第2章：贝叶斯决策理论
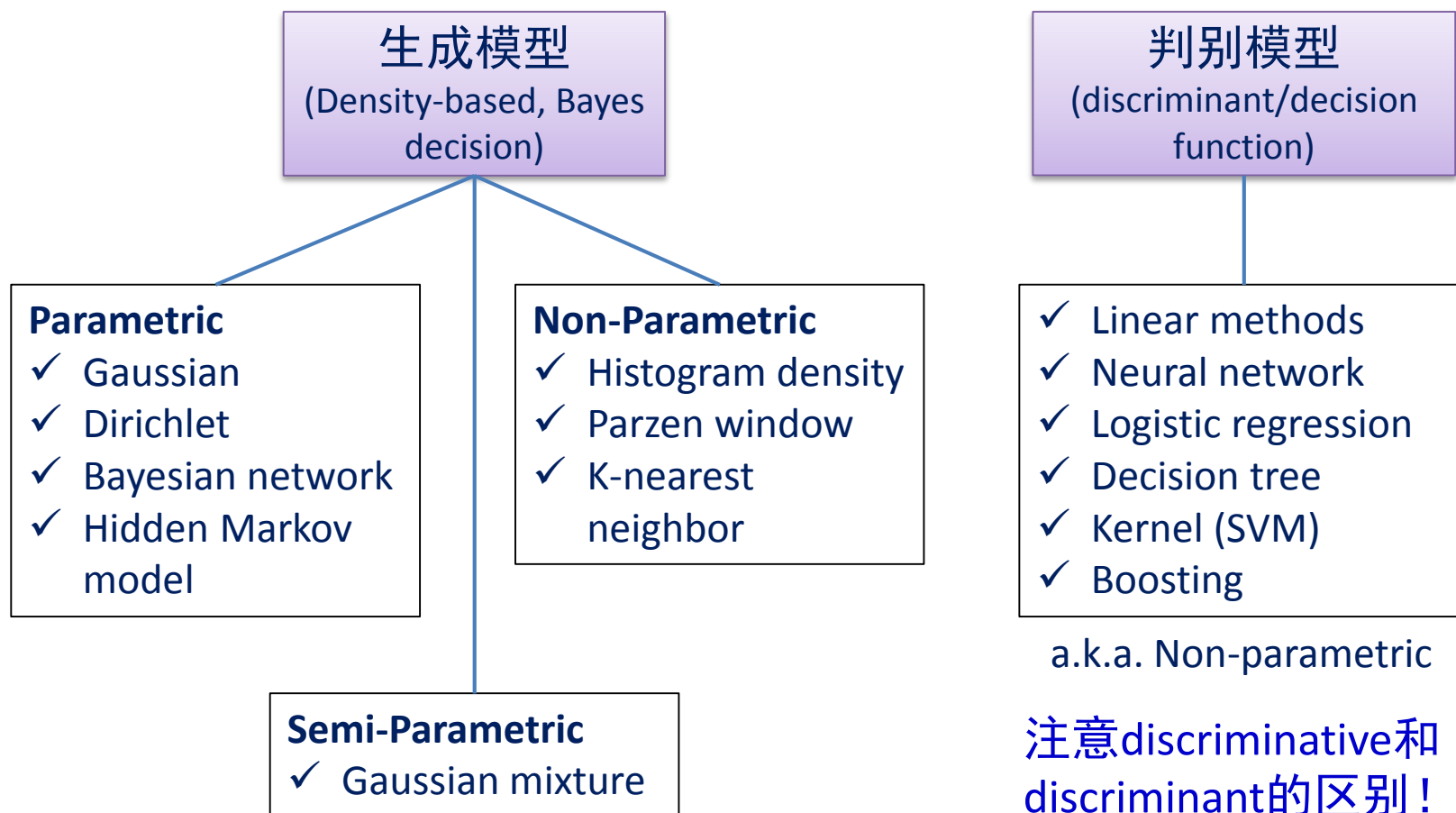
刘成林(liucl@nlpr.ia.ac.cn)

2019年9月11日

助教：王铁强(tieqiang.wang@nlpr.ia.ac.cn)
陈卓(zhuo.chen@nlpr.ia.ac.cn)
吴金文(jinwen.wu@nlpr.ia.ac.cn)

中国科学院大学
University of Chinese Academy of Sciences

统计模式识别方法

生成模型
(Density-based, Bayes decision)

判别模型
(discriminant/decision function)

Parametric
- ✓ Gaussian
- ✓ Dirichlet
- ✓ Bayesian network
- ✓ Hidden Markov model

Non-Parametric
- ✓ Histogram density
- ✓ Parzen window
- ✓ K-nearest neighbor

- ✓ Linear methods
- ✓ Neural network
- ✓ Logistic regression
- ✓ Decision tree
- ✓ Kernel (SVM)
- ✓ Boosting

a.k.a. Non-parametric

Semi-Parametric
- ✓ Gaussian mixture

注意discriminative和
discriminant的区别！
Discriminative model/learning
Discriminant function/analysis

University of Chinese Academy of Sciences

提 纲

- 问题表示
- 最小错误率决策：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

University of Chinese Academy of Sciences

导论：问题表示

- 类别： $\omega_i, i = 1, ..., c$

- 特征矢量 $\mathbf{x} = [x_1, ..., x_d] \in R^d$

- 先验概率 $P(\omega_i)$ $\sum\limits_{i=1}^{c} P(\omega_i) = 1$

- 概率密度函数(条件概率) $p(\mathbf{x} \mid \omega_i)$

- 后验概率

$$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i) P(\omega_i)}{p(\mathbf{x})} = \frac{p(\mathbf{x} \mid \omega_i) P(\omega_i)}{\sum\limits_{j=1}^{c} p(\mathbf{x} \mid \omega_j) P(\omega_j)}$$

$$\sum\limits_{i=1}^{c} P(\omega_i \mid \mathbf{x}) = 1$$

最小错误率决策：2类的例子

- Salmon (ω_1) and sea bass (ω_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 ω_1 if $P(\omega_1) > P(\omega_2)$, otherwise ω_2
  - Minimum error decision

$$P(error) = \begin{cases} P(\omega_2) & \text{if we decide } \omega_1 \\ P(\omega_1) & \text{if we decide } \omega_2 \end{cases}$$
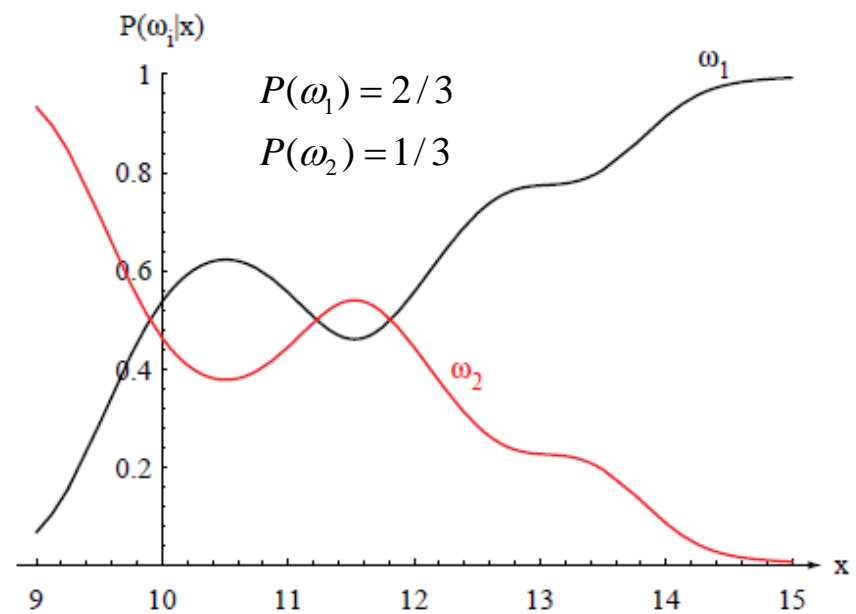
  - 教室门口判断性别的例子：错误率？

2类的例子

- 有传感器(特征)的情况
  Decision based on posterior probabilities

$$P(\omega_1) = 2/3$$
$$P(\omega_2) = 1/3$$
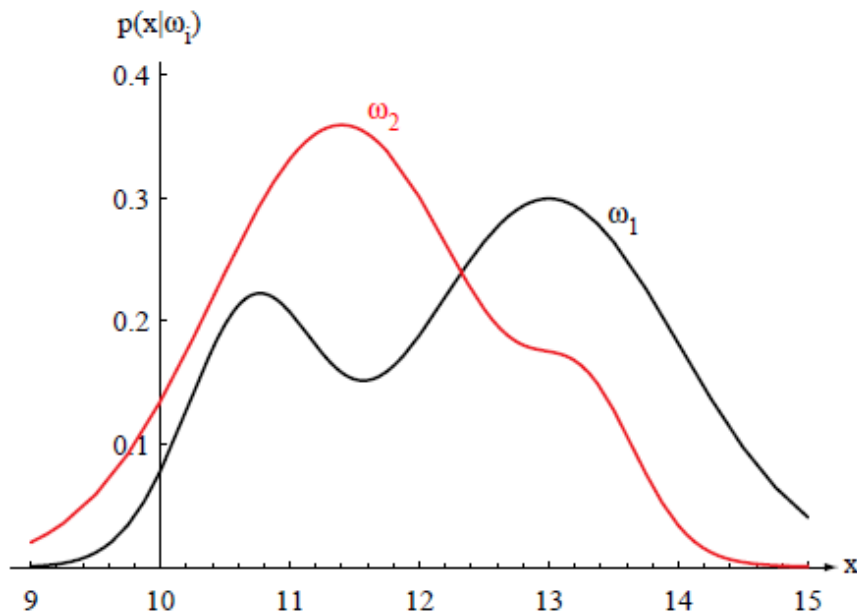
x轴：一维特征空间

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

- 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 ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; otherwise decide ω_2

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

- Evidence (a.k.a. likelihood)

Decide ω_1 if $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$; otherwise decide ω_2

– see $$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i)P(\omega_i)}{p(\mathbf{x})}$$

教室门口判断性别的例子：用什么传感器(**x**)?

University of Chinese Academy of Sciences

最小风险决策：贝叶斯决策的一般形式

- 决策代价(loss)
  - True class ω_j, decided as α_i $\quad \lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$
    - 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)

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

- Minimum risk decision: 2-class case
  - Condition risk

$$
\begin{aligned}
R(\alpha_1|\mathbf{x}) &= \lambda_{11}P(\omega_1|\mathbf{x}) + \lambda_{12}P(\omega_2|\mathbf{x}) \\
R(\alpha_2|\mathbf{x}) &= \lambda_{21}P(\omega_1|\mathbf{x}) + \lambda_{22}P(\omega_2|\mathbf{x})
\end{aligned}
$$

  - 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 ω_1 if

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

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

University of Chinese Academy of Sciences

最小错误率分类

- Zero-one loss

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

$$\begin{aligned} R(\alpha_i|\mathbf{x}) &= \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x}) \\ &= \sum_{j\neq i} P(\omega_j|\mathbf{x}) \\ &= 1 - P(\omega_i|\mathbf{x}) \end{aligned}$$
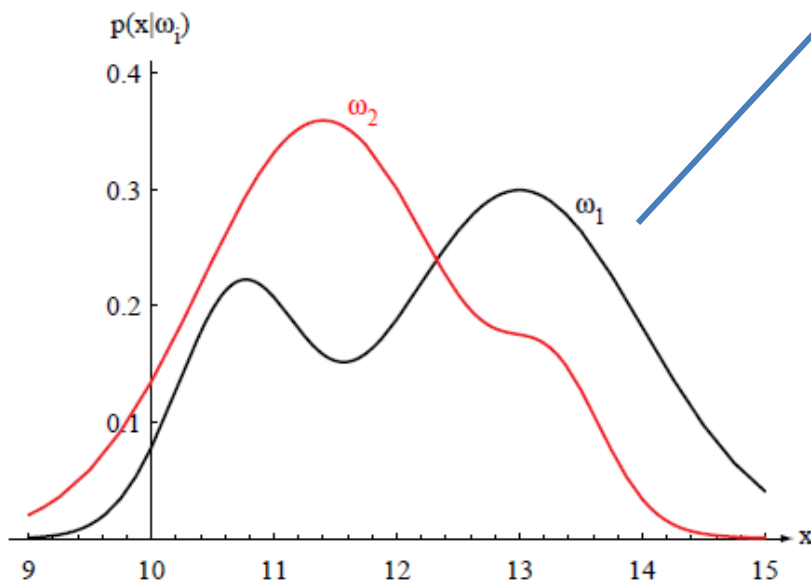
- Minimum error decision: Maximum a posteriori (MAP)

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

University of Chinese Academy of Sciences

- 2-class case
  - decide ω_1 if

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

0-1 loss

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}$$
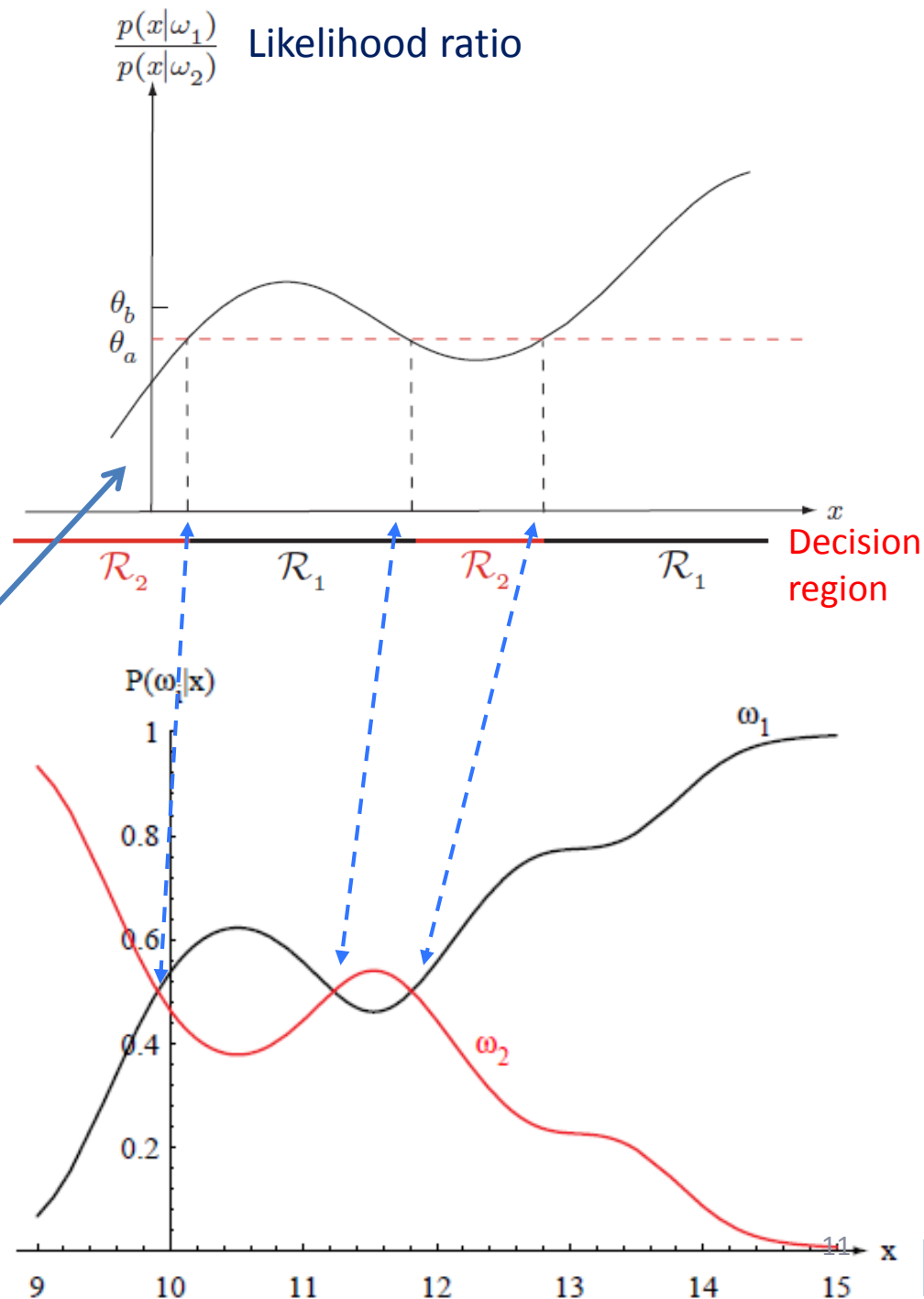
带拒识的决策

- 为什么要拒识？错误识别可能带来严重后果
  - 比如医疗诊断，金额识别

Ambiguity Rejection

6/0?

7/1?   4/9?

6/5?   42/32/312?

Distance Rejection

In fact, the Tories made it worse now for the sick and needy than Labour had to make it in 1950. And as a percentage of social serice expenditure, health had fallen from 28.5 to 23.1 percent.

English is outlier for a digit recognizer

経営不振に陥っているソニーのパソコン事業は国内 投資ファンド が買い取り、開発と製造をてがける新会社バイオをつくった。ソニーの国内向け通販サイトと同社の直営店を通じ、個人から注文をとってきた。

These are outlier for an English recognizer

带拒识的决策

- Formulation (Problem 13, Chapter 2)
  - C+1 classes

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

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

$$\Longrightarrow \quad R_i(\mathbf{x}) = \begin{cases} \lambda_s[1 - P(\omega_i \mid \mathbf{x})], & i = 1, \ldots, c \\ \lambda_r, & \text{reject} \end{cases}$$

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

University of Chinese Academy of Sciences

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 $\quad \arg\max_i g_i(\mathbf{x})$
  - E.g., conditional risk $\quad g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x})$
  - Posterior probability $\quad g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x})$
  - Likelihood $\quad g_i(\mathbf{x}) = p(\mathbf{x} \mid \omega_i)P(\omega_i)$

$$g_i(\mathbf{x}) = \log p(\mathbf{x} \mid \omega_i) + \log P(\omega_i)$$
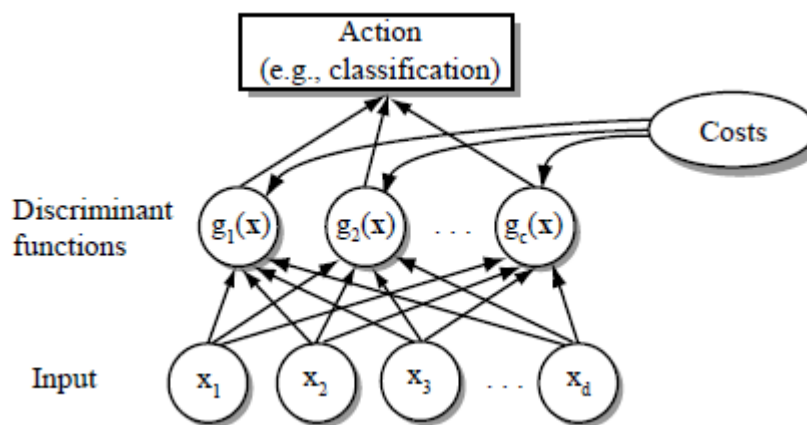
- 决策面(Decision surface)
  - 特征空间中二类判别函数相等的点的集合

$$g(\mathbf{x}) \equiv g_1(\mathbf{x}) - g_2(\mathbf{x}) \qquad g(\mathbf{x}) = P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x}) = 0$$

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

有判别函数就可以分类了，为什么还来求决策面？
答：加深对特征空间和分类器性质的理解
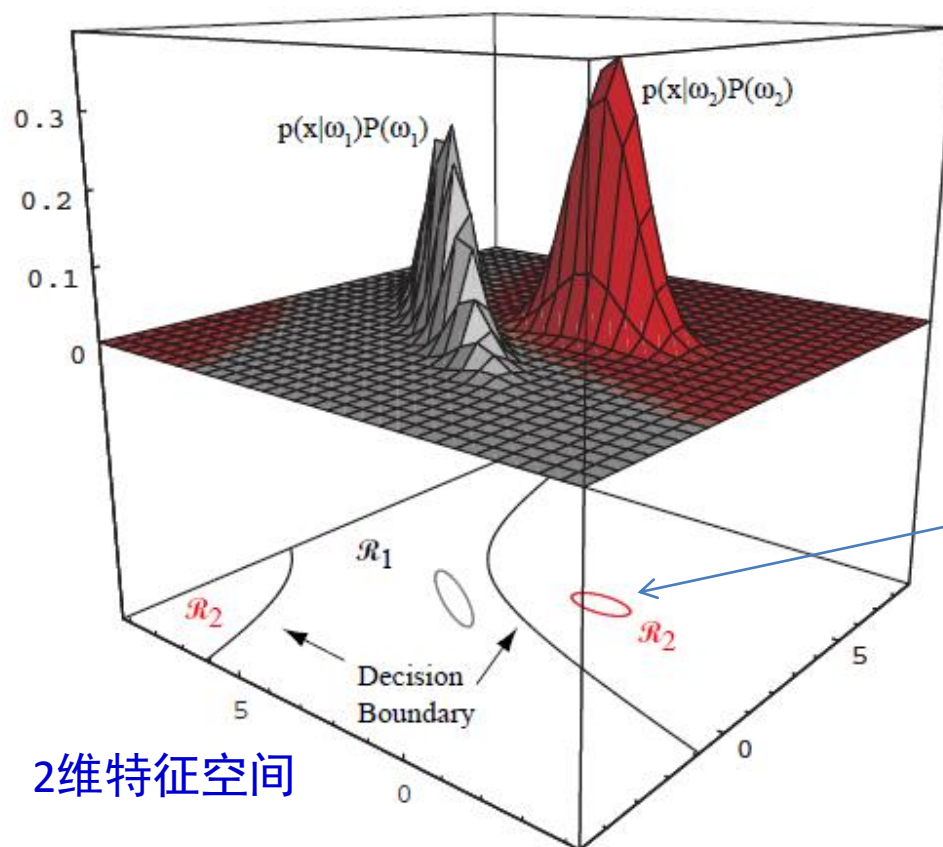
正态分布下的一个例子

Density 1/e ellipse

2维特征空间

贝叶斯决策用于模式分类

- Bayes决策的关键
  - 类条件概率密度估计
  - 先验概率：从训练样本估计或假设等概率
  - 决策代价[$λ_{ij}$]，一般为0-1代价
- 分类器设计
  - 收集训练样本
  - 用每一类的样本估计类条件概率密度 $p(\mathbf{x}\,|\,\omega_i)$
  - 估计类先验概率
  - 模型参数集：$\{p(\mathbf{x}\,|\,\omega_i, \theta_i),\, P(\omega_i)\},\, i=1,...,M$
- 分类过程
  - 计算测试样本属于每一类的后验概率
  - 最大后验概率/最小风险决策

概率密度估计方法

- 参数法：假定概率密度函数形式

$$p(\mathbf{x} \mid \omega_i) = p(\mathbf{x} \mid \theta_i)$$

  - Distribution: Gaussian, Gamma, Bernouli
  - Estimation: maximum-likelihood (ML), Bayesian estimation

- 非参数法：可以表示任意概率分布
  - Parzen window, k-NN
  - 需要保存所有或大部分样本

- Semi-parametric
  - Distribution: Gaussian mixture (GM)
  - Estimation: expectation-maximization (EM)

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 μ
- Variance σ^2 (standard deviation σ)

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

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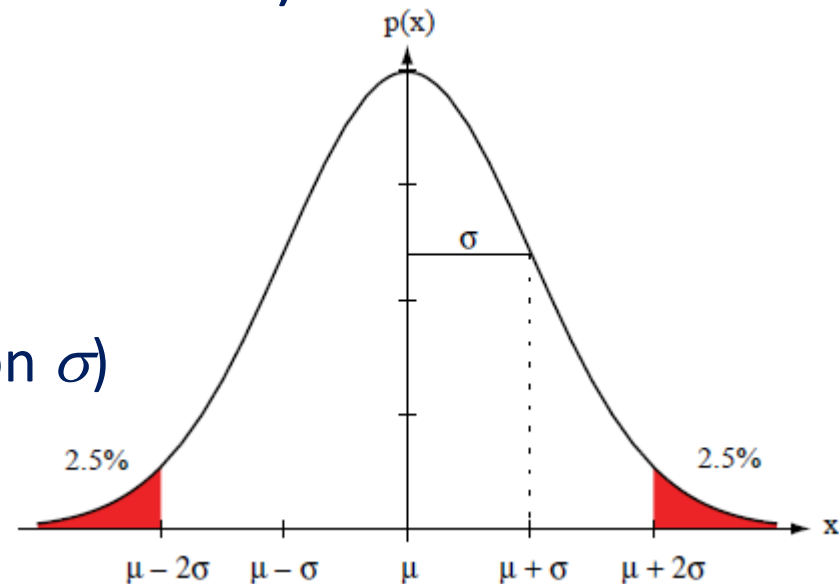
$$H(p(x)) = -\int p(x)\ \ln p(x)\ dx$$

- 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

University of Chinese Academy of Sciences

18

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

  - 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*, $\sigma_{ij} = 0$

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

  - 等密度点轨迹：hyperellipsoid
  - Mahalanobis distance

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

2D space
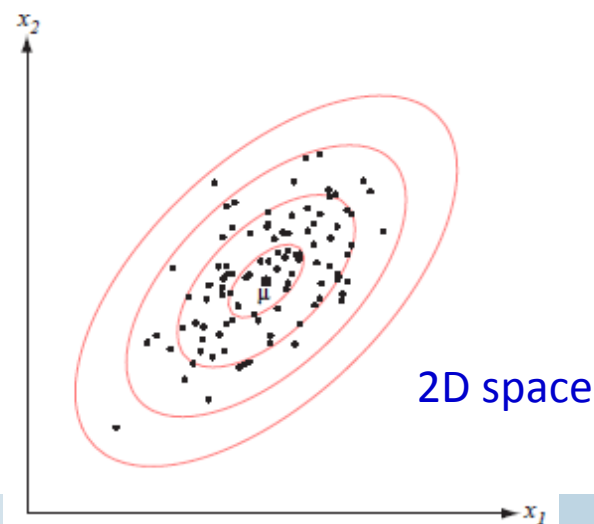
Break

协方差矩阵的性质

- 实对称矩阵

- Eigenvalues & eigenvecters (本征值，本征向量)

$$\Sigma \phi_i = \lambda_i \phi_i \qquad \Phi = [\phi_1 \phi_2 \cdots \phi_d] \qquad \Lambda = diag[\lambda_1, \lambda_2, \cdots, \lambda_d]$$

    – Orthonormal

$$\Phi^T \Phi = I$$
$$\Phi^T = \Phi^{-1}$$
$$\Longleftrightarrow \qquad \phi_i^t \phi_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

    – 矩阵表示

$$\Sigma \Phi = \Phi \Lambda \quad \leftrightarrow \quad \Sigma = \Phi \Lambda \Phi^T \quad \leftrightarrow \quad \Sigma = \sum_{i=1}^{d} \lambda_i \phi_i \phi_i^T$$

- 矩阵对角化

$$\Phi^T \Sigma \Phi = \Lambda = \begin{bmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \lambda_d \end{bmatrix} \quad \Longleftrightarrow \quad \phi_i^T \Sigma \phi_i = \lambda_i$$

- 应用：Principal component analysis (PCA)
  - 一种降维(特征提取)方法
  - 将随机矢量投影到低维子空间，使子空间投影的重建误差最小
  - 选择本征值最大的m (m<d)个本征向量作为子空间的基(basis)

线性空间中正交变换($\Phi^T \Phi = I$) 不影响欧氏距离

$$\sum_{j=1}^{d} [(\mathbf{x} - \mu)^T \phi_j]^2 = \| \mathbf{x} - \mu \|^2$$

子空间投影 $\quad \mu + \sum_{j=1}^{m} [(\mathbf{x} - \mu)^T \phi_j] \phi_j$

投影重建误差

$$r_E = \| \mathbf{x} - \mu \|^2 - \sum_{j=1}^{m} [(\mathbf{x} - \mu)^T \phi_j]^2 = \sum_{j=m+1}^{d} [(\mathbf{x} - \mu)^T \phi_j]^2$$
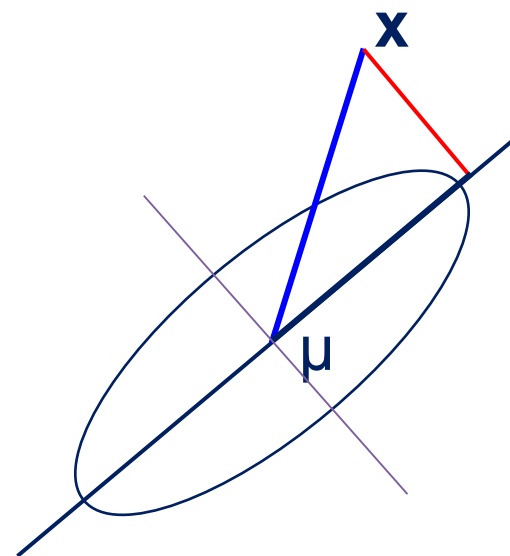
期望 $\quad \mathcal{E}(r_E) = \mathcal{E} \left\{ \sum_{j=m+1}^{d} [(\mathbf{x} - \mu)^T \phi_j]^2 \right\} = \mathcal{E} \left\{ \sum_{j=m+1}^{d} \phi_j^T (\mathbf{x} - \mu)(\mathbf{x} - \mu)^T \phi_j \right\}$

$$= \sum_{j=m+1}^{d} \phi_j^T \mathcal{E}[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T] \phi_j = \sum_{j=m+1}^{d} \phi_j^T \Sigma \phi_j = \sum_{j=m+1}^{d} \lambda_j$$

X

μ

University of Chinese Academy of Sciences

线性变换的高斯分布

- 线性变换 $\mathbf{y} = \mathbf{A}^t \mathbf{x}$
  - $A^t A = 1$: 正交变换(坐标轴旋转)
  - 变换后的分布仍为正态分布

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

  - Diagonalization

    $$A = \Phi$$

    $$A^t \Sigma A = \Lambda$$

  - Whitening transform

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

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

    $$= \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = I$$
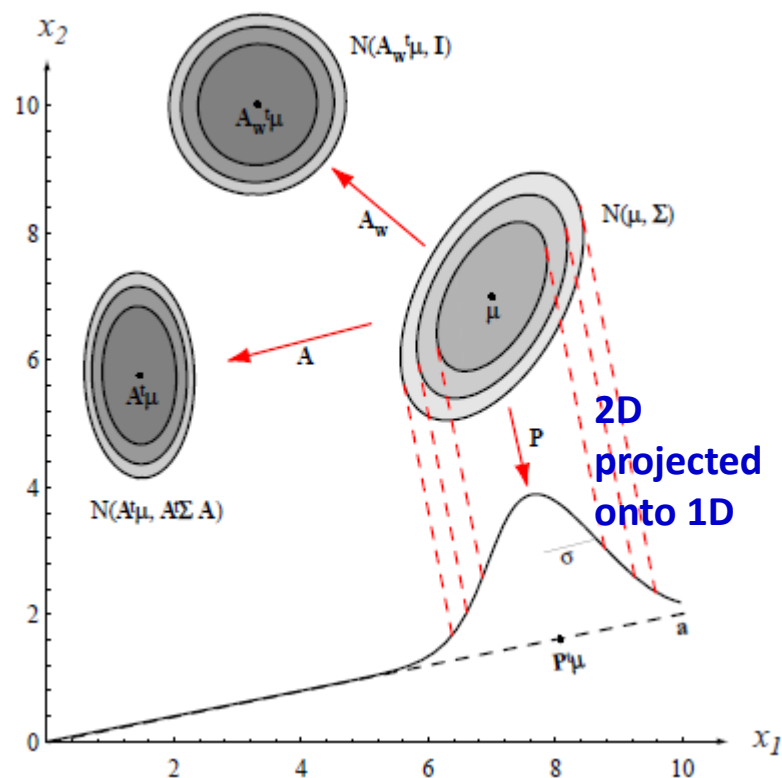
相关知识：特征提取/降维

- Dimensionality reduction
- Feature extraction
  - Feature generation: original data $\mathbf{d} \rightarrow \mathbf{x}$
  - Linear feature extraction $\mathbf{x} = \mathbf{A}^T \mathbf{d}$
- Feature selection (for reduction and performance)
  - Feature subset selection: a learning/optimization problem
- Feature transform (for extraction or reduction)
  - Linear transform $\mathbf{y} = \mathbf{A}^T \mathbf{x}$
  - Nonlinear transform: may increase dimensionality, e.g. kernel PCA, kernel LDA
- Handcrafted feature
- Feature learning
  - Automatic feature generation, e.g. convolutional neural network (CNN)

University of Chinese Academy of Sciences

高斯密度下的判别函数

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

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

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

  – Quadratic discriminant function (QDF)

  – 在不同covariance假设条件下得到一些特殊形式

- Case 1: $\Sigma_i = \sigma^2 I$
  (去掉与类别无关项)

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

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

  - 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

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

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

University of Chinese Academy of Sciences

- Case 1: $\Sigma_i = \sigma^2 I$ (continued)
  - 二类决策面(判别函数相等的点构成)

$$g_i(\mathbf{x}) = g_j(\mathbf{x})$$

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

W: 法向量
平面与μ_1-μ_2垂直

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

位置移向先验
概率小的类别

  - 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

University of Chinese Academy of Sciences
中国科学院大学
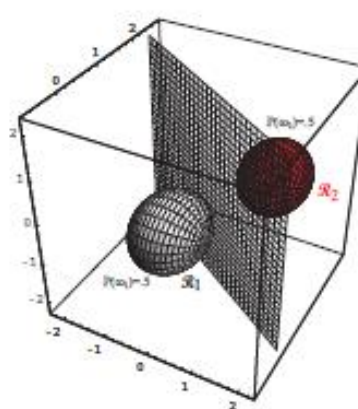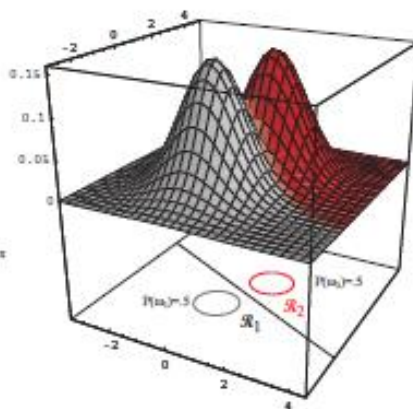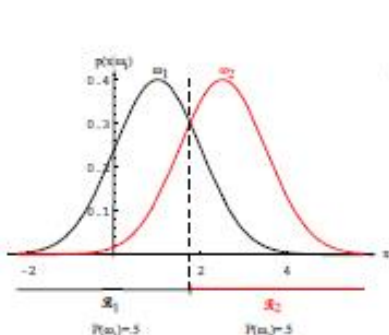
– 当先验概率变化，决策面发生平移

移向概率
小的类别

$P(\omega_1)$=0.8

$P(\omega_1)$=0.99

University of Chinese Academy of Sciences
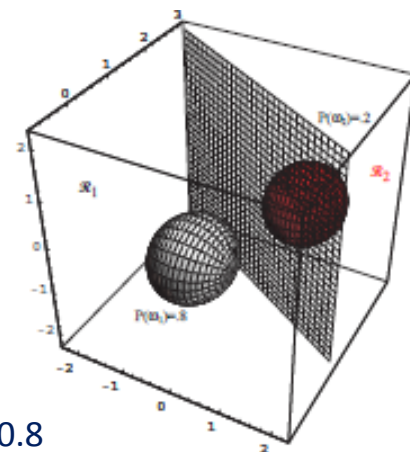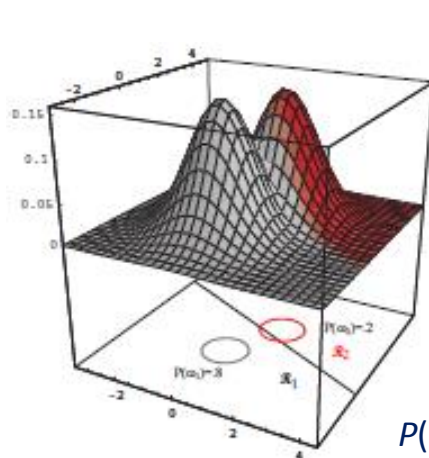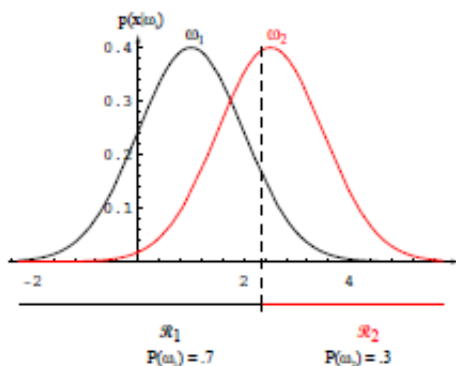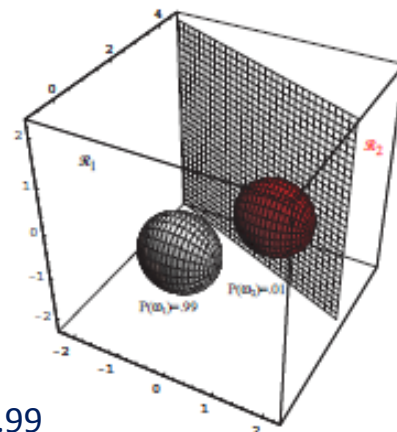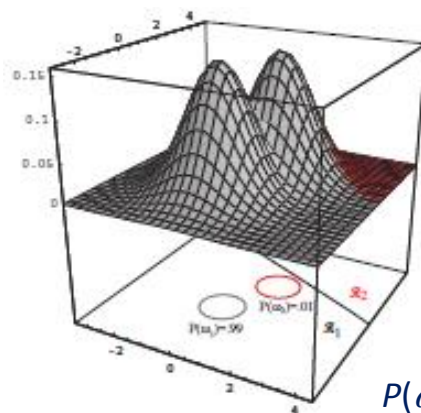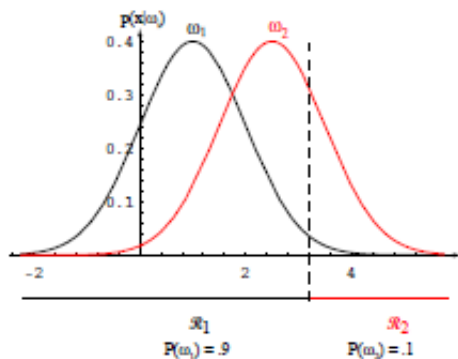
- Case 2: $\Sigma_i = \Sigma$

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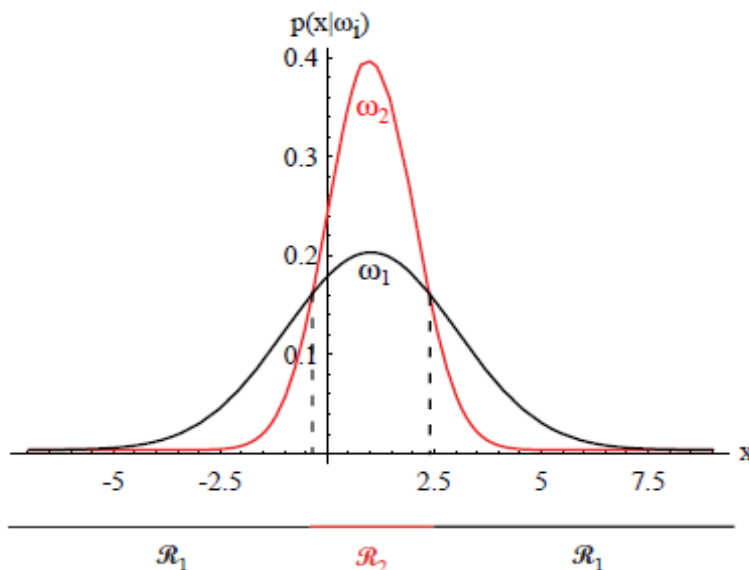
$$\Longrightarrow \quad g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) + \ln P(\omega_i)$$

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

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

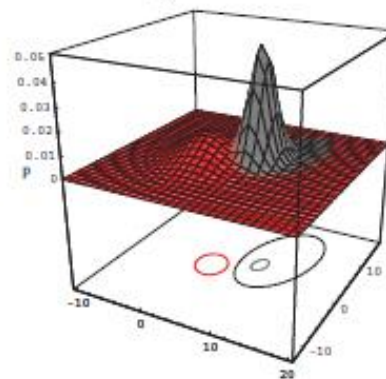$$\mathbf{w}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i \qquad w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i)$$
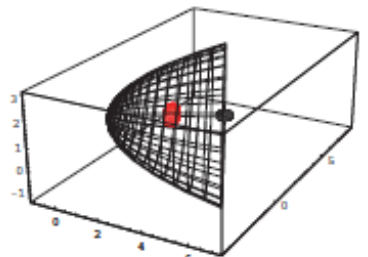
  – 二类决策面 $g_i(\mathbf{x}) = g_j(\mathbf{x})$

$$\Longrightarrow \quad \mathbf{w}^t(\mathbf{x} - \mathbf{x}_0) = 0 \qquad \mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

- 注意跟μ_1-μ_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$(\mu_1 + \mu_2)/2$
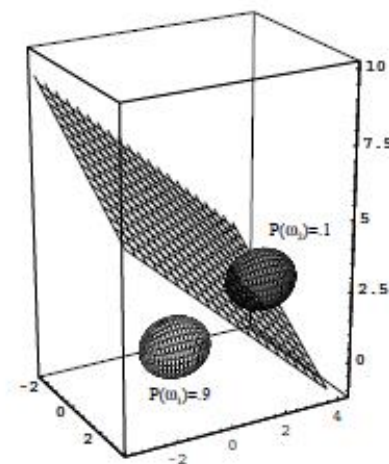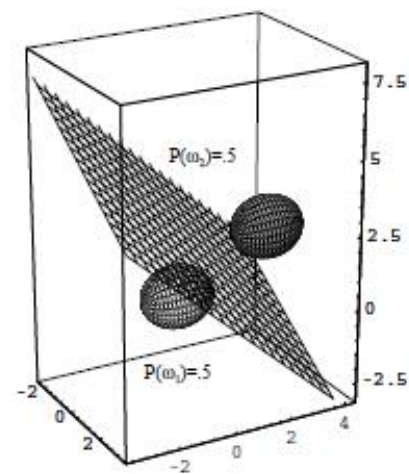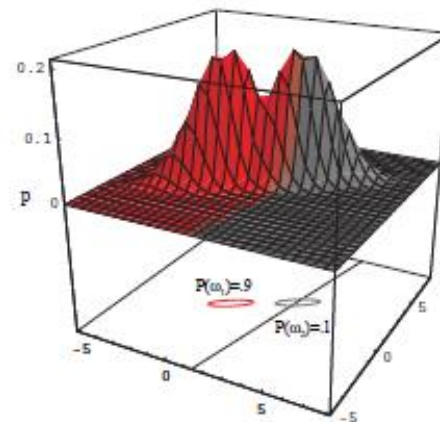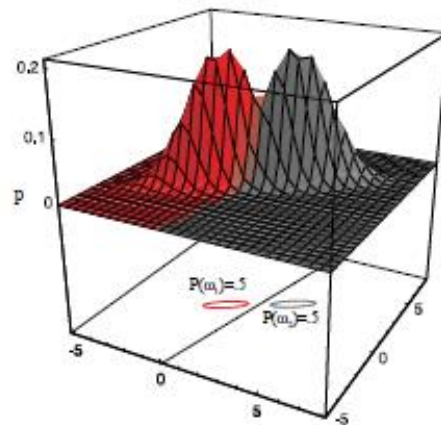
$$P(\omega_1) = P(\omega_2) \qquad\qquad P(\omega_1) \neq P(\omega_2)$$

- Case 3: Σ_i= arbitrary

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

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

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

$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t \Sigma_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln |\Sigma_i| + \ln P(\omega_i)$$
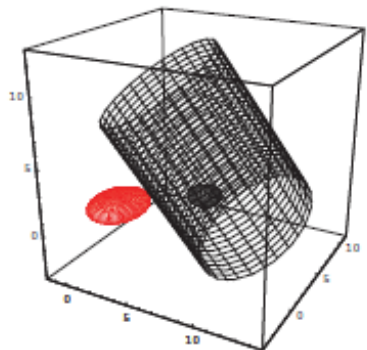
- 二类决策面：$g_1(\mathbf{x})=g_2(\mathbf{x})$, hyperquadratics
  - 等均值的情况下，1D的例子

2D的例子
（z轴是概率密度）
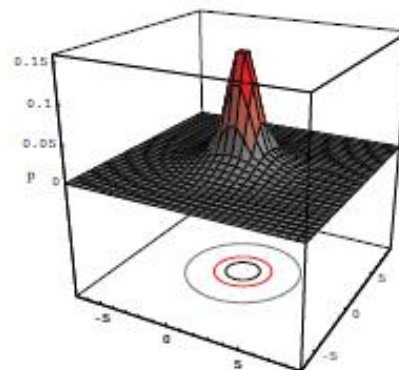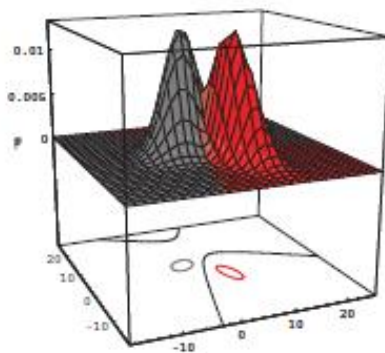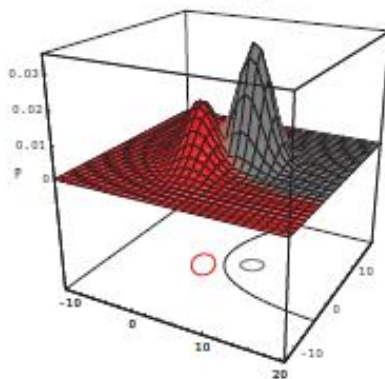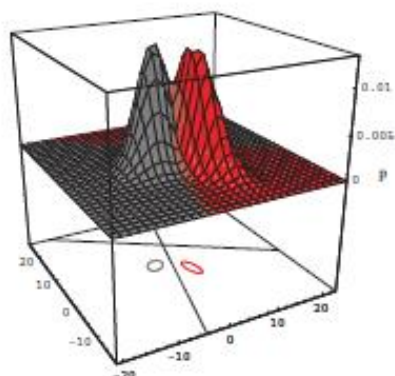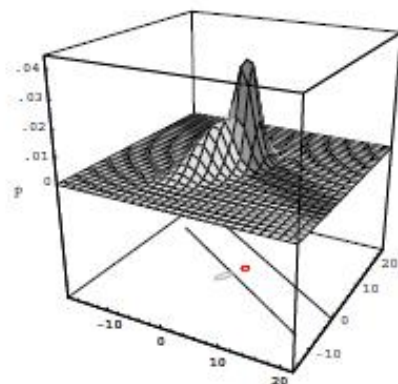
3D的例子
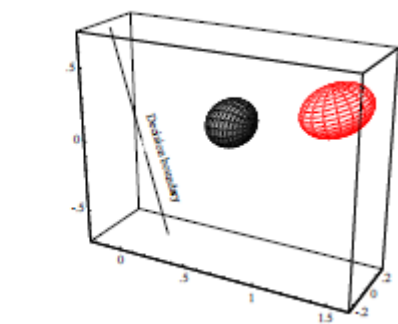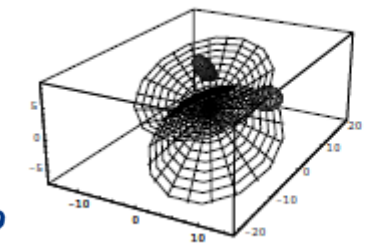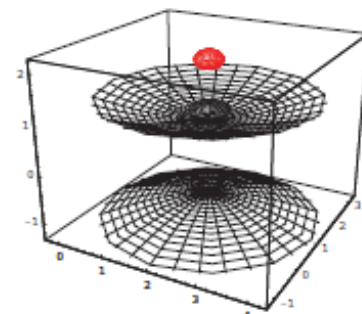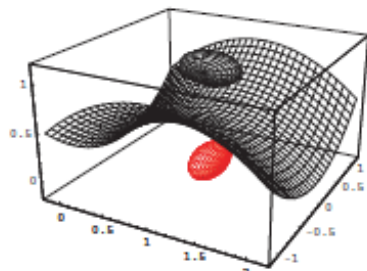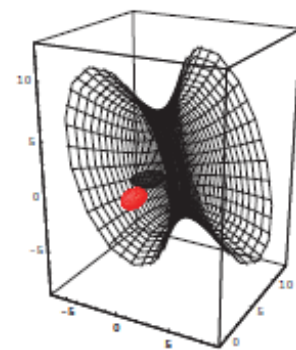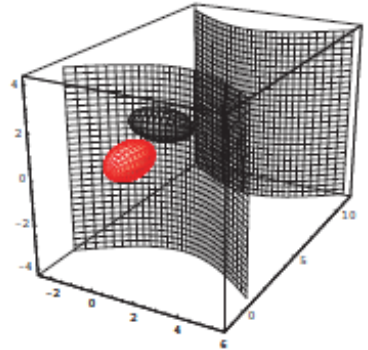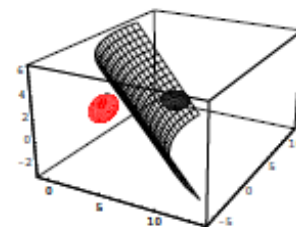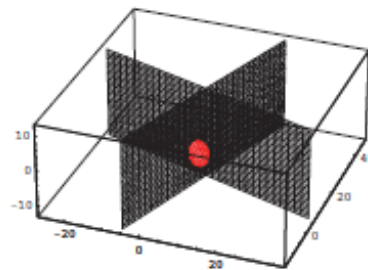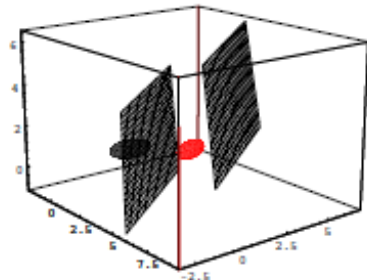
2D，4类的例子
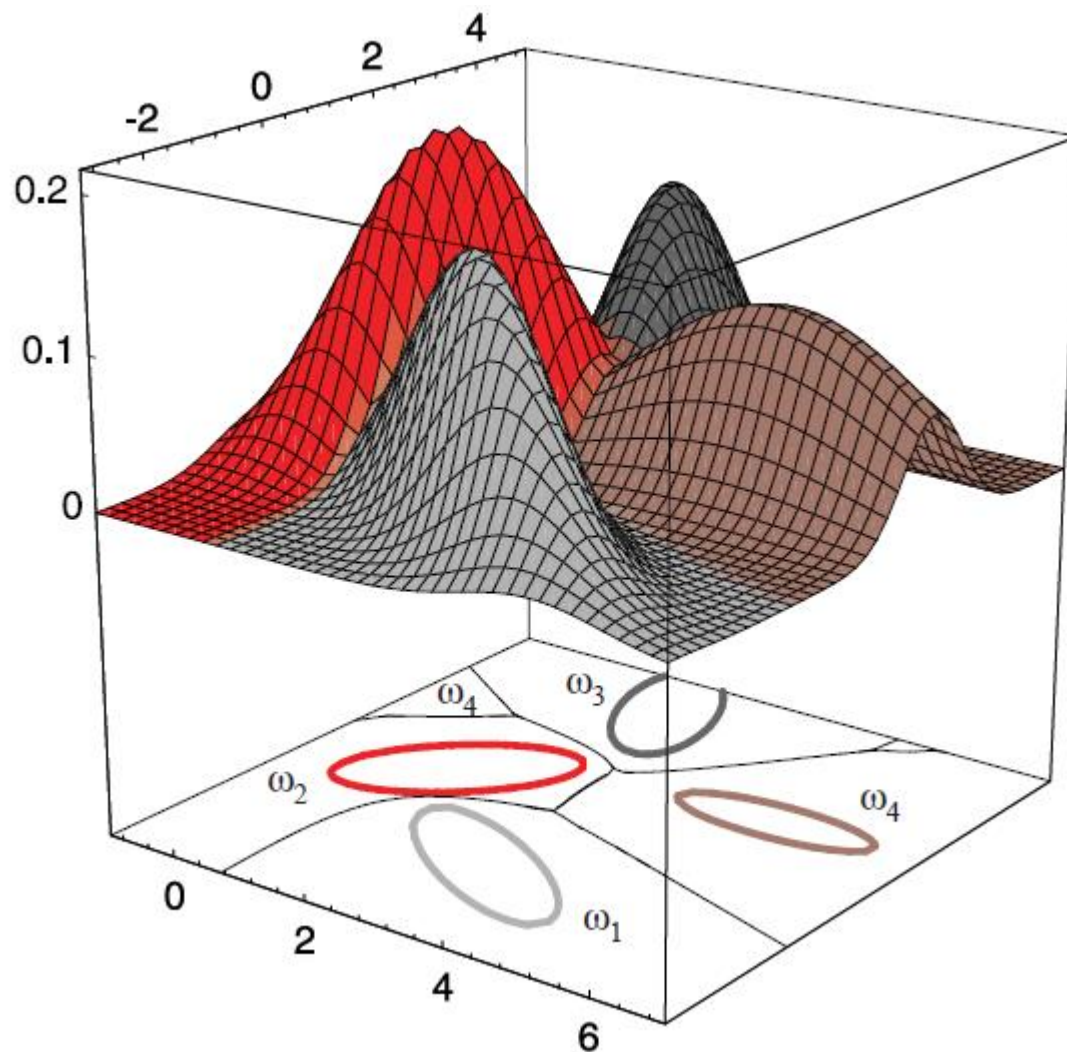
- 一个具体例子
  - 2类，2D $P(\omega_1) = P(\omega_2) = 0.5$

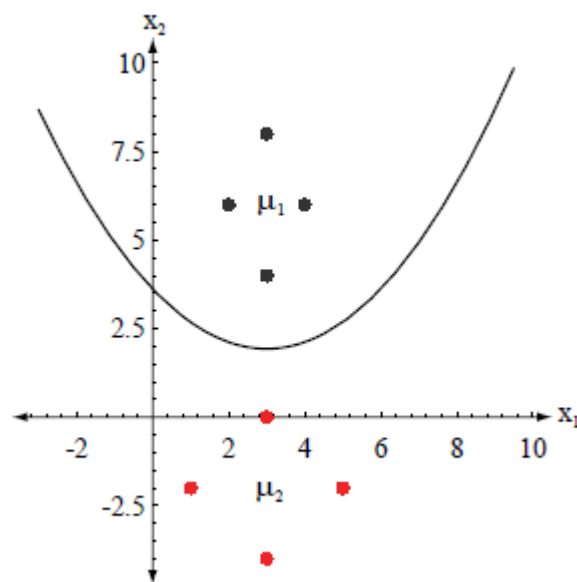$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \qquad \Sigma_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

$$\mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \qquad \Sigma_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

- 决策面 $g_1(\mathbf{x})=g_2(\mathbf{x})$
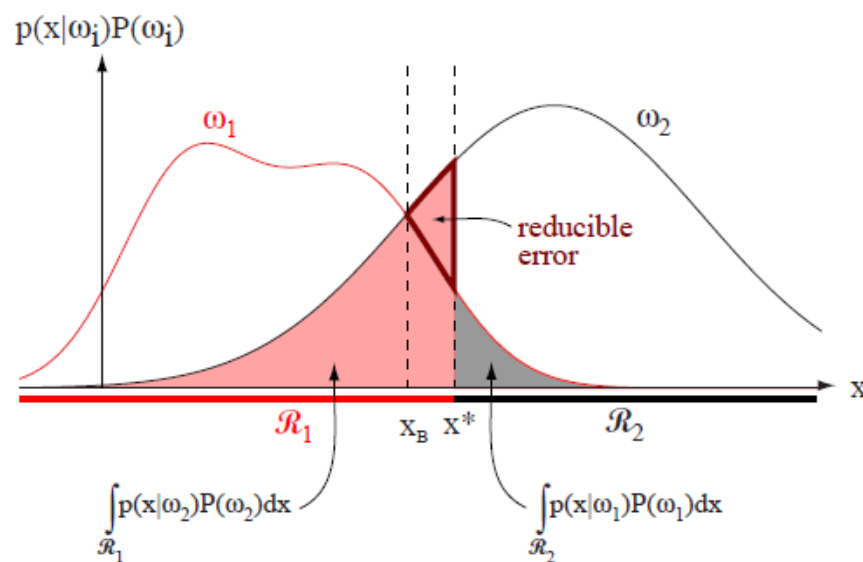
$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2$$

分类错误率

- ## 2类的情况

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

- ## 一般情况

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

决策面为x_B时为最小错误率分类

University of Chinese Academy of Sciences
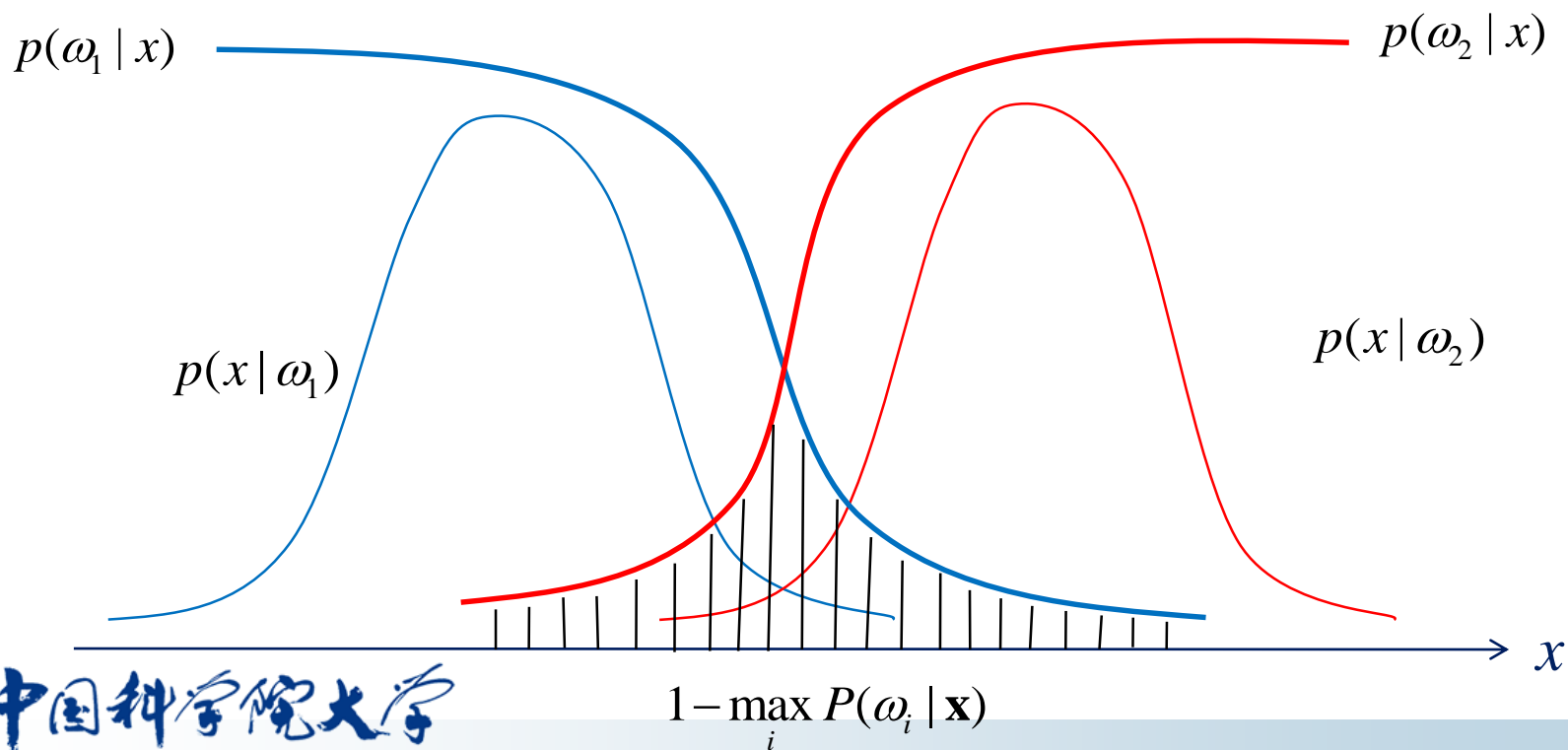
- 最大后验概率决策(0-1 loss)的情况

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

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

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

$p(\omega_1 \mid x)$

$p(\omega_2 \mid x)$

$p(x \mid \omega_1)$

$p(x \mid \omega_2)$

x

$1 - \max_i P(\omega_i \mid \mathbf{x})$

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：在不同的特征空间才有可能！

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

University of Chinese Academy of Sciences