

# 第2章：贝叶斯决策理论

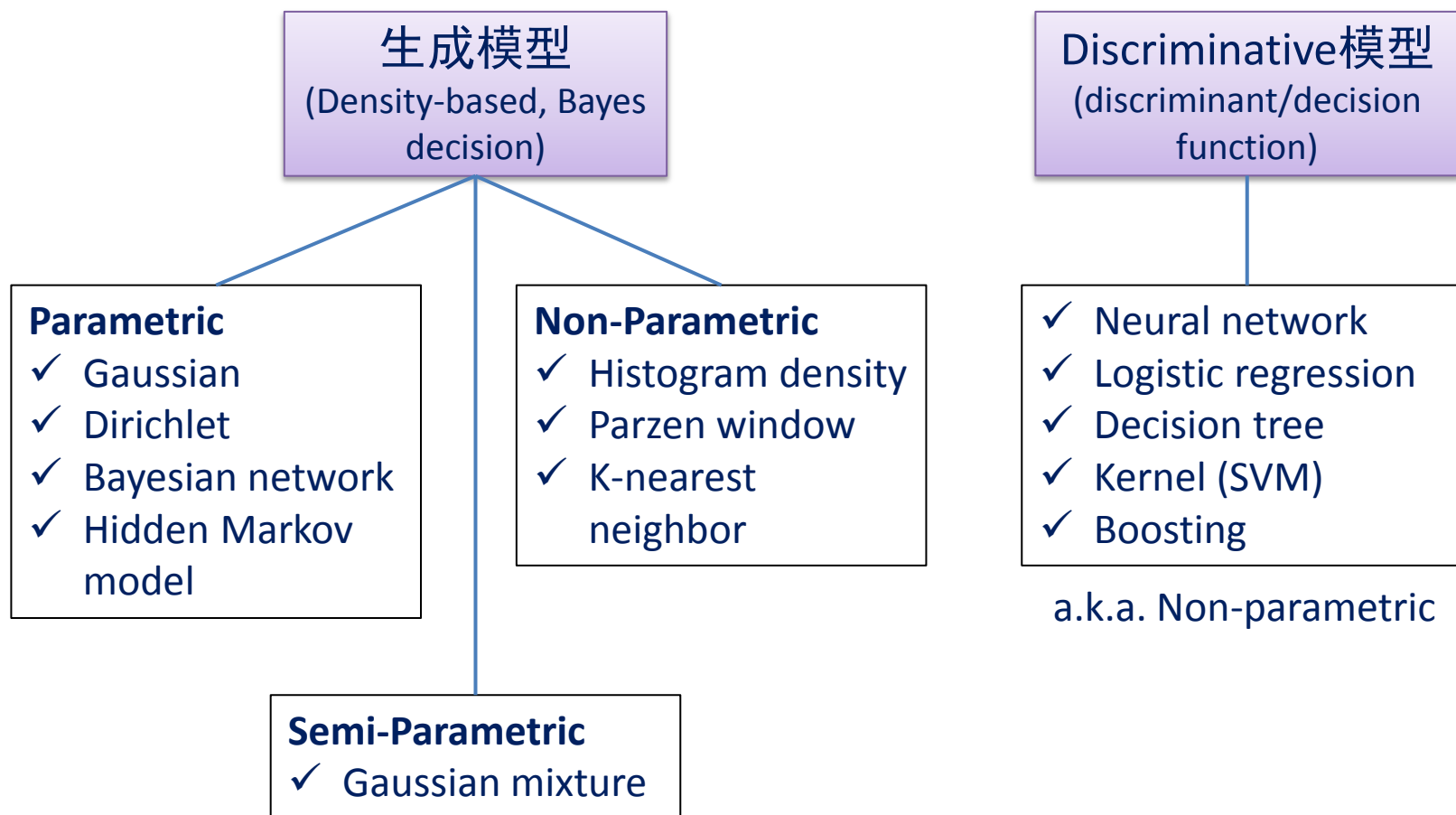
刘成林([liucl@nlpr.ia.ac.cn](mailto:liucl@nlpr.ia.ac.cn))

2015年9月23日

助教：杨学行([xhyang@nlpr.ia.ac.cn](mailto:xhyang@nlpr.ia.ac.cn))

吴一超([yichao.wu@nlpr.ia.ac.cn](mailto:yichao.wu@nlpr.ia.ac.cn))

# 统计模式识别方法



# 提纲

- 导论：2类的例子
- 最小风险决策
- 判别函数和决策面
- 高斯概率密度
- 高斯密度下的判别函数
- 分类错误率

# 导论：问题表示

- 类别:  $\omega_i, i = 1, \dots, c$
- 特征矢量  $\mathbf{x} = [x_1, \dots, x_d] \in R^d$
- 先验概率  $P(\omega_i) \quad \sum_{i=1}^c P(\omega_i) = 1$
- 概率密度函数(条件概率)  $p(\mathbf{x} | \omega_i)$
- 后验概率

$$P(\omega_i | \mathbf{x}) = \frac{p(\mathbf{x} | \omega_i)P(\omega_i)}{p(\mathbf{x})} = \frac{p(\mathbf{x} | \omega_i)P(\omega_i)}{\sum_{j=1}^c p(\mathbf{x} | \omega_j)P(\omega_j)}$$

$$\sum_{i=1}^c P(\omega_i | \mathbf{x}) = 1$$

# 2类的例子

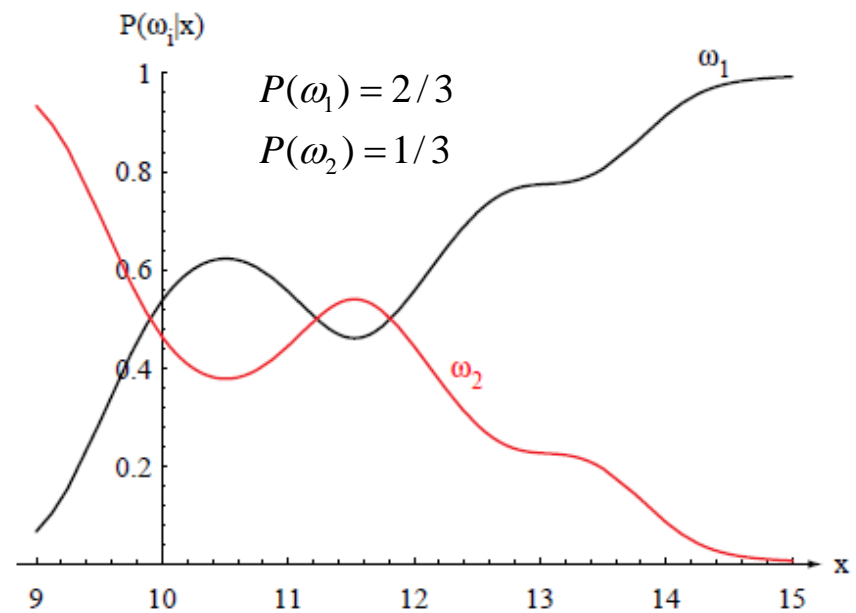
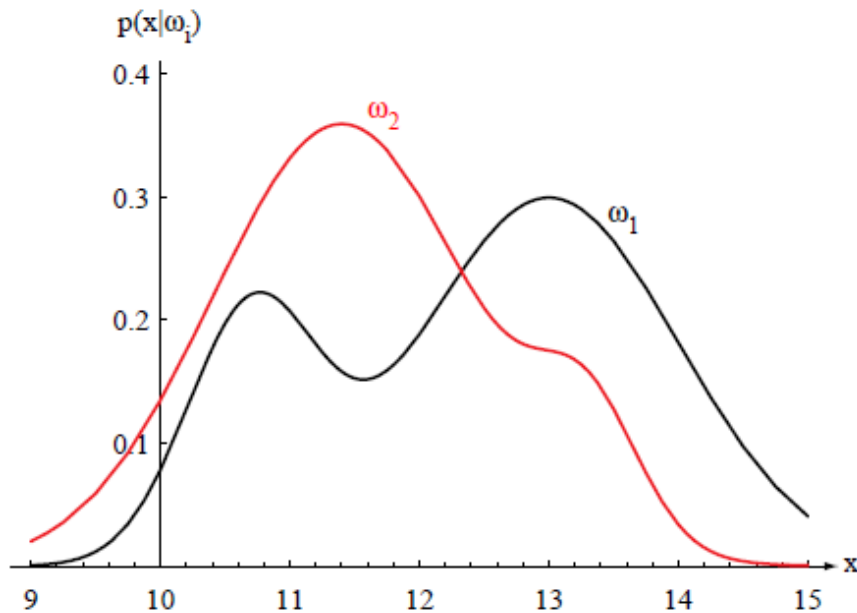
- Salmon ( $\omega_1$ ) and sea bass ( $\omega_2$ )
- If we have only prior probability
  - 例如，教室门口判断进来的是男生还是女生，没有任何传感器
  - Decide  $\omega_1$  if  $P(\omega_1) > P(\omega_2)$ , otherwise  $\omega_2$
  - Minimum error decision

$$P(error) = \begin{cases} P(\omega_2) & \text{if we decide } \omega_1 \\ P(\omega_1) & \text{if we decide } \omega_2 \end{cases}$$

- 教室门口判断性别的例子：错误率？

# 2类的例子

- Decision based on posterior probabilities



$$P(\omega_i | \mathbf{x}) = \frac{p(\mathbf{x} | \omega_i)P(\omega_i)}{p(\mathbf{x})} = \frac{p(\mathbf{x} | \omega_i)P(\omega_i)}{\sum_{j=1}^c p(\mathbf{x} | \omega_j)P(\omega_j)}$$

- Decision based on posterior probabilities

$$P(error|x) = \begin{cases} P(\omega_1|x) & \text{if we decide } \omega_2 \\ P(\omega_2|x) & \text{if we decide } \omega_1. \end{cases}$$

Decide  $\omega_1$  if  $P(\omega_1|x) > P(\omega_2|x)$ ; otherwise decide  $\omega_2$

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

- Evidence (a.k.a. likelihood)

Decide  $\omega_1$  if  $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$ ; otherwise decide  $\omega_2$

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

# 最小风险决策

- 决策代价(loss)

- True class  $\omega_j$ , decided as  $\alpha_i$   $\lambda_{ij} = \lambda(\alpha_i | \omega_j)$

- 有时2类代价相差很大，比如医疗诊断的场合、工业检测、自动商店判断性别

- Condition risk

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

- Overall (expected) risk

$$R = \int R(\alpha(\mathbf{x}) | \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

- Minimum risk decision (Bayes decision)

$$\arg \min_i R(\alpha_i | x)$$



- Minimum risk decision: 2-class case

- Condition risk

$$R(\alpha_1 | \mathbf{x}) = \lambda_{11}P(\omega_1 | \mathbf{x}) + \lambda_{12}P(\omega_2 | \mathbf{x})$$

$$R(\alpha_2 | \mathbf{x}) = \lambda_{21}P(\omega_1 | \mathbf{x}) + \lambda_{22}P(\omega_2 | \mathbf{x})$$

- Decision rule

$$R(\alpha_1 | x) < R(\alpha_2 | x) \leftrightarrow (\lambda_{21} - \lambda_{11})P(\omega_1 | \mathbf{x}) > (\lambda_{12} - \lambda_{22})P(\omega_2 | \mathbf{x})$$

- Equivalently, decide  $\omega_1$  if

$$(\lambda_{21} - \lambda_{11})\underline{p(\mathbf{x}|\omega_1)P(\omega_1)} > (\lambda_{12} - \lambda_{22})p(\mathbf{x}|\omega_2)P(\omega_2)$$

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

# 最小错误率分类

- Zero-one loss

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \quad i, j = 1, \dots, c$$

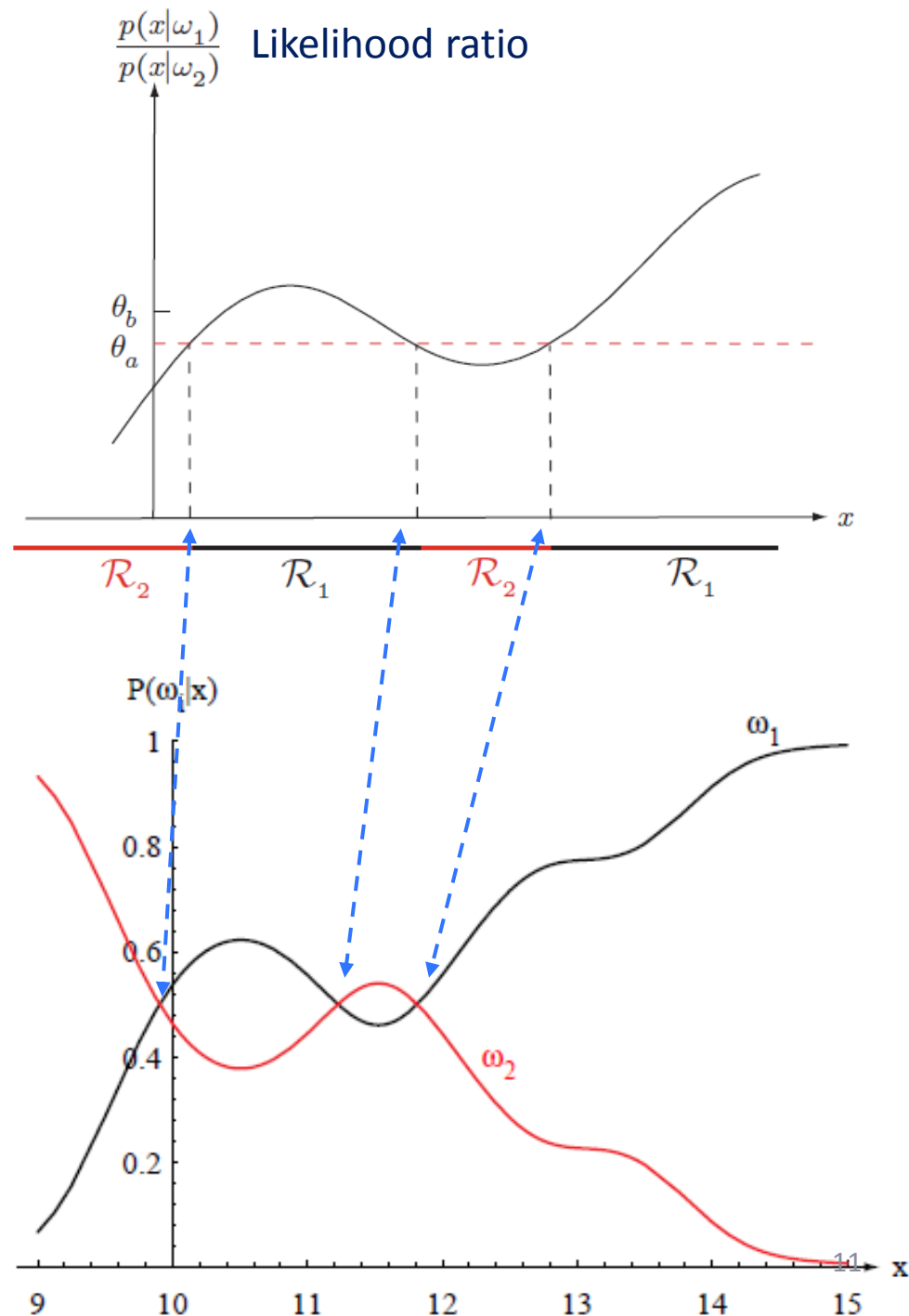
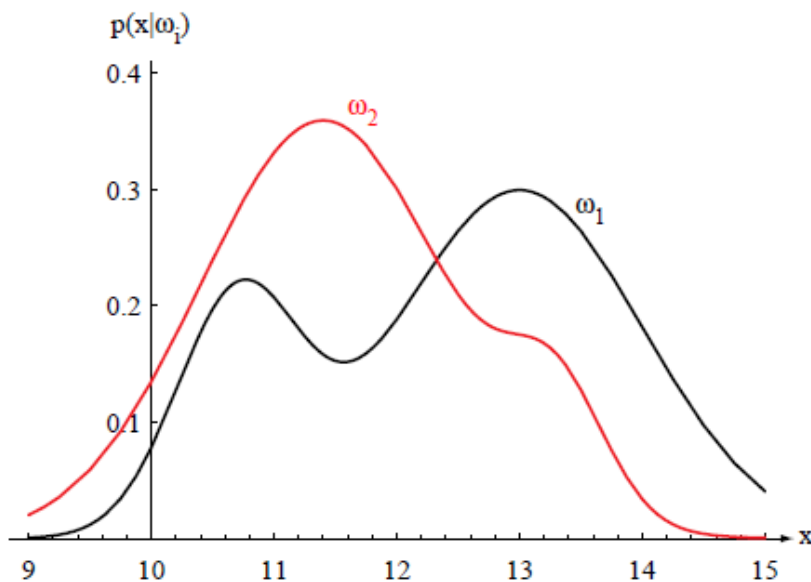
$$\begin{aligned} R(\alpha_i|\mathbf{x}) &= \sum_{j=1}^c \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x}) \\ &= \sum_{j \neq i} P(\omega_j|\mathbf{x}) \\ &= 1 - P(\omega_i|\mathbf{x}) \end{aligned}$$

- Minimum error decision: Maximum a posteriori (MAP)

Decide  $\omega_i$  if  $P(\omega_i|\mathbf{x}) > P(\omega_j|\mathbf{x})$  for all  $j \neq i$

- 2-class case
  - decide  $\omega_1$  if

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \boxed{\quad} \frac{P(\omega_2)}{P(\omega_1)}$$



# 带拒识的决策

- (Problem 13, Chapter 2)

- C+1 classes

$$\lambda(\alpha_i | \omega_j) = \begin{cases} 0, & i = j \\ \lambda_s, & i \neq j \\ \lambda_r, & \text{reject} \end{cases} \quad \lambda_r < \lambda_s$$

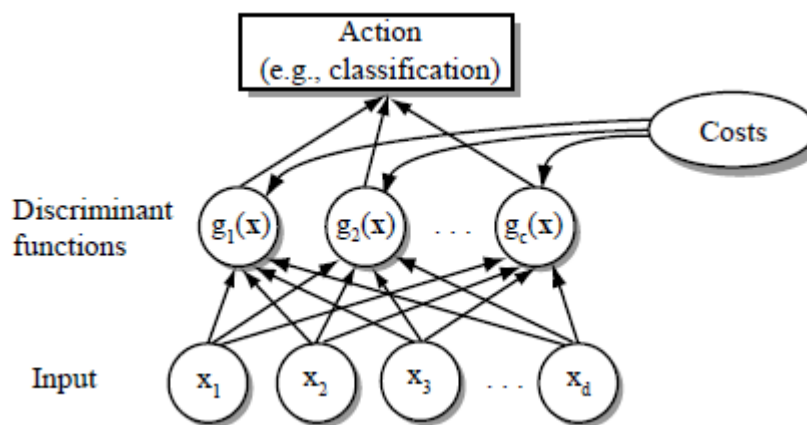
$$R(\alpha_i | \mathbf{x}) = \sum_{j=1}^c \lambda(\alpha_i | \omega_j) P(\omega_j | \mathbf{x})$$

$$\Rightarrow R_i(\mathbf{x}) = \begin{cases} \lambda_s [1 - P(\omega_i | \mathbf{x})], & i = 1, \dots, c \\ \lambda_r, & \text{reject} \end{cases}$$

$$\arg \min_i R_i(\mathbf{x}) = \begin{cases} \arg \max_i P(\omega_i | \mathbf{x}), & \text{if } \max_i P(\omega_i | \mathbf{x}) > 1 - \lambda_r / \lambda_s \\ \text{reject}, & \text{otherwise} \end{cases}$$

# 判别函数、决策面

- 判别函数(Discriminant Function)
  - 表征模式属于每一类的广义似然度  $g_i(\mathbf{x})$ ,  $i=1, \dots, c$
  - 分类决策  $\arg \max_i g_i(\mathbf{x})$
  - E.g., conditional risk  $g_i(\mathbf{x}) = -R(\alpha_i | \mathbf{x})$
  - Posterior probability  $g_i(\mathbf{x}) = P(\omega_i | \mathbf{x})$
  - Likelihood  $g_i(\mathbf{x}) = p(\mathbf{x} | \omega_i)P(\omega_i)$   
 $g_i(\mathbf{x}) = \log p(\mathbf{x} | \omega_i) + \log P(\omega_i)$

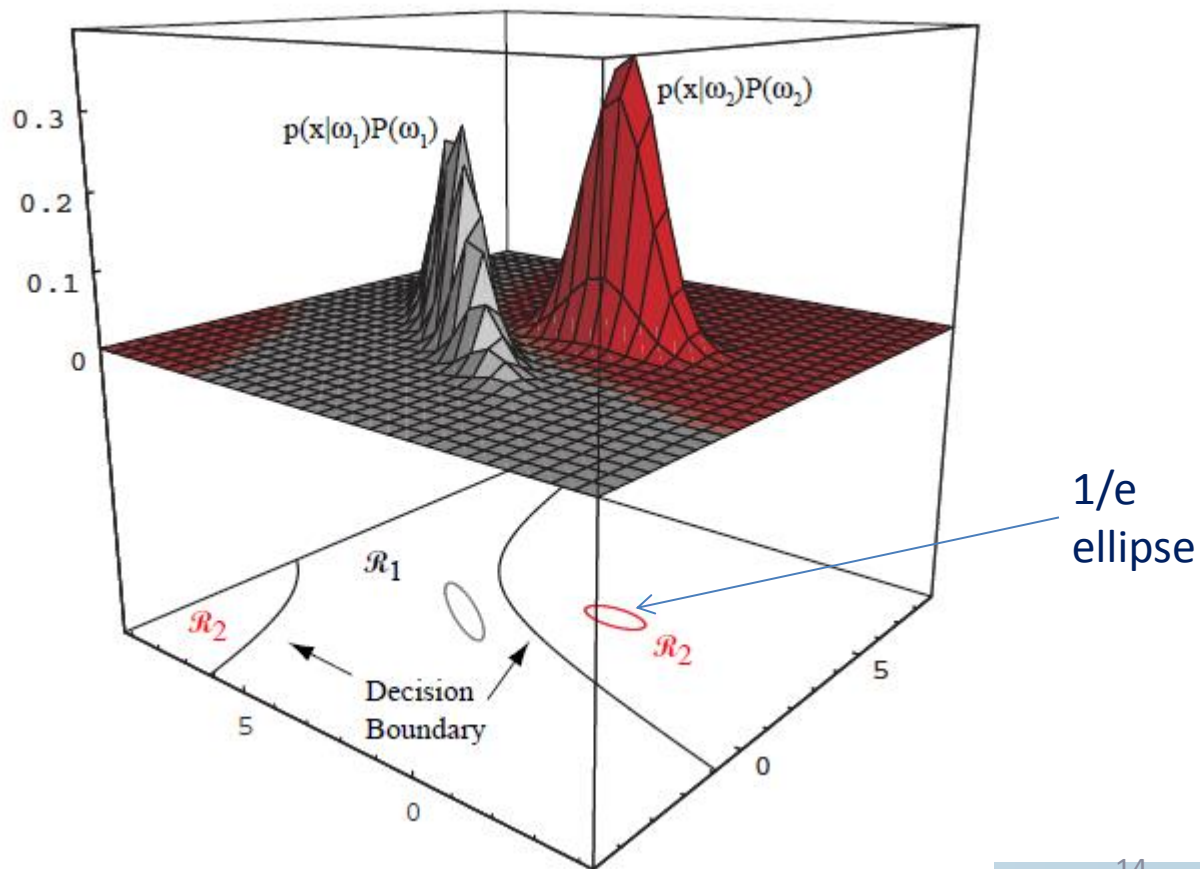


- 决策面(Decision surface)
  - 特征空间中二类判别函数相等的点的集合

$$g(\mathbf{x}) \equiv g_1(\mathbf{x}) - g_2(\mathbf{x}) \quad g(\mathbf{x}) = P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x})$$

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

正态分布下的  
一个例子



# Break

# 高斯密度函数

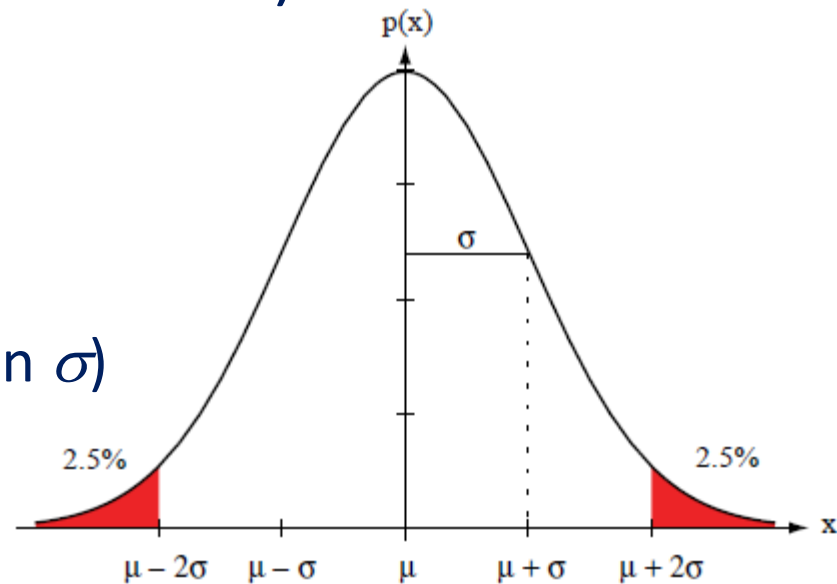
- Gaussian density (normal distribution)

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

- Mean  $\mu$
- Variance  $\sigma^2$  (standard deviation  $\sigma$ )

$$\mu \equiv \mathcal{E}[x] = \int_{-\infty}^{\infty} xp(x) dx$$

$$\sigma^2 \equiv \mathcal{E}[(x - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx$$



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

- 在给定均值和方差的所有分布中，正态分布的熵最大(Problem 20, Chapter 2)
- 根据Central Limit Theorem，大量独立随机变量之和趋近正态分布
- 实际环境中，很多类别的特征分布趋近正态分布



- Multivariate normal density  $p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- 公式要牢记

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

- Mean  $\boldsymbol{\mu} \equiv \mathcal{E}[\mathbf{x}] = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x} \quad \mu_i = \mathcal{E}[x_i]$

- Covariance matrix

$$\boldsymbol{\Sigma} \equiv \mathcal{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t] = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t p(\mathbf{x}) d\mathbf{x}$$

$$\sigma_{ij} = \mathcal{E}[(x_i - \mu_i)(x_j - \mu_j)]$$

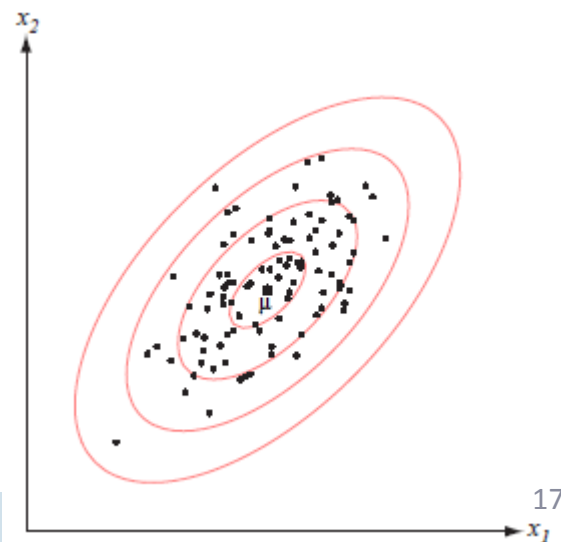
If  $x_i$  and  $x_j$  are statistically independent,  $\sigma_{ij} = 0$

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{bmatrix}$$

- 等密度点轨迹: hyperellipsoid

- Mahalanobis distance

$$r^2 = (\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$



- Covariance matrix eigenvalues & eigenvectors

$$\Sigma = \Phi \Lambda \Phi^t \quad \Phi = [\phi_1 \phi_2 \cdots \phi_d] \quad \Lambda = \text{diag}[\lambda_1, \lambda_2, \cdots, \lambda_d]$$

- Orthonormal  $\Phi^t \Phi = I \quad \Phi^t = \Phi^{-1}$

- 线性变换  $y = A^t x$

- $A^t A = 1$ : 正交变换(坐标轴旋转)

- 变换后的分布仍为正态分布

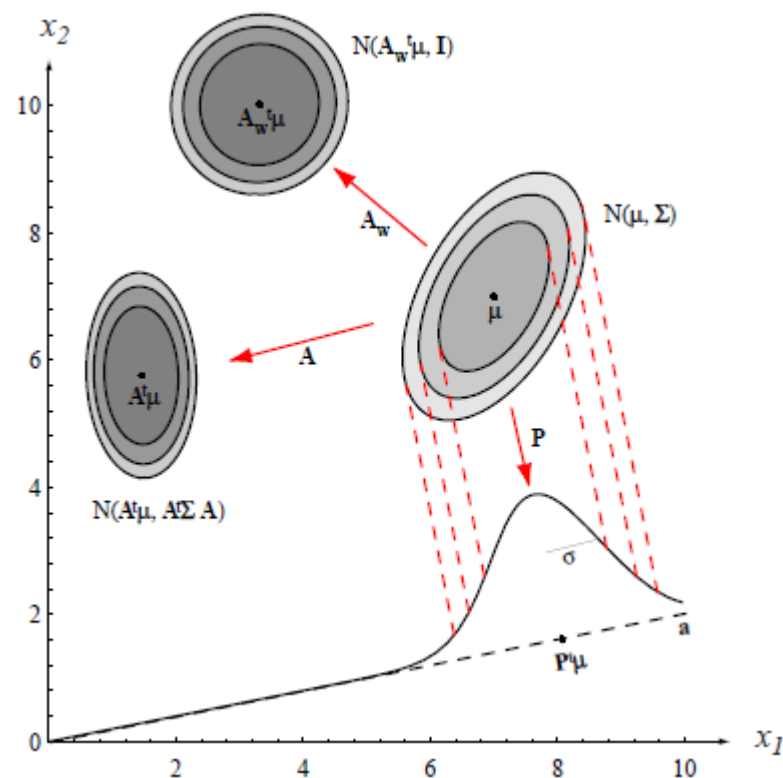
$$p(y) \sim N(A^t \mu, A^t \Sigma A)$$

- Whitening transform

$$A_w = \Phi \Lambda^{-1/2}$$

$$A_w^t \Sigma A_w = \Lambda^{-1/2} \Phi^t \Sigma \Phi \Lambda^{-1/2}$$

$$= \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = I$$



# 高斯密度下的判别函数

- 判别函数  $g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$

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

$$g_i(\mathbf{x}) = -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^t \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i)$$

- Quadratic discriminant function (QDF)
- 在不同covariance假设条件下得到一些特殊形式

- Case 1:  $\Sigma_i = \sigma^2 I$

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln P(\omega_i)$$

- Euclidean distance  $\|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = (\mathbf{x} - \boldsymbol{\mu}_i)^t(\mathbf{x} - \boldsymbol{\mu}_i)$

- 展开二次式  $(\mathbf{x} - \boldsymbol{\mu}_i)^t(\mathbf{x} - \boldsymbol{\mu}_i)$

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}[\mathbf{x}^t\mathbf{x} - 2\boldsymbol{\mu}_i^t\mathbf{x} + \boldsymbol{\mu}_i^t\boldsymbol{\mu}_i] + \ln P(\omega_i)$$

- 忽略与类别无关项，得到线性判别函数

$$g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0}$$

$$\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i \quad w_{i0} = \frac{-1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \ln P(\omega_i)$$

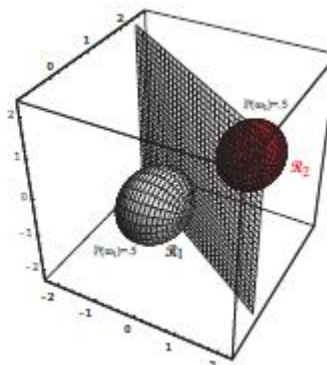
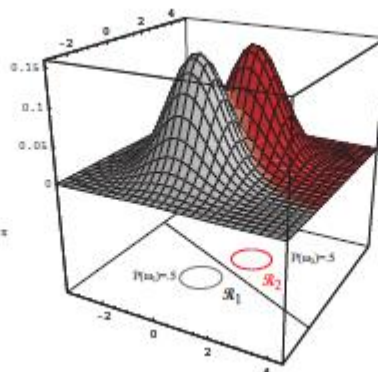
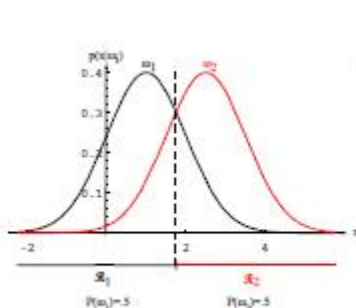
- 二类决策面  $g_i(\mathbf{x}) = g_j(\mathbf{x})$

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x}_0) = 0 \quad \mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$$

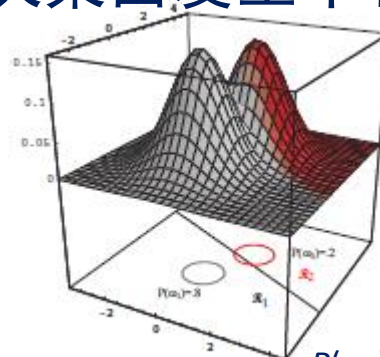
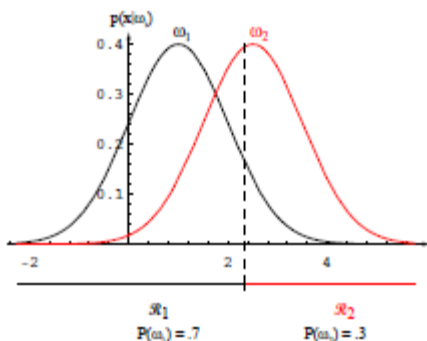
$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

## – 1D, 2D, 3D的情况

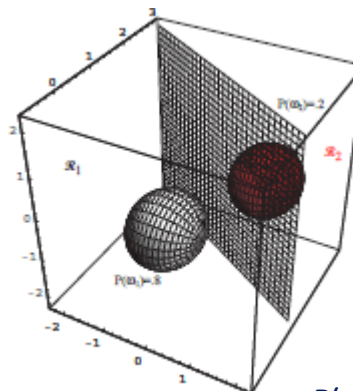
- 当 $P(\omega_1) = P(\omega_2)$ ，决策面为二类均值的等分面



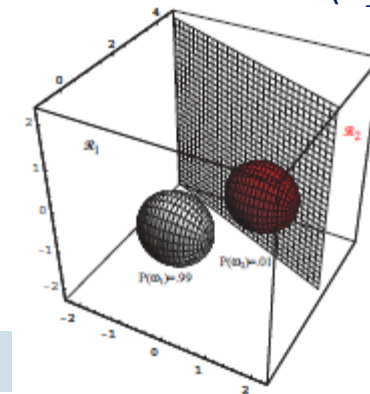
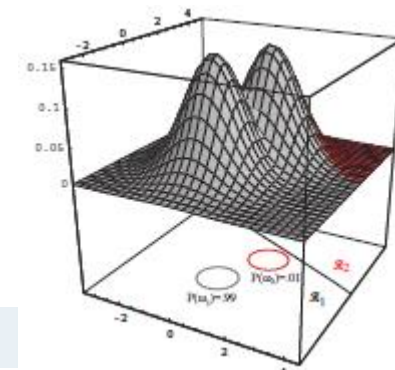
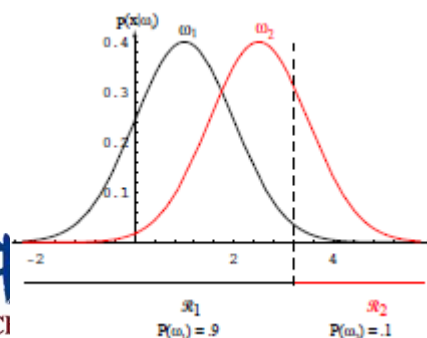
## – 当先验概率变化，决策面发生平移



$P(\omega_1) = 0.8$



$P(\omega_1) = 0.99$



- Case 1:  $\Sigma_i = \Sigma$

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

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

– 展开二次式  $(\mathbf{x} - \mu_i)^t \Sigma^{-1}(\mathbf{x} - \mu_i)$

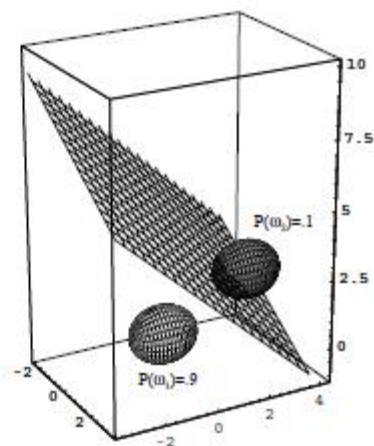
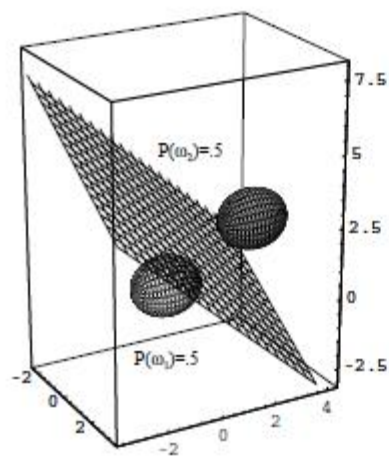
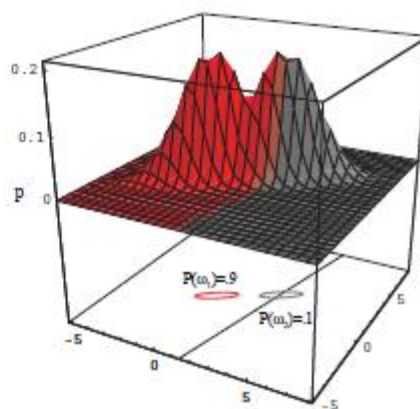
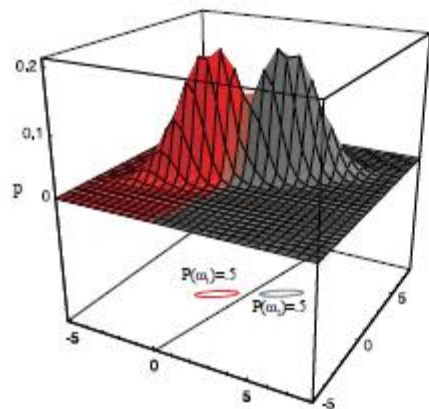
线性判别函数!  $g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + w_{i0}$

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

– 二类决策面  $\mathbf{w}^t(\mathbf{x} - \mathbf{x}_0) = 0 \quad \mathbf{w} = \Sigma^{-1}(\mu_i - \mu_j)$

$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln [P(\omega_i)/P(\omega_j)]}{(\mu_i - \mu_j)^t \Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)$$

- 注意跟 $\mu_1 - \mu_2$ 的关系，决策面不一定与之垂直
- 当 $P(\omega_1) = P(\omega_2)$ ，决策面经过 $(\mu_1 + \mu_2)/2$





- Case 3:  $\Sigma_i =$  arbitrary

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

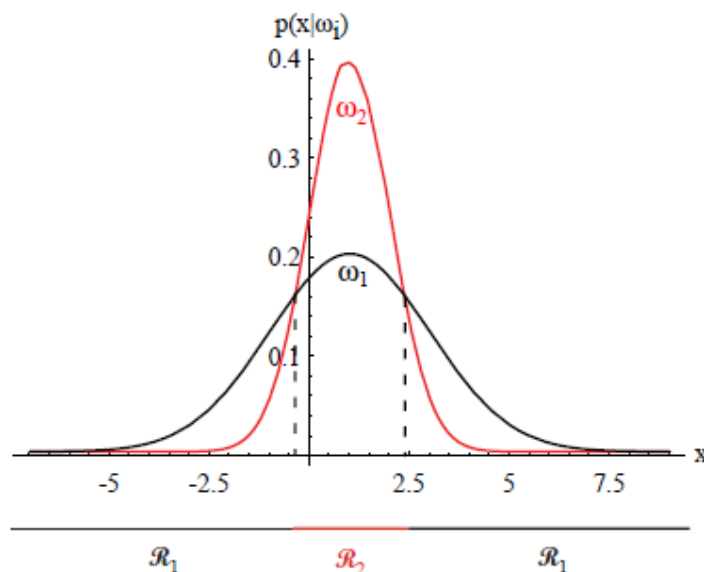
$$g_i(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t \mathbf{x} + w_{i0}$$

$$\mathbf{W}_i = -\frac{1}{2} \boldsymbol{\Sigma}_i^{-1} \quad \mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i$$

$$w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

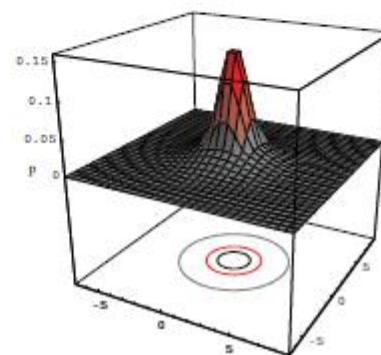
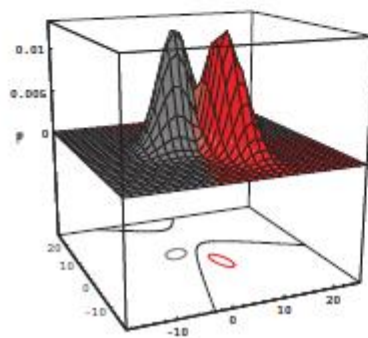
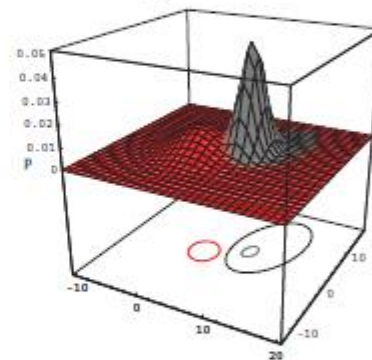
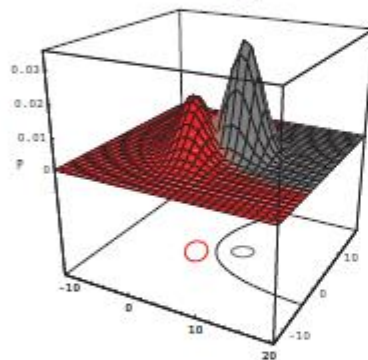
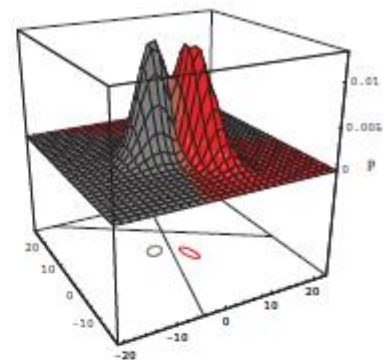
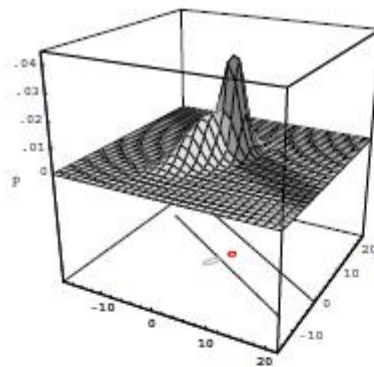
– 二类决策面:  $g_1(\mathbf{x}) = g_2(\mathbf{x})$ , hyperquadrics

- 等均值的情况下, 1D的例子

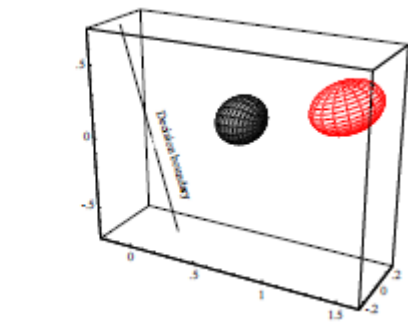
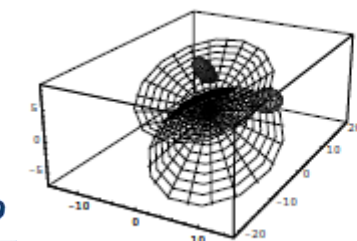
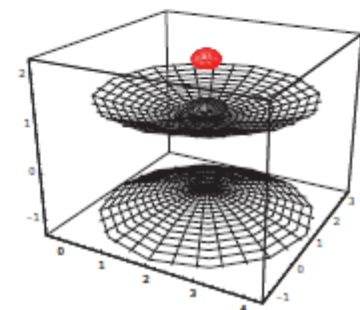
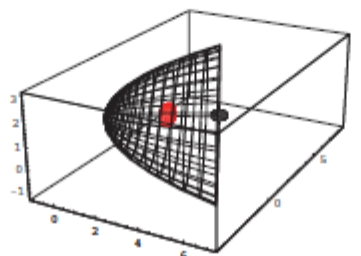
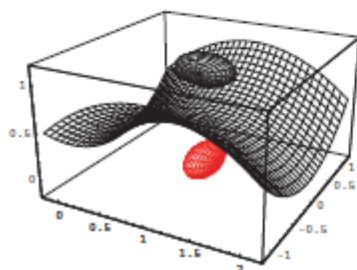
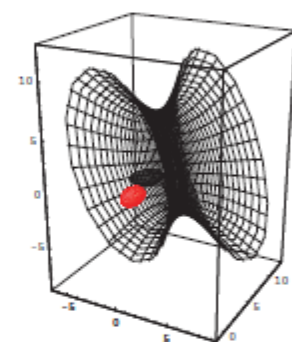
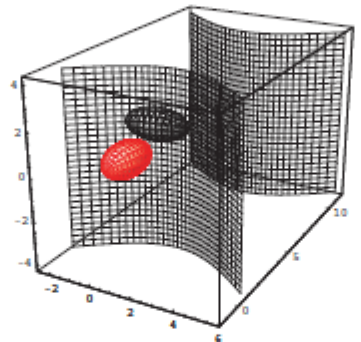
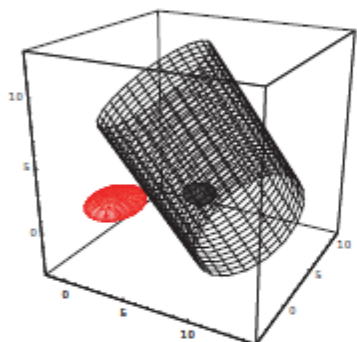
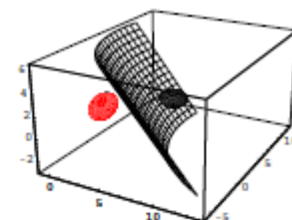
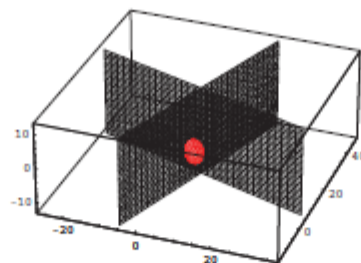
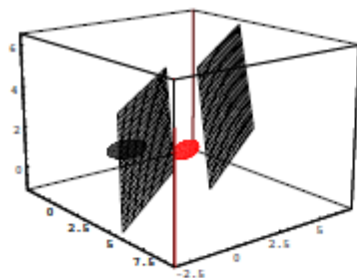




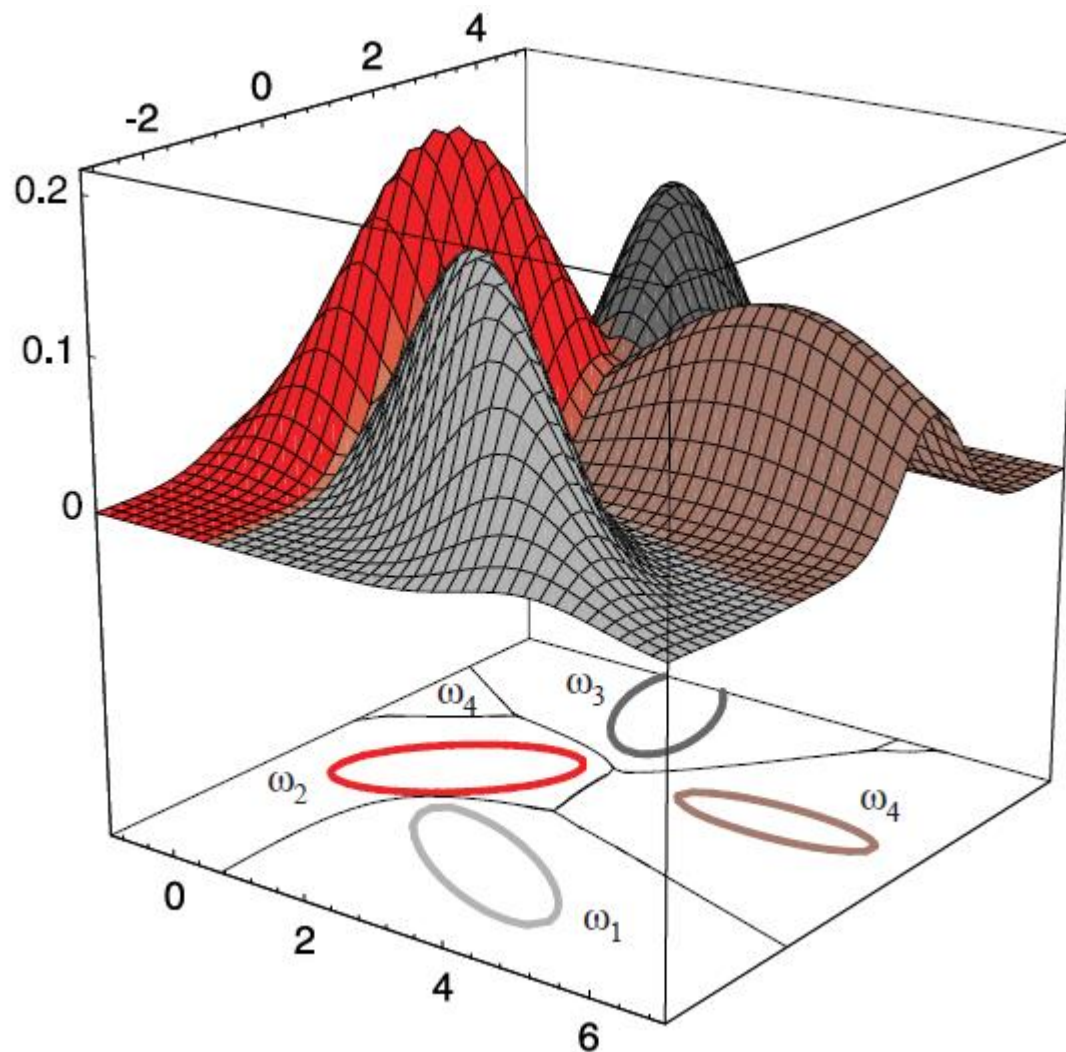
## 2D的例子



## 3D的例子



## 2D, 4类的例子



- 一个具体例子

- 2类, 2D  $P(\omega_1) = P(\omega_2) = 0.5$

$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix}$$

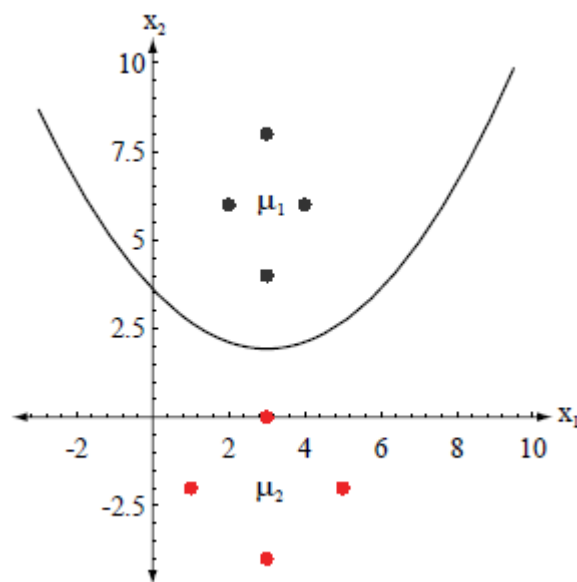
$$\Sigma_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

$$\mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

$$\Sigma_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

- 决策面  $g_1(\mathbf{x}) = g_2(\mathbf{x})$

$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2$$



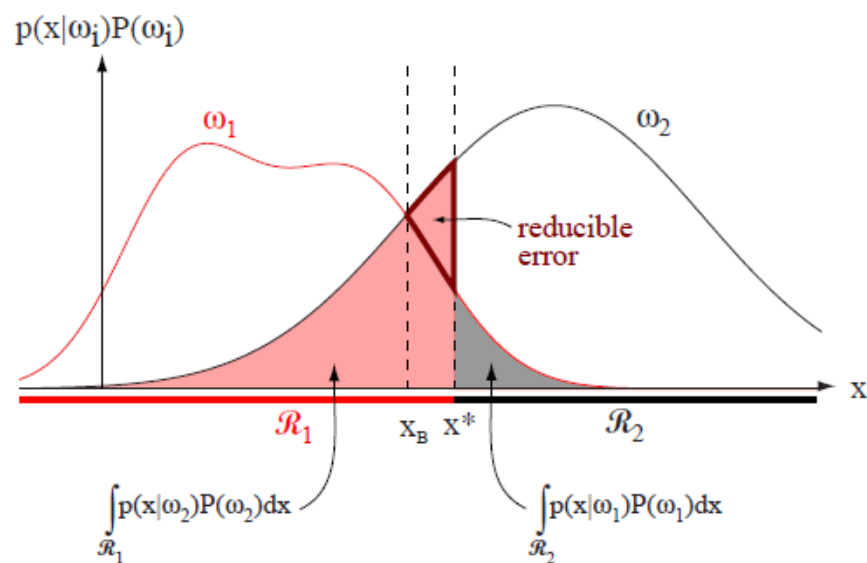
# 分类错误率

- 2类情况

$$\begin{aligned}P(error) &= P(x \in \mathcal{R}_2, \omega_1) + P(x \in \mathcal{R}_1, \omega_2) \\&= P(x \in \mathcal{R}_2 | \omega_1)P(\omega_1) + P(x \in \mathcal{R}_1 | \omega_2)P(\omega_2) \\&= \int_{\mathcal{R}_2} p(x|\omega_1)P(\omega_1) dx + \int_{\mathcal{R}_1} p(x|\omega_2)P(\omega_2) dx.\end{aligned}$$

- 一般情况

$$\begin{aligned}P(correct) &= \sum_{i=1}^c P(x \in \mathcal{R}_i, \omega_i) \\&= \sum_{i=1}^c P(x \in \mathcal{R}_i | \omega_i)P(\omega_i) \\&= \sum_{i=1}^c \int_{\mathcal{R}_i} p(x|\omega_i)P(\omega_i) dx\end{aligned}$$

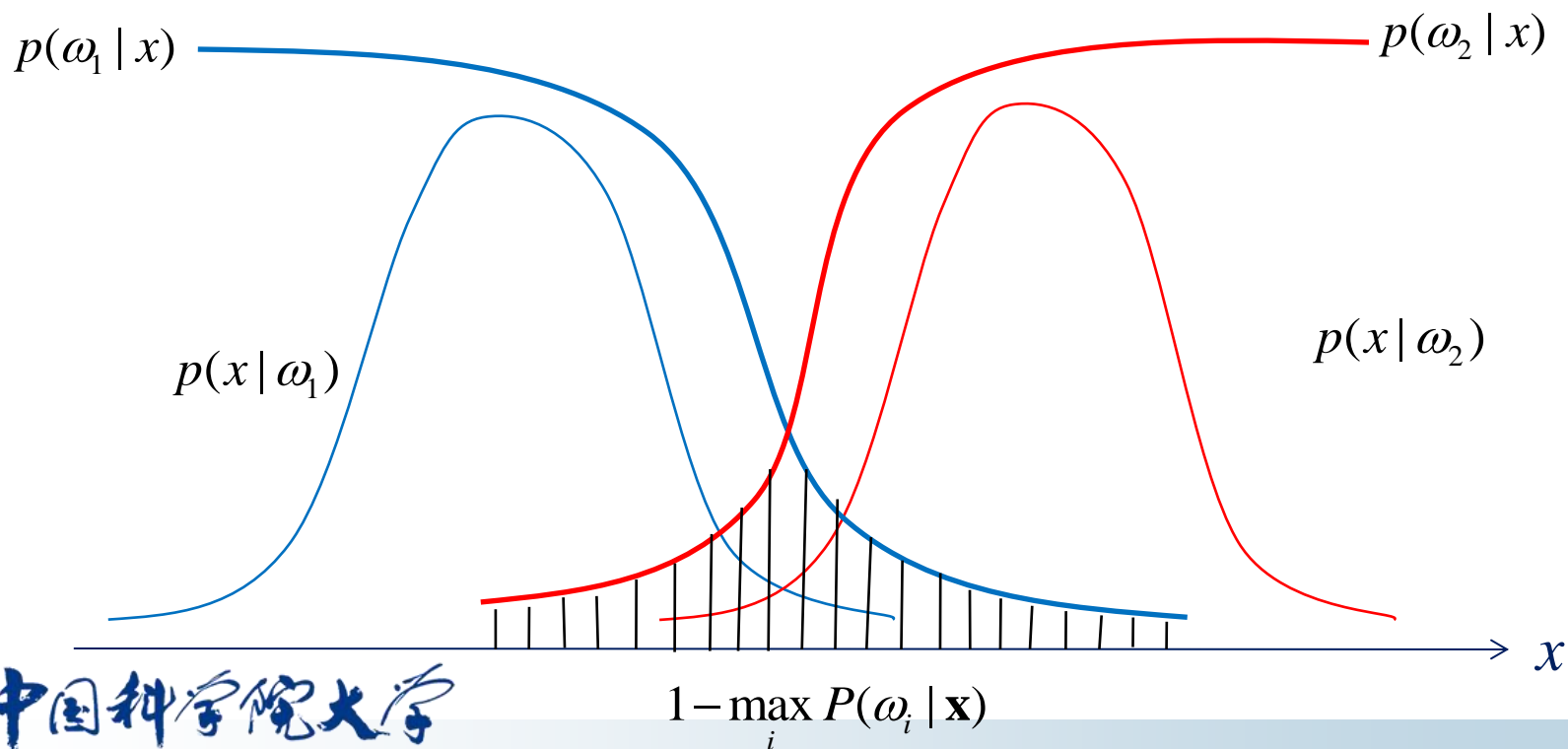


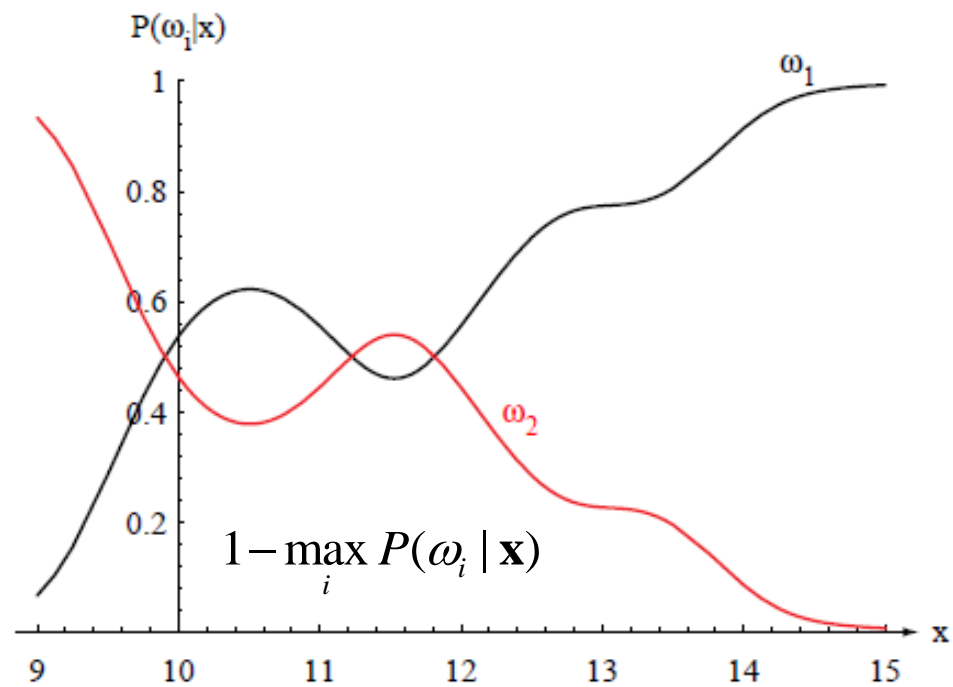
- 最大后验概率决策(0-1 loss)的情况

$$P(\text{correct}) = \int_{\mathbf{x}} \max_i P(\mathbf{x} | \omega_i) P(\omega_i) d\mathbf{x}$$

$$= \int_{\mathbf{x}} \max_i P(\omega_i | \mathbf{x}) P(\mathbf{x}) d\mathbf{x}$$

$$P(\text{error}) = \int_{\mathbf{x}} \left[ 1 - \max_i P(\omega_i | \mathbf{x}) \right] P(\mathbf{x}) d\mathbf{x}$$







# 讨论

- 贝叶斯分类器(基于贝叶斯决策的分类器)是最优的吗？
  - 最小风险、最大后验概率决策
  - 最优的条件：概率密度、风险能准确估计
  - 具体的参数法、非参数法是贝叶斯分类器的近似，实际中难以达到最优
- 判别模型：回避了概率密度估计，以较小复杂度估计后验概率或判别函数
- 什么方法能胜过贝叶斯分类器：在不同的特征空间！



# 下次课内容

- 第2章
  - 离散变量的贝叶斯决策
  - 复合模式分类
- 第3章
  - 最大似然参数估计
  - 贝叶斯估计