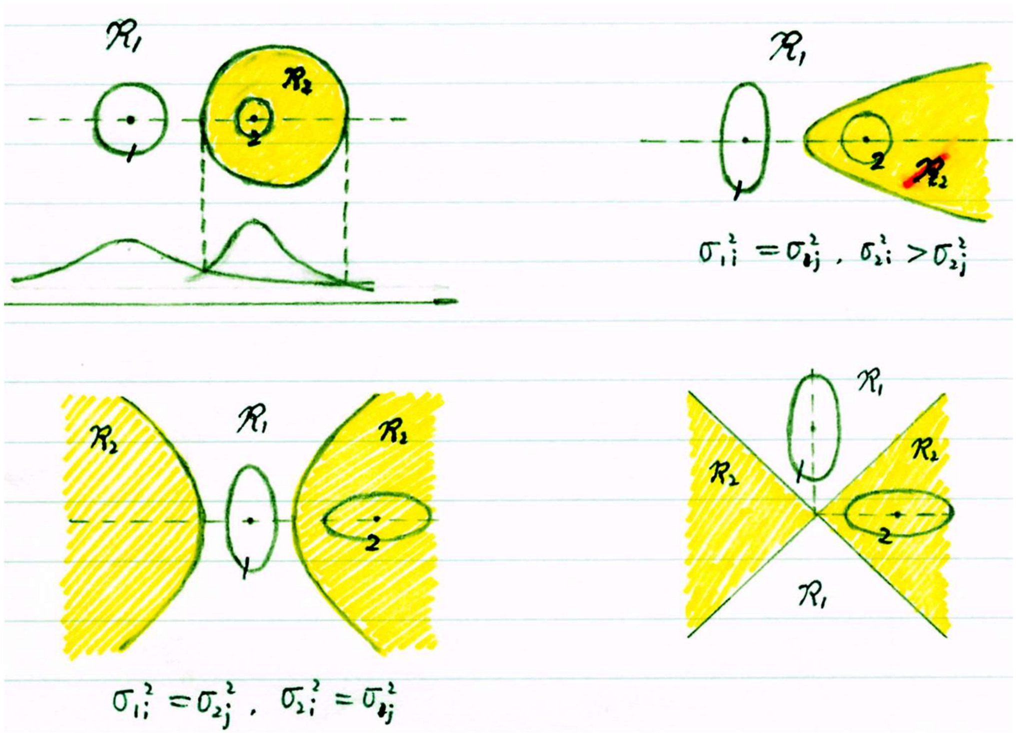
决策面 $g_i(x) = g_j(x)$,

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

为超二次曲面

举例：二维常见情况，

x_1, x_2 相互独立，$P(\omega_i) = P(\omega_j)$，$\sigma_{1j}^2$，$\sigma_{2j}^2$，$\sigma_{1i}^2$，$\sigma_{2i}^2$ 已知



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。