Chapter 2

Bayesian Decision Theory– 贝叶斯决策论

要点：

- 重点掌握贝叶斯决策论、最小误差率分类规则、分类器与判别函数、正态密度、正态分布的判别函数
- 了解贝叶斯决策论(离散性特征)

• 2.1 引 言

中山大學

■ 贝叶斯决策是统计模式识别的基本方法, 采用概率的形式来描述，它的前提是:

(1). 各类别的总体概率分布是已知的.

(2). 要决策分类的类别数是一定的.

➢ 例如：对于鲑鱼与鲈鱼的2类问题，如果用 ω 表示类别状态，那么当 $\omega = \omega_1$ 时是鲈鱼，当 $\omega = \omega_2$ 时是鲑鱼。由于每次出现的类别不确定，可以假设ω是一个用概率来描述的随机变量。

➢ 在不知道更多信息的情况下，每次出现鲈鱼的先验概率为$P(\omega_1)$，而鲑鱼的先验概率为 $P(\omega_2)$ ，其中 $P(\omega_1)+P(\omega_2)=1$

先验概率反映了在鱼没有出现之前，我们拥有可能出现鱼的类别的先验知识。

- 仅根据先验信息的判定准则

若 $P(\omega_1) > P(\omega_2)$, 则事件 ω_1 成立;
反之, 则 ω_2 成立。
错误的概率是它们之中较小的那个.
但通常不这样做!

- 利用类条件概率密度: $P(x|\omega_1)$ 及 $P(x|\omega_2)$
描述了两种鱼类外观上光泽度的差异。
其中，x为光泽度指标。

- 类条件概率密度为类别状态为ω时的x的概率密度函数

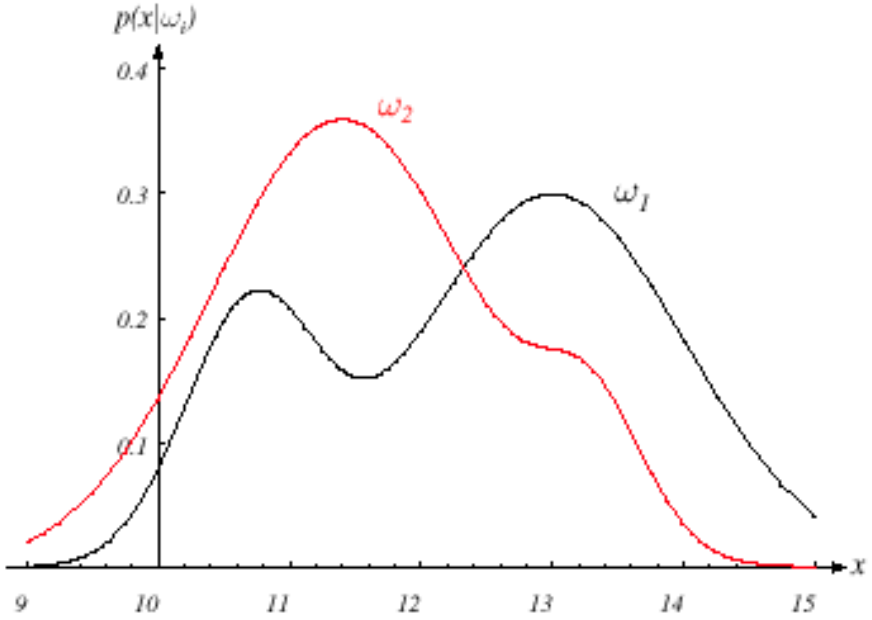


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

注：假定的类条件概率密度函数图,显示了模式处于类别 ω_i 时观察某个特定特征值 x 的概率密度.如果 x 代表了鱼的长度,那么这两条曲线可描述两种鱼的长度区别.概率函数已归一化,因此每条曲线下的面积为1

■贝叶斯公式:

处于类别 ω_i 并具有特征值 x的模式的联合概率密度可写成两种形式:

$$p(\omega_i, x) = P(\omega_i \mid x)p(x) = p(x \mid \omega_i)P(\omega_i)$$

于是，可以导出**贝叶斯公式**:

$$P(\omega_i \mid x) = \frac{p(x \mid \omega_i)P(\omega_i)}{p(x)} \qquad (1)$$

其中 $P(\omega_i \mid x)$ 称为状态的后验概率.

混合概率密度函数: $p(x) = \sum_{j=1}^{2} p(x \mid \omega_j)P(\omega_j)$

$$\text{后验概率}P(\omega_i \mid x) = \frac{\text{似然函数}p(x \mid \omega_i) \times \text{先验概率}P(\omega_i)}{\text{证据因子}p(x)}$$

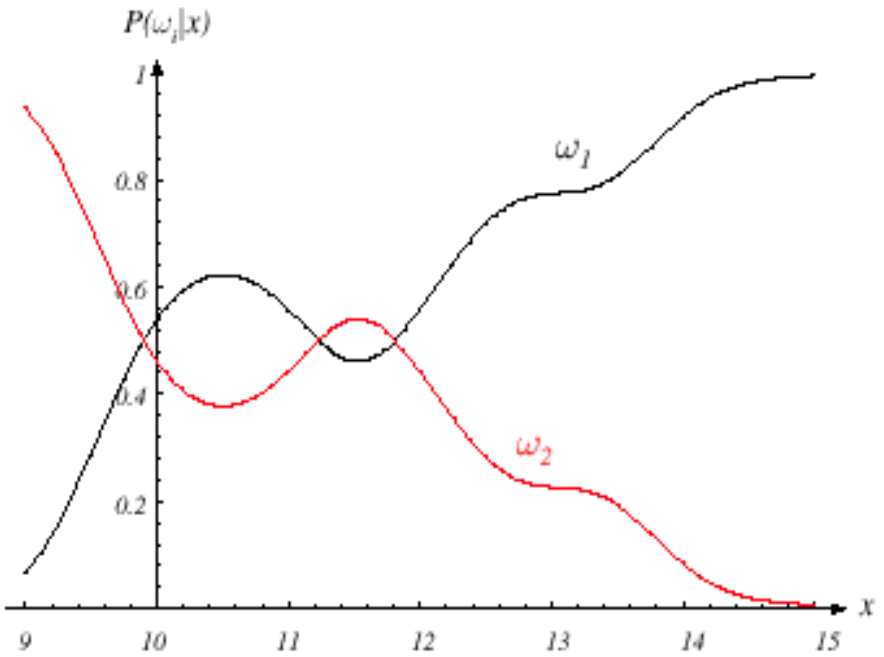


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

在先验概率 $P(w_1)=2/3, P(w_2)=1/3$ 及图2-1给出的后验概率图.此情况下,假定一个模式具有特征值 $x=14$,那么它属于 ω_2类的概率约为0.08, 属于ω_1的概率约为0.92.在每个x 处的后验概率之和为1.0

- **基于后验概率的决策准则**
  (x 表示观察值)

若 $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ ⇒ **类别判定** ω_1

若 $P(\omega_1 \mid x) < P(\omega_2 \mid x)$ ⇒ **类别判定** ω_2

- **决策后所导致的错误率**

$P(error \mid x) = P(\omega_1 \mid x)$ ⇐ **若判定** ω_2

$P(error \mid x) = P(\omega_2 \mid x)$ ⇐ **若判定** ω_1

- 最小化错误概率条件下的贝叶斯决策规则

为了追求最小的错误率，采取如下判定准则：

若 $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ ，则判定类别为 ω_1 ；

反之，判为 ω_2 。

- 可以证明, 依从这样的准则可以获得最小错误率：

$$P(error \mid x) = \min[P(\omega_1 \mid x), P(\omega_2 \mid x)]$$

我们称该准则为“贝叶斯决策准则”。

- 平均错误率：

$$P(error) = \int_{-\infty}^{\infty} P(error, x)dx = \int_{-\infty}^{\infty} P(error \mid x)\, p(x)dx$$

- 根据贝叶斯公式，由于p(x)为标量，则可以采用等价判定准则：

  若 $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$，则判定类别为 ω_1；

  反之，判为 ω_2 。

$$P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{p(x)}$$

• 2.2 贝叶斯决策论-连续性特征

中山大學

■ 概 述

1. 允许利用多于一个的特征
2. 允许多于两种类别状态的情形
3. 允许有其它行为而不仅是判定类别。
4. 引入损失函数代替误差概率。

■ 考察损失函数对判定准则的影响

令$\{\omega_1, \omega_2, \cdots, \omega_c\}$表示一系列类别状态。

令$\{\alpha_1, \alpha_2, \cdots, \alpha_a\}$表示一系列可能采取的行动（或决策）。

令 $\lambda(\alpha_i \mid \omega_j)$表示当实际状态为 ω_j 时，采取α_i 的行为会带来的风险。

那么，特征x与行动α_i 相关联的损失为： $R(\alpha_i \mid x) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$

因此，$R(\alpha_i \mid x)$ 称为条件风险。

- 借助 $R(\alpha_i \mid x)$ 可以提供一个总风险的优化过程，即遇到特征x，我们可以选择最小化风险的行为来使预期的损失达到最小。
- 假设对于特征x，决策的行为是 $\alpha(x)$ ，则总风险可表示为：

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

为了最小化总风险，对所有 $i=1,2,\ldots\ldots a$ 计算条件风险

$$R(\alpha_i \mid x)=\sum_{j=1}^{c}\lambda(\alpha_i \mid \omega_j)P(\omega_j \mid x) \qquad (12)$$

选择行为α_i，使得 $R(\alpha_i \mid x)$最小化。最小化后的总风险值称为贝叶斯风险，记为 R^*，它是可获得的最优结果。

■ 两类分类问题

行为α_1对应类别判决ω_1，α_2则对应ω_2。为了简化符号，令

$$\lambda_{i,j} = \lambda(\alpha_i \mid \omega_j)$$

那么可得两种行为的损失函数

$$R(\alpha_1 \mid x) = \lambda_{1,1}P(\omega_1 \mid x) + \lambda_{1,2}P(\omega_2 \mid x)$$
$$R(\alpha_2 \mid x) = \lambda_{2,1}P(\omega_1 \mid x) + \lambda_{2,2}P(\omega_2 \mid x)$$

■ 决策

- 按照贝叶斯决策规则，为了使得条件风险最小，

  如果 $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$ 则判为 ω_1

  相反，则判为 ω_2

- 用后验概率来表示，等价规则为

  如果 $(\lambda_{2,1} - \lambda_{1,1})P(\omega_1 \mid x) > (\lambda_{1,2} - \lambda_{2,2})P(\omega_2 \mid x)$

  则判为 ω_1 否则，判决为 ω_2

  通常：$(\lambda_{2,1} - \lambda_{1,1}) > 0$ $(\lambda_{1,2} - \lambda_{2,2}) > 0$ ？

- 结合贝叶斯公式，用先验概率与条件密度来表示后验概率，等价规则为

  如果 $(\lambda_{2,1} - \lambda_{1,1})P(x \mid \omega_1)P(\omega_1) > (\lambda_{1,2} - \lambda_{2,2})P(x \mid \omega_2)P(\omega_2)$

  则判为 ω_1 否则，判决为 ω_2

■ 决策

➢等价规则为

如果 $$\frac{P(x|\omega_1)}{P(x|\omega_2)} > \frac{(\lambda_{1,2}-\lambda_{2,2})}{(\lambda_{2,1}-\lambda_{1,1})}\frac{P(\omega_2)}{P(\omega_1)} \qquad (18)$$

则判为 ω_1 ； 否则，判决为 ω_2

注意公式(18)的右边是与x无关的常数，因此可以视为左边的似然比超过某个阈值，则判为 ω_1

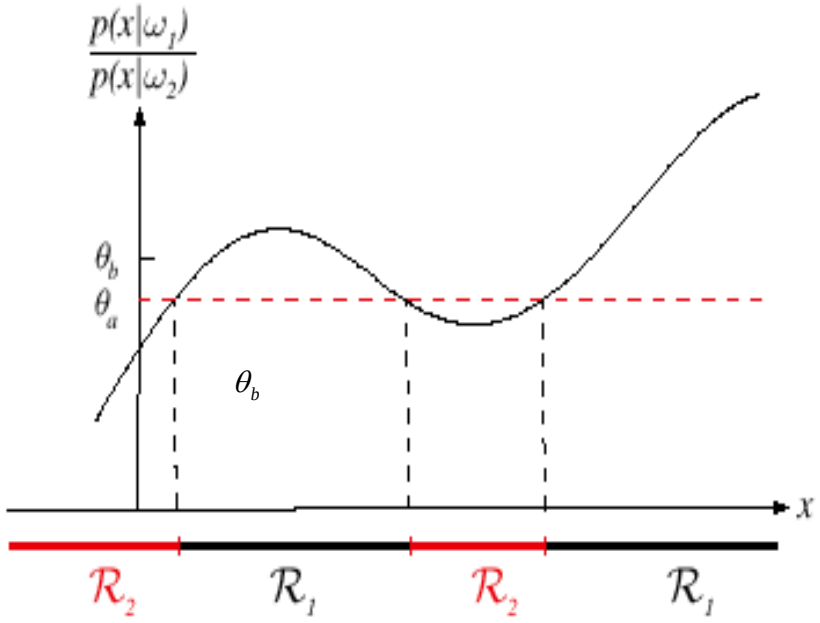


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, we get the larger threshold θ_b, and hence $\mathcal{R}_1$ becomes smaller. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

左图说明，如果引入一个0-1损失或分类损失，那么判别边界将由阈值 θ_a 决定；而如果损失函数将模式ω_2判为ω_1 的惩罚大于反过来情况，将得到较大的阈值θ_b使得R_1变小

• 2.3 最小误差率分类

- 当损失函数简化到所谓的“对称损失”或“0-1损失”函数

$$\lambda(\alpha_i \mid \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

$$i, j = 1,2,\ldots\ldots c$$

- 这个损失函数将0损失赋给一个正确的判决，而将一个单位损失赋给任何一种错误判决，因此所有误判都是等价的。与这个损失函数对应的风险就是平均误差概率。

$$R(\alpha_i \mid x) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$

$$= \sum_{j \neq i} P(\omega_j \mid x)$$

$$= 1 - P(\omega_i \mid x)$$

因此, 最小化风险, 就是最大化后验概率 $P(\omega_i \mid x)$, 即最小误差率的分类准则。

对于$i \neq j$, 若 $P(\omega_i \mid x) > P(\omega_j \mid x)$,则判定类别为 ω_i ；
反之，判为 ω_j 。

2.3.1 极小极大化准则（先验概率未知情形）

- 有时我们需要设计在整个先验概率范围内都能很好操作的分类器。一种合理的设计方法就是使先验概率取任何一种值时所引起的总风险的最坏情况尽可能小，也就是说最小化最大可能的风险。

- 我们以R_1表示分类器判为ω_1时的特征空间的区域，同样的有R_2和 ω_2，总风险的形式可表示为

$$R = \int_{R_1} \lambda_{1,1} P(\omega_1) p(x \mid \omega_1) + \lambda_{1,2} P(\omega_2) p(x \mid \omega_2)) dx \quad \text{判为}\omega_1$$

$$+ \int_{R_2} \lambda_{2,1} P(\omega_1) p(x \mid \omega_1) + \lambda_{2,2} P(\omega_2) p(x \mid \omega_2)) dx \quad \text{判为}\omega_2$$

结合公式 $P(\omega_2)=1-P(\omega_1)$ 与 $\int_{R1} p(x|\omega_1)dx=1-\int_{R2} p(x|\omega_1)dx$

可以得到

$$R(P(\omega_1))=\lambda_{2,2}+(\lambda_{1,1}-\lambda_{2,2})\int_{R1} p(x|\omega_1)dx+$$

$$P(\omega_1)\left[(\lambda_{1,1}-\lambda_{2,2})+(\lambda_{2,1}-\lambda_{1,1})\int_{R2} p(x|\omega_1)dx-(\lambda_{1,2}-\lambda_{2,2})\int_{R1} p(x|\omega_2)dx\right]$$

作业:计算

等式表明一旦判别边界确定后，总风险与 $P(\omega_1)$ 成线形关系。如果能找到一个边界使比例为0，那么风险将与先验概率独立。这就是极小极大化求解。

风险

$$R_{mm}=\lambda_{2,2}+(\lambda_{1,2}-\lambda_{2,2})\int_{R_1} p(x|\omega_2)dx$$

$$=\lambda_{1,1}+(\lambda_{2,1}-\lambda_{1,1})\int_{R_2} p(x|\omega_1)dx$$

2.3.2 Neyman-Pearson准则

- 最小化某个约束的风险（资源有限的情形）。
- 对某个给定的i，最小化在约束条件 $\int R(\alpha_1 \mid x)dx < 常数$ 的总风险。
- 例如：将鲈鱼误判为鲑鱼的误差率不得超过1%。

• 2.4 分类器与判别函数

2.4.1 多类情况

有许多方式来表述模式分类器，用的最多的是一种判别函数 $g_i(x)$ 若对于所有的 $j \neq i$ 都有

$$g_i(x) > g_j(x)$$

则分类器将这个特征向量x判给 ω_i

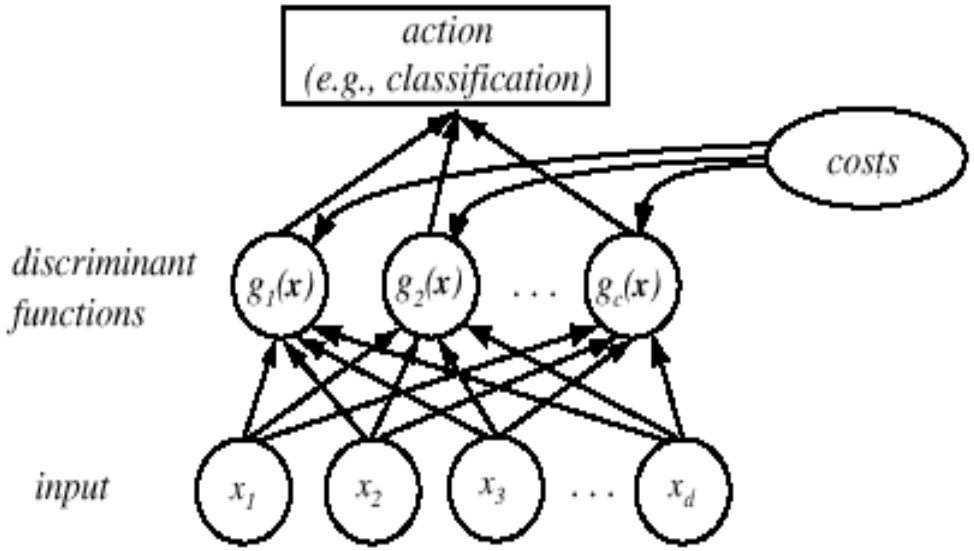


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction of the flow of information, though frequently the arrows are omitted when the direction of flow is self-evident. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

上图为包含d个输入c个判别函数的系统。确定哪个判别函数值最大，并相应地对输入作分类。

- 不同情况下的分类器的表示方式
  - 一般风险的情况下为 $g_i(x) = -R(\alpha_i \mid x)$
  - 最小误差概率情况下 $g_i(x) = P(\omega_i \mid x)$
  - 其它一些较常见的形式 $g_i(x) = p(\omega_i \mid x)P(\omega_i)$

$$g_i(x) = P(\omega_i \mid x) = \frac{p(x \mid \omega_i)P(\omega_i)}{\sum_j p(x \mid \omega_j)P(\omega_j)}$$

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

- 尽管判别函数可写成各种不同的形式，但是判决规则是相同的。每种判决规则都是将特征空间划分c个判决区域，$R_1,\ldots\ldots R_c$ 如果对于所有的 $j \neq i$ ，有 $g_i(x) > g_j(x)$ 那么x属于 R_i 。 要求我们将x分给 ω_i 。此区域由判决边界来分割，其判决边界即判决空间中使判决函数值最大的曲面。如图

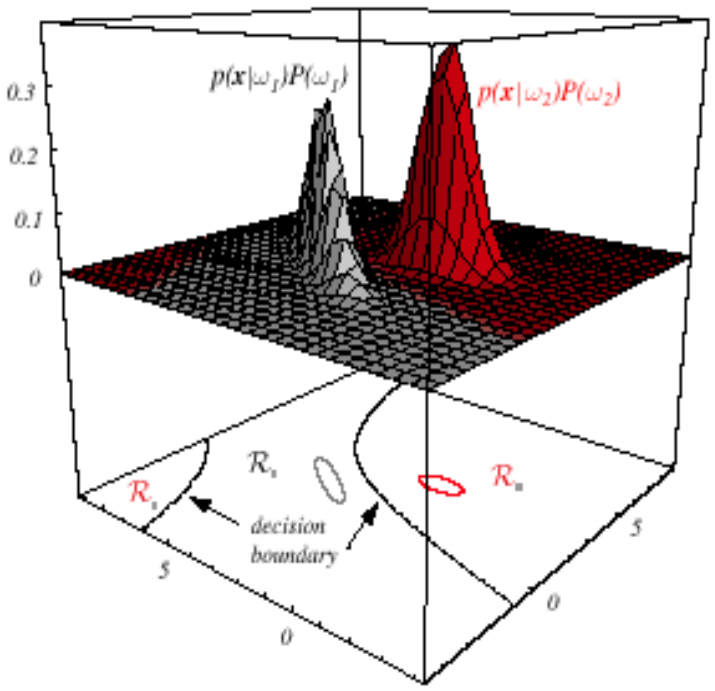


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

在这个二维的两类问题的分类器中，概率密度为高斯分布。判别边界由两个双曲面构成，因此判决区域R_2并非是简单连通的。椭圆轮廓线标记出1/e乘以概率密度的峰值。

2.4.2 两类情况（二分分类器-dichotomizer）

对于二分分类器，可以定义一个简单判别函数

$$g(x) = g_1(x) - g_2(x)$$

则如果 $g(x) > 0$ ，则将x判给 ω_1，否则给 ω_2 。

- 最小误差概率情况下 $g(x) = P(\omega_1 \mid x) - P(\omega_2 \mid x)$

或： $$g(x) = \ln \frac{p(x \mid \omega_1)}{p(x \mid \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

• 2.5 正态密度

• **单变量密度函数**

单变量正态分布

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

容易计算其期望值与方差

$$\mu=E(x)=\int_{-\infty}^{+\infty}xp(x)dx$$

$$\sigma^2=E\left[(x-\mu)^2\right]=\int_{-\infty}^{+\infty}(x-\mu)^2\,p(x)dx$$

$$p(x)\sim N(\mu,\sigma^2)$$

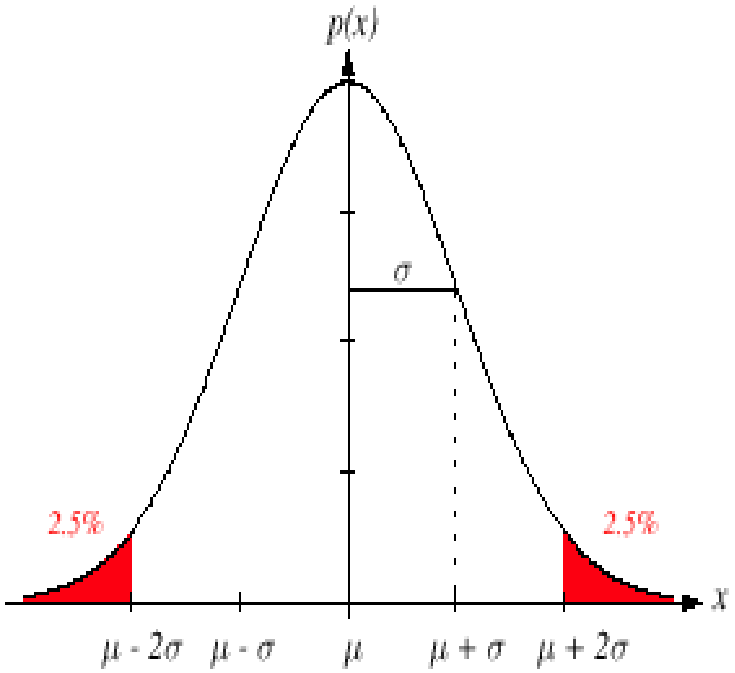


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x-\mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

单变量正态分布大约有95%的区域在 $|x-\mu| \leq 2\sigma$ **范围内，如图**

此分布的峰值为 $p(\mu) = 1/\sqrt{2\pi}\sigma$

- **正态分布与熵之间的关系**

熵的定义

$$H(p(x)) = -\int p(x)\ln p(x)dx$$

单位为奈特； 若换为 $\log_2$,单位为比特。熵是一个非负的量用来描述一种分布中随机选取的样本点的不确定性。可以证明正态分布在所有具有给定均值和方差的分布中具有最大熵。并且，如中心极限定理所述，大量的小的，独立的随机分布的总和等效为高斯分布。

- **多元密度函数**

多元正态密度

$$p(x)=\frac{1}{(2\pi)^{d/2}\left|\sum\right|^{1/2}}\exp\left[(-\frac{1}{2}(x-\mu)^{t}\sum{}^{-1}(x-\mu)\right]$$

其中x是一个d维列向量，μ是d维均值向量，$\sum$是 $d\times d$ 的协方差矩阵，$|\sum|$ 和 $\sum^{-1}$分别是其行列式的值和逆。

$p(x)\sim N(\mu,\Sigma)$ **形式上有：**

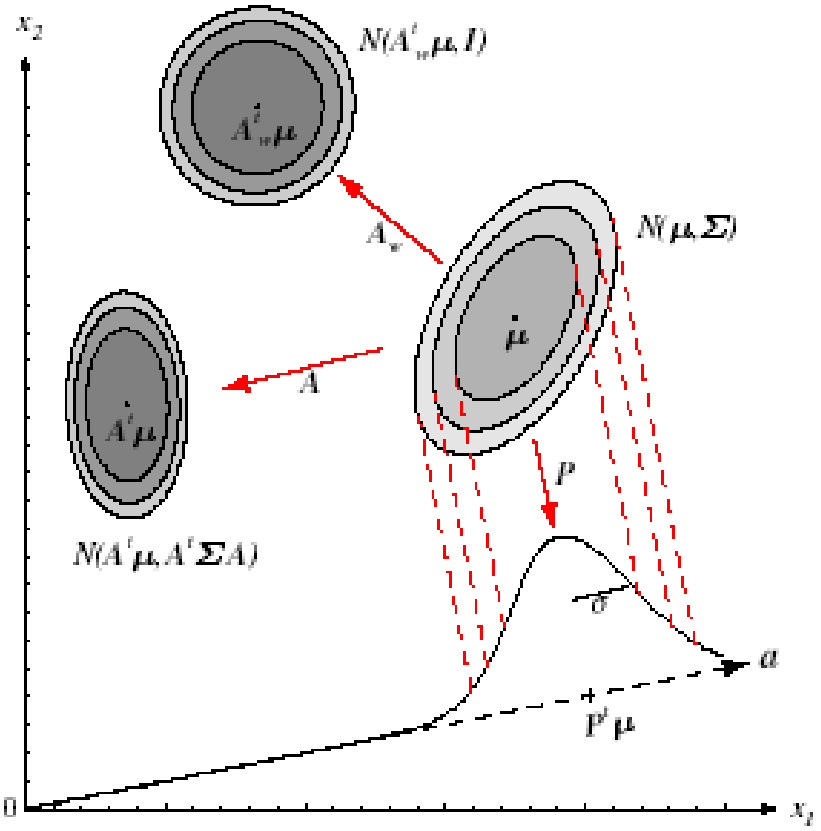
$$\mu\equiv\varepsilon[x]=\int xp(x)dx$$

$$\Sigma\equiv\varepsilon[(x-\mu)(x-\mu)']=\int(x-\mu)(x-\mu)'\,p(x)dx$$

- 协方差矩阵Σ通常是对称的且半正定。我们将严格限定 $\sum$ 是正定的。对角线元素σ_{ii}是相应的x_i方差； 非对角线元素σ_{ij}是x_i和 x_j的协方差。如果x_i和 x_j统计独立，则 $\sigma_{ij}=0$。如果所有的非对 角线元素为0，那么p(x)变成了x中各元素的单变量正态密度函数的内积。

- 服从正态分布的随机变量的线性组合，不管这些随机变量是独立还是非独立的，也是一个正态分布。(这是个非常有用的结论）

  特别地，如果 $p(x)\sim N(\mu,\Sigma)$ ，A是一d*k的矩阵且 $y=A^t x$ 是一k维向量，则 $p(x)\sim N(A^t\mu, A^t\Sigma A)$



$$p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ \mathbf{y} = \mathbf{A}^t\mathbf{x} \Rightarrow p(\mathbf{y}) \sim N(\mathbf{A}^t\boldsymbol{\mu}, \mathbf{A}^t\boldsymbol{\Sigma}\mathbf{A})$$

白化(Whitening) 变换

- $\mathbf{\Phi}$: 其列向量是 $\mathbf{\Sigma}$**的正交特征向量.**
- $\mathbf{\Lambda}$**:** 与特征值对应的对角矩阵.
- 白化(Whitening) 变换

$$\mathbf{A}_w = \mathbf{\Phi}\mathbf{\Lambda}^{-1/2}$$

$$\mathbf{A}_w^t \mathbf{\Sigma} \mathbf{A}_w = \mathbf{I}$$

• 2.6 正态分布的判别函数

中山大學

最小误差概率分类可通过判别函数获得

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

如果已知 $p(x \mid \omega_i) \sim N(\mu_i, \Sigma_i)$ **那么**

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

其中：

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

情况1： $\Sigma_i = \sigma^2 I$

- 这种情况发生在各特征统计独立，且每个特征具有相同的 σ^2 方差时。此时的协方差阵是对角阵，仅仅是 σ^2 与单位阵*I* 的乘积。几何上它与样本落于相等大小的超球体聚类中的情况相对应，第*i*类的聚类以均值向量 μ_i 为中心。

- 省略掉其它无关紧要的附加常量，可得到简单的判决函数

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

展开后我们得到

$$g_i(x) = -\frac{1}{2\sigma^2}\left[x^t x - 2\mu_i^t x + \mu_i^t \mu_i\right] + \ln P(\omega_i)$$

省略附加常量，等价于线性判决函数

$$g_i(x) = w_i^t x + w_{i0}$$

其中 $$w_i = \frac{1}{\sigma_i^2}\mu_i$$

且 $$w_{i0} = \frac{-1}{2\sigma^2}\mu_i^t \mu_i + \ln P(\omega_i)$$

称 w_{i0} 为第 *i*个方向的阈值或者偏置。

- 使用线性判别函数的分类器称为“线性机器”。这类分类器有许多有趣的理论性质，其中一些将在第5章中详细讨论。此处只需注意到一个线性机器的判定面是一些超平面，它们是由两类问题中可获得最大后验概率的线性方程 $g_i(x) = g_j(x)$ 来确定。

- 在以上的例子中，该方程可写为

$$w^t(x - x_0) = 0$$

其中 $w = \mu_i - \mu_j$

且 $$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)$$

此方程定义了一个通过 x_0 且与向量 w 正交的超平面。由于 $w = \mu_i - \mu_j$，将 R_i 与 R_j 分开的超平面与两中心点的连线垂直。若 $P(\omega_i) = P(\omega_j)$则上式右边第二项为零，因此超平面垂直平分两中心点的连线。如图

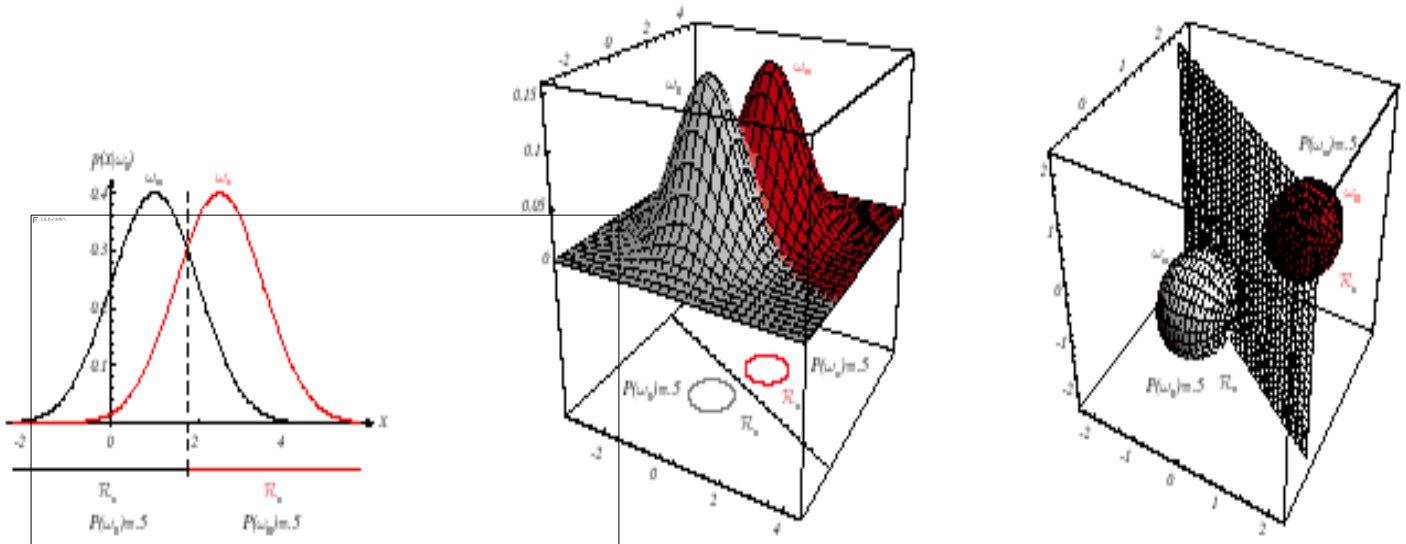
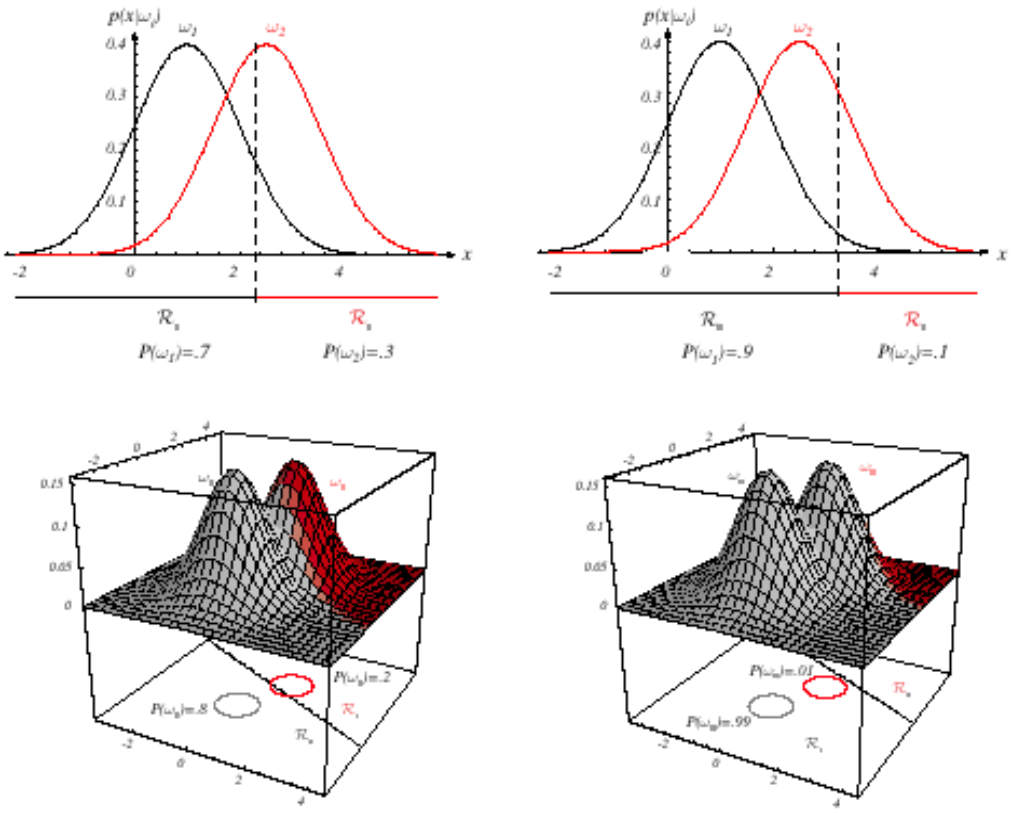


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-, and three-dimensional examples, we indicate $p(\mathbf{x}|\omega_i)$ and the boundaries for the case $P(\omega_1) = P(\omega_2)$. In the three-dimensional case, the grid plane separates $\mathcal{R}_1$ from $\mathcal{R}_2$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

如果两种分布的协方差矩阵相等且与单位阵成比例，那么它们呈d维球状分布，其判决边界是一个d-1维归一化超平面，垂直于两个中心的连线。在这些一维，二维及三维的例子中，是假设在 $P(\omega_i) = P(\omega_j)$ 的情况下来显示 $p(x \mid \omega_i)$ 和判决边界的。

- 如果所有c类的先验概率 $P(\omega_i)$相等，那么$\ln P(\omega_i)$ 项就成了另一可省略的附加常量。此种情况下，最优判决规则可简单陈述如下：

  为将某特征向量 x 归类，通过测量每一个 x 到 c 个均值向量中的每一个欧氏距离，并将 x 归为离它最近的那一类中。这样一个分类器被称为 “最小距离分类器”。如果每个均值向量被看成是其所属模式类的一个理想原型或模板，那么本质上是一个模板匹配技术。



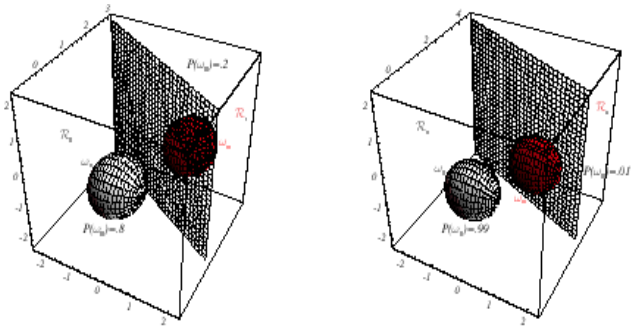


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

如图：随着先验概率的改变，判决边界也随之改变；对于差别较大的离散先验概率而言，判决边界不会落于这些一维，二维 及三维球状高斯分步的中心点之间。

情况2 ： $\sum_i = \sum$

- 第二类简单的情况是所有类的协方差阵都相等，但各自的均值向量是任意的。几何上，这种情况对应于样本落在相同大小和相同形状的超椭球体聚类中，第 i类的聚类中心在向量 μ_i 附近。此时的判决函数可从

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

简化为

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \sum^{-1}(x-\mu_j) + \ln P(\omega_i)$$

- 将二次型展开后，可再次得到线性判决函数

$$g_i(x) = w_i^t x + w_{i0} \qquad \text{其中} \quad w_i = \sum^{-1}\mu_i$$

$$w_{i0} = \frac{-1}{2\sigma^2}\mu_i^t\Sigma^{-1}\mu_i + \ln P(\omega_i)$$

由于判决函数是线性的，判决边界同样是超平面

$$w^t(x - x_0) = 0$$

其中 $$w = \sum^{-1}(\mu_i - \mu_j)$$

且 $$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\mu_i - \mu_j)^t \sum^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)$$

如果先验概率相等，其判决面与均值连线相交于中点；若不等，最优边界超平面将远离可能性较大的均值。如图

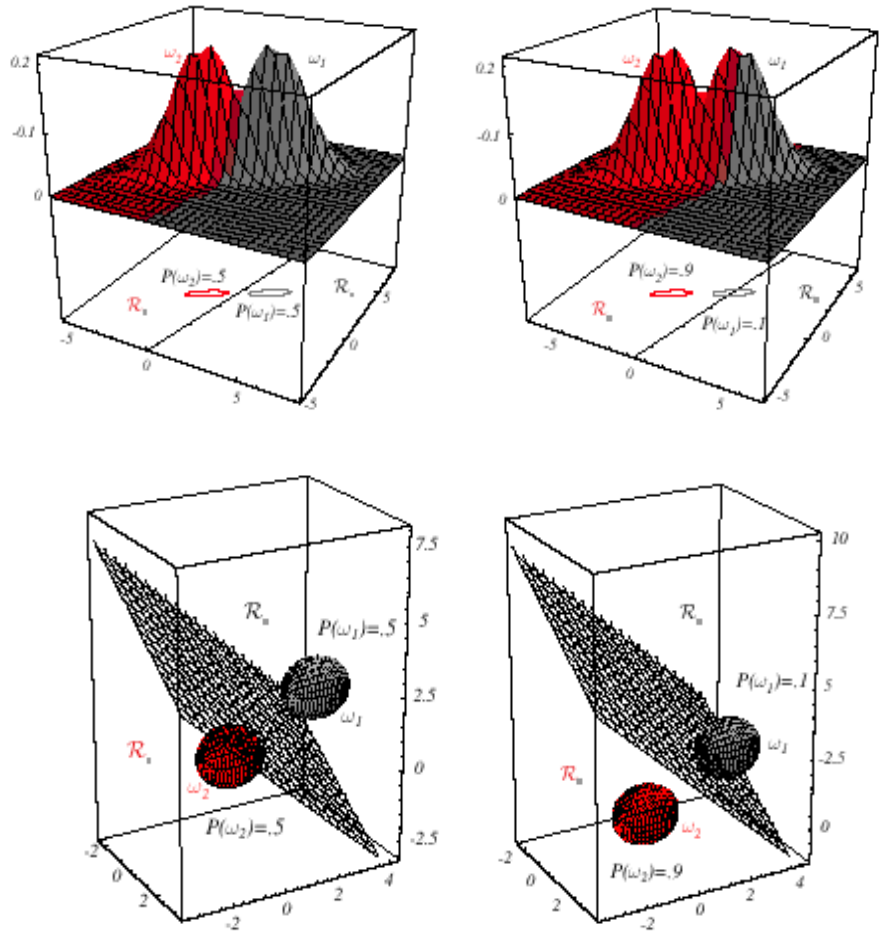


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

相等但非对称的高斯分布概率密度（由二维平面和三维椭球面表示）及判决区域。判决超平面未必和均值连线垂直正交。

情况3 ： $\sum_i =$ 任意

在一般的多元正态分布的情况下，每一类的协方差是不同，其判决函数显然也是二次型

$$g_i(x) = x^t W_i x + w_i^t x + w_{i0}$$

其中

$$W_i = -\frac{1}{2}\sum_i^{-1}$$

$$w_i = \sum_i^{-1}\mu_i$$

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

在两类问题中，其对应的判决面是超二次曲面。

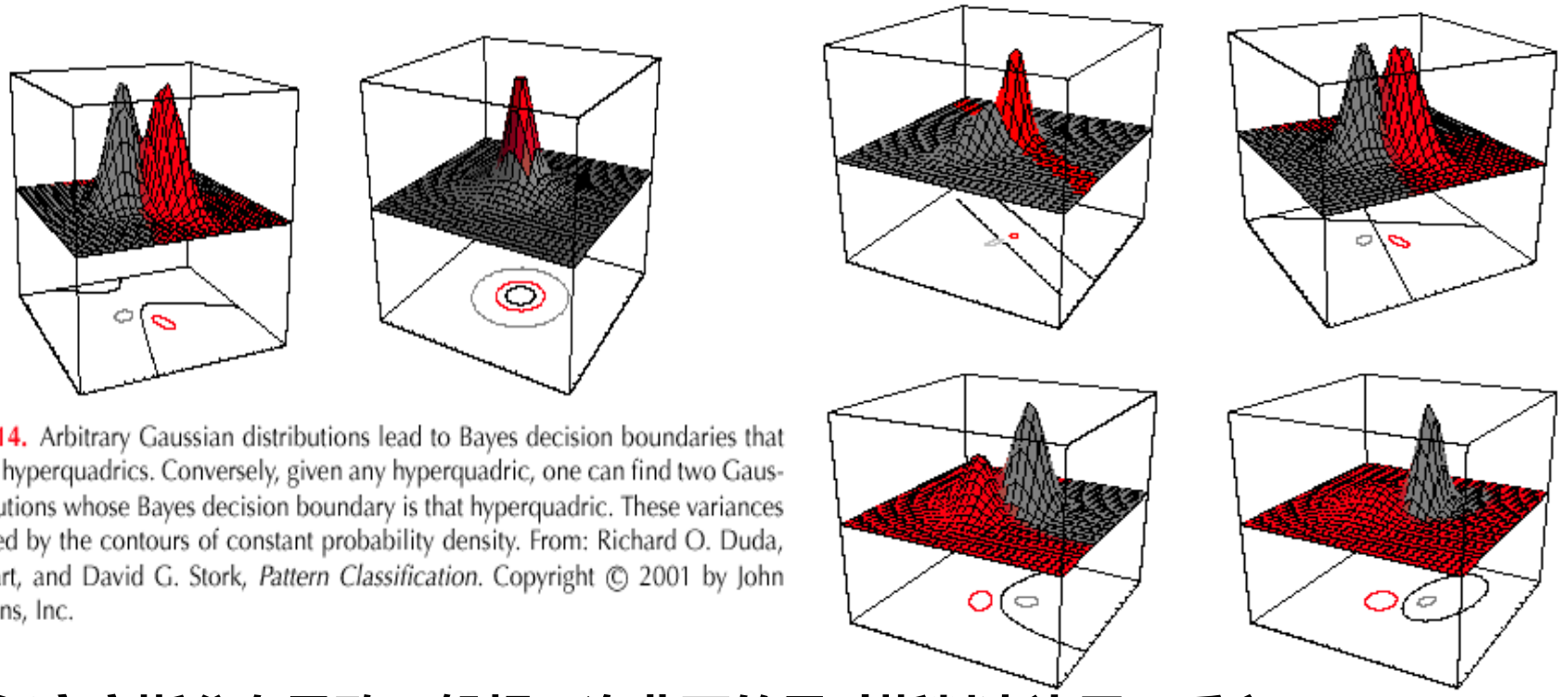
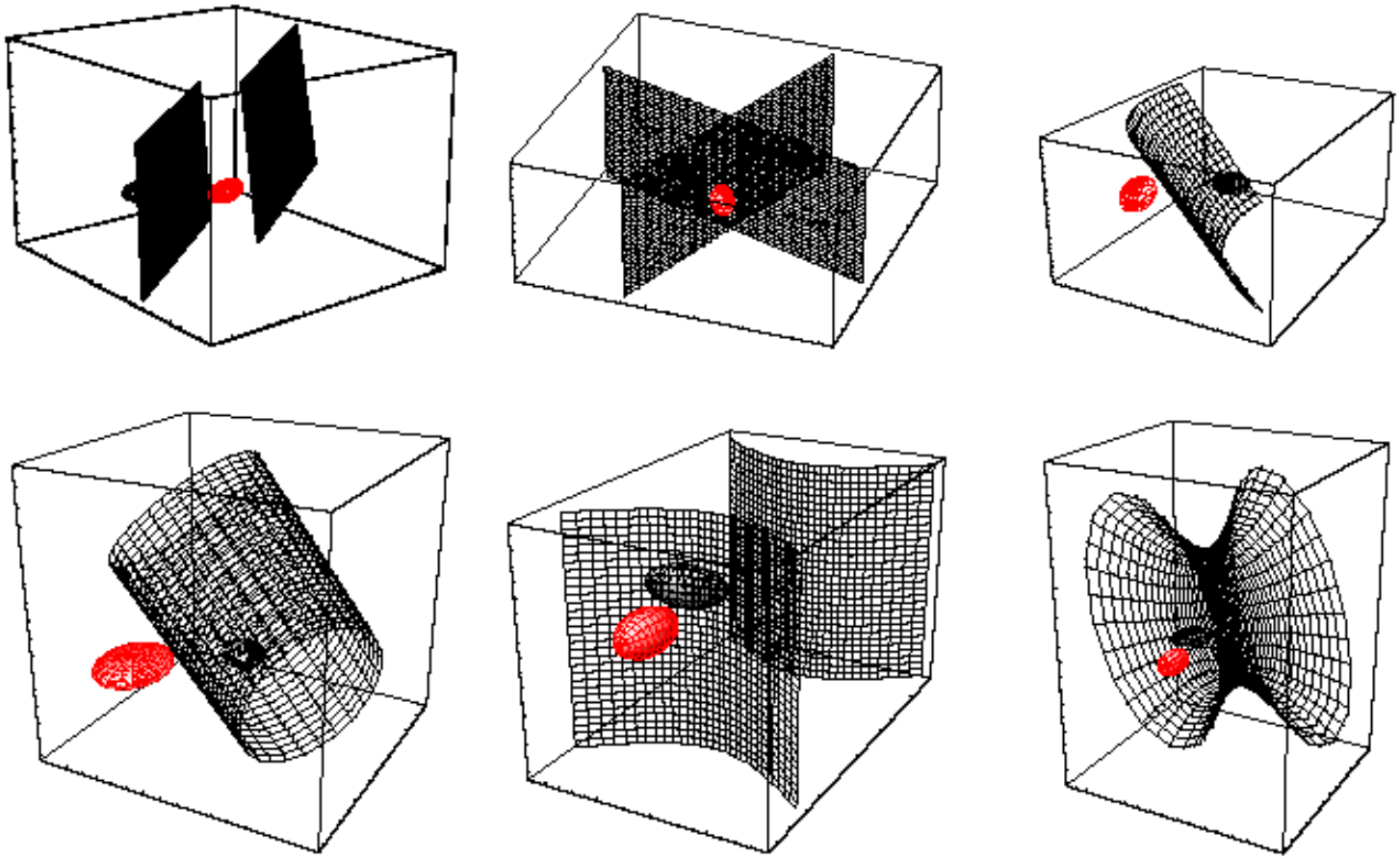
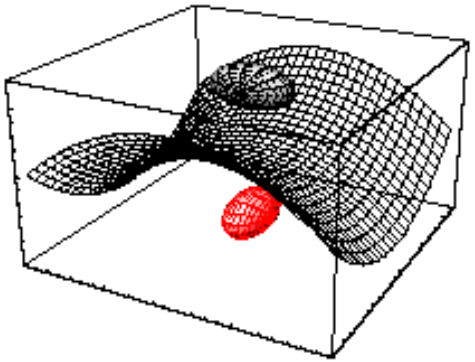
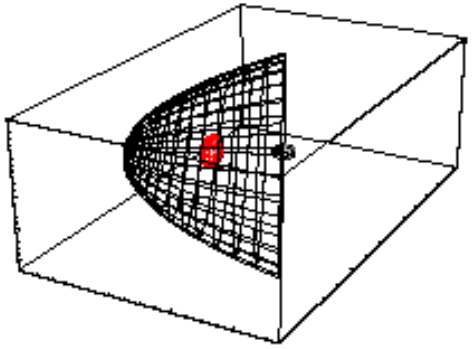


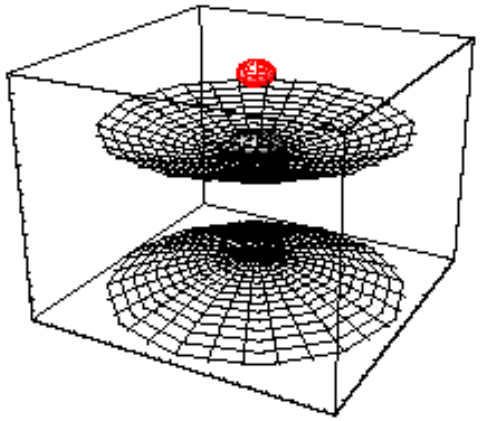
FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability density. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

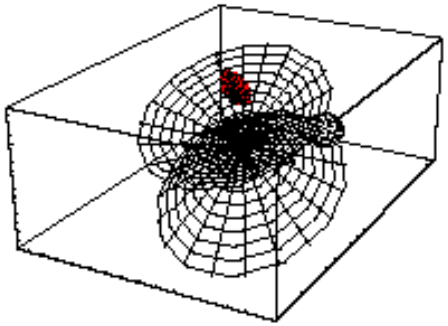
任意高斯分布导致一般超二次曲面的贝叶斯判决边界。反之，给定任意超二次曲面，就能求出两个高斯分布，其贝叶斯判别边界就是该超二次曲面。它们的方差由常概率密度的围线表示

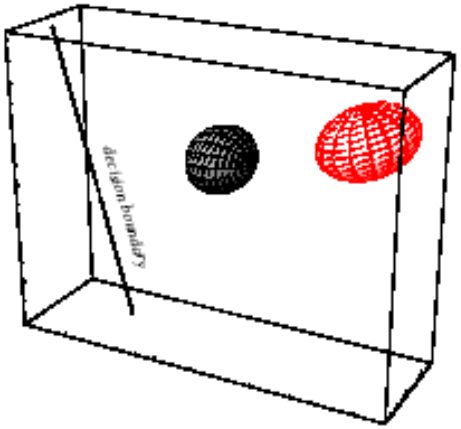


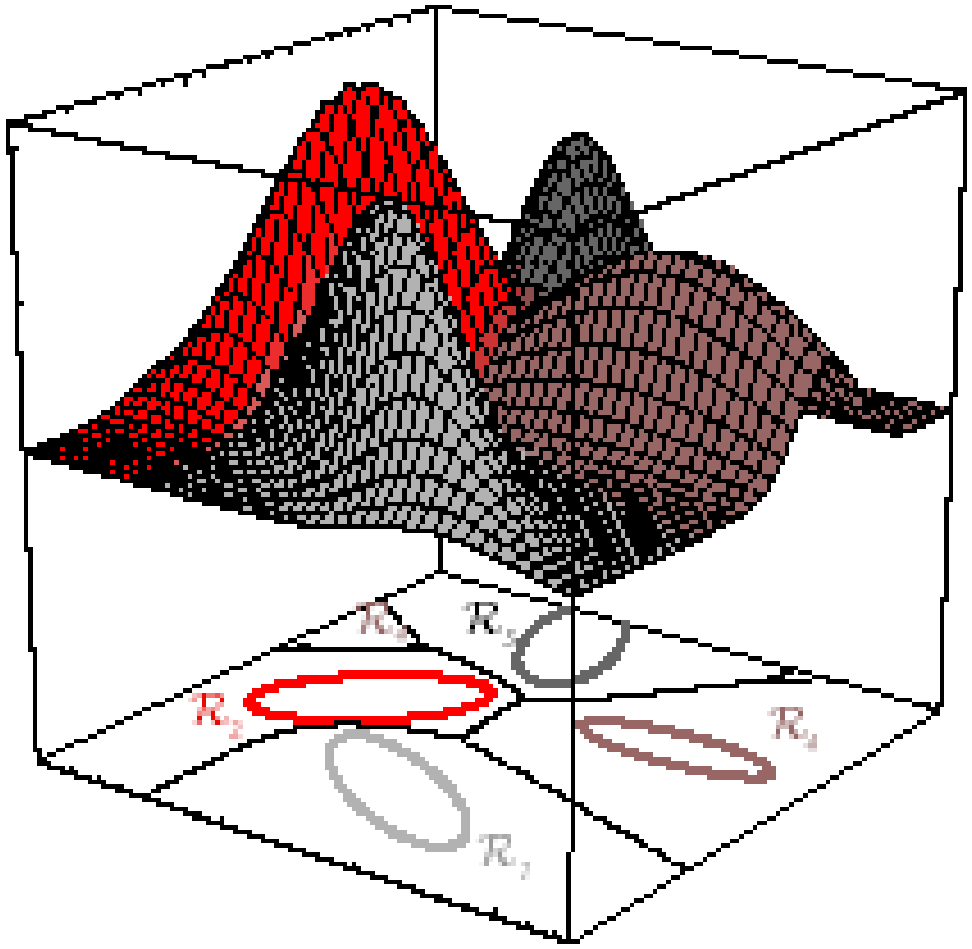




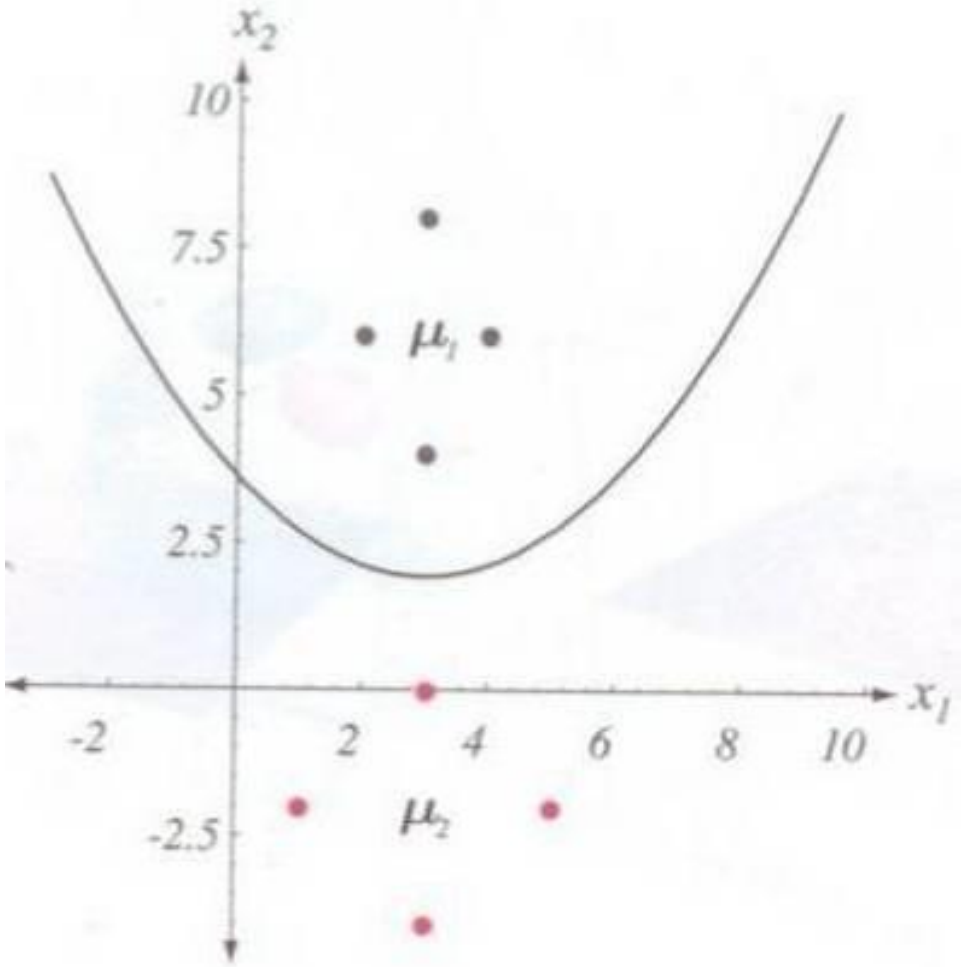





decision boundary


R_3
R_2
R_1
R_4

P42 例1



P42 例3

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

$$\boldsymbol{\Sigma}_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

$$P(\omega_1) = P(\omega_2) = 0.5$$

decision boundary

$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2, \text{ not passing } \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

2.9 贝叶斯决策论 -离散特征

- 到目前为止所讨论的特征向量 x 可以为 d 维欧氏空间中的任意一点。但是，在许多实际应用中，x中的元素可能是二进制，三进制或者更高的离散整数值，以至于 x 可以被认为是m个离散值$v_1,\ldots\ldots v_m$中的一个。在这种情况下，$p(x|\omega_j)$ 变得奇异化，积分形式

$$\int p(x|\omega_j)dx$$

转变为求和形式 $$\sum_x P(x|\omega_j)$$

其它方面与连续的情况基本相同，这里不一一赘述。

概率密度函数 $p(\square)$ 换成 概率分布函数 $P(\square)$

2.9.1 独立的二值特征

考虑两类问题，其中特征向量的元素为二值的，并且条件独立。

令 $x=(x_1,\ldots,x_d)'$ ，其中 x_i 可能为0或1，且

$$p_i=\Pr[x_i=1|\omega_1]$$

且

$$q_i=\Pr[x_i=1|\omega_2]$$

假设条件独立，可将x元素的概率写为 $P(x|\omega_i)$ ，即

$$P(x|\omega_1)=\prod_{i=1}^{d}p_i^{x_i}(1-p_i)^{1-x_i}$$

且

$$P(x|\omega_2)=\prod_{i=1}^{d}q_i^{x_i}(1-q_i)^{1-x_i}$$

那么似然比为

$$\frac{P(x|\omega_1)}{P(x|\omega_2)}=\prod_{i=1}^{d}\left(\frac{p_i}{q_i}\right)^{x_i}\left(\frac{1-p_i}{1-q_i}\right)^{1-x_i}$$

由公式 $g(x)=\ln\frac{p(x|\omega_1)}{p(x|\omega_2)}+\ln\frac{P(\omega_1)}{P(\omega_2)}$ 得判决函数

$$g(x)=\sum_{i=1}^{d}[x_i\ln\frac{p_i}{q_i}+(1-x_i)\ln\frac{1-p_i}{1-q_i}]+\ln\frac{P(\omega_1)}{P(\omega_2)}$$

注意判决函数对 x_i 是线性的，可改写为

$$g(x)=\sum_{i=1}^{d}w_ix_i+w_0$$

其中

$$w_i=\ln\frac{p_i(1-q_i)}{q_i(1-p_i)}\qquad i=1,\ldots,d$$

且

$$w_0=\sum_{i=1}^{d}\ln\frac{1-p_i}{1-q_i}+\ln\frac{P(\omega_1)}{P(\omega_2)}$$

若 $g(x)>0$ 判别为 ω_1 ；否则为 ω_2

- g(x)可以看作是x的各分量的加权组合。

$$g(x)=\sum_{i=1}^{d} w_i x_i + w_0$$

- 注意权重W_i 的意义。

$$w_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} \qquad i=1,\ldots,d$$

- 特征独立的条件产生线性分类器，而如果特征不独立将产生复杂的分类器。

Example: 三维二值特征的贝叶斯决策

$$P(\omega_1) = P(\omega_2) = 0.5$$

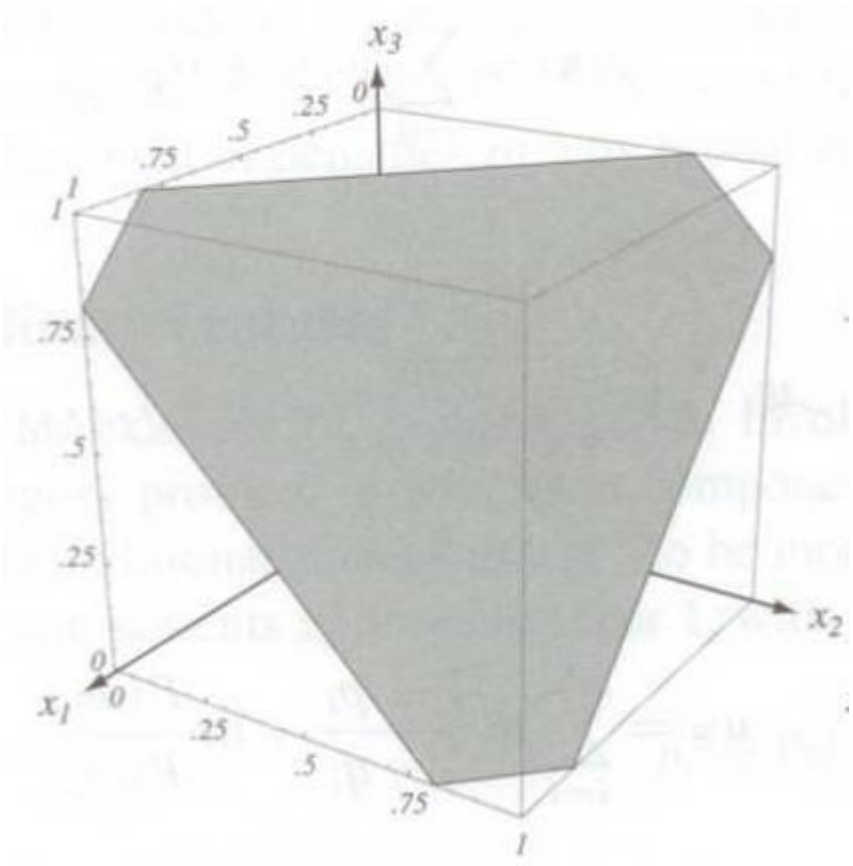
$$p_i = 0.8,\ q_i = 0.5,\ i = 1,2,3$$

$$w_i = \ln \frac{0.8(1-0.5)}{0.5(1-0.8)}$$

$$= 1.3863$$

$$w_0 = \sum_{i=1}^{3} \ln \frac{1-0.8}{1-0.5} + \ln \frac{0.5}{0.5}$$

$$= -2.75$$



Example: 三维二值特征的贝叶斯决策

$$P(\omega_1) = P(\omega_2) = 0.5$$

$$p_i = 0.8,\ q_i = 0.5,\ i = 1,2$$

$$p_3 = q_3 = 0.5$$

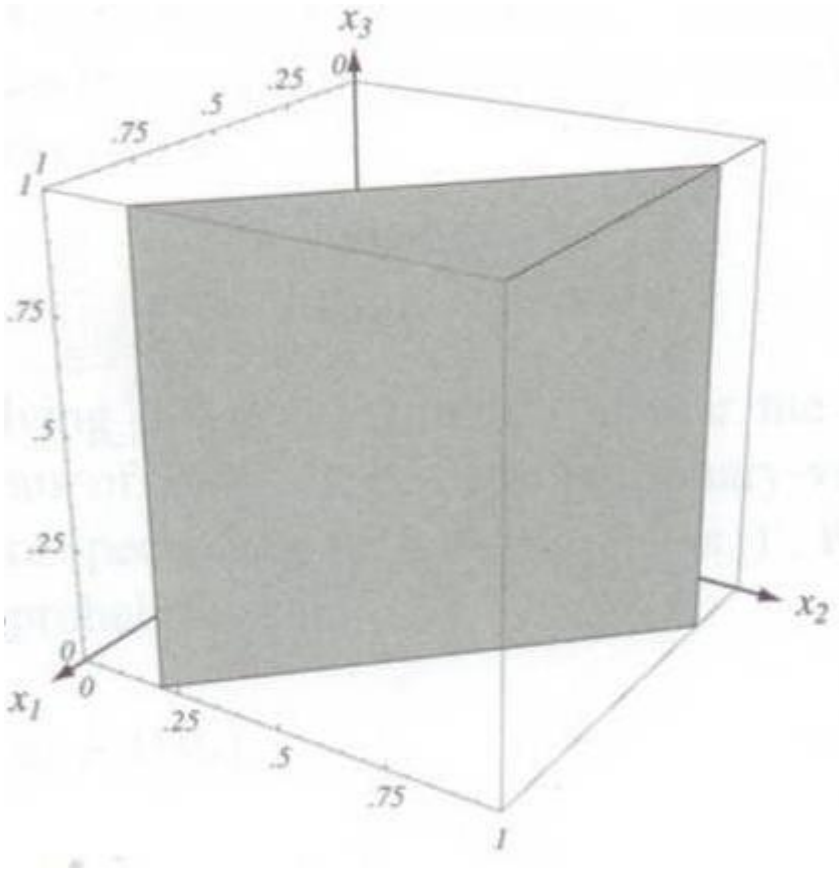
$$w_i = \ln \frac{0.8(1-0.5)}{0.5(1-0.8)}$$

$$= 1.3863, \quad i = 1,2$$

$$w_3 = 0$$

$$w_0 = \sum_{i=1}^{2} \ln \frac{1-0.8}{1-0.5} + \ln \frac{0.5}{0.5}$$

$$= -1.83$$



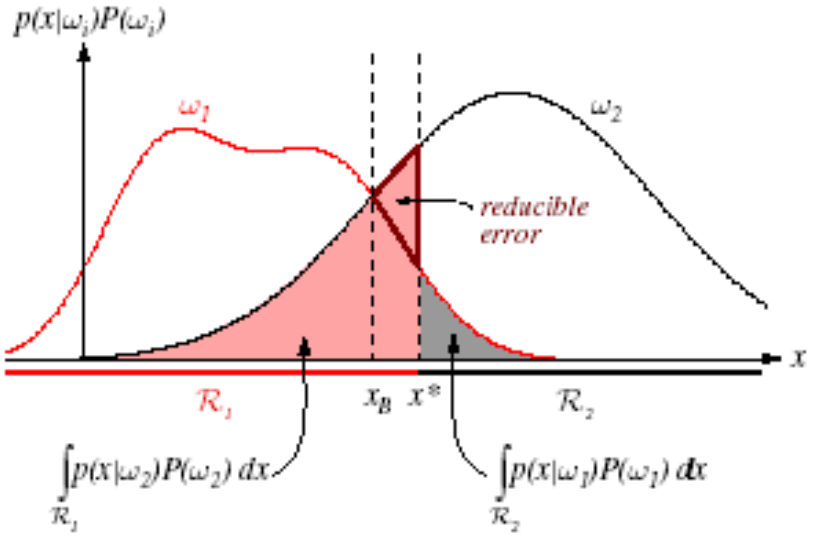
· 2.7 误差概率和误差积分

中山大學

- **二分分类器: 考虑以非最优方式将空间分成两个区域R_i 与 R_j，则误差概率为:**

$$\begin{aligned} P(error) &= P(x \in R_2, \omega_1) + P(x \in R_1, \omega_2) \\ &= P(x \in R_2 \mid \omega_1)P(\omega_1) + P(x \in R_1 \mid \omega_2)P(\omega_2) \\ &= \int_{R_2} p(x \mid \omega_1)P(\omega_1)dx + \int_{R_1} p(x \mid \omega_2)P(\omega_2)dx \end{aligned}$$

上式的值与判决点的取值有关.



- 多类情况

- 正确分类的概率

$$P(correct) = \sum_{i=1}^{c} \int_{R_i} p(\mathbf{x} \mid \omega_i) P(\omega_i) d\mathbf{x}$$

- 贝叶斯分类器通过选择对所有x使得被积函数最大化的区域使正确分类的概率最大化。
  - 没有其他分类方法能产生更小的分类概率。

中山大學

2.8 正态密度的错误上界

- 在高斯函数的情况下，整个误差率计算过程相当复杂。
  - 特别是高维情形。
  - 判决区域可能不连续。
- 在两类情况下，一般错误积分公式可近似的给出一个误差率的上界。

$$P(error)=\int_{-\infty}^{\infty}P(error,x)dx=\int_{-\infty}^{\infty}P(error\,|\,x)\,p(x)dx$$

$$P(error\,|\,x)=\min[P(\omega_1\,|\,\mathbf{x}),P(\omega_2\,|\,\mathbf{x})]$$

Chernoff 界

$$\min[a,b] \le a^{\beta} b^{1-\beta} \quad \text{for } a,b \ge 0 \text{ and } 0 \le \beta \le 1$$

$$P(error \mid x) = \min[P(\omega_1 \mid \mathbf{x}), P(\omega_2 \mid \mathbf{x})]$$

$$P(error) = \int_{-\infty}^{\infty} P(error, x)dx$$

$$= \int_{-\infty}^{\infty} P(error \mid x)\, p(x)dx$$

$$= \int_{-\infty}^{+\infty} \min[P(\omega_1 \mid \mathbf{x}), P(\omega_2 \mid \mathbf{x})]\, p(x)dx$$

$$= \int_{-\infty}^{+\infty} \min[\, p(x \mid \omega_1)P(\omega_1),\, p(x \mid \omega_2)P(\omega_2)]dx$$

$$P(error) \le P^{\beta}(\omega_1)P^{1-\beta}(\omega_2) \int p^{\beta}(\mathbf{x} \mid \omega_1)\, p^{1-\beta}(\mathbf{x} \mid \omega_2)d\mathbf{x}$$

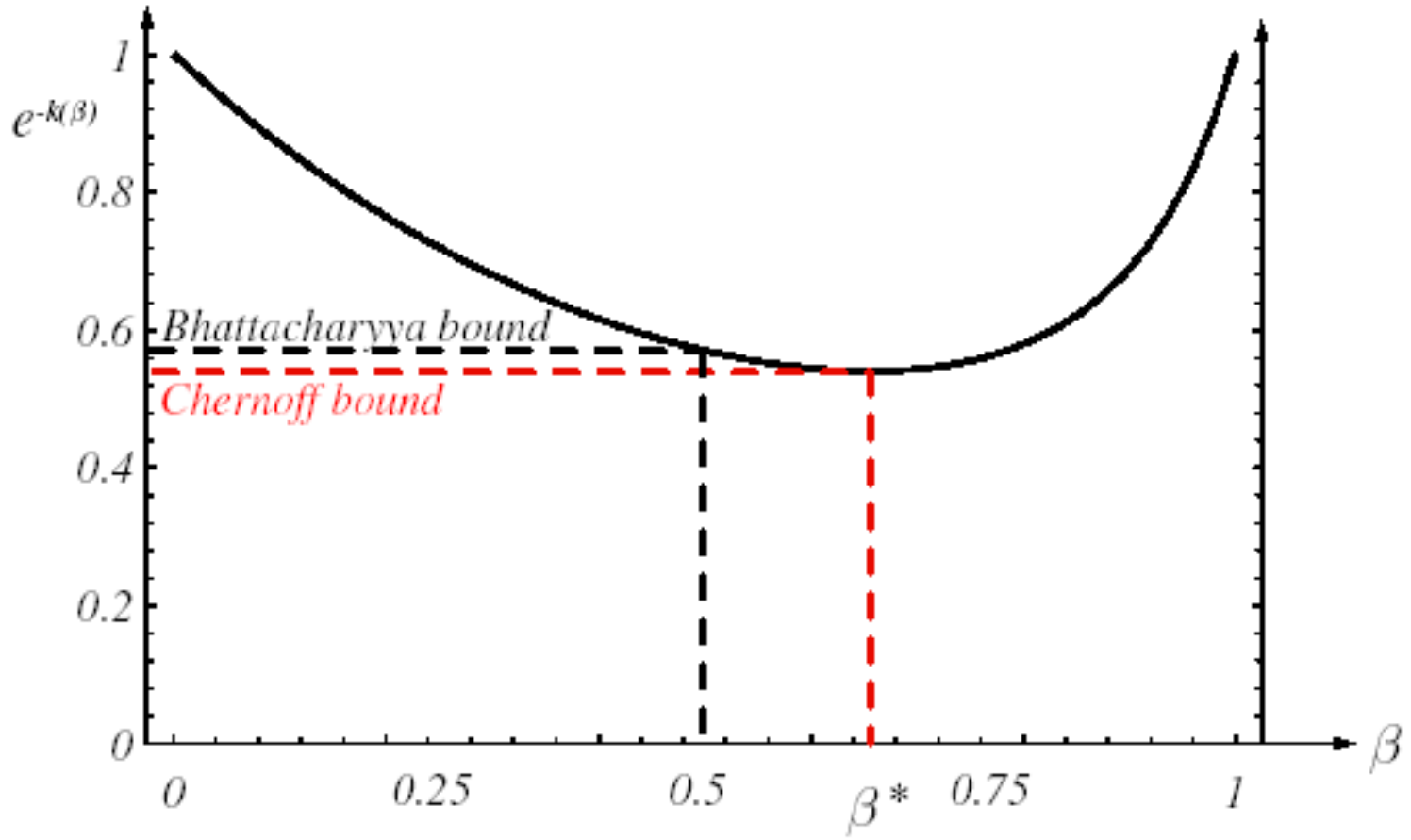
对于正态密度, $\int p^{\beta}(\mathbf{x} \mid \omega_1)\, p^{1-\beta}(\mathbf{x} \mid \omega_2)d\mathbf{x} = e^{-k(\beta)}$

习题36，作业！

其中: $k(\beta) = \dfrac{\beta(\beta-1)}{2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^t[(1-\beta)\boldsymbol{\Sigma}_1 + \beta\boldsymbol{\Sigma}_2]^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$

$$+ \frac{1}{2}\ln \frac{\left|(1-\beta)\boldsymbol{\Sigma}_1 + \beta\boldsymbol{\Sigma}_2\right|}{\left|\boldsymbol{\Sigma}_1\right|^{1-\beta}\left|\boldsymbol{\Sigma}_2\right|^{\beta}}$$

Chernoff Bound 0.66
Bhattacharyya Bound 0.5



Bhattacharyya Bound

set $\beta = 1/2$

$$P(error) \le \sqrt{P(\omega_1)P(\omega_2)} \int \sqrt{p(\mathbf{x}\,|\,\omega_1)\,p(\mathbf{x}\,|\,\omega_2)}\,d\mathbf{x}$$

$$= \sqrt{P(\omega_1)P(\omega_2)}\,e^{-k(1/2)}$$

$$k(1/2) = \frac{1}{8}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^t \left[\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}\right]^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$

$$+ \frac{1}{2}\ln \frac{\left|\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}\right|}{\sqrt{|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|}}$$

Example: 在高斯分布下的错误率的界

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

$$\boldsymbol{\Sigma}_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

$$P(\omega_1) = P(\omega_2) = 0.5$$

Example:在高斯分布下的错误率的界

- Bhattacharyya 界
  - $k(1/2) = 4.06$
  - $P(error) < 0.0087$
- Chernoff 界
  - 0.0016380 通过数值查找
- 错误率估计
  - 0.0021
  - 对高维不实用

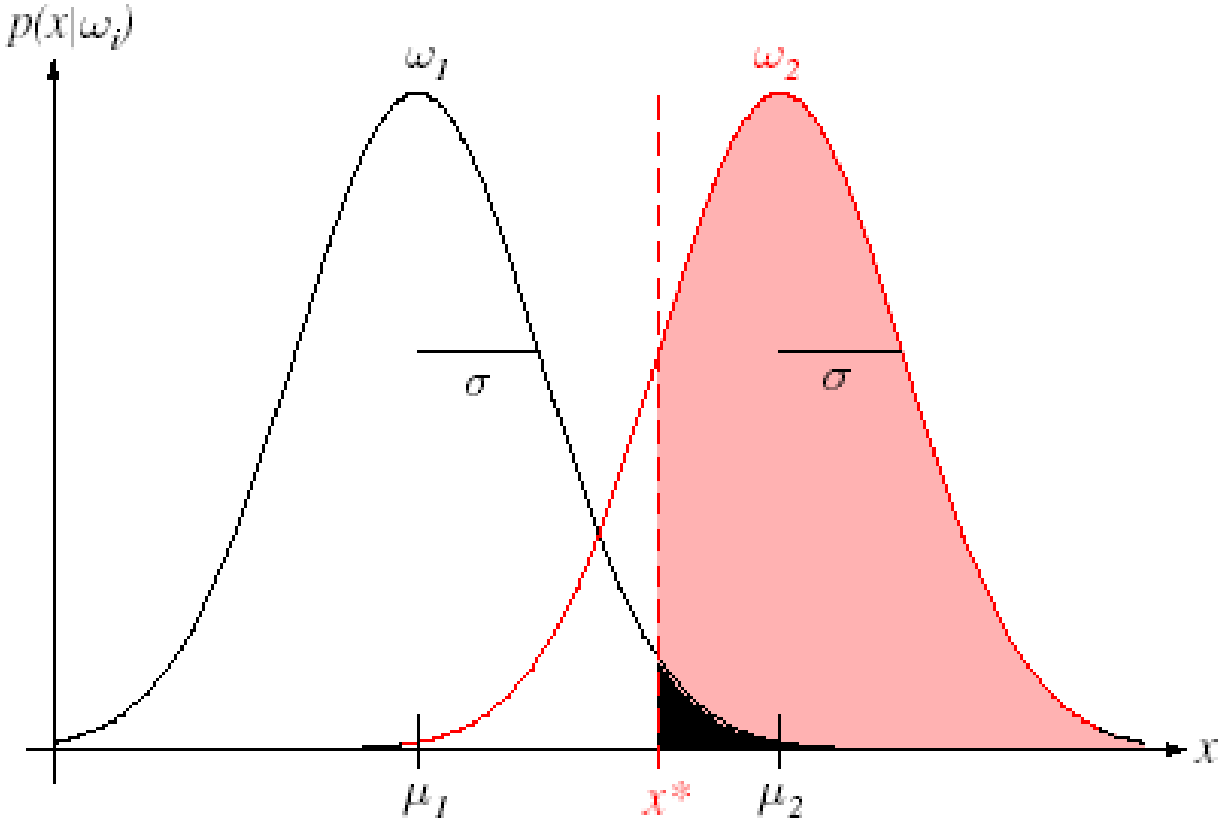
- 信号检测理论

- 在检测器某点 上有个内部信号x：
  - 当外部信号出现时，有均值 μ_2
  - 当外部信号不出现时，有均值 μ_1

$$p(x|\omega_i) \sim N(\mu_i, \sigma^2)$$

- 检测器分类器将利用一个阈值x*来判定是否存在外部脉冲。

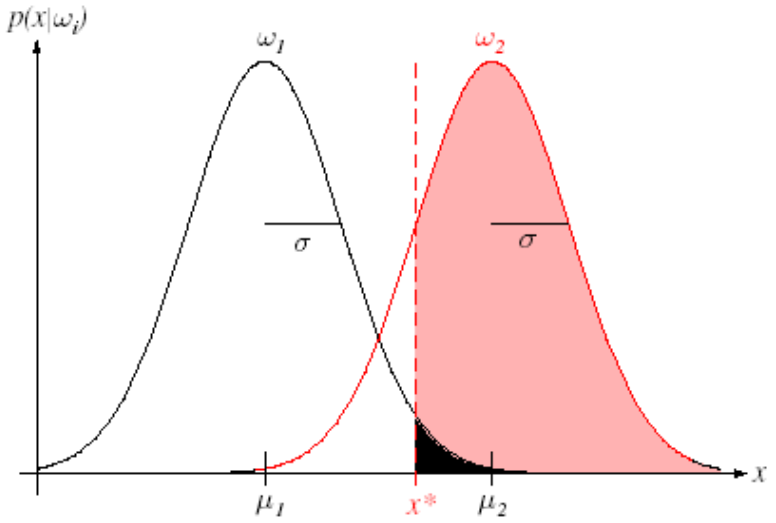
- 信号检测理论



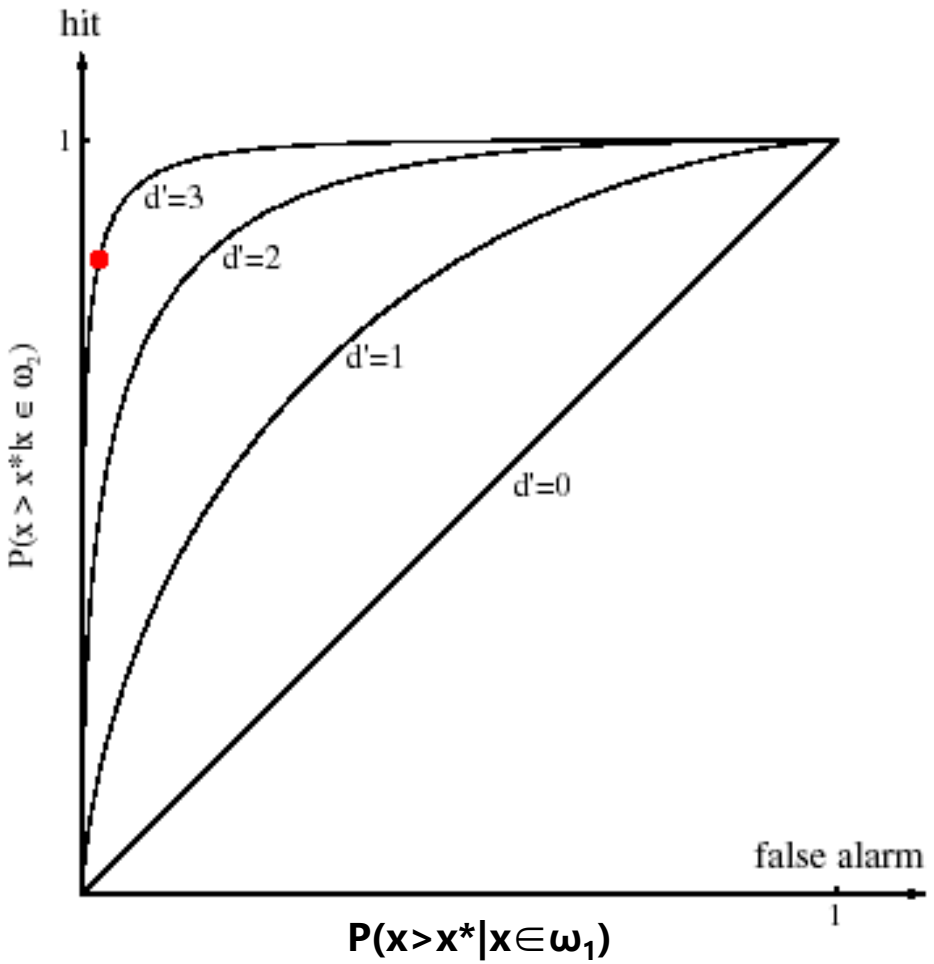
$$\text{判别能力 } d' = \frac{|\mu_2 - \mu_1|}{\sigma}$$

四种概率：

- 一次击中: $P(x>x^*|x \text{ in } \omega_2)$ 识别率
- 一次虚警: $P(x>x^*|x \text{ in } \omega_1)$ 错误接受率
- 一次漏检: $P(x<x^*|x \text{ in } \omega_2)$ 错误拒绝率
- 一次正确拒绝: $P(x<x^*|x \text{ in } \omega_1)$



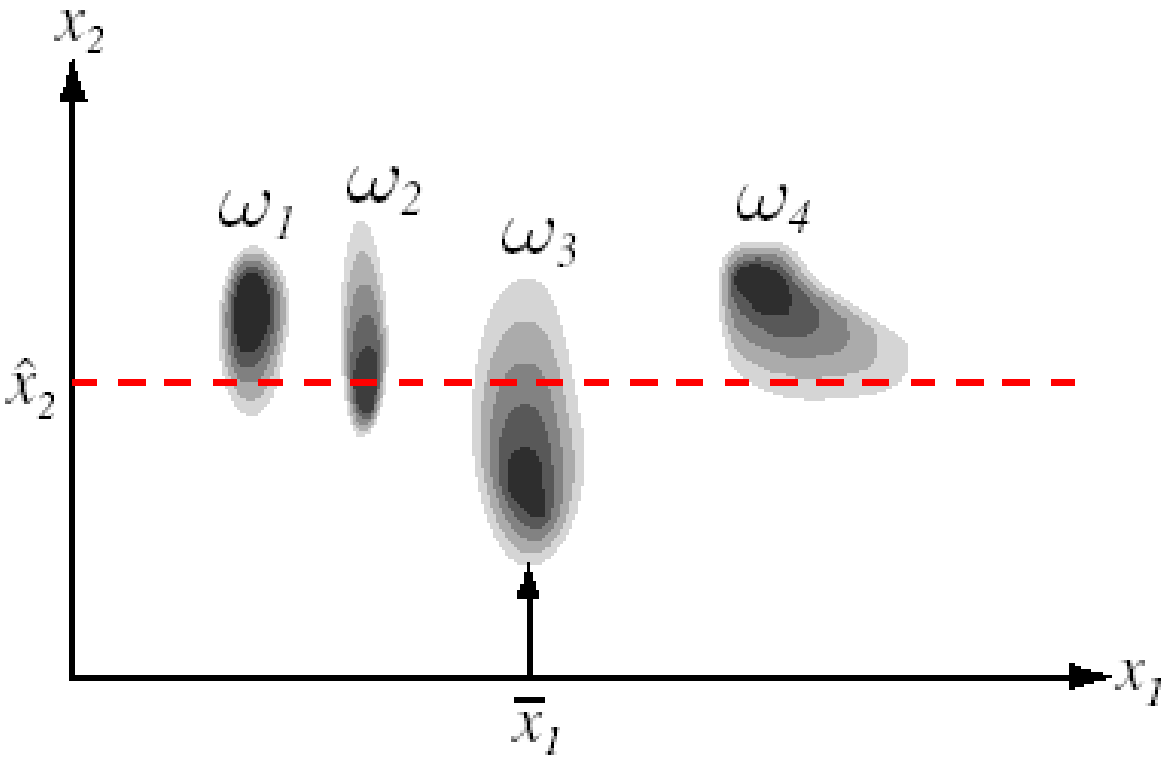
Receiver Operating Characteristic (ROC)



2.10 丢失特征和噪声特征

- 丢失特征
  - 考虑训练集数据未受损，测试集数据受损情形。

丢失特征举例:



丢失特征情形下的决策：

$$\mathbf{x}=[\mathbf{x}_g,\mathbf{x}_b]$$

$$P(\omega_i\mid\mathbf{x}_g)=\frac{P(\omega_i,\mathbf{x}_g)}{P(\mathbf{x}_g)}=\frac{\int p(\omega_i,\mathbf{x}_g,\mathbf{x}_b)d\mathbf{x}_b}{\int p(\mathbf{x})d\mathbf{x}_b}$$

$$=\frac{\int P(\omega_i\mid\mathbf{x}_g,\mathbf{x}_b)p(\mathbf{x}_g,\mathbf{x}_b)dx_b}{\int p(\mathbf{x})d\mathbf{x}_b}$$

$$=\frac{\int g_i(\mathbf{x})p(\mathbf{x})d\mathbf{x}_b}{\int p(\mathbf{x})d\mathbf{x}_b}$$

其中: $g_i(x)=g_i(x_g,x_b)=P(\omega_i\mid x_g,x_b)$ 是判别函数

- 噪声特征

假设用x_t来表示观测到的x_b特征量的真实值。

噪声模型: $p(\mathbf{x}_b \mid \mathbf{x}_t)$, 假设 x_b 独立于 ω_i 和 $\mathbf{x}_g$

$$P(\omega_i \mid \mathbf{x}_g, \mathbf{x}_b) = \frac{\int p(\omega_i, \mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t) dx_t}{p(\mathbf{x}_g, \mathbf{x}_b)}$$

$$p(\omega_i, \mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t) = P(\omega_i \mid \mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t) p(\mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t)$$

$$= P(\omega_i \mid \mathbf{x}_g, \mathbf{x}_t) p(\mathbf{x}_b \mid \mathbf{x}_g, \mathbf{x}_t) p(\mathbf{x}_g, \mathbf{x}_t), \quad p(\mathbf{x}_b \mid \mathbf{x}_g, \mathbf{x}_t) = p(\mathbf{x}_b \mid \mathbf{x}_t)$$

$$P(\omega_i \mid \mathbf{x}_g, \mathbf{x}_b) = \frac{\int p(\omega_i \mid \mathbf{x}_g, \mathbf{x}_t) p(\mathbf{x}_g, \mathbf{x}_t) p(\mathbf{x}_b \mid \mathbf{x}_t) dx_t}{\int p(\mathbf{x}_g, \mathbf{x}_t) p(\mathbf{x}_b \mid \mathbf{x}_t) dx_t}$$

$$= \frac{\int g_i(\mathbf{x}) p(\mathbf{x}) p(\mathbf{x}_b \mid \mathbf{x}_t) dx_t}{\int p(\mathbf{x}) p(\mathbf{x}_b \mid \mathbf{x}_t) dx_t}, \quad \mathbf{x} = [\mathbf{x}_g, \mathbf{x}_t]$$

2.11 贝叶斯置信网

中山大學

用有向无环图来表示事件的因果依赖关系的网络图。

图 2-24　由节点(大写的黑体字母标记)和与它们相关的离散状态(小写字母)所组成的置信网。因此节点 A 具有状态 $a_1, a_2, a_3, \cdots$，简单记为 a，节点 B 具有状态 $b_1, b_2, \cdots$，记为 b，等等。节点之间的连线代表条件概率，比如，P(c|a)可由一个元素为 $P(c_i|a_j)$ 的矩阵来描述

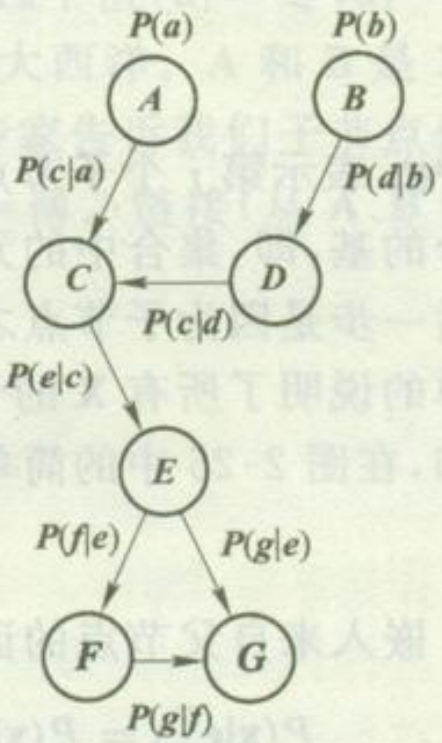
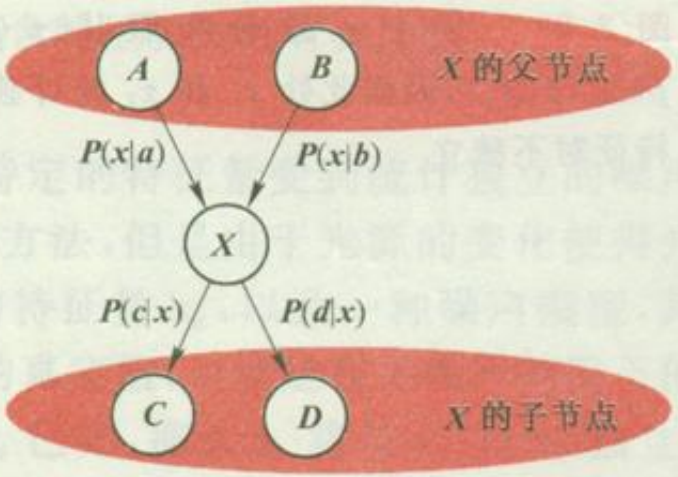
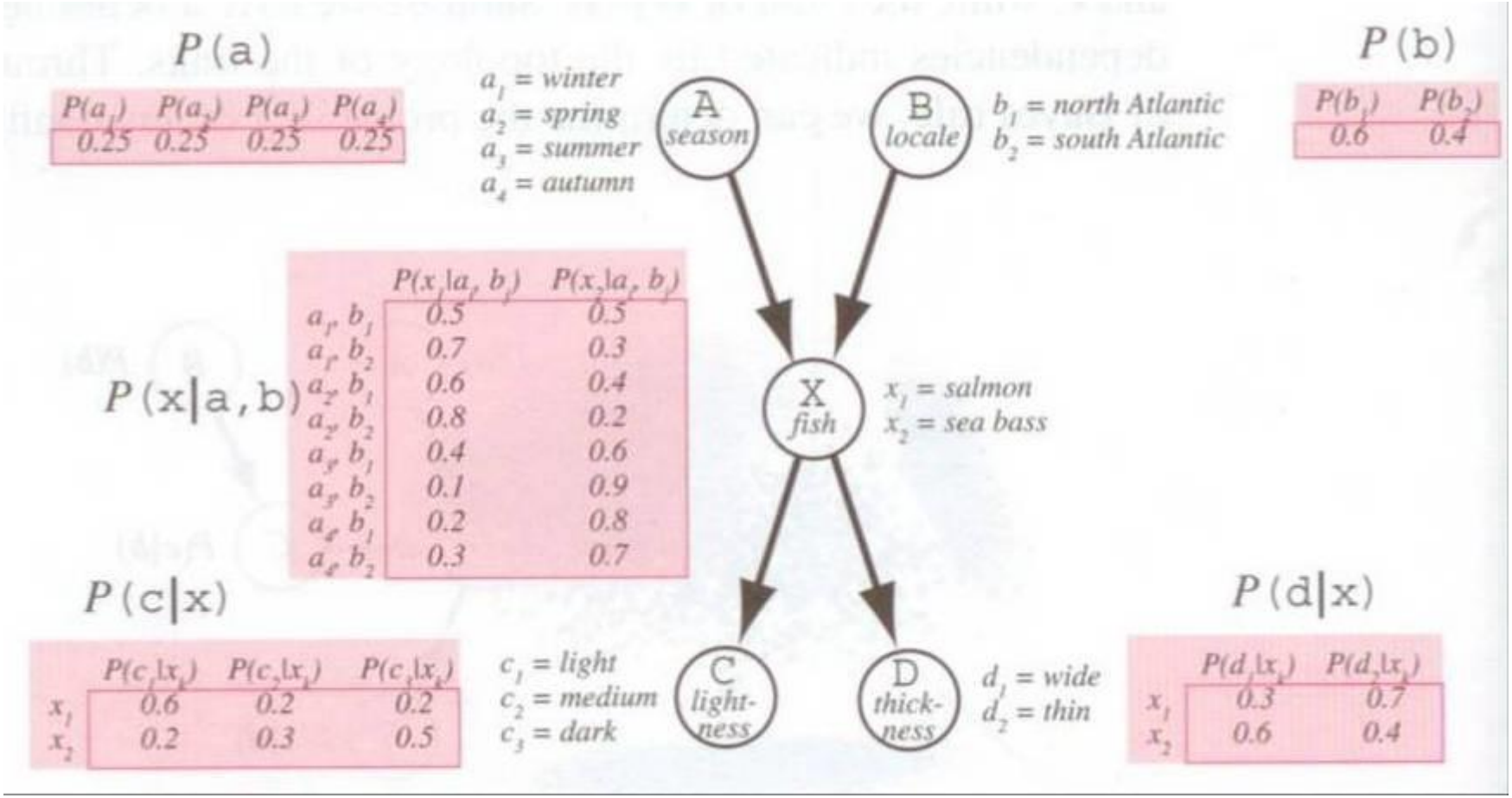


图 2-25　一个置信网的一部分，由节点 X，具有变量值（x_1，x_2，…），以及其父节点（A 和 B）和子节点（C 和 D）组成



Example: 鱼分类置信网



$$p(a_3, b_1, x_2, c_3, d_2) = P(a_3)P(b_1)P(x_2 \mid a_3, b_1)P(c_3 \mid x_2)P(d_2 \mid x_2)$$
$$= 0.25 \times 0.6 \times 0.4 \times 0.5 \times 0.4 = 0.012$$

2.12 复合贝叶斯决策论及上下文

中山大學

- 以鱼分类为例，来看上下文关系；
- 复合判决与序贯复合判决；

2.12 复合贝叶斯决策论及上下文

$$\boldsymbol{\omega} = \left(\omega(1), \cdots, \omega(n)\right)^t$$

$\omega(i)$ 从c种类别 $\omega_1, \cdots, \omega_c$中取一个值。

$$\mathbf{X} = \left(\mathbf{x}_1, \cdots, \mathbf{x}_n\right)$$

$$P(\boldsymbol{\omega} \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid \boldsymbol{\omega}) P(\boldsymbol{\omega})}{p(\mathbf{X})} = \frac{p(\mathbf{X} \mid \boldsymbol{\omega}) P(\boldsymbol{\omega})}{\sum_{\omega} p(\mathbf{X} \mid \boldsymbol{\omega}) P(\boldsymbol{\omega})}$$

简化

$$p(\mathbf{X} \mid \boldsymbol{\omega}) = \prod_{i=1}^{n} p(\mathbf{x}_i \mid \omega(i))$$

• 本章小结

- 贝叶斯决策的基本思想非常简单。为了最小化风险，总是选择那些能够最小化条件风险的行为。贝叶斯公式允许我们通过先验概率 $P(\omega_j)$ 和条件密度 $P(x \mid \omega_j)$ 来计算后验概率。

- 若内在的分布为多元的高斯分布，判别边界将是超二次型，其形状与位置取决于先验概率及该分布的均值与协方差。