模式识别

中国科学技术大学 汪增福

- 第一章 绪论
- 第二章 统计模式识别中的几何方法
- ✓ 第三章 统计模式识别中的概率方法
- 第四章 分类器的错误率
- 第五章 统计模式识别中的聚类方法
- 第六章 结构模式识别中的句法方法
- 第七章 总结

第三章 统计模式识别中的概率方法

- 本章主要内容

  主要讨论如何根据被观测模式的统计特性建立适当的判决规则使分类器的错误率最小。

  - 问题概述
  - 最小错误概率判决准则
  - 最小风险判决规则
  - **Neyman-Pearson**判决规则
  - 最小最大判决规则
  - 类条件概率密度的参数估计
  - 类条件概率密度的非参数估计

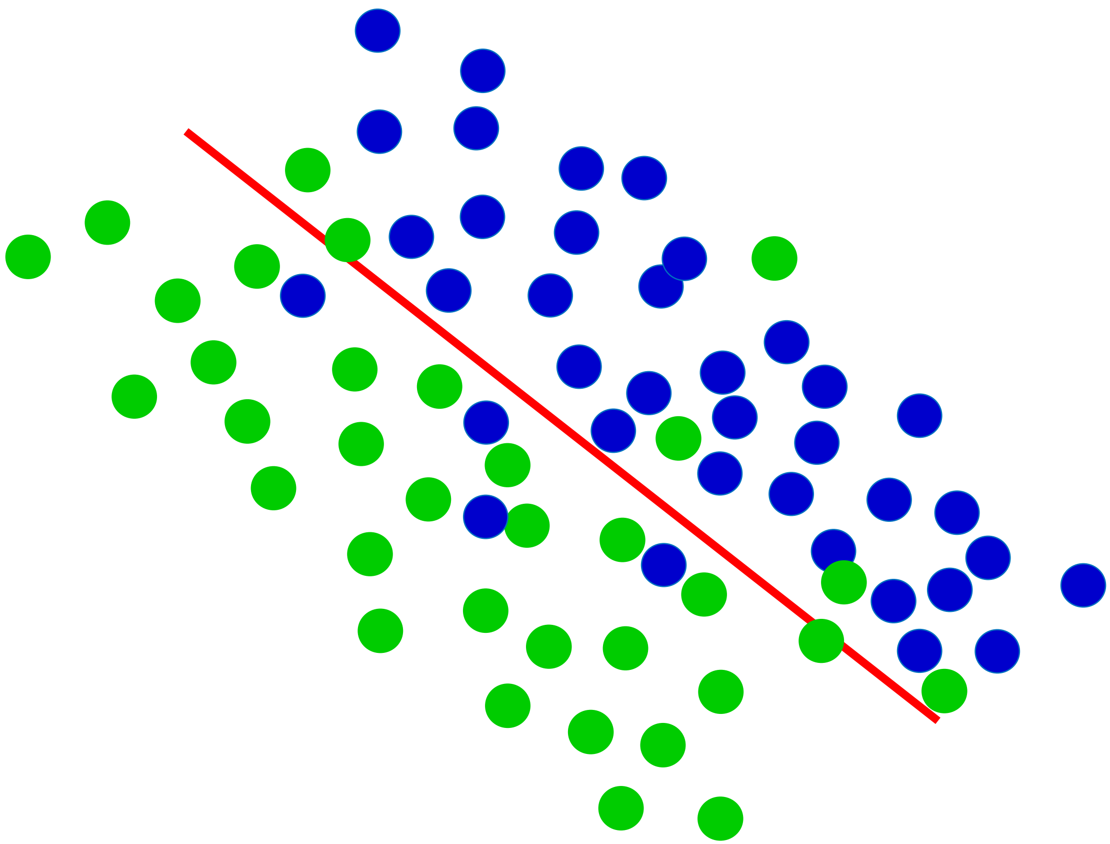
§ 3.1 问题概述

混叠现象

- 特征选择不合适
- 不确定性

目标：实现没有错误的分类

➡ 使使分类错误率最小



§ 3.1 问题概述

相关概念

§ 3.1 问题概述

相关概念

先验概率　　事件发生之前就已知道的概率称之为先验概率。

姐姐照片的全体　ω_1　$P(\omega_1)$

妹妹照片的全体　ω_2　$P(\omega_2)$

$$P(\omega_1)+P(\omega_2)=1$$

一般地，如果有N个互斥的类别，则有：

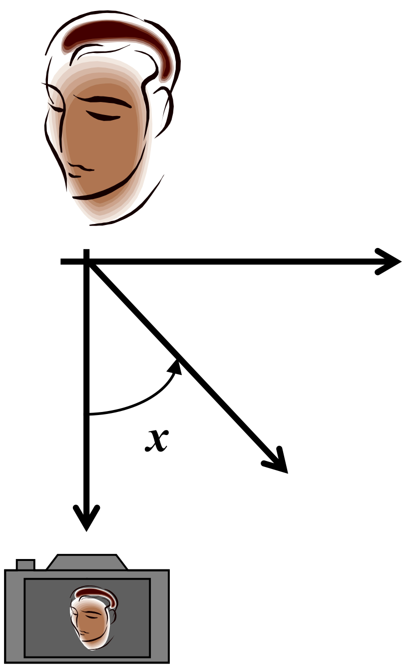
$$P(\omega_1)+P(\omega_2)+\Lambda\ ++P(\omega_N)=1$$

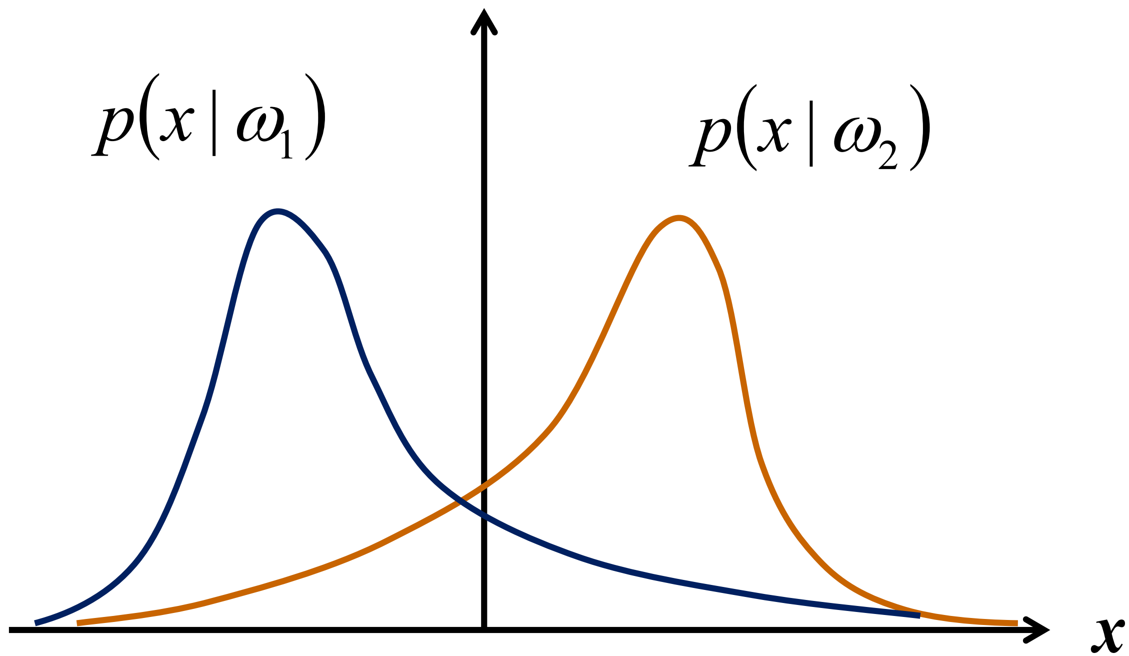
§ 3.1 问题概述

相关概念

类条件概率密度 $p(X\,|\,\omega)$ 似然函数

在属于类别ω的条件下，观测样本作为X出现的概率密度函数。





§ 3.1 问题概述

相关概念

后验概率 $P(\omega \mid \boldsymbol{X})$

在观测样本$\boldsymbol{X}$被观测的情况下，$\boldsymbol{X}$属于类别ω的概率。

若干约定

- 用大写英文字母表示概率
  - 先验概率 $P(\omega)$
  - 后验概率 $P(\omega \mid \boldsymbol{X})$
- 用小写英文字母表示概率密度
  - 全概率密度 $p(\boldsymbol{X})$
  - 类条件概率密度 $p(\boldsymbol{X} \mid \omega)$

有何关系？

§ 3.1 问题概述

相关概念

全概率公式

$$p(\boldsymbol{X})=\sum_{j=1}^{C} p(\boldsymbol{X}\mid\omega_j)P(\omega_j)$$

函数形式及参数已知或可以估计

贝叶斯公式

$$P(\omega_j\mid\boldsymbol{X})=\frac{p(\boldsymbol{X}\mid\omega_j)P(\omega_j)}{p(\boldsymbol{X})}$$

已知或未知

$$=\frac{p(\boldsymbol{X}\mid\omega_j)P(\omega_j)}{\sum_{j=1}^{C} p(\boldsymbol{X}\mid\omega_j)P(\omega_j)}$$

已知

分类判决的依据

§ 3.1 问题概述

相关概念

由以往的经验得到

先验概率：事件尚未发生，要求事件发生的可能性大小。

由因求果　　执果寻因

后验概率：事件已经发生，追寻事件发生的原因。

评估由某个因素引发事件的可能性的大小。

考虑了一个事实之后的条件概率

后验概 率的计算以先验概率为基础

§ 3.2 最小错误概率判决准则

已知条件

类别	先验概率	类条件概率密度
ω_1	$P(\omega_1)$	$p(\boldsymbol{X} \mid \omega_1)$
ω_2	$P(\omega_2)$	$p(\boldsymbol{X} \mid \omega_2)$

问题：设计一个具有最小误分概率的两类分类器

求解： 设观测样本为 X， 则X的全概率密度为

$$p(\boldsymbol{X}) = p(\boldsymbol{X} \mid \omega_1)P(\omega_1) + p(\boldsymbol{X} \mid \omega_2)P(\omega_2)$$

X属于两个类别的后验概率为

$$\begin{cases} P(\omega_1 \mid \boldsymbol{X}) = \dfrac{p(\boldsymbol{X} \mid \omega_1)P(\omega_1)}{p(\boldsymbol{X})} \\ P(\omega_2 \mid \boldsymbol{X}) = \dfrac{p(\boldsymbol{X} \mid \omega_2)P(\omega_2)}{p(\boldsymbol{X})} \end{cases} \Rightarrow \text{最大后验概率判决}$$

§ 3.2 最小错误概率判决准则

求解（续）：

最大后验概率判决规则

$$\begin{cases} P(\omega_1 \mid \boldsymbol{X}) > P(\omega_2 \mid \boldsymbol{X}) \quad \Rightarrow \quad \boldsymbol{X} \in \omega_1 \\ P(\omega_2 \mid \boldsymbol{X}) > P(\omega_1 \mid \boldsymbol{X}) \quad \Rightarrow \quad \boldsymbol{X} \in \omega_2 \end{cases}$$

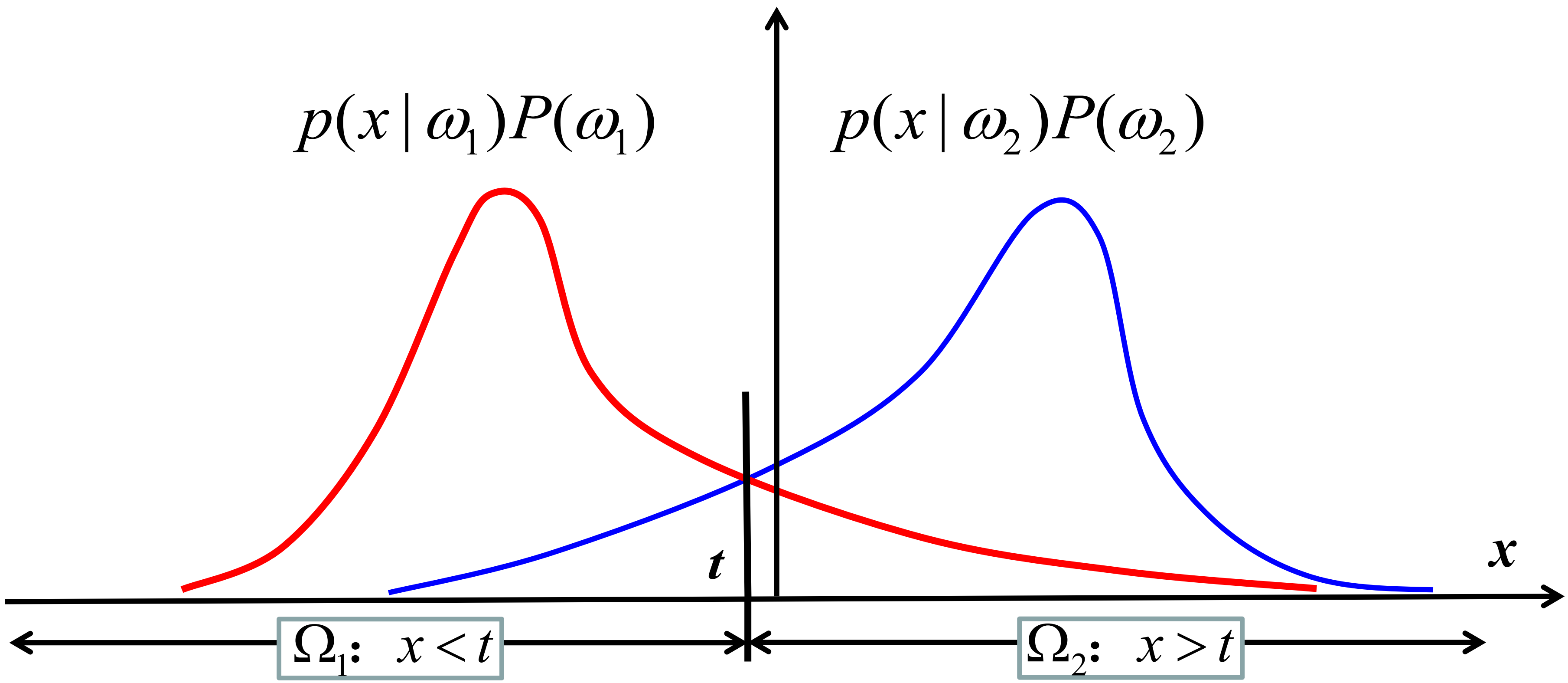
$$\begin{cases} p(\boldsymbol{X} \mid \omega_1)P(\omega_1) > p(\boldsymbol{X} \mid \omega_2)P(\omega_2) \quad \Rightarrow \quad \boldsymbol{X} \in \omega_1 \\ p(\boldsymbol{X} \mid \omega_2)P(\omega_2) > p(\boldsymbol{X} \mid \omega_1)P(\omega_1) \quad \Rightarrow \quad \boldsymbol{X} \in \omega_2 \end{cases}$$

可以证明，最大后验概率判决规则给出的判决结果的错误概率是最小的。

§ 3.2 最小错误概率判决准则

一维两类问题示例

$$
\begin{cases}
p(X|\omega_1)P(\omega_1) > p(X|\omega_2)P(\omega_2) \Rightarrow X \in \omega_1 \\
p(X|\omega_2)P(\omega_2) > p(X|\omega_1)P(\omega_1) \Rightarrow X \in \omega_2
\end{cases}
$$

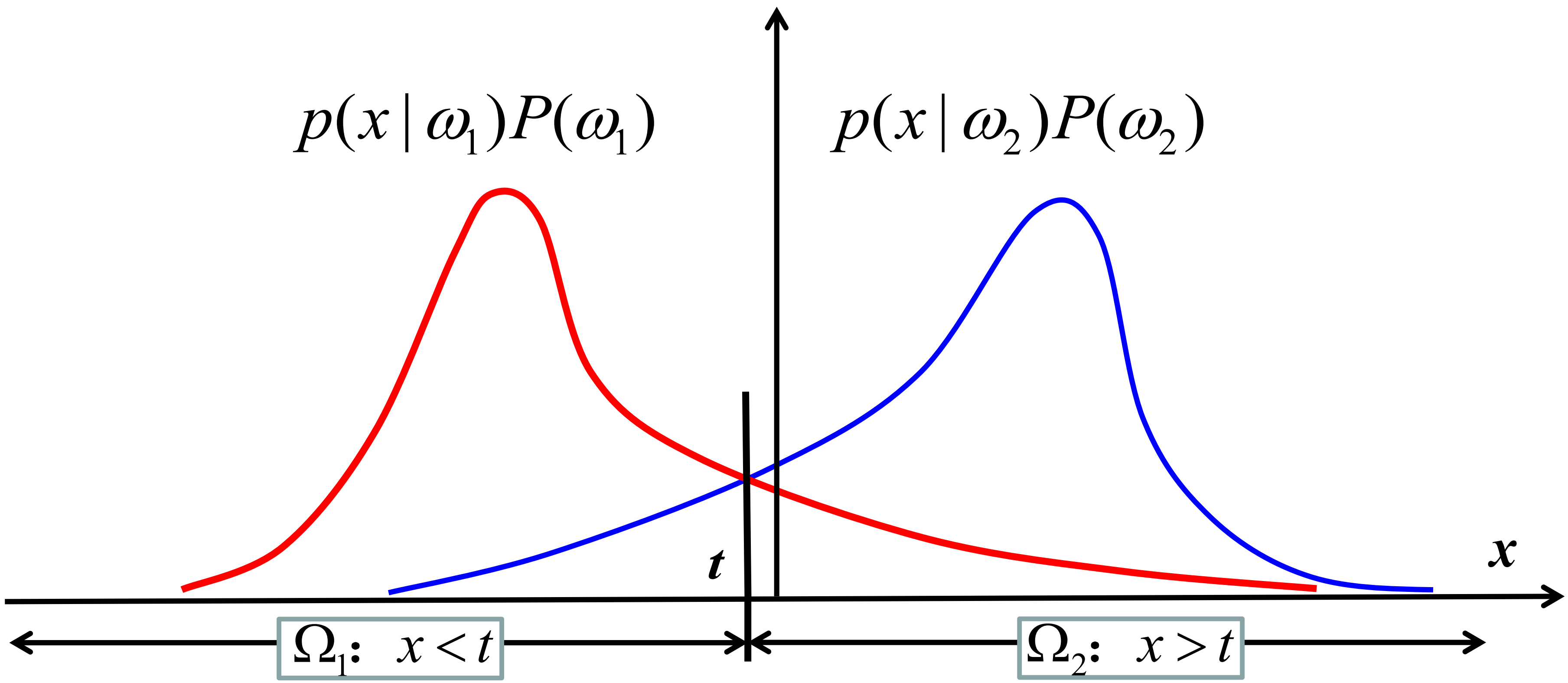


§ 3.2 最小错误概率判决准则

一维两类问题示例

误判情况　1. ω_1类别的样本落入Ω_2中

　　　　　2. ω_2类别的样本落入Ω_1中



§ 3.2 最小错误概率判决准则

一维两类问题的分类错误率

误判情况　1. ω_1类别的样本落入Ω_2中

　　　　　2. ω_2类别的样本落入Ω_1中

总的分类错误概率

$$P(e) = P(x\text{在}\Omega_2\text{中} \mid \omega_1) + P(x\text{在}\Omega_1\text{中} \mid \omega_2)$$

$$= \int_{\Omega_2} \mathrm{P}(\omega_1)\mathrm{p}(\mathrm{x} \mid \omega_1)\mathrm{dx} + \int_{\Omega_1} \mathrm{P}(\omega_2)\mathrm{p}(\mathrm{x} \mid \omega_2)\mathrm{dx}$$

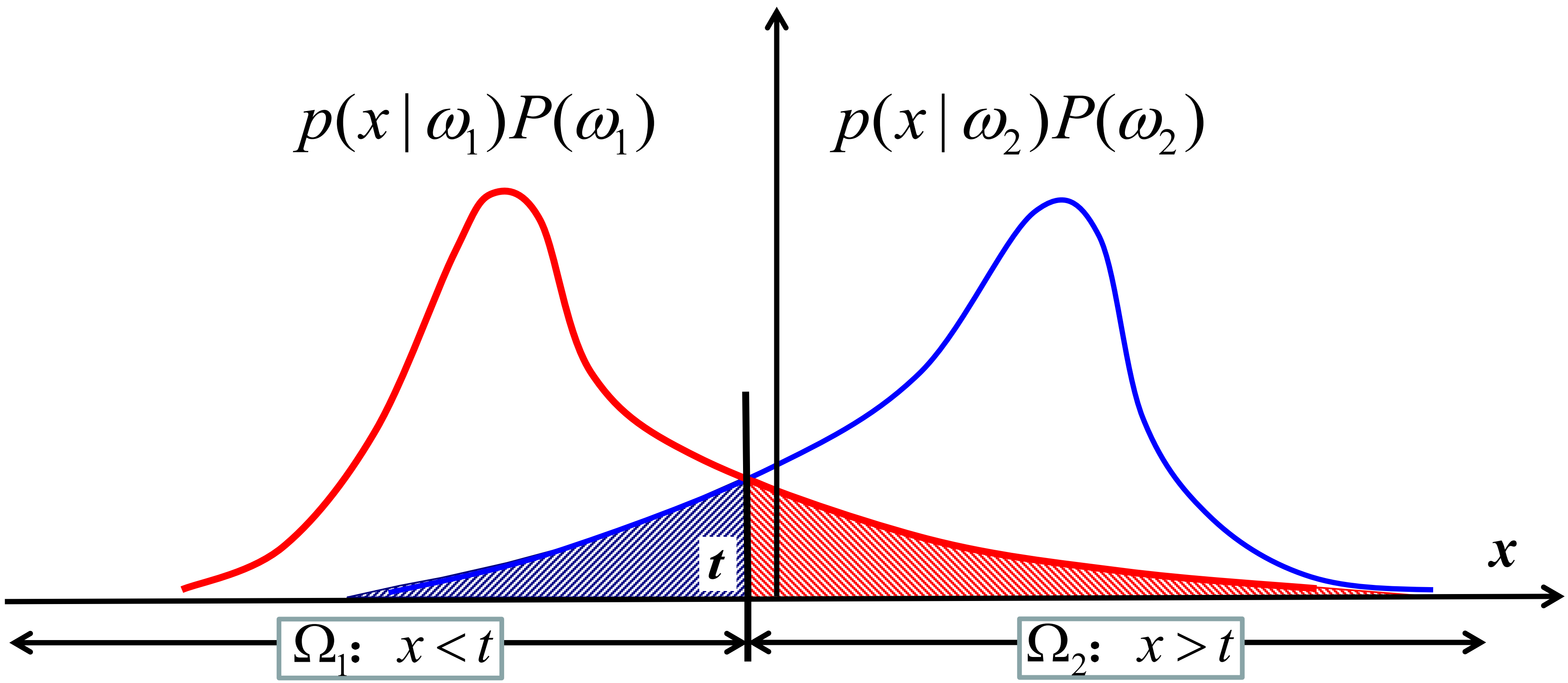
$$= \mathrm{P}(\omega_1)\boxed{\int_{\Omega_2} \mathrm{p}(\mathrm{x} \mid \omega_1)\mathrm{dx}} + \mathrm{P}(\omega_2)\boxed{\int_{\Omega_1} \mathrm{p}(\mathrm{x} \mid \omega_2)\mathrm{dx}}$$

$$= \mathrm{P}(\omega_1)\mathrm{P}_1(\mathrm{e}) + \mathrm{P}(\omega_2)\mathrm{P}_2(\mathrm{e})$$

§ 3.2 最小错误概率判决准则

一维两类问题示例：分类错误率

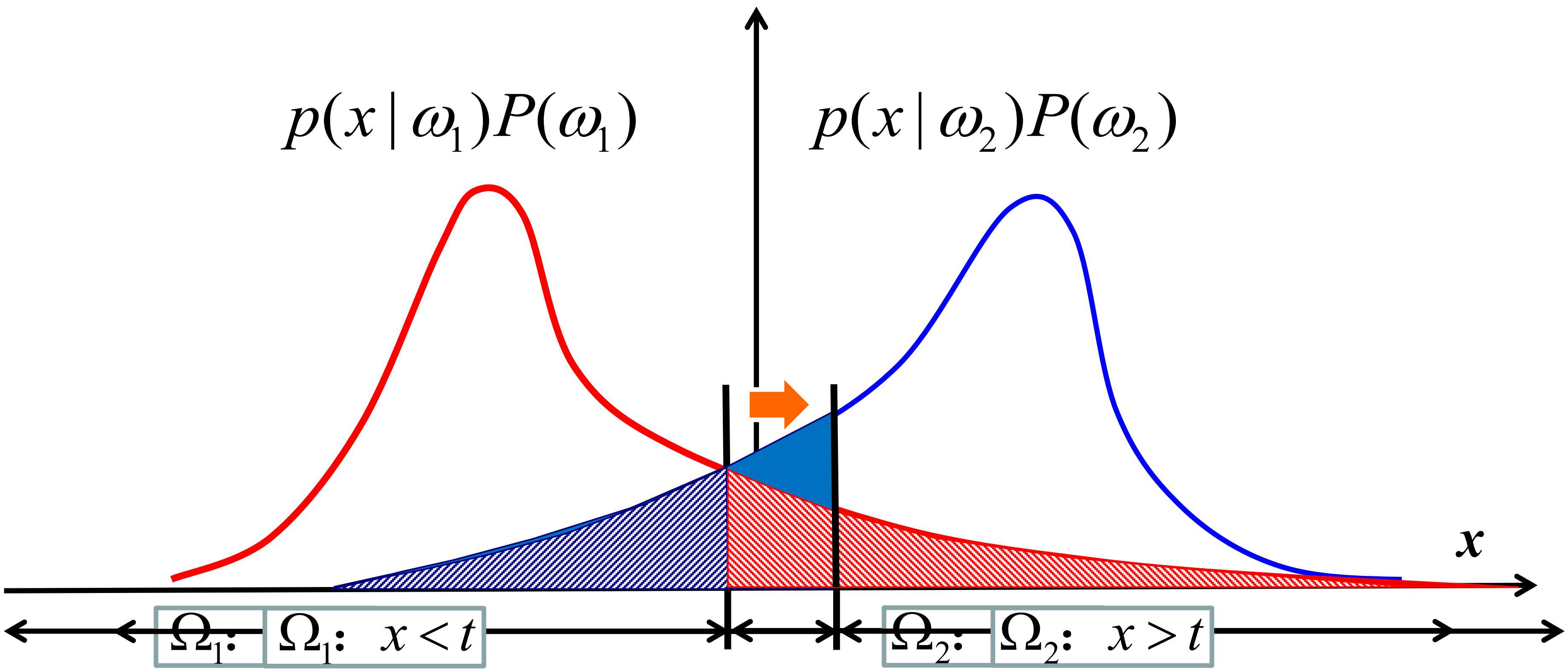
$$P(e)=\int_{\Omega_2}\mathrm{P}(\omega_1)\mathrm{p}(\mathrm{x}\,|\,\omega_1)\mathrm{dx}+\int_{\Omega_1}\mathrm{P}(\omega_2)\mathrm{p}(\mathrm{x}\,|\,\omega_2)\mathrm{dx}$$



§ 3.2 最小错误概率判决准则

一维两类问题示例：分类错误率

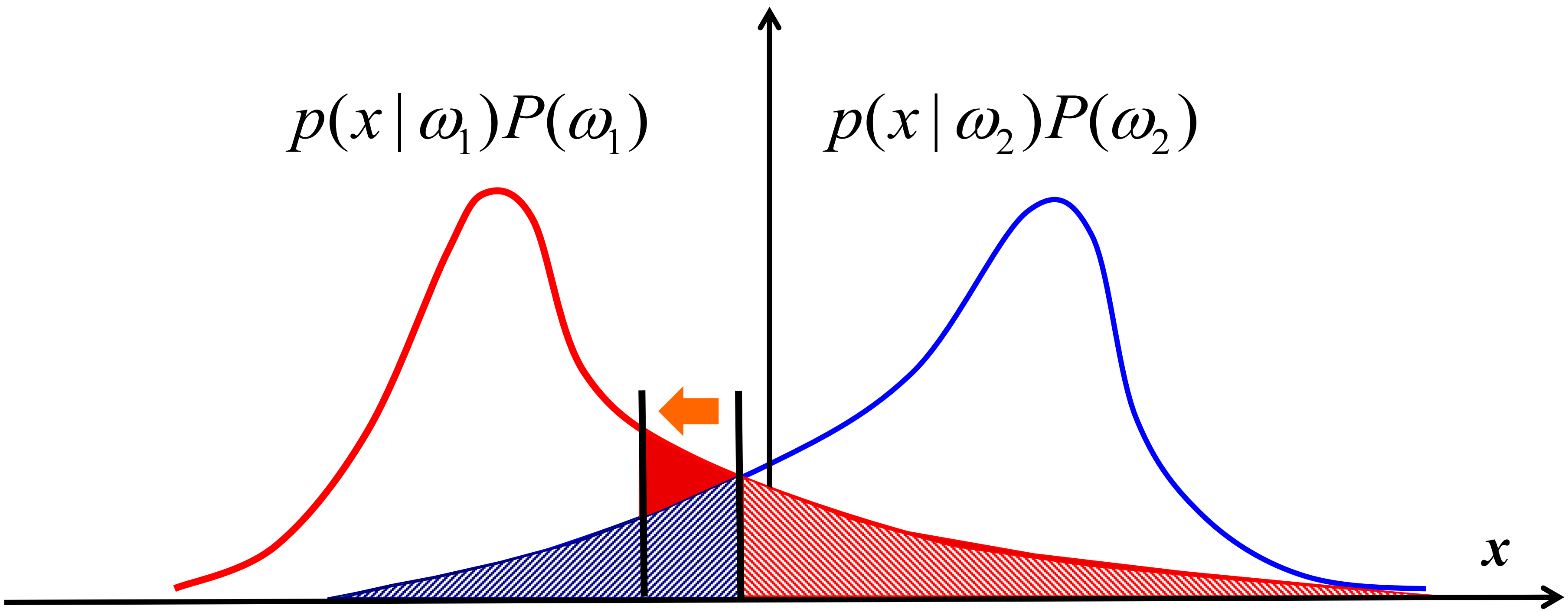
$$P(e) = \int_{\Omega_2} \mathrm{P}(\omega_1)\mathrm{p}(\mathrm{x} \mid \omega_1)\mathrm{dx} + \int_{\Omega_1} \mathrm{P}(\omega_2)\mathrm{p}(\mathrm{x} \mid \omega_2)\mathrm{dx}$$



§ 3.2 最小错误概率判决准则

一维两类问题示例：分类错误率

$$P(e) = \int_{\Omega_2} \mathrm{P}(\omega_1)\mathrm{p}(\mathrm{x} \mid \omega_1)\mathrm{dx} + \int_{\Omega_1} \mathrm{P}(\omega_2)\mathrm{p}(\mathrm{x} \mid \omega_2)\mathrm{dx}$$



§ 3.2 最小错误概率判决准则

多类情况下的最大后验概率判决规则

$$P(\omega_i \mid \boldsymbol{X}) > \underset{j=1,2,...N}{Maximum}\{P(\omega_j \mid \boldsymbol{X})\} \Rightarrow \boldsymbol{X} \in \omega_i$$

$$p(x \mid \omega_i)P(\omega_i) > \underset{j=1,2,...N}{Maximum}\{p(x \mid \omega_j)P(\omega_j)\} \Rightarrow \boldsymbol{X} \in \omega_i$$

小结：

- 优势：最大后验概率判决的错误概率最小
- 缺点：没有考虑判决的风险

判决错误概率最小化 ➡ 判决风险最小化

§ 3.3 最小风险判决规则

- 从导弹或飞机来袭的实例谈起

漏警

实际情况	判决结果
敌方飞机或导弹来袭	敌方飞机或导弹来袭
敌方飞机或导弹来袭	敌方飞机或导弹未袭
敌方飞机或导弹未袭	敌方飞机或导弹未袭
敌方飞机或导弹未袭	敌方飞机或导弹来袭

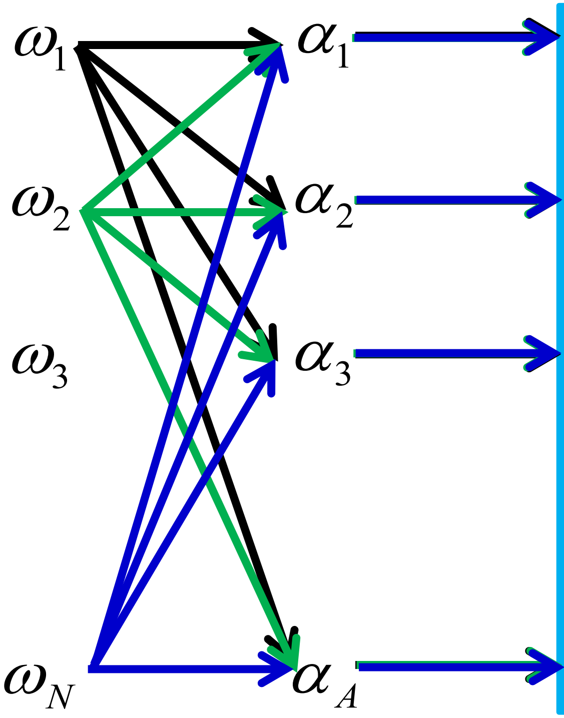
不同判决导致的损失程度一般也不同！

虚警

§ 3.3 最小风险判决规则

N类分类问题的风险评估

类别　　判决　　风险



$$\begin{bmatrix} L(\alpha_1 \mid \omega_1) & L(\alpha_1 \mid \omega_2) & \cdots & L(\alpha_1 \mid \omega_N) \\ L(\alpha_2 \mid \omega_1) & L(\alpha_2 \mid \omega_2) & \cdots & L(\alpha_2 \mid \omega_N) \\ L(\alpha_3 \mid \omega_1) & L(\alpha_3 \mid \omega_2) & \cdots & L(\alpha_3 \mid \omega_N) \\ L(\alpha_A \mid \omega_1) & L(\alpha_A \mid \omega_2) & \cdots & L(\alpha_A \mid \omega_N) \end{bmatrix}$$

一般而言，$A \geq N$

$\mathscr{L} \longrightarrow$ 风险矩阵

§ 3.3 最小风险判决规则

N类分类问题的风险评估

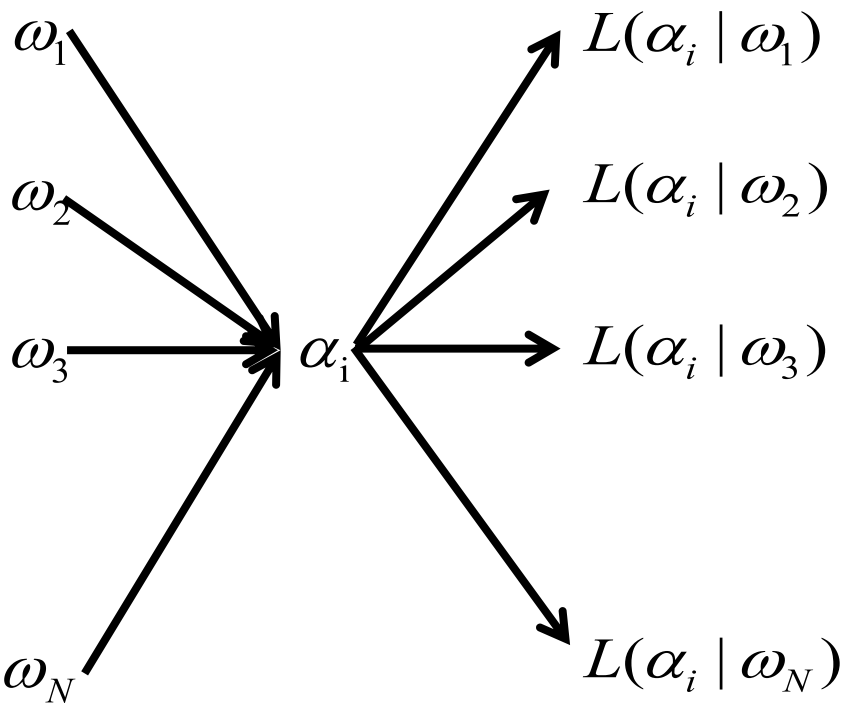
决策表构建

决策 \ 损失 \ 状态	ω_1	ω_2	$\cdots$	ω_j	$\cdots$	ω_N
α_1	$L(\alpha_1,\omega_1)$	$L(\alpha_1,\omega_2)$	$\cdots$	$L(\alpha_1,\omega_j)$	$\cdots$	$L(\alpha_1,\omega_N)$
α_2	$L(\alpha_2,\omega_1)$	$L(\alpha_2,\omega_2)$	$\cdots$	$L(\alpha_2,\omega_j)$	$\cdots$	$L(\alpha_2,\omega_N)$
			$\cdots$		$\cdots$	
α_{i}	$L(\alpha_i,\omega_1)$	$L(\alpha_i,\omega_2)$	$\cdots$	$L(\alpha_i,\omega_j)$	$\cdots$	$L(\alpha_i,\omega_N)$
			$\cdots$		$\cdots$	
α_{A}	$L(\alpha_A,\omega_1)$	$L(\alpha_A,\omega_2)$	$\cdots$	$L(\alpha_A,\omega_j)$	$\cdots$	$L(\alpha_A,\omega_N)$

§ 3.3 最小风险判决规则

条件平均风险

类别　　判决　　风险



观测样本为 X

$$P(\omega_j \mid \boldsymbol{X}),\ j = 1,2,\ldots N$$

判决 α_i 的条件平均风险

$$R(\alpha_i \mid \boldsymbol{X}) = \sum_{j=1}^{N} L(\alpha_i \mid \omega_j) P(\omega_j \mid \boldsymbol{X})$$

$$\sum_{j=1}^{N} P(\omega_j \mid \boldsymbol{X}) = 1$$

$$P(\omega_j \mid \boldsymbol{X})$$

$$= \frac{p(\boldsymbol{X} \mid \omega_j) P(\omega_j)}{p(\boldsymbol{X})}$$

$$= \frac{p(\boldsymbol{X} \mid \omega_j) P(\omega_j)}{\sum_{j=1}^{N} p(\boldsymbol{X} \mid \omega_j) P(\omega_j)}$$

§ 3.3 最小风险判决规则

最小风险判决规则

$$R(\alpha_i \mid \boldsymbol{X}) = \underset{j=1,2,\ldots A}{Minimum}\{R(\alpha_j \mid \boldsymbol{X})\} \Rightarrow \quad \boldsymbol{X} \in \omega_i$$

算法步骤：

Step1. 对观测样本X，计算

$$P(\omega_j \mid \boldsymbol{X}) = \frac{p(\boldsymbol{X} \mid \omega_j)P(\omega_j)}{\sum_{j=1}^{N} p(\boldsymbol{X} \mid \omega_j)P(\omega_j)}, j = 1,2,\ldots N$$

Step2. 计算各判决的条件平均风险

$$R(\alpha_j \mid \boldsymbol{X}) = \sum_{k=1}^{N} L(\alpha_j \mid \omega_k)P(\omega_k \mid \boldsymbol{X}), j = 1,2,\ldots N$$

Step3. 将观测样本X判属于使条件平均风险最小化的判决所对应的类别。 。

中国科学技术大学
University of Science and Technology of China

§ 3.3 最小风险判决规则

最小错误概率判决 VS 最小风险判决

作如下考察：

正确判决，损失为0；
错误判决，损失为1。 } 0-1损失函数

定义

$$L(\alpha_i \mid \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

$$\Rightarrow \quad L(\alpha_i \mid \omega_j) = 1 - \delta_{ij}$$

$$R(\alpha_i \mid \boldsymbol{X}) = \sum_{j=1}^{N} L(\alpha_i \mid \omega_j) P(\omega_j \mid \boldsymbol{X})$$

$$= \sum_{j=1}^{N} (1 - \delta_{ij}) P(\omega_j \mid \boldsymbol{X})$$

$$= 1 - P(\omega_i \mid \boldsymbol{X})$$

代入最小风险判决中： $R(\alpha_i \mid \boldsymbol{X}) = \underset{j=1,2,\ldots A}{Minimum} \{R(\alpha_j \mid \boldsymbol{X})\} \Rightarrow \boldsymbol{X} \in \omega_i$

$$P(\omega_i \mid \boldsymbol{X}) > \underset{j=1,2,\ldots N}{Maximum} \{P(\omega_j \mid \boldsymbol{X})\} \Rightarrow \boldsymbol{X} \in \omega_i$$

一个特例！

$\Rightarrow$ 最大后验概率判决

§ 3.4 最大似然判决规则

贝叶斯统计判决规则的似然比表现形式

- 最小错误概率判决规则的似然比表现形式 ω_1/ω_2

最大后验概率判决规则

$$\begin{cases} p(\boldsymbol{X}|\omega_1)P(\omega_1) > p(\boldsymbol{X}|\omega_2)P(\omega_2) \Rightarrow \boldsymbol{X} \in \omega_1 \\ p(\boldsymbol{X}|\omega_2)P(\omega_2) > p(\boldsymbol{X}|\omega_1)P(\omega_1) \Rightarrow \boldsymbol{X} \in \omega_2 \end{cases}$$

两端同除 $p(\boldsymbol{X}|\omega_2)P(\omega_1)$

$$\begin{cases} \dfrac{p(\boldsymbol{X}|\omega_1)}{p(\boldsymbol{X}|\omega_2)} > \dfrac{P(\omega_2)}{P(\omega_1)} \Rightarrow \boldsymbol{X} \in \omega_1 \\ \dfrac{p(\boldsymbol{X}|\omega_1)}{p(\boldsymbol{X}|\omega_2)} < \dfrac{P(\omega_2)}{P(\omega_1)} \Rightarrow \boldsymbol{X} \in \omega_2 \end{cases}$$

定义似然比 $l_{12}(\boldsymbol{X}) = \dfrac{p(\boldsymbol{X}|\omega_1)}{p(\boldsymbol{X}|\omega_2)}$ 和判决阈值 $\theta_{12} = \dfrac{P(\omega_2)}{P(\omega_1)}$

$$\boxed{\begin{cases} l_{12}(\boldsymbol{X}) > \theta_{12} \Rightarrow \boldsymbol{X} \in \omega_1 \\ l_{12}(\boldsymbol{X}) < \theta_{12} \Rightarrow \boldsymbol{X} \in \omega_2 \end{cases}}$$

§ 3.4 最大似然判决规则

贝叶斯统计判决规则的似然比表现形式

- 最小风险判决规则的似然比表现形式　　ω_1 / ω_2

  最小条件平均风险判决规则

$$\begin{cases} R(\alpha_1=\omega_1 \mid \boldsymbol{X}) < R(\alpha_2=\omega_2 \mid \boldsymbol{X}) \Rightarrow \boldsymbol{X} \in \omega_1 \\ R(\alpha_2=\omega_2 \mid \boldsymbol{X}) < R(\alpha_1=\omega_1 \mid \boldsymbol{X}) \Rightarrow \boldsymbol{X} \in \omega_2 \end{cases}$$

$$R(\alpha_1=\omega_1 \mid \boldsymbol{X}) = \sum_{j=1}^{2} L(\alpha_1 \mid \omega_j) P(\omega_j \mid \boldsymbol{X})$$

$$= L(\alpha_1 \mid \omega_1) P(\omega_1 \mid \boldsymbol{X}) + L(\alpha_1 \mid \omega_2) P(\omega_2 \mid \boldsymbol{X})$$

$$R(\alpha_2=\omega_2 \mid \boldsymbol{X}) = \sum_{j=1}^{2} L(\alpha_2 \mid \omega_j) P(\omega_j \mid \boldsymbol{X})$$

$$= L(\alpha_2 \mid \omega_1) P(\omega_1 \mid \boldsymbol{X}) + L(\alpha_2 \mid \omega_2) P(\omega_2 \mid \boldsymbol{X})$$

$$\begin{cases} (L(\alpha_2 \mid \omega_1) - L(\alpha_1 \mid \omega_1)) P(\omega_1 \mid \boldsymbol{X}) > (L(\alpha_1 \mid \omega_2) - L(\alpha_2 \mid \omega_2)) P(\omega_2 \mid \boldsymbol{X}) \Rightarrow \boldsymbol{X} \in \omega_1 \\ (L(\alpha_2 \mid \omega_1) - L(\alpha_1 \mid \omega_1)) P(\omega_1 \mid \boldsymbol{X}) < (L(\alpha_1 \mid \omega_2) - L(\alpha_2 \mid \omega_2)) P(\omega_2 \mid \boldsymbol{X}) \Rightarrow \boldsymbol{X} \in \omega_2 \end{cases}$$

§ 3.4 最大似然判决规则

贝叶斯统计判决规则的似然比表现形式

- 最小风险判决规则的似然比表现形式 ω_1 / ω_2

  最小条件平均风险判决规则

$$\begin{cases} (L(\alpha_2 | \omega_1) - L(\alpha_1 | \omega_1))P(\omega_1 | \boldsymbol{X}) > (L(\alpha_1 | \omega_2) - L(\alpha_2 | \omega_2))P(\omega_2 | \boldsymbol{X}) \Rightarrow \boldsymbol{X} \in \omega_1 \\ (L(\alpha_2 | \omega_1) - L(\alpha_1 | \omega_1))P(\omega_1 | \boldsymbol{X}) < (L(\alpha_1 | \omega_2) - L(\alpha_2 | \omega_2))P(\omega_2 | \boldsymbol{X}) \Rightarrow \boldsymbol{X} \in \omega_2 \end{cases}$$

$$\begin{cases} \dfrac{P(\omega_1 | \boldsymbol{X})}{P(\omega_2 | \boldsymbol{X})} > \dfrac{L(\alpha_1 | \omega_2) - L(\alpha_2 | \omega_2)}{L(\alpha_2 | \omega_1) - L(\alpha_1 | \omega_1)} \Rightarrow \boldsymbol{X} \in \omega_1 \\ \dfrac{P(\omega_1 | \boldsymbol{X})}{P(\omega_2 | \boldsymbol{X})} < \dfrac{L(\alpha_1 | \omega_2) - L(\alpha_2 | \omega_2)}{L(\alpha_2 | \omega_1) - L(\alpha_1 | \omega_1)} \Rightarrow \boldsymbol{X} \in \omega_2 \end{cases}$$

$$\boxed{\frac{P(\omega_1 | \boldsymbol{X})}{P(\omega_2 | \boldsymbol{X})} = \frac{\dfrac{p(\boldsymbol{X} | \omega_1)P(\omega_1)}{p(\boldsymbol{X})}}{\dfrac{p(\boldsymbol{X} | \omega_2)P(\omega_2)}{p(\boldsymbol{X})}} = \frac{p(\boldsymbol{X} | \omega_1)P(\omega_1)}{p(\boldsymbol{X} | \omega_2)P(\omega_2)}}$$

$$\begin{cases} \dfrac{p(\boldsymbol{X} | \omega_1)}{p(\boldsymbol{X} | \omega_2)} > \dfrac{(L(\alpha_1 | \omega_2) - L(\alpha_2 | \omega_2))P(\omega_2)}{(L(\alpha_2 | \omega_1) - L(\alpha_1 | \omega_1))P(\omega_1)} \Rightarrow \boldsymbol{X} \in \omega_1 \\ \dfrac{p(\boldsymbol{X} | \omega_1)}{p(\boldsymbol{X} | \omega_2)} < \dfrac{(L(\alpha_1 | \omega_2) - L(\alpha_2 | \omega_2))P(\omega_2)}{(L(\alpha_2 | \omega_1) - L(\alpha_1 | \omega_1))P(\omega_1)} \Rightarrow \boldsymbol{X} \in \omega_2 \end{cases}$$

§ 3.4 最大似然判决规则

贝叶斯统计判决规则的似然比表现形式

- 最小风险判决规则的似然比表现形式　　ω_1 / ω_2

$$\begin{cases} \dfrac{p(\boldsymbol{X} \mid \omega_1)}{p(\boldsymbol{X} \mid \omega_2)} > \dfrac{(L(\alpha_1 \mid \omega_2) - L(\alpha_2 \mid \omega_2))P(\omega_2)}{(L(\alpha_2 \mid \omega_1) - L(\alpha_1 \mid \omega_1))P(\omega_1)} \Rightarrow \boldsymbol{X} \in \omega_1 \\ \dfrac{p(\boldsymbol{X} \mid \omega_1)}{p(\boldsymbol{X} \mid \omega_2)} < \dfrac{(L(\alpha_1 \mid \omega_2) - L(\alpha_2 \mid \omega_2))P(\omega_2)}{(L(\alpha_2 \mid \omega_1) - L(\alpha_1 \mid \omega_1))P(\omega_1)} \Rightarrow \boldsymbol{X} \in \omega_2 \end{cases}$$

定义似然比和判决阈值

$$l_{12} = \frac{p(\boldsymbol{X} \mid \omega_1)}{p(\boldsymbol{X} \mid \omega_2)} \qquad \theta_{12} = \frac{(L(\alpha_1 \mid \omega_2) - L(\alpha_2 \mid \omega_2))P(\omega_2)}{(L(\alpha_2 \mid \omega_1) - L(\alpha_1 \mid \omega_1))P(\omega_1)}$$

$$\begin{cases} l_{12}(\boldsymbol{X}) > \theta_{12} \Rightarrow \boldsymbol{X} \in \omega_1 \\ l_{12}(\boldsymbol{X}) < \theta_{12} \Rightarrow \boldsymbol{X} \in \omega_2 \end{cases}$$

最小风险判决规则的似然比形式

➡ 此形式与最小错误概率判决规则的似然比形式完全一样！

➡ 统称为最大似然判决规则

§ 3.4 最大似然判决规则

贝叶斯统计判决规则的似然比表现形式

- N类情况下的最大似然比判决规则

$$L_{ij}=\frac{p(\boldsymbol{X}\mid\omega_i)}{p(\boldsymbol{X}\mid\omega_j)},\qquad i,j=1,2,\ldots N,\ j\neq i$$

$$L_{ij}(\boldsymbol{X})>\theta_{ij},\forall j\neq i,\ j=1,2,\ldots N\ \Rightarrow\ \boldsymbol{X}\in\omega_i$$

最小错误概率判决

$$\theta_{ij}=\frac{P(\omega_j)}{P(\omega_i)},\quad i,j=1,2,\ldots N,\ j\neq i$$

最小风险判决

$$\theta_{ij}=\frac{(L(\alpha_i\mid\omega_j)-L(\alpha_j\mid\omega_j))P(\omega_j)}{(L(\alpha_j\mid\omega_i)-L(\alpha_i\mid\omega_i))P(\omega_i)},\quad i,j=1,2,\ldots N,\ j\neq i$$

§ 3.5 拒绝判决

有时候，选择拒绝判决可能是更明智之举。

- 在N类判决问题中引入拒绝判决

拒绝判决 α_{N+1} $L(\alpha_{N+1} \mid \omega_j)$

$$R(\alpha_i \mid \boldsymbol{X}) = \sum_{j=1}^{N} L(\alpha_i \mid \omega_j) P(\omega_j \mid \boldsymbol{X}), \quad i = 1,2,\ldots N+1$$

拒绝判决发生的条件

$$R(\alpha_{N+1} \mid \boldsymbol{X}) < R(\alpha_i \mid \boldsymbol{X}),\ \forall i = 1,2,\ldots N \ \Rightarrow \ \boldsymbol{X} \in \alpha_{N+1}$$

若错误判决的损失 λ_F 拒绝判决的损失 λ_R

$$R(\alpha_{N+1} \mid X) = \sum_{j=1}^{N} L(\alpha_{N+1} \mid \omega_j) P(\omega_j \mid X) = \sum_{j=1}^{N} \lambda_R P(\omega_j \mid X) = \lambda_R \sum_{j=1}^{N} P(\omega_j \mid X) = \lambda_R$$

$$R(\alpha_i \mid X) = \sum_{j=1}^{N} L(\alpha_i \mid \omega_j) P(\omega_j \mid X) = \sum_{j=1, j\neq i}^{N} \lambda_F P(\omega_j \mid X) = \lambda_F \sum_{j=1, j\neq i}^{N} P(\omega_j \mid X)$$

$i \neq N+1$

$$= \lambda_F \left(\sum_{j=1}^{N} P(\omega_j \mid X) - P(\omega_i \mid X)\right) = \lambda_F (1 - P(\omega_i \mid X))$$

§ 3.5 拒绝判决

拒绝判决发生的条件

$$R(\alpha_{N+1} \mid \boldsymbol{X}) < R(\alpha_i \mid \boldsymbol{X}),\ \forall i = 1,2,\ldots N \ \Rightarrow \ \boldsymbol{X} \in \alpha_{N+1}$$

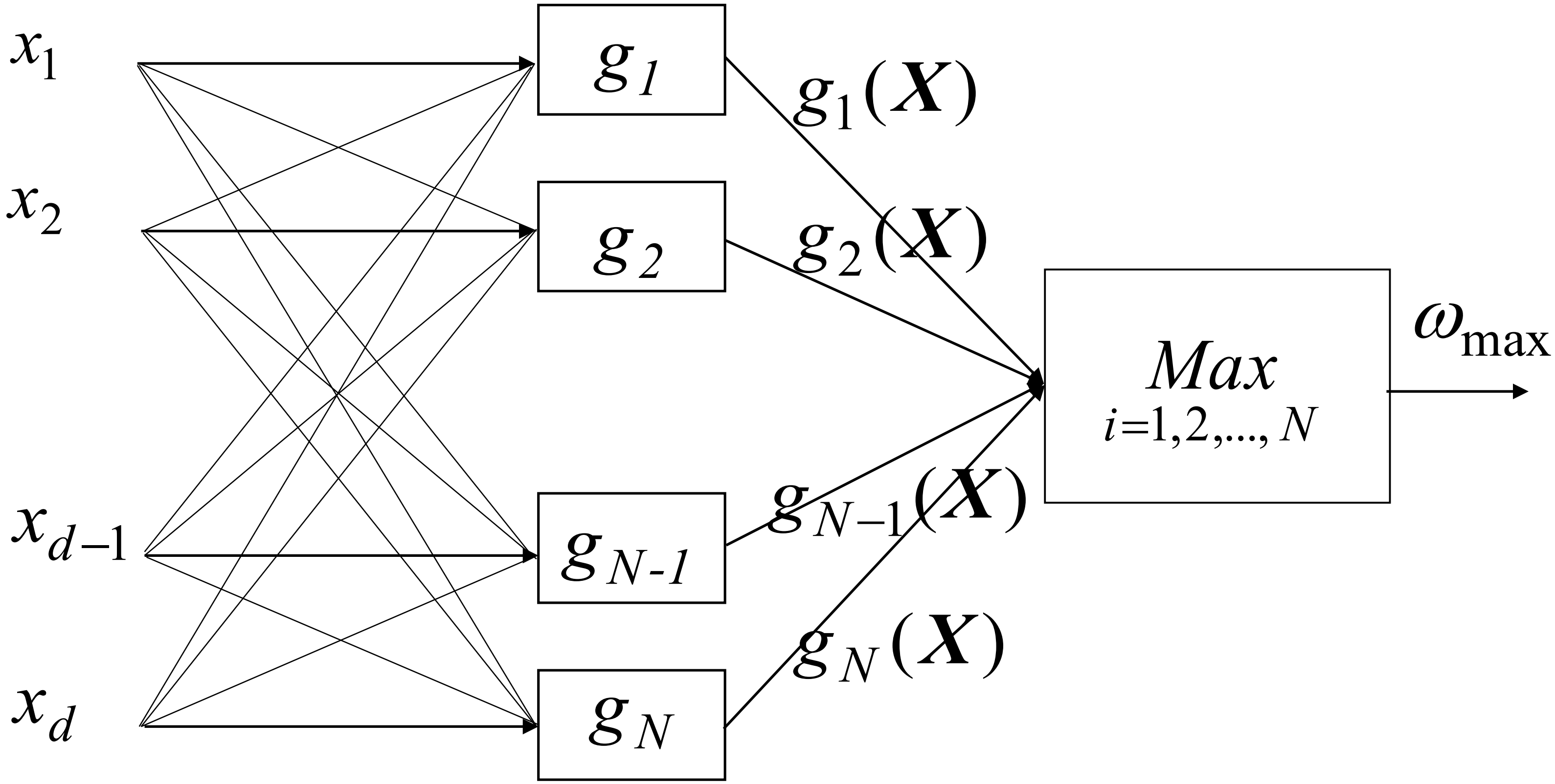
$$R(\alpha_{N+1} \mid \boldsymbol{X}) = \lambda_R \qquad R(\alpha_i \mid \boldsymbol{X}) = \lambda_F(1 - P(\omega_i \mid \boldsymbol{X})), i \neq N+1$$

$$\lambda_R < \lambda_F(1 - P(\omega_i \mid \boldsymbol{X})), \forall i = 1,2,\ldots N \ \Rightarrow \ \boldsymbol{X} \in \alpha_{N+1}$$

$$P(\omega_i \mid \boldsymbol{X}) < (1 - \frac{\lambda_R}{\lambda_F}), \forall i = 1,2,\ldots N \ \Rightarrow \ \boldsymbol{X} \in \alpha_{N+1}$$

所设置的错误判决损失和拒绝判决损失两者之间需要满足一定的条件。

§ 3.6 贝叶斯分类器的一般结构



最大后验概率判决

$$g_i(\boldsymbol{X}) = P(\omega_i \mid \boldsymbol{X})$$

最小风险判决

$$g_i(\boldsymbol{X}) = -R(\alpha_i \mid \boldsymbol{X})$$

§ 3.6 贝叶斯分类器的一般结构

判别函数的选择

- 一个判别函数乘上一个正常数后仍然有资格作为判别函数。
- 一个判别函数加上一个正常数后仍然有资格作为判别函数。
- 一个单调增加函数作用于判别函数后形成的新的函数仍然有资格作为判别函数。

$$g_i(\boldsymbol{X}) = p(\boldsymbol{X} \mid \omega_i)P(\omega_i)$$

$$g_i(\boldsymbol{X}) = \log p(\boldsymbol{X} \mid \omega_i) + \log P(\omega_i)$$

$$g_i(\boldsymbol{X}) = \frac{p(\boldsymbol{X} \mid \omega_i)P(\omega_i)}{\sum_{j=1}^{N} p(\boldsymbol{X} \mid \omega_j)P(\omega_j)}$$

§ 3.7 Neyman-Pearson判决规则

问题的提出

最小错误概率判决规则

$$\begin{cases} \boxed{P(\omega_1 \mid \boldsymbol{X})} > P(\omega_2 \mid \boldsymbol{X}) \quad \Rightarrow \quad \boldsymbol{X} \in \omega_1 \\ \boxed{P(\omega_2 \mid \boldsymbol{X})} > P(\omega_1 \mid \boldsymbol{X}) \quad \Rightarrow \quad \boldsymbol{X} \in \omega_2 \end{cases}$$

最小风险判决规则

$$\begin{cases} \boxed{\dfrac{P(\omega_1 \mid \boldsymbol{X})}{P(\omega_2 \mid \boldsymbol{X})}} > \dfrac{L(\alpha_1 \mid \omega_2) - L(\alpha_2 \mid \omega_2)}{L(\alpha_2 \mid \omega_1) - L(\alpha_1 \mid \omega_1)} \Rightarrow \quad \boldsymbol{X} \in \omega_1 \\ \boxed{\dfrac{P(\omega_1 \mid \boldsymbol{X})}{P(\omega_2 \mid \boldsymbol{X})}} < \dfrac{L(\alpha_1 \mid \omega_2) - L(\alpha_2 \mid \omega_2)}{L(\alpha_2 \mid \omega_1) - L(\alpha_1 \mid \omega_1)} \Rightarrow \quad \boldsymbol{X} \in \omega_2 \end{cases}$$

$$P(\omega_j \mid \boldsymbol{X}) = \frac{\boxed{p(\boldsymbol{X} \mid \omega_j)} \boxed{P(\omega_j)}}{\sum_{j=1}^{N} p(\boldsymbol{X} \mid \omega_j) P(\omega_j)}, \; j = 1,2 \longrightarrow \text{未知}$$

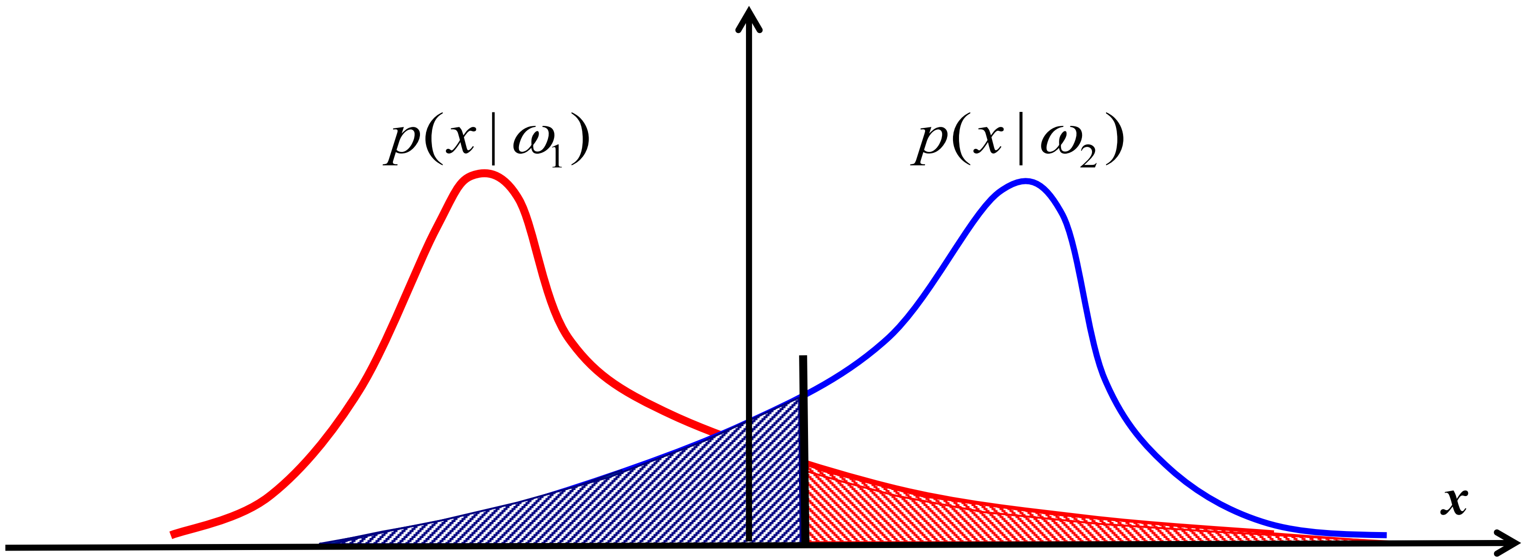
§ 3.7 Neyman-Pearson判决规则

问题的提出

$$P(e)=\int_{\Omega_2}P(\omega_1)p(x|\omega_1)dx+\int_{\Omega_1}P(\omega_2)p(x|\omega_2)dx$$

$$=P(\omega_1)\boxed{\int_{\Omega_2}p(x|\omega_1)dx}+P(\omega_2)\boxed{\int_{\Omega_1}p(x|\omega_2)dx}$$

两种分类错误的危害程度是不一样的！



§ 3.7 Neyman-Pearson判决规则

- 在保持一类分类错误概率不变的条件下，使另一类分类错误概率最小。

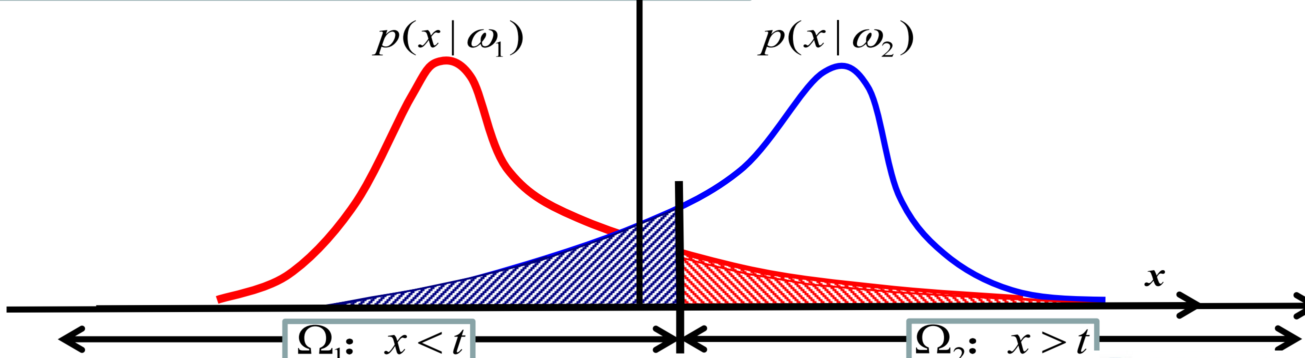
用Lagrange乘子法求解

目标函数　　　约束条件

$$J = \int_{\Omega_2} p(x|\omega_1)dx + \lambda\left(\int_{\Omega_1} p(x|\omega_2)dx - \alpha\right)$$

$$= 1 - \int_{\Omega_1} p(x|\omega_1)dx + \lambda\left(\int_{\Omega_1} p(x|\omega_2)dx - \alpha\right)$$

$$\boxed{= (1-\lambda\alpha) + \int_{\Omega_1} (\lambda p(x|\omega_2) - p(x|\omega_1))dx} \xrightarrow{\text{确定}\Omega_1} \text{极小化}$$



§ 3.7 Neyman-Pearson判决规则

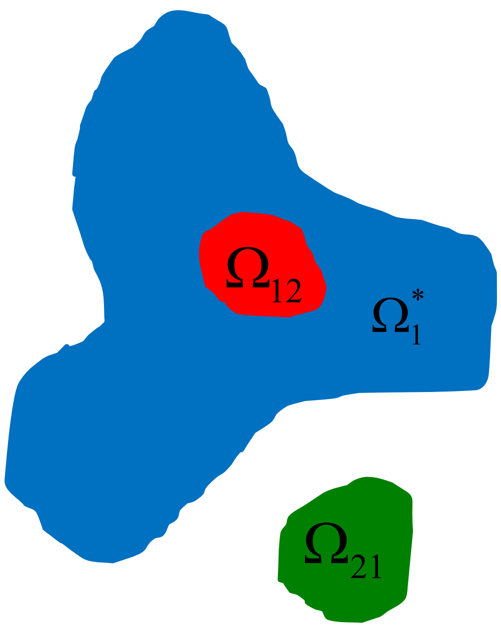
$$J=(1-\lambda\alpha)+\int_{\Omega_1}(\lambda p(x\,|\,\omega_2)-p(x\,|\,\omega_1))dx$$

➡ 设使*J*极小化的 Ω_1为 Ω_1^*

结论：极值解 Ω_1^* 由满足$\lambda p(x\,|\,\omega_2)-p(x\,|\,\omega_1)<0$ 的全体*x*所组成。

证明：用反证法。设另有 $\Omega_1^\#$，使得后项积分有更小的值。

则$\Omega_1^\#$一定可以表示成 $\Omega_1^\#=(\Omega_1^*-\Omega_{12})\cup\Omega_{21}$

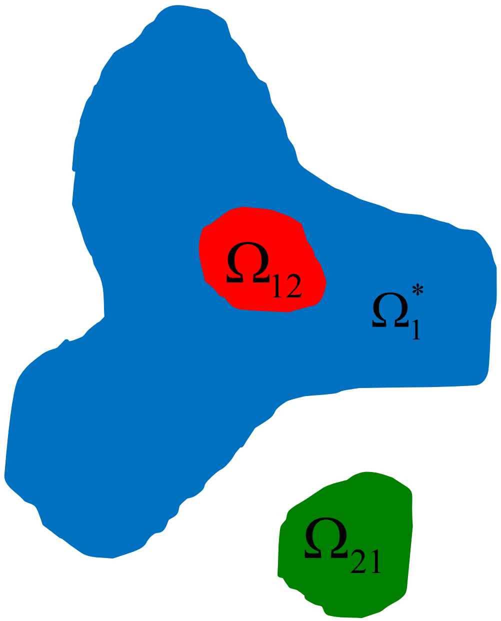


$$\begin{cases}\Omega_{12}: \lambda p(x\,|\,\omega_2)-p(x\,|\,\omega_1)<0\\ \Omega_{21}: \lambda p(x\,|\,\omega_2)-p(x\,|\,\omega_1)>0\end{cases}$$

当判决区域取为$\Omega_1^\#$时

$$J^\#=(1-\lambda\alpha)+\int_{\Omega_1^\#}(\lambda p(x\,|\,\omega_2)-p(x\,|\,\omega_1))dx$$

§ 3.7 Neyman-Pearson判决规则



$$
\begin{aligned}
J^\# &= (1-\lambda\alpha) + \int_{\Omega_1^\#} (\lambda p(x|\omega_2) - p(x|\omega_1))dx \\
&= (1-\lambda\alpha) + \int_{\Omega_1^* - \Omega_{12}} (\lambda p(x|\omega_2) - p(x|\omega_1))dx \\
&\quad + \int_{\Omega_{21}} (\lambda p(x|\omega_2) - p(x|\omega_1))dx \\
&= \boxed{(1-\lambda\alpha) + \int_{\Omega_1^*} (\lambda p(x|\omega_2) - p(x|\omega_1))dx} \\
&\quad \boxed{- \int_{\Omega_{12}} (\lambda p(x|\omega_2) - p(x|\omega_1))dx} \\
&\quad \boxed{+ \int_{\Omega_{21}} (\lambda p(x|\omega_2) - p(x|\omega_1))dx} \\
&\geq J^*
\end{aligned}
$$

$$\Omega_1^\# = (\Omega_1^* - \Omega_{12}) \cup \Omega_{21}$$

Ω_{12}：$\lambda p(x|\omega_2) - p(x|\omega_1) < 0$

Ω_{21}：$\lambda p(x|\omega_2) - p(x|\omega_1) > 0$

与假设矛盾。

§ 3.7 Neyman-Pearson判决规则

Neyman-Pearson判决规则可表为

$$\begin{cases} \lambda p(x|\omega_2) - p(x|\omega_1) < 0 \Rightarrow & x \in \omega_1 \\ \lambda p(x|\omega_2) - p(x|\omega_1) > 0 \Rightarrow & x \in \omega_2 \end{cases}$$

$$\begin{cases} \dfrac{p(x|\omega_1)}{p(x|\omega_2)} > \lambda \Rightarrow & x \in \omega_1 \\ \dfrac{p(x|\omega_1)}{p(x|\omega_2)} < \lambda \Rightarrow & x \in \omega_2 \end{cases}$$

特殊情况下可求解析解

试探法求数值解

λ的确定

$$\lambda \rightarrow \Omega_1 \rightarrow \int p(x|\omega_2)dx = \alpha ?$$

§ 3.7 Neyman-Pearson判决规则

λ的确定

特殊情况举例：

- 特征空间为一维空间
- 类条件概率密度函数为单峰函数

$$J=(1-\lambda\alpha)+\int_{-\infty}^{T}(\lambda p(x\mid\omega_2)-p(x\mid\omega_1))dx$$

$$\frac{\partial J}{\partial T}=p(T\mid\omega_1)-\lambda p(T\mid\omega_2)=0 \qquad \frac{\partial J}{\partial \lambda}=-\alpha+\int_{-\infty}^{T}p(x\mid\omega_2)dx=0$$

$$\lambda=\frac{p(T\mid\omega_1)}{p(T\mid\omega_2)} \qquad \int_{-\infty}^{T}p(x\mid\omega_2)dx=\alpha$$

§ 3.8 最小最大判决规则

主要思想来源于对策论，是一种保守但稳妥的判决规则。

追求的境界：最坏情况下结果最好，其它情况下结果可用。

生产条件相同的两农户的实例：

无度的消费为挨饿受冻埋下了伏笔！

A农户：激进型消费理念，有时风光，有时凄惨；

B农户：保守型消费理念，未雨绸缪，消费有度。

生活虽谈不上奢侈，但也不是很拘谨，并且能有盈余；
到了收成差一点的年头，因为有储备，生活也过得去。

§ 3.8 最小最大判决规则

考虑N类别分类问题

$$\omega_j, j = 1,2,\ldots N$$

$$\alpha_j, j = 1,2,\ldots A$$

在很多情况下，先验概率不仅未知，并且可能是随时间变化的。

$$R(\alpha_i \mid \boldsymbol{X}) = \sum_{j=1}^{N} L(\alpha_i \mid \omega_j) P(\omega_j \mid \boldsymbol{X})$$

问题点：

1. 不知道如何使用
2. 不能反映总的平均风险

$\because \boldsymbol{X}$是一个随机向量

$\therefore\ \alpha_i, i = 1,2,\ldots A$也是一个随机向量 $\Rightarrow\ \alpha(\boldsymbol{X})$

总的平均风险

$$\overline{R} = \int_{\boldsymbol{E}_d} R(\alpha(\boldsymbol{X}) \mid \boldsymbol{X}) p(\boldsymbol{X}) d\boldsymbol{X} \Rightarrow \text{最小化}$$

$$\overline{R} = \int_{\Omega_1} R(\alpha_1 \mid \boldsymbol{X}) p(\boldsymbol{X}) d\boldsymbol{X} + \int_{\Omega_2} R(\alpha_2 \mid \boldsymbol{X}) p(\boldsymbol{X}) d\boldsymbol{X} + \Lambda + \int_{\Omega_N} R(\alpha_N \mid \boldsymbol{X}) p(\boldsymbol{X}) d\boldsymbol{X}$$

$\therefore$ 总的平均风险与类别划分结果有关

§ 3.8 最小最大判决规则

当先验概率变化时如何确定一个合理的分类规则？

ω_1/ω_2 两类问题

$$\overline{R}=\int_{\Omega_1}R(\alpha_1\mid \boldsymbol{X})p(\boldsymbol{X})d\boldsymbol{X}+\int_{\Omega_2}R(\alpha_2\mid \boldsymbol{X})p(\boldsymbol{X})d\boldsymbol{X}$$

假定判决数等于类别数，则

$$\begin{aligned}R(\alpha_i\mid \boldsymbol{X})&=L(\alpha_i\mid\omega_1)P(\omega_1\mid \boldsymbol{X})+L(\alpha_i\mid\omega_2)P(\omega_2\mid \boldsymbol{X})\\&=L(\alpha_i\mid\omega_1)\frac{p(\boldsymbol{X}\mid\omega_1)P(\omega_1)}{p(\boldsymbol{X})}+L(\alpha_i\mid\omega_2)\frac{p(\boldsymbol{X}\mid\omega_2)P(\omega_2)}{p(\boldsymbol{X})}\qquad i=1,2\end{aligned}$$

$$\begin{aligned}\overline{R}&=\int_{\Omega_1}[L(\alpha_1\mid\omega_1)p(\boldsymbol{X}\mid\omega_1)P(\omega_1)+L(\alpha_1\mid\omega_2)p(\boldsymbol{X}\mid\omega_2)P(\omega_2)]d\boldsymbol{X}\\&+\int_{\Omega_2}[L(\alpha_2\mid\omega_1)p(\boldsymbol{X}\mid\omega_1)P(\omega_1)+L(\alpha_2\mid\omega_2)p(\boldsymbol{X}\mid\omega_2)P(\omega_2)]d\boldsymbol{X}\end{aligned}$$

代入 $P(\omega_2)=1-P(\omega_1)$

按积分项整理

$$\begin{aligned}\overline{R}&=L(\alpha_1\mid\omega_1)P(\omega_1)\int_{\Omega_1}p(\boldsymbol{X}\mid\omega_1)d\boldsymbol{X}+[L(\alpha_1\mid\omega_2)-L(\alpha_1\mid\omega_2)P(\omega_1)]\int_{\Omega_1}p(\boldsymbol{X}\mid\omega_2)d\boldsymbol{X}\\&+L(\alpha_2\mid\omega_1)P(\omega_1)\int_{\Omega_2}p(\boldsymbol{X}\mid\omega_1)d\boldsymbol{X}+[L(\alpha_2\mid\omega_2)-L(\alpha_2\mid\omega_2)P(\omega_1)]\int_{\Omega_2}p(\boldsymbol{X}\mid\omega_2)d\boldsymbol{X}\end{aligned}$$

§ 3.8 最小最大判决规则

$$\overline{R} = L(\alpha_1 | \omega_1)P(\omega_1)\int_{\Omega_1} p(\boldsymbol{X} | \omega_1)d\boldsymbol{X} + [L(\alpha_1 | \omega_2) - L(\alpha_1 | \omega_2)P(\omega_1)]\int_{\Omega_1} p(\boldsymbol{X} | \omega_2)d\boldsymbol{X}$$
$$+ L(\alpha_2 | \omega_1)P(\omega_1)\int_{\Omega_2} p(\boldsymbol{X} | \omega_1)d\boldsymbol{X} + [L(\alpha_2 | \omega_2) - L(\alpha_2 | \omega_2)P(\omega_1)]\int_{\Omega_2} p(\boldsymbol{X} | \omega_2)d\boldsymbol{X}$$

$$\int_{\Omega_1} p(\boldsymbol{X} | \omega_1)d\boldsymbol{X} + \int_{\Omega_2} p(\boldsymbol{X} | \omega_1)d\boldsymbol{X} = 1 \Rightarrow \int_{\Omega_1} p(\boldsymbol{X} | \omega_1)d\boldsymbol{X} = 1 - \int_{\Omega_2} p(\boldsymbol{X} | \omega_1)d\boldsymbol{X}$$

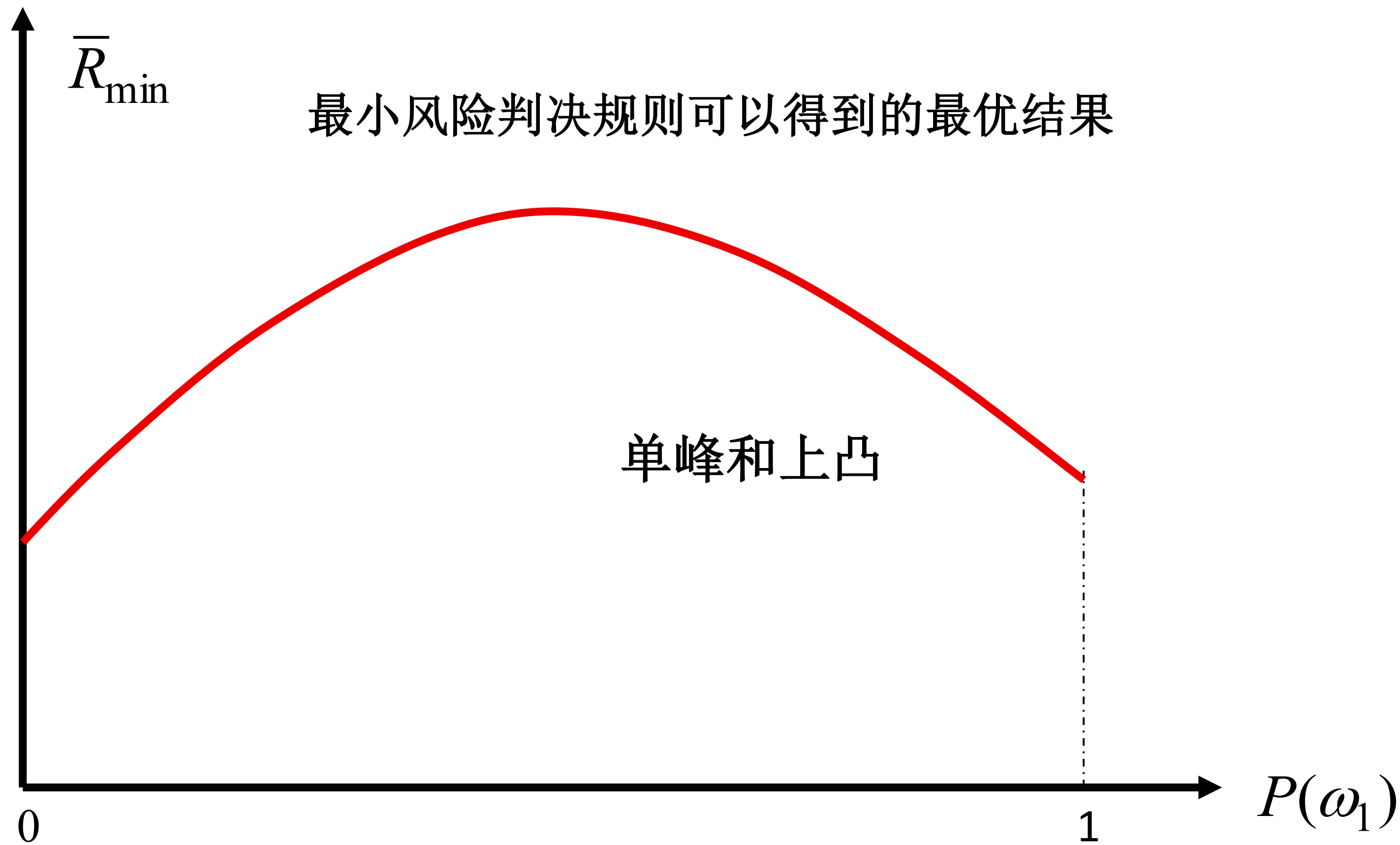
$$\int_{\Omega_1} p(\boldsymbol{X} | \omega_2)d\boldsymbol{X} + \int_{\Omega_2} p(\boldsymbol{X} | \omega_2)d\boldsymbol{X} = 1 \Rightarrow \int_{\Omega_2} p(\boldsymbol{X} | \omega_2)d\boldsymbol{X} = 1 - \int_{\Omega_1} p(\boldsymbol{X} | \omega_2)d\boldsymbol{X}$$

$$\overline{R} = L(\alpha_1 | \omega_1)P(\omega_1)[1 - \int_{\Omega_2} p(\boldsymbol{X} | \omega_1)d\boldsymbol{X}] + [L(\alpha_1 | \omega_2) - L(\alpha_1 | \omega_2)P(\omega_1)]\int_{\Omega_1} p(\boldsymbol{X} | \omega_2)d\boldsymbol{X}$$
$$+ L(\alpha_2 | \omega_1)P(\omega_1)\int_{\Omega_2} p(\boldsymbol{X} | \omega_1)d\boldsymbol{X} + [L(\alpha_2 | \omega_2) - L(\alpha_2 | \omega_2)P(\omega_1)][1 - \int_{\Omega_1} p(\boldsymbol{X} | \omega_2)d\boldsymbol{X}]$$
$$= \underbrace{L(\alpha_2 | \omega_2) + [L(\alpha_1 | \omega_2) - L(\alpha_2 | \omega_2)]\int_{\Omega_1} p(\boldsymbol{X} | \omega_2)d\boldsymbol{X}}_{\Rightarrow a} +$$
$$+ \underbrace{\{[L(\alpha_1 | \omega_1) - L(\alpha_2 | \omega_2)] + [L(\alpha_2 | \omega_1) - L(\alpha_1 | \omega_1)]\int_{\Omega_2} p(\boldsymbol{X} | \omega_1)d\boldsymbol{X} - [L(\alpha_1 | \omega_2) - L(\alpha_2 | \omega_2)]\int_{\Omega_1} p(\boldsymbol{X} | \omega_2)d\boldsymbol{X}\}}_{\Rightarrow b} P(\omega_1)$$

$$\therefore\ \overline{R} = a + bP(\omega_1)$$

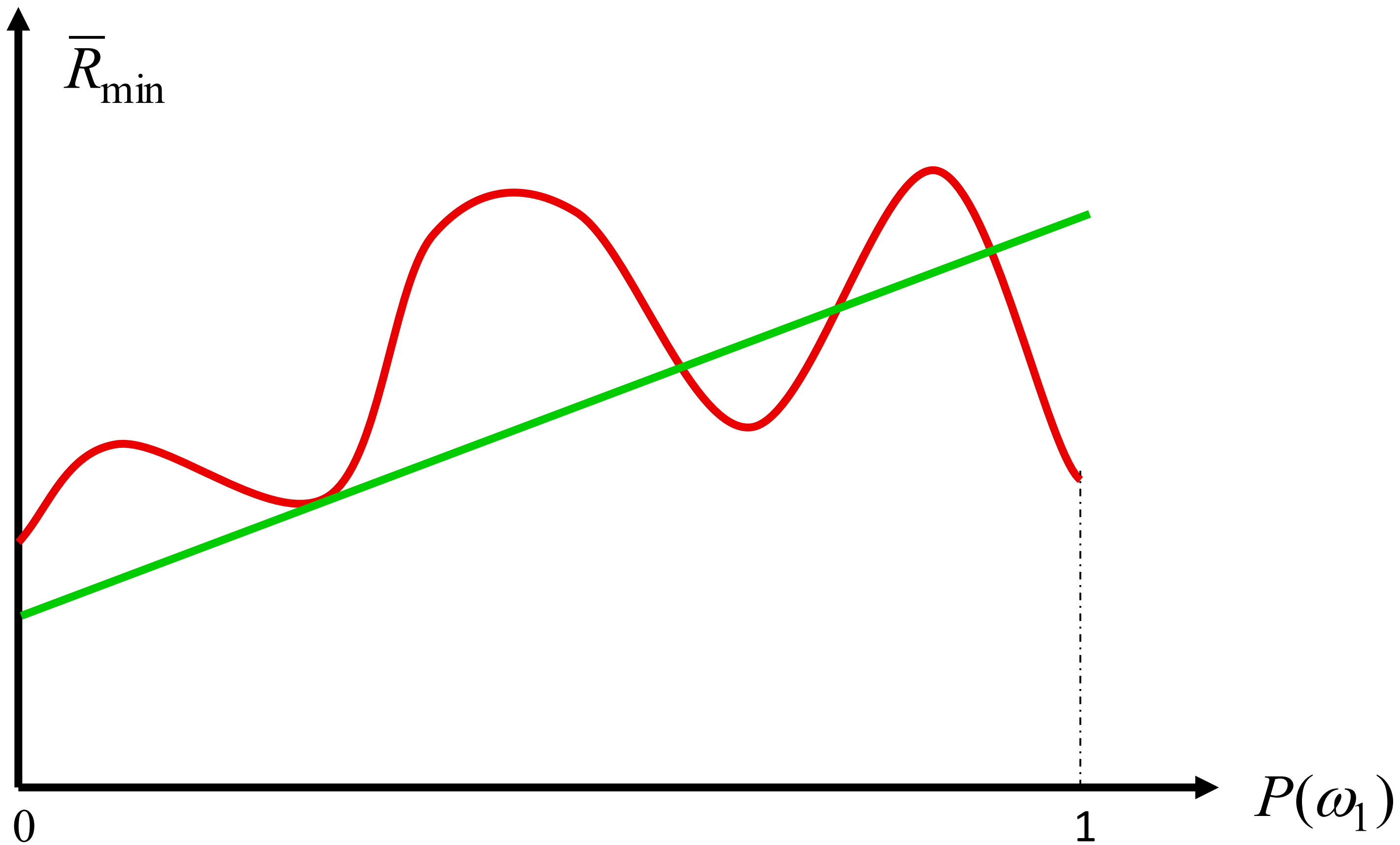
§ 3.8 最小最大判决规则

$\overline{R} = a + bP(\omega_1)$ 最小化总平均风险 $\overline{R}_{\min} = a + bP(\omega_1)$



§ 3.8 最小最大判决规则

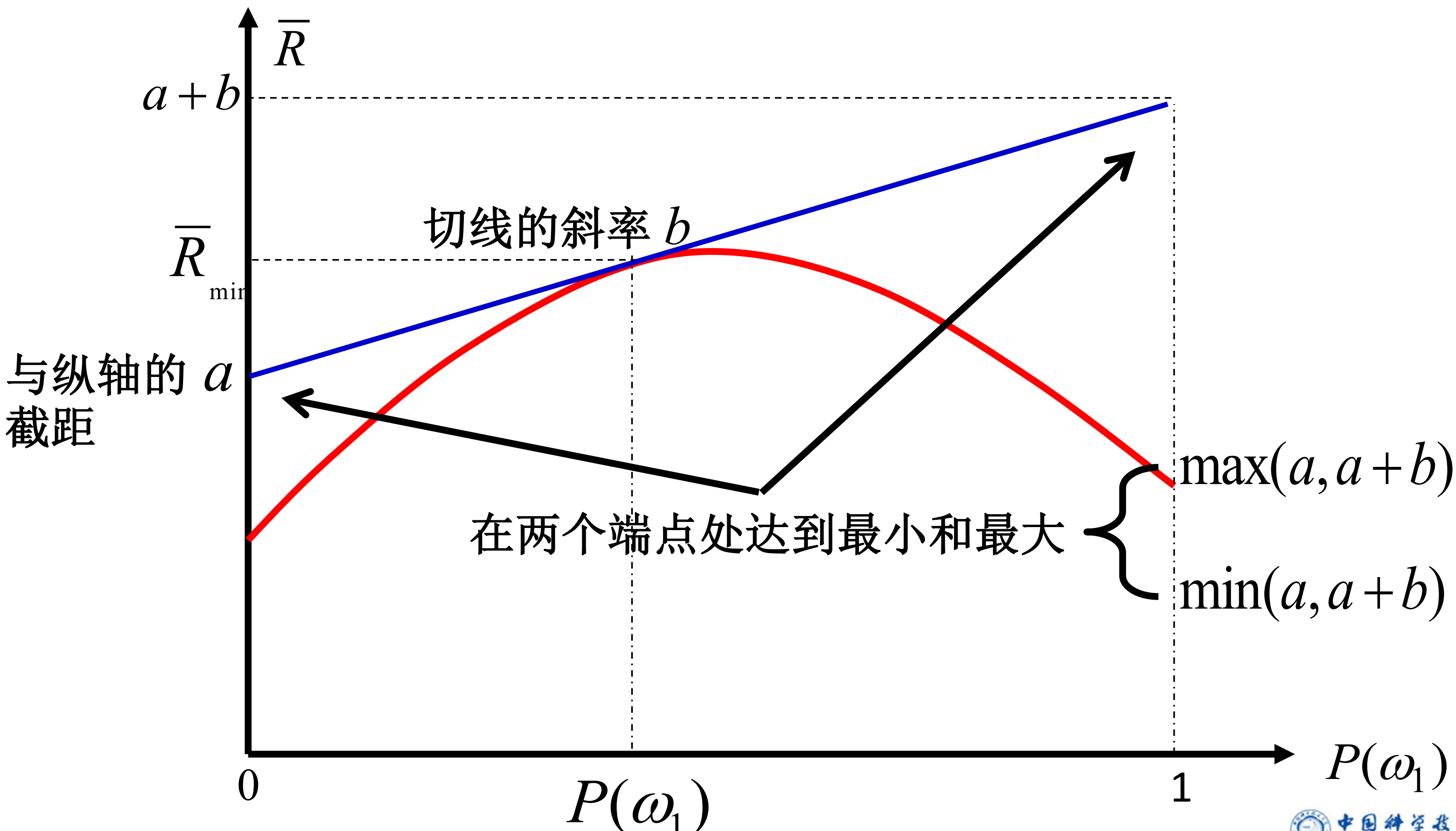
单峰性和上凸性的证明



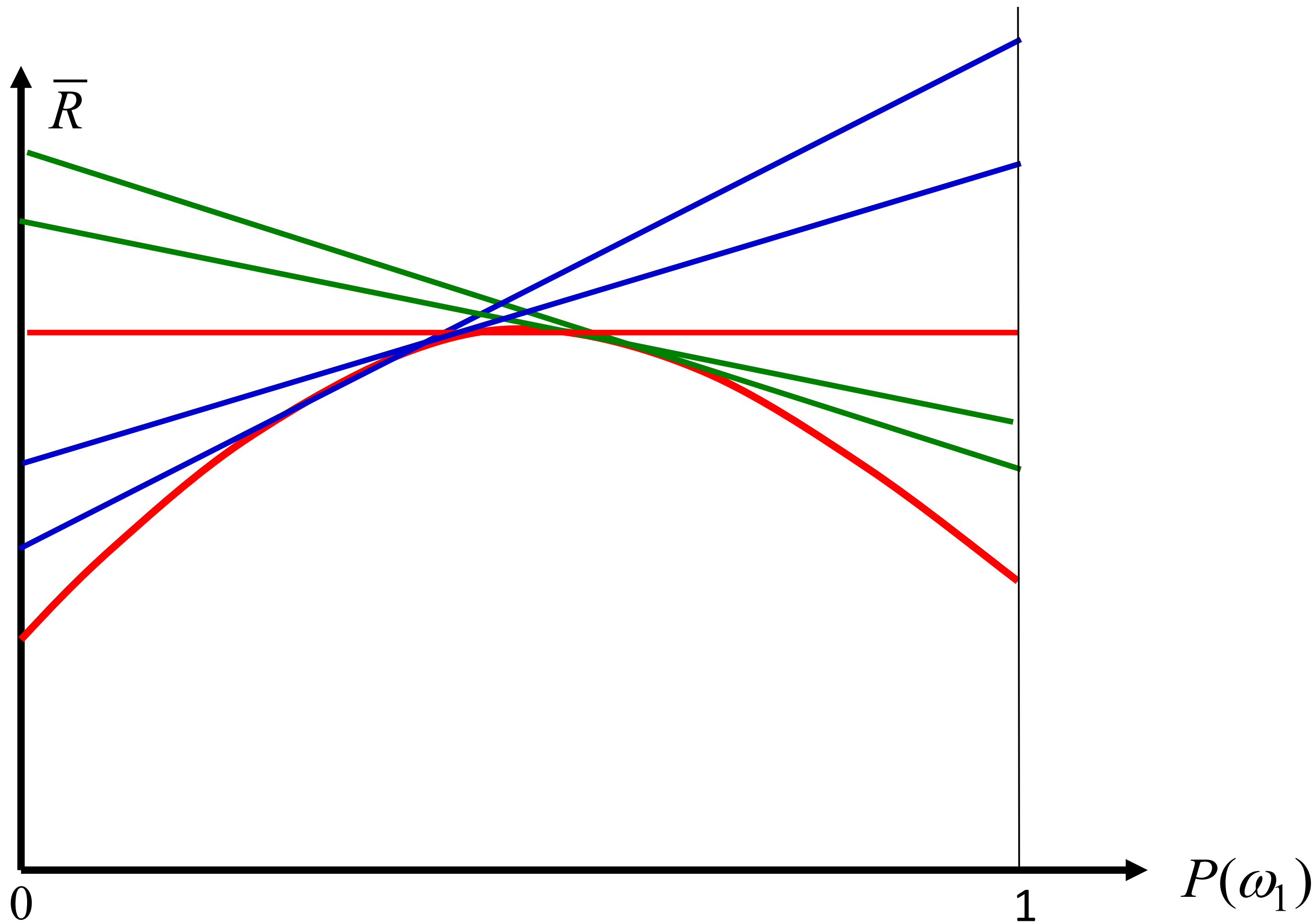
§ 3.8 最小最大判决规则

最小总平均风险

$$\overline{R}_{\min} = a + bP(\omega_1) \quad \Rightarrow \quad \frac{\partial \overline{R}_{\min}}{\partial P(\omega_1)} = b$$



§ 3.8 最小最大判决规则



结论：使最大的总的平均风险最小化的分类判决由关系曲线最大值所对应的分类判决所给出。

§ 3.8 最小最大判决规则

Step1. 确定关系曲线的极大值点 $(P_{\min,\max}(\omega_1),\overline{R}_{\min,\max})$；

$$\frac{\partial \overline{R}_{\min}}{\partial P(\omega_1)}=b=0$$

Step2. 根据极大值点，如下构造分类器：

$$\begin{cases}\dfrac{p(\boldsymbol{X}\mid\omega_1)}{p(\boldsymbol{X}\mid\omega_2)}>\dfrac{(L(\alpha_1\mid\omega_2)-L(\alpha_2\mid\omega_2))(1-P_{\min,\max}(\omega_1))}{(L(\alpha_2\mid\omega_1)-L(\alpha_1\mid\omega_1))P_{\min,\max}(\omega_1)} \Rightarrow \boldsymbol{X}\in\omega_1\\[2ex] \dfrac{p(\boldsymbol{X}\mid\omega_1)}{p(\boldsymbol{X}\mid\omega_2)}<\dfrac{(L(\alpha_1\mid\omega_2)-L(\alpha_2\mid\omega_2))(1-P_{\min,\max}(\omega_1))}{(L(\alpha_2\mid\omega_1)-L(\alpha_1\mid\omega_1))P_{\min,\max}(\omega_1)} \Rightarrow \boldsymbol{X}\in\omega_2\end{cases}$$

§ 3.8 最小最大判决规则

问题点：

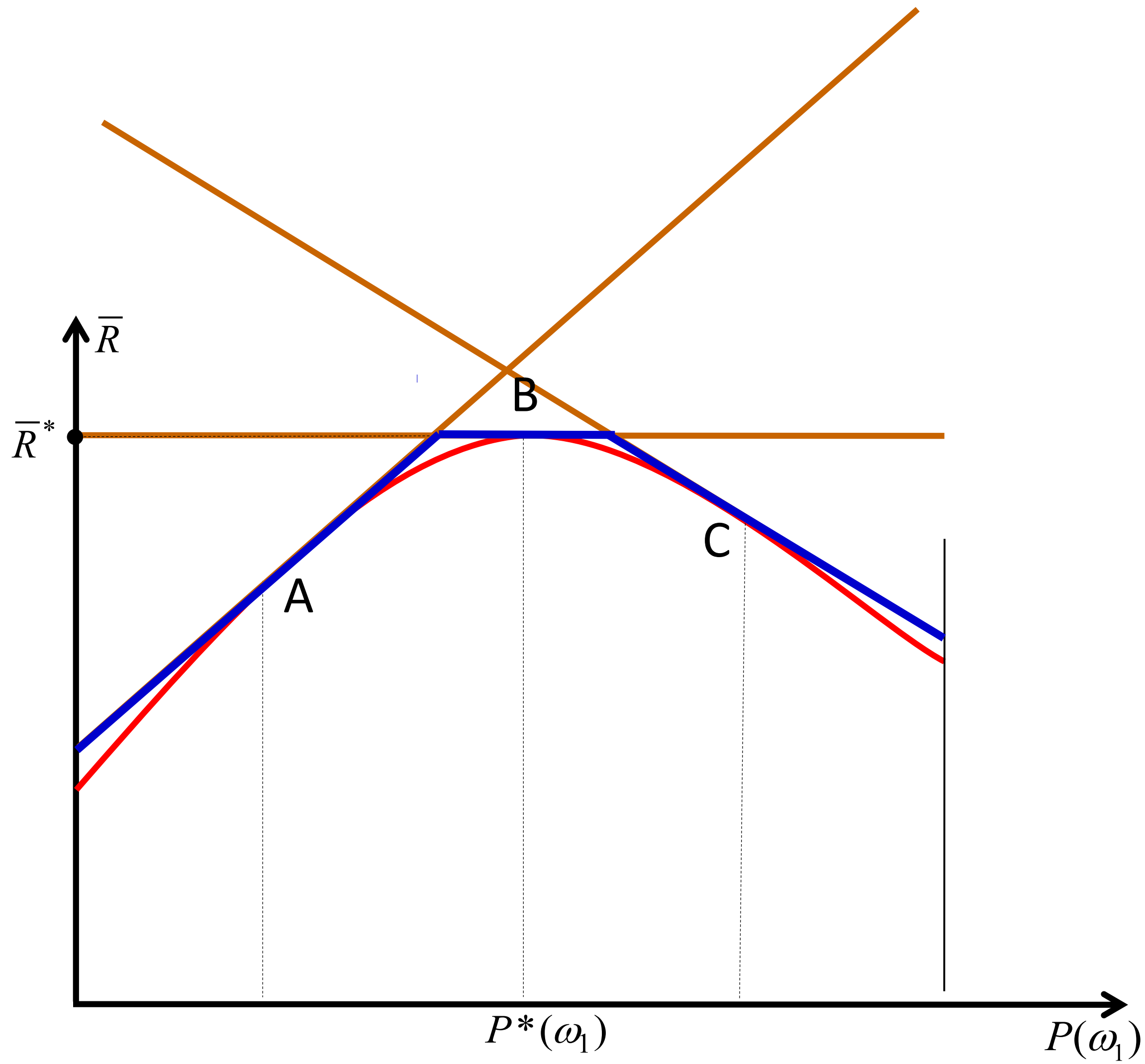
使最大的可能风险最小化是一种过于保守的判决，往往导致分类器性能下降过多。

改进思路：分段线性化

- 用和关系曲线上特定的三个点**A**、**B**、**C**相切的三条折线对最小总平均风险进行近似（三条折线所在的区间形成相互连接但互不交叠的三个区间）。
- 对实际的 $P(\omega_1)$进行估计，根据其落入的区间，选择使用相应的判决规则。

分段线性化处理可以大幅度提高分类器的性能。

§ 3.8 最小最大判决规则



§ 3.9 正态分布下的分类器设计

问题：正态分布下N类别贝叶斯分类器的设计

判别函数 $g_i(\boldsymbol{X}) = p(\boldsymbol{X} \mid \omega_i)P(\omega_i), \quad i = 1,2,\ldots N$

在正态分布下，有：

$$g_i(\boldsymbol{X}) = \frac{P(\omega_i)}{(2\pi)^{\frac{d}{2}}\left|\boldsymbol{\Sigma}_i\right|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_i)\right\}, \qquad i = 1,2,\ldots N$$

改造之，有：

$$g_i(\boldsymbol{X}) = \ln P(\omega_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right| - -\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_i), \qquad i = 1,2,\ldots N$$

进一步进行简化，有：

$$g_i(\boldsymbol{X}) = \ln P(\omega_i) - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right| - \frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_i), \qquad i = 1,2,\ldots N$$

§ 3.9 正态分布下的分类器设计

$$g_i(\boldsymbol{X}) = \ln P(\omega_i) - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right| - \frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_i), \quad \mathrm{i} = 1,2,\ldots N$$

某些特殊情况下的分类判决结果

（1）$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$

$$\left|\boldsymbol{\Sigma}_i\right| = \sigma^{2d} \qquad \boldsymbol{\Sigma}_i^{-1} = \frac{1}{\sigma^2}\boldsymbol{I}$$

样本到类别均值向量的欧氏距离

$$g_i(\boldsymbol{X}) = \ln \boxed{P(\omega_i)} - \frac{1}{2\sigma^2}\boxed{(\boldsymbol{X}-\boldsymbol{\mu}_i)^T(\boldsymbol{X}-\boldsymbol{\mu}_i)}, \quad \mathrm{i} = 1,2,\ldots N$$

当各类别的先验概率相等时，判决样本和欧氏意义下相距最近的类别均值向量具有相同的类别。

此时，分类判决的决策面具有什么形式？

$$g_i(\boldsymbol{X}) = \ln P(\omega_i) - \frac{1}{2\sigma^2}(\boldsymbol{X}^T\boldsymbol{X} - 2\boldsymbol{\mu}_i^T\boldsymbol{X} + \boldsymbol{\mu}_i^T\boldsymbol{\mu}_i), \qquad i = 1,2,\ldots N$$

$$g_i(\boldsymbol{X}) = \frac{1}{\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{X} - \frac{1}{2\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i + \ln P(\omega_i), \qquad i = 1,2,\ldots N$$

§ 3.9 正态分布下的分类器设计

$$g_i(\boldsymbol{X}) = \boxed{\frac{1}{\sigma^2}\boldsymbol{\mu}_i^T} \boldsymbol{X} \boxed{-\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i + \ln P(\omega_i)}, \qquad i = 1,2,\ldots N$$

$$\boldsymbol{W}_i \qquad\qquad w_{i,N+1}$$

$$g_i(\boldsymbol{X}) = \boldsymbol{W}_i^T\boldsymbol{X} + w_{i,N+1}, \qquad i = 1,2,\ldots N$$ —线性函数

若类别ω_i和ω_j相邻，则相应的决策面为

$$g_i(\boldsymbol{X}) - g_j(\boldsymbol{X}) = 0$$

$$\frac{1}{\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{X} - \frac{1}{2\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i + \ln P(\omega_i) - \frac{1}{\sigma^2}\boldsymbol{\mu}_j^T\boldsymbol{X} + \frac{1}{2\sigma^2}\boldsymbol{\mu}_j^T\boldsymbol{\mu}_j - \ln P(\omega_j)$$

$$= \frac{1}{\sigma^2}(\boldsymbol{\mu}_i^T - \boldsymbol{\mu}_j^T)\boldsymbol{X} - \frac{1}{2\sigma^2}(\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i - \boldsymbol{\mu}_j^T\boldsymbol{\mu}_j) + \ln\frac{P(\omega_i)}{P(\omega_j)}$$

$$= \frac{1}{\sigma^2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T\boldsymbol{X} - \frac{1}{\sigma^2}\left[\frac{1}{2}(\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i - \boldsymbol{\mu}_j^T\boldsymbol{\mu}_j) - \sigma^2\ln\frac{P(\omega_i)}{P(\omega_j)}\right] = 0$$

§ 3.9 正态分布下的分类器设计

$$\frac{1}{\sigma^2}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T \boldsymbol{X}-\frac{1}{\sigma^2}[\frac{1}{2}(\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i-\boldsymbol{\mu}_j^T\boldsymbol{\mu}_j)-\sigma^2\ln\frac{P(\omega_i)}{P(\omega_j)}]=0$$

$$(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T \boldsymbol{X}-[\frac{1}{2}(\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i-\boldsymbol{\mu}_j^T\boldsymbol{\mu}_j)-\sigma^2\ln\frac{P(\omega_i)}{P(\omega_j)}]=0$$

$$\because \quad \boldsymbol{\mu}_i^T\boldsymbol{\mu}_i-\boldsymbol{\mu}_j^T\boldsymbol{\mu}_j=(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)$$

$$(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T \boldsymbol{X}-[\frac{1}{2}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)-\frac{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}\sigma^2\ln\frac{P(\omega_i)}{P(\omega_j)}]=0$$

$$(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T \boldsymbol{X}-(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T[\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)-\frac{\sigma^2\ln\frac{P(\omega_i)}{P(\omega_j)}}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)]=0$$

$$(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\left[\boldsymbol{X}-[\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)-\frac{\sigma^2\ln\frac{P(\omega_i)}{P(\omega_j)}}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)]\right]=0$$

§ 3.9 正态分布下的分类器设计

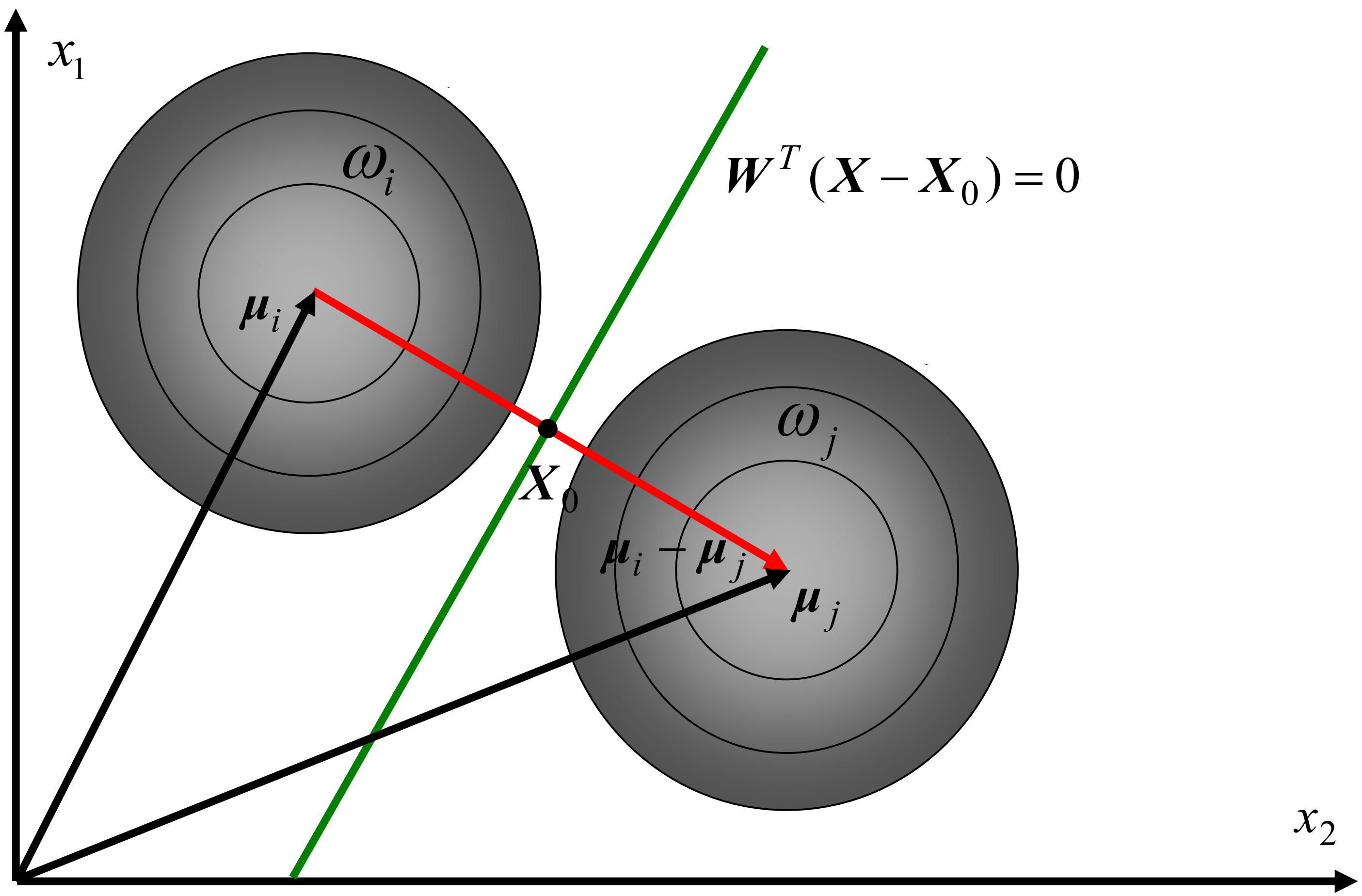
$$\boxed{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}^T\left[X-\boxed{[\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)-\frac{\sigma^2\ln\frac{P(\omega_i)}{P(\omega_j)}}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)]}\right]=0$$

$$\boldsymbol{W} \qquad\qquad \boldsymbol{X}_0$$

$$\boldsymbol{W}^T(\boldsymbol{X}-\boldsymbol{X}_0)=0$$

过点X_0并与权向量X正交的超平面方程

§ 3.9 正态分布下的分类器设计

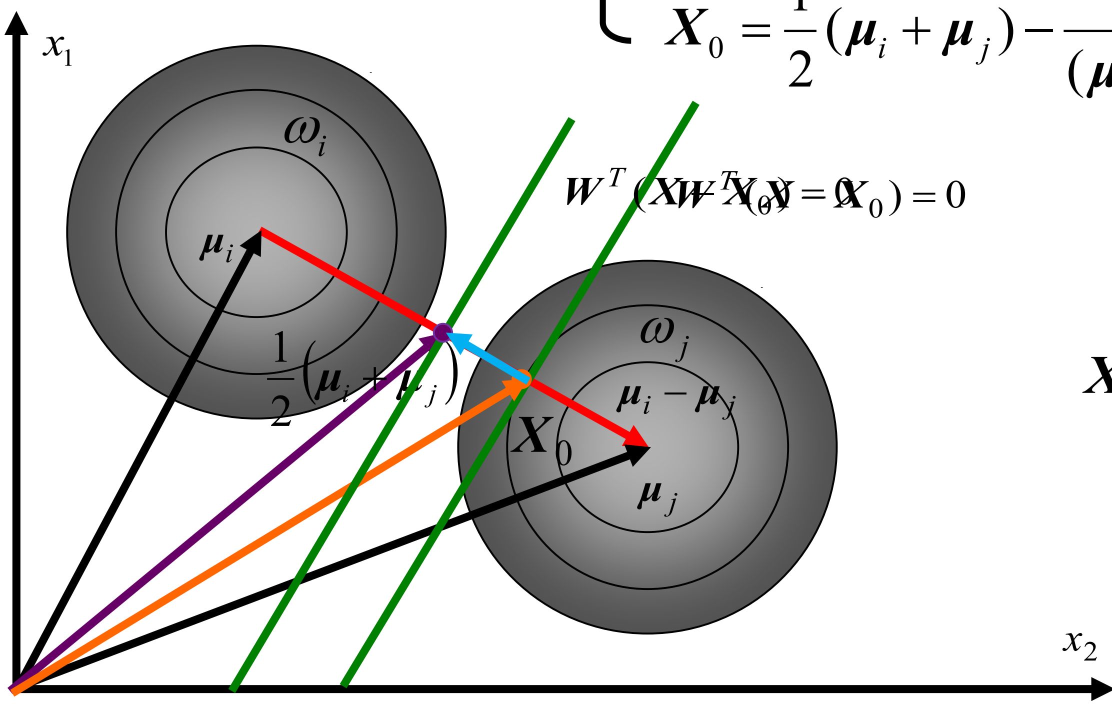


§ 3.9 正态分布下的分类器设计

结论： **1.** 若 $P(\omega_i)=P(\omega_j)$，则 $X_0=\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)$ 。

2. 若 $P(\omega_i)\neq P(\omega_j)$，则 $X_0=\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)-\alpha(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$。

$$\begin{cases} \boldsymbol{W}^T(\boldsymbol{X}-\boldsymbol{X}_0)=0 \\ X_0=\dfrac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)-\dfrac{\sigma^2\ln\dfrac{P(\omega_i)}{P(\omega_j)}}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j) \end{cases}$$



$$X_0-\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)=-\alpha(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$$

§ 3.9 正态分布下的分类器设计

$$g_i(X) = \ln P(\omega_i) - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(X-\mu_i)^T \Sigma_i^{-1}(X-\mu_i), \quad \mathrm{i} = 1,2,\dots N$$

（2）$\Sigma_i = \Sigma \neq$ **对角阵**

样本到类别均值向量的马氏距离

$$g_i(X) = \ln P(\omega_i) - \frac{1}{2}(X-\mu_i)^T \Sigma^{-1}(X-\mu_i), \qquad \mathrm{i} = 1,2,\dots \mathrm{N}$$

当各类别的先验概率相等时，判决样本和马氏意义下相距最近的类别均值向量具有相同的类别。

用于分类判决的决策面方程

若类别ω_i和ω_j相邻，则相应的决策面为

$$g_i(X) - g_j(X) = 0$$

$$\ln P(\omega_i) - \frac{1}{2}(X-\mu_i)^T \Sigma^{-1}(X-\mu_i) - \ln P(\omega_j) + \frac{1}{2}(X-\mu_j)^T \Sigma^{-1}(X-\mu_j) = 0$$

§ 3.9 正态分布下的分类器设计

$$\ln P(\omega_i) - \frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_i) - \ln P(\omega_j) + \frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_j) = 0$$

$$\frac{1}{2}\left[(\boldsymbol{X}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_i) - (\boldsymbol{X}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_j)\right] - \ln\frac{P(\omega_i)}{P(\omega_j)} = 0$$

$\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$ 项对消

$$\because \quad (\boldsymbol{X}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_i) - (\boldsymbol{X}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_j)$$

$$= -\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} - \boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j$$

$$= -2\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + 2\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} - \boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j$$

$$= -2(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i - \boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j$$

和 $(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\Sigma^{-1}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j) = \boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{\mu}_i - \boldsymbol{\mu}_j^T\Sigma^{-1}\boldsymbol{\mu}_j$

$$\therefore \quad \frac{1}{2}\left[(\boldsymbol{X}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_i) - (\boldsymbol{X}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_j)\right]$$

$$= -(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \frac{1}{2}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\Sigma^{-1}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j) = -(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{X}-\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)\right)$$

§ 3.9 正态分布下的分类器设计

$$-(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j))-\ln\frac{P(\omega_i)}{P(\omega_j)}=0$$

$$(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j))+\ln\frac{P(\omega_i)}{P(\omega_j)}=0$$

$$(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j))+\frac{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}\ln\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}=0$$

$$(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{X}-\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)+\frac{\ln\frac{P(\omega_i)}{P(\omega_j)}}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)\right)=0$$

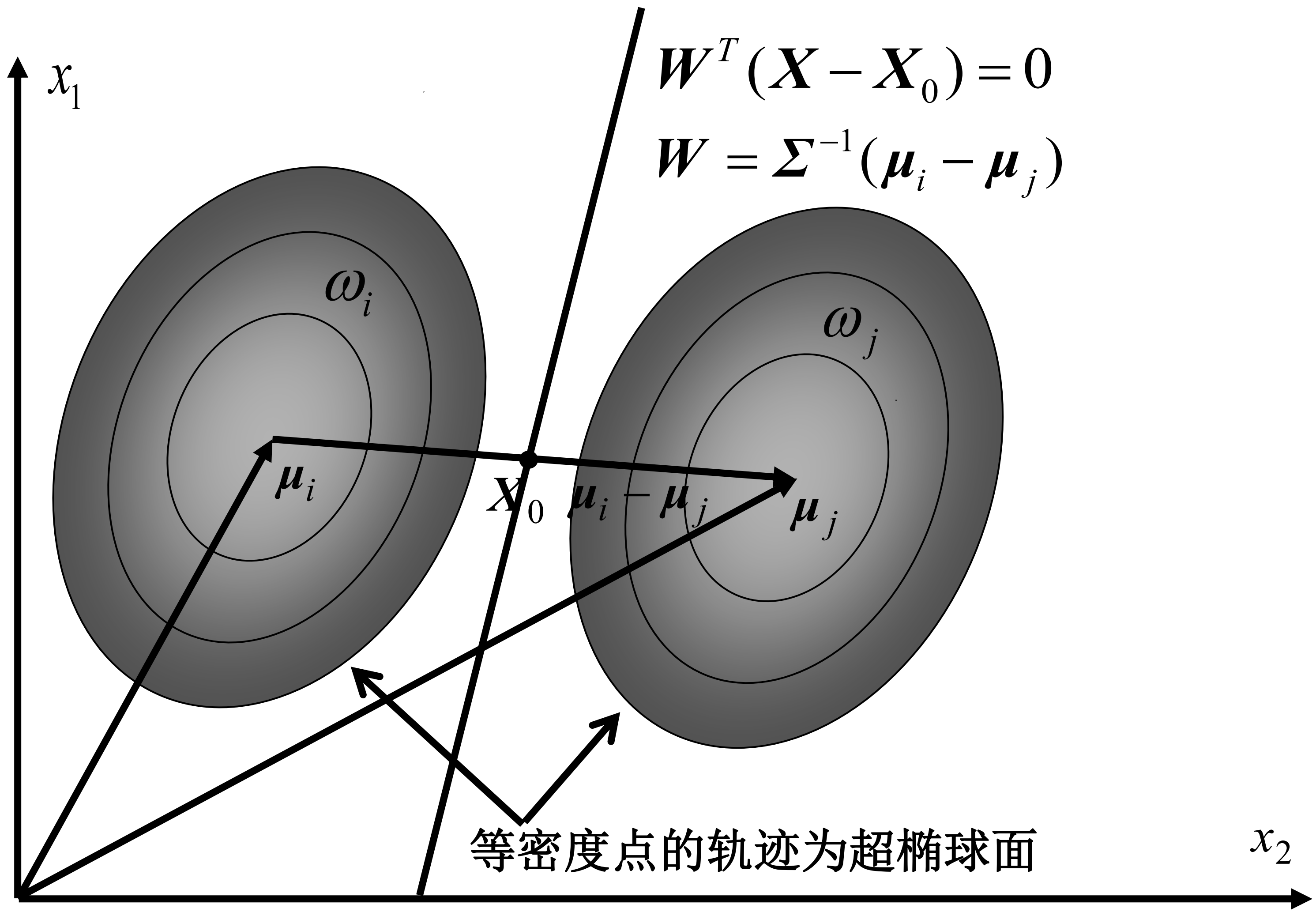
§ 3.9 正态分布下的分类器设计

$$(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{X}-\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)+\frac{\ln\dfrac{P(\omega_i)}{P(\omega_j)}}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)\right)=0$$

$$\underbrace{\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)\right)^T}_{\boldsymbol{W}}\left(\boldsymbol{X}-\underbrace{\left(\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)-\frac{\ln\dfrac{P(\omega_i)}{P(\omega_j)}}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)\right)}_{\boldsymbol{X}_0}\right)=0$$

$$\Rightarrow\quad \boldsymbol{W}^T(\boldsymbol{X}-\boldsymbol{X}_0)=0$$

§ 3.9 正态分布下的分类器设计



§ 3.9 正态分布下的分类器设计

$$g_i(\boldsymbol{X}) = \ln P(\omega_i) - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right| - \frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_i), \quad \mathrm{i} = 1,2,\ldots N$$

（3） $\boldsymbol{\Sigma}_i$**任意**

$$g_i(\boldsymbol{X}) = \ln P(\omega_i) - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right| - \frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_i), \qquad i = 1,2,\ldots N$$

$$g_i(\boldsymbol{X}) = -\frac{1}{2}\boldsymbol{X}^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X} + \boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X} - \frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i) - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right|$$

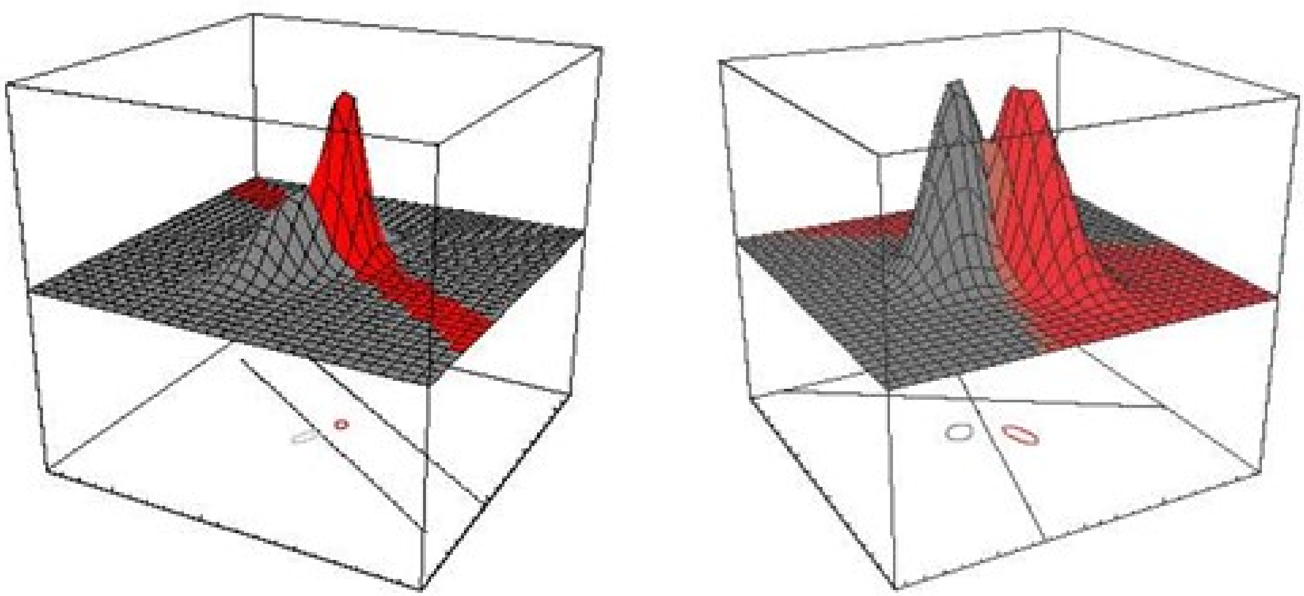
$$g_i(\boldsymbol{X}) = \boldsymbol{X}^T(\underbrace{-\frac{1}{2}\boldsymbol{\Sigma}_i}_{\boldsymbol{A}_i})^{-1}\boldsymbol{X} + (\underbrace{\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i}_{\boldsymbol{W}_i})^T\boldsymbol{X} + \underbrace{\ln P(\omega_i) - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right| - \frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i}_{w_{i.N+1}}$$

$$g_i(\boldsymbol{X}) = \boldsymbol{X}^T\boldsymbol{A}_i^{-1}\boldsymbol{X} + \boldsymbol{W}_i^T\boldsymbol{X} + w_{i.N+1} \qquad i = 1,2,\ldots N$$

§ 3.9 正态分布下的分类器设计

$$g_i(\boldsymbol{X}) = \boldsymbol{X}^T \boldsymbol{A}_i^{-1} \boldsymbol{X} + \boldsymbol{W}_i^T \boldsymbol{X} + w_{i.N+1} \qquad i = 1,2,\ldots N$$

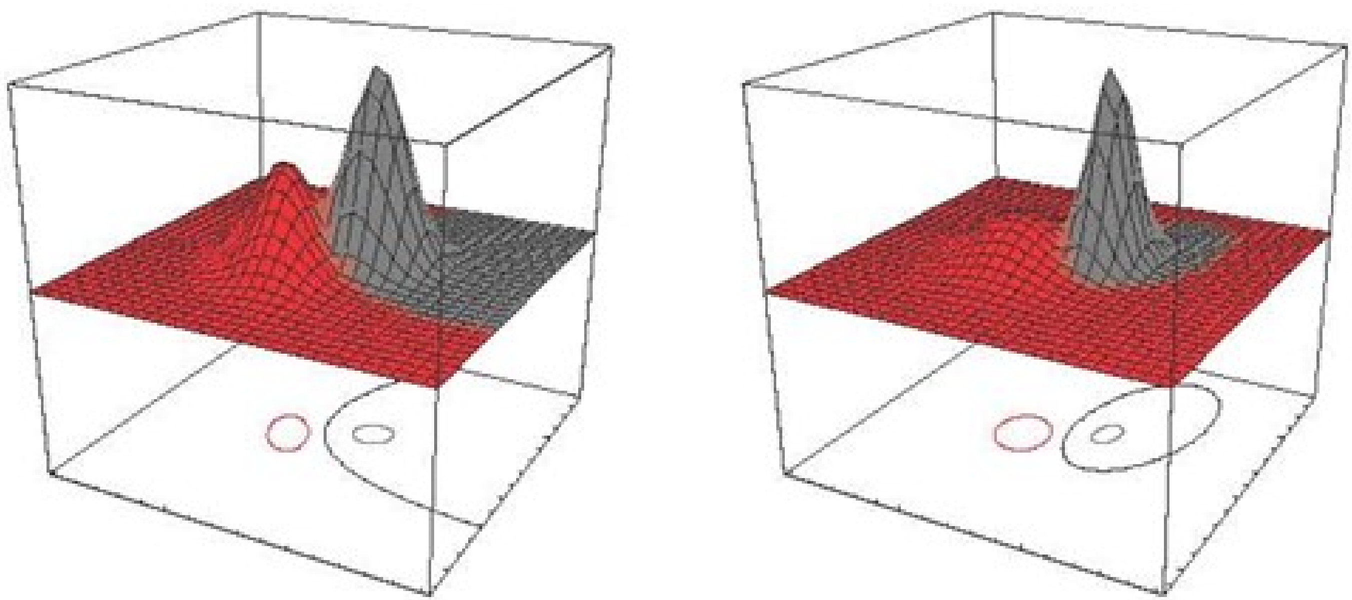
- 决策面一般是一个超二次曲面。
  - ✓ 超球面、超椭球面、超抛物面、超双曲面
- 特殊条件下，超二次曲面可退化为成对出现的超平面。



§ 3.9 正态分布下的分类器设计

$$g_i(\boldsymbol{X}) = \boldsymbol{X}^T \boldsymbol{A}_i^{-1} \boldsymbol{X} + \boldsymbol{W}_i^T \boldsymbol{X} + w_{i.N+1} \qquad i = 1,2,\ldots N$$

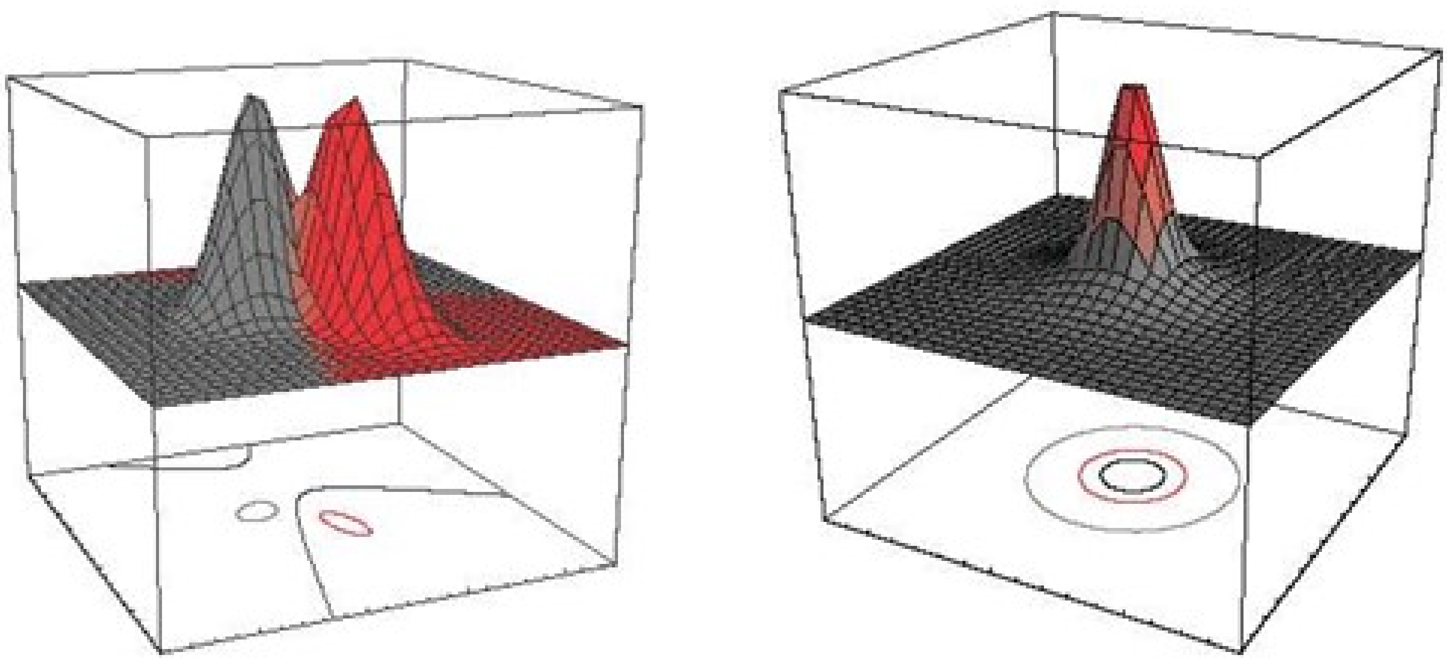
- 决策面一般是一个超二次曲面。
  - ✓ 超球面、超椭球面、超抛物面、超双曲面
- 特殊条件下，超二次曲面可退化为成对出现的超平面。



§ 3.9 正态分布下的分类器设计

$$g_i(\boldsymbol{X}) = \boldsymbol{X}^T \boldsymbol{A}_i^{-1} \boldsymbol{X} + \boldsymbol{W}_i^T \boldsymbol{X} + w_{i.N+1} \qquad i = 1,2,\ldots N$$

- 决策面一般是一个超二次曲面。
  - ✓ 超球面、超椭球面、超抛物面、超双曲面
- 特殊条件下，超二次曲面可退化为成对出现的超平面。



§ 3.10 类条件概率密度的参数估计

类条件概率密度函数的估计

已知条件：有监督的训练样本集合

输出：所估计的类条件概率密度

手段和方法：

- 参数估计
  - 函数形式已知，但参数未知
  - 最大似然估计
  - 贝叶斯估计和贝叶斯学习
- 非参数估计
  - Parzen窗估计
  - Kn-近邻估计
  - 函数形式和参数均未知

§ 3.10 类条件概率密度的参数估计

类条件概率密度函数的最大似然估计

已知条件：类别数（S），训练样本及标签。

输出：估计参数。

方法和步骤：

Step 1. 化整为零，分而治之；

Step 2. 独立作战，各个击破。

独立性假设

$$\mathscr{X}=\{X_1,X_2,\ldots,X_n\}$$

S个样本子集， S个子问题

$$\mathscr{X}_1=\{\boldsymbol{X}_{1,1},\boldsymbol{X}_{1,2},\ldots,\boldsymbol{X}_{1,n_1}\} \Rightarrow p(\boldsymbol{X}\mid\omega_1,\boldsymbol{\theta}_1)$$

$$\mathscr{X}_2=\{\boldsymbol{X}_{2,1},\boldsymbol{X}_{2,2},\ldots,\boldsymbol{X}_{2,n_2}\} \Rightarrow p(\boldsymbol{X}\mid\omega_2,\boldsymbol{\theta}_2)$$

....

$$\mathscr{X}_i=\{\boldsymbol{X}_{i,1},\boldsymbol{X}_{i,2},\ldots,\boldsymbol{X}_{i,n_i}\} \Rightarrow p(\boldsymbol{X}\mid\omega_i,\boldsymbol{\theta}_i)$$

....

$$\mathscr{X}_s=\{\boldsymbol{X}_{S,1},\boldsymbol{X}_{S,2},\ldots,\boldsymbol{X}_{S,n_S}\} \Rightarrow p(\boldsymbol{X}\mid\omega_S,\boldsymbol{\theta}_S)$$

§ 3.10 类条件概率密度的参数估计

类条件概率密度函数的最大似然估计

$\mathscr{X}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_n\} \Rightarrow p(\boldsymbol{X}\mid\boldsymbol{\theta})$

假设$\boldsymbol{X}_k, k=1,2,\ldots n$按照类条件概率密度 $p(\boldsymbol{X}\mid\boldsymbol{\theta})$ 从总体中独立抽取，则

$$p(\mathscr{X}\mid\boldsymbol{\theta})=p(\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_n\mid\boldsymbol{\theta})=\prod_{k=1}^{n}p(\boldsymbol{X}_k\mid\boldsymbol{\theta})$$ ——以$\boldsymbol{\theta}$为参数的似然函数

样本子集确定后，上述函数仅为θ的函数。

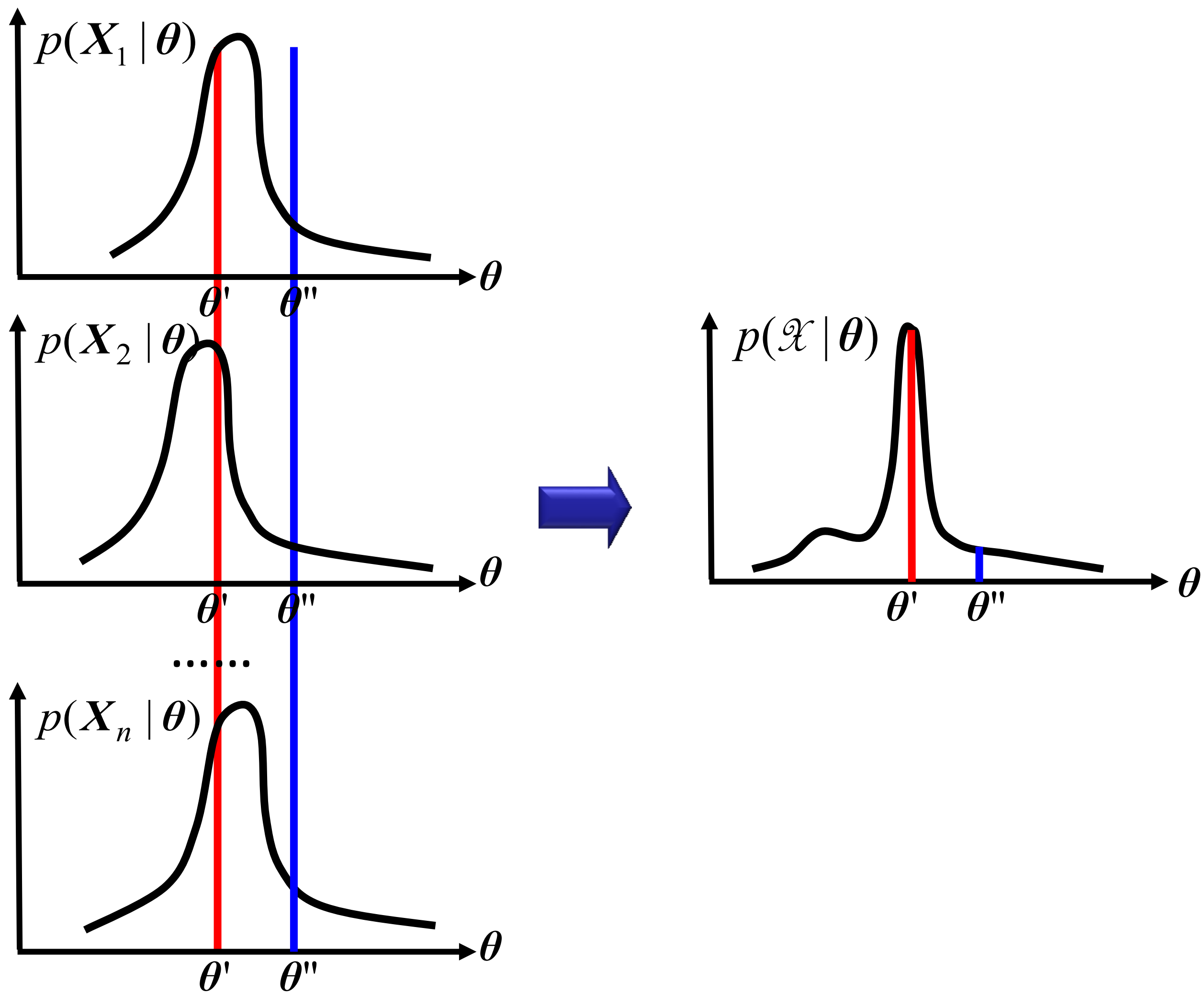
固定θ，可计算似然函数的值，其值大小反映了θ值的似然程度。

对不同的θ，计算似然函数的值，可得到不同θ的取值可能性的估计。

取使似然函数的值最大化的θ作为参数θ的估计。

§ 3.10 类条件概率密度的参数估计

类条件概率密度函数的最大似然估计



§ 3.10 类条件概率密度的参数估计

类条件概率密度函数的最大似然估计

$$\mathscr{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$$

$$p(\mathscr{X} \mid \boldsymbol{\theta}) = p(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \mid \boldsymbol{\theta}) = \prod_{k=1}^{n} p(\boldsymbol{X}_k \mid \boldsymbol{\theta})$$

$$\boldsymbol{\mathcal{L}}(\boldsymbol{\theta}) = \log p(\mathscr{X} \mid \boldsymbol{\theta}) = \log \prod_{k=1}^{n} p(\boldsymbol{X}_k \mid \boldsymbol{\theta}) = \sum_{k=1}^{n} \log p(\boldsymbol{X}_k \mid \boldsymbol{\theta})$$

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{\mathcal{L}}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log p(\mathscr{X} \mid \boldsymbol{\theta}) = \sum_{k=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{X}_k \mid \boldsymbol{\theta}) = 0$$

求解上述方程，得到参数θ的最大似然估计。

§ 3.10 类条件概率密度的参数估计

类条件概率密度函数的最大似然估计

- 正态分布的参数估计举例

条件：$\mathscr{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$，均值向量μ未知。

$$\mathcal{L}(\boldsymbol{\mu}) = \ln p(\mathscr{X} \mid \boldsymbol{\mu}) = \ln \prod_{k=1}^{n} p(\boldsymbol{X}_k \mid \boldsymbol{\mu}) = \sum_{k=1}^{n} \ln p(\boldsymbol{X}_k \mid \boldsymbol{\mu})$$

$$= \sum_{k=1}^{n} \ln \left[\frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\boldsymbol{X}_k - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}_k - \boldsymbol{\mu}) \right\} \right]$$

$$\nabla_{\boldsymbol{\mu}} \mathcal{L}(\boldsymbol{\mu}) = \frac{\partial}{\partial \boldsymbol{\mu}} \sum_{k=1}^{n} \ln \left[\frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\boldsymbol{X}_k - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}_k - \boldsymbol{\mu}) \right\} \right]$$

$$= \sum_{k=1}^{n} -\frac{\partial}{\partial \boldsymbol{\mu}} \left[\frac{1}{2} (\boldsymbol{X}_k - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}_k - \boldsymbol{\mu}) \right] = \sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}_k - \boldsymbol{\mu}) = 0$$

$$\Rightarrow \quad \hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{X}_k$$

§ 3.10 类条件概率密度的参数估计

类条件概率密度函数的最大似然估计

- 正态分布的参数估计举例

条件：$\mathscr{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$，协方差矩阵Σ未知。

$$\nabla_{\Sigma} \mathcal{L}(\Sigma) = \frac{\partial}{\partial \Sigma} \sum_{k=1}^{n} \ln\left[\frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{X}_k - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{X}_k - \boldsymbol{\mu})\right\}\right]$$

$$= \sum_{k=1}^{n}\left\{-\frac{1}{2|\Sigma|}\frac{\partial|\Sigma|}{\partial \Sigma} - \frac{\partial}{\partial \Sigma}\left[\frac{1}{2}(\boldsymbol{X}_k - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{X}_k - \boldsymbol{\mu})\right]\right\}$$

$$= \sum_{k=}^{n}\left\{-\frac{1}{2}\Sigma^{-1} + \frac{1}{2}(\boldsymbol{X}_k - \boldsymbol{\mu})(\boldsymbol{X}_k - \boldsymbol{\mu})^T \Sigma^{-1}\Sigma^{-1}\right\}$$

$$\Rightarrow \hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{X}_k - \boldsymbol{\mu})(\boldsymbol{X}_k - \boldsymbol{\mu})^T$$

$$\frac{\partial|\Sigma|}{\partial \Sigma} = |\Sigma|(\Sigma^{-1})^T = |\Sigma|(\Sigma^T)^{-1}$$

§ 3.10 类条件概率密度的参数估计

类条件概率密度函数的贝叶斯估计

适用范围：概率密度函数形式已知但参数未知且不固定。

方法：用类似于最小风险判决的方法来估计未知随机参数。

假设：$\boldsymbol{\theta}$ 取值的参数空间Θ是一个连续空间

另用 $\lambda(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta})$ 标记真实参数为 $\boldsymbol{\theta}$, 得到的估计量为 $\hat{\boldsymbol{\theta}}$ 时承担的风险

条件平均风险 $$R(\hat{\boldsymbol{\theta}} \mid \boldsymbol{X}) = \int_{\Theta} \lambda(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{X}) \, d\boldsymbol{\theta}$$

总的平均风险（贝叶斯风险）

$$\overline{R} = \int_{E_d} R(\hat{\boldsymbol{\theta}} \mid \boldsymbol{X}) p(\boldsymbol{X}) d\boldsymbol{X}$$

$$= \int_{E_d} p(\boldsymbol{X}) \int_{\Theta} \lambda(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{X}) d\boldsymbol{\theta} d\boldsymbol{X}$$

显然，贝叶斯风险与 $\lambda(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta})$ 的选择有关！

§ 3.10 类条件概率密度的参数估计

类条件概率密度函数的贝叶斯估计

选择二次函数为风险函数 $\lambda(\hat{\theta}|\theta)=(\theta-\hat{\theta})^2$

贝叶斯估计量的定理

如果选择风险函数 $\lambda(\hat{\theta}|\theta)=(\theta-\hat{\theta})^2$, 则

$$\hat{\theta}=E(\theta|X)=\int_{\Theta}\theta p(\theta|X)\,d\theta$$

[证明]

$$R(\hat{\theta}|X)=\int_{\Theta}\lambda(\hat{\theta}|\theta)p(\theta|X)\,d\theta=\int_{\Theta}(\theta-\hat{\theta})^2 p(\theta|X)\,d\theta$$

$$=\int_{\Theta}(\theta-E(\theta|X)+E(\theta|X)-\hat{\theta})^2 p(\theta|X)\,d\theta$$

$$=\int_{\Theta}(\theta-E(\theta|X))^2 p(\theta|X)\,d\theta+\int_{\Theta}(E(\theta|X)-\hat{\theta})^2 p(\theta|X)\,d\theta$$

$$+2\int_{\Theta}(\theta-E(\theta|X))(E(\theta|X)-\hat{\theta})p(\theta|X)\,d\theta$$

§ 3.10 类条件概率密度的参数估计

类条件概率密度函数的贝叶斯估计

[证明]（续）

$$R(\hat{\theta}\mid X)=\int_{\Theta}(\theta-E(\theta\mid X))^2 p(\theta\mid X)\,d\theta+\int_{\Theta}(E(\theta\mid X)-\hat{\theta})^2 p(\theta\mid X)\,d\theta$$

$$+2\int_{\Theta}(\theta-E(\theta\mid X))(E(\theta\mid X)-\hat{\theta})p(\theta\mid X)\,d\theta$$

$$\int_{\Theta}(\theta-E(\theta\mid X))(E(\theta\mid X)-\hat{\theta})p(\theta\mid X)\,d\theta$$

$$=(E(\theta\mid X)-\hat{\theta})\int_{\Theta}(\theta-E(\theta\mid X))p(\theta\mid X)\,d\theta$$

$$=(E(\theta\mid X)-\hat{\theta})\left[\int_{\Theta}\theta p(\theta\mid X)\,d\theta-E(\theta\mid X)\int_{\Theta}p(\theta\mid X)\,d\theta\right]$$

$$=(E(\theta\mid X)-\hat{\theta})(E(\theta\mid X)-E(\theta\mid X))=0$$

§ 3.10 类条件概率密度的参数估计

类条件概率密度函数的贝叶斯估计

[证明]（续）

$$R(\hat{\boldsymbol{\theta}} \mid \boldsymbol{X}) = \int_{\boldsymbol{\Theta}} (\boldsymbol{\theta} - E(\boldsymbol{\theta} \mid \boldsymbol{X}))^2 p(\boldsymbol{\theta} \mid \boldsymbol{X})\, d\boldsymbol{\theta} + \int_{\boldsymbol{\Theta}} (E(\boldsymbol{\theta} \mid \boldsymbol{X}) - \hat{\boldsymbol{\theta}})^2 p(\boldsymbol{\theta} \mid \boldsymbol{X})\, d\boldsymbol{\theta}$$

易见：第一项非负并且其取值与 $\hat{\boldsymbol{\theta}}$ 无关；

第二项也非负，但其取值与 $\hat{\boldsymbol{\theta}}$ 有关。

故：欲使贝叶斯风险最小化，需选择估计量使第二项最小化。

即 $\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta} \mid \boldsymbol{X}) = \int_{\boldsymbol{\Theta}} \boldsymbol{\theta} p(\boldsymbol{\theta} \mid \boldsymbol{X})\, d\boldsymbol{\theta}$ 证毕

结论：以上定理给出了估计待求参数的方法。

$$p(\boldsymbol{\theta} \mid \boldsymbol{X}) \Rightarrow p(\boldsymbol{\theta} \mid \mathscr{X}) = \frac{p(\mathscr{X} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathscr{X})} = \frac{p(\mathscr{X} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} p(\mathscr{X} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

§ 3.10 类条件概率密度的参数估计

类条件概率密度函数的贝叶斯学习

不经过参数估计，直接根据样本集合推断总体的概率分布。

采取分而治之的策略，只需要做：

$$\mathscr{X}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_n\} \Rightarrow p(\boldsymbol{X}\mid\mathscr{X})$$

求解思路 $\Rightarrow$ $p(\boldsymbol{\theta}\mid\mathscr{X})$（$\boldsymbol{\theta}$的后验概率密度）

$p(\boldsymbol{X},\boldsymbol{\theta})$ 联合概率密度

收敛?

$$p(\boldsymbol{X})=\int_{\Theta} p(\boldsymbol{X},\boldsymbol{\theta})\,d\boldsymbol{\theta}$$

$$p(\boldsymbol{X}\mid\mathscr{X})=\int_{\Theta} p(\boldsymbol{X},\boldsymbol{\theta}\mid\mathscr{X})\,d\boldsymbol{\theta}$$

$$p(\boldsymbol{X},\boldsymbol{\theta})=p(\boldsymbol{X}\mid\boldsymbol{\theta})\,p(\boldsymbol{\theta}) \Rightarrow p(\boldsymbol{X},\boldsymbol{\theta}\mid\mathscr{X})=p(\boldsymbol{X}\mid\boldsymbol{\theta},\mathscr{X})\,p(\boldsymbol{\theta}\mid\mathscr{X})$$

$$\boxed{p(\boldsymbol{X}\mid\mathscr{X})}=\int_{\Theta} p(\boldsymbol{X}\mid\boldsymbol{\theta},\mathscr{X})\,p(\boldsymbol{\theta}\mid\mathscr{X})\,d\boldsymbol{\theta}=\int_{\Theta} p(\boldsymbol{X}\mid\boldsymbol{\theta})\,\boxed{p(\boldsymbol{\theta}\mid\mathscr{X})}\,d\boldsymbol{\theta}$$

§ 3.10 类条件概率密度的参数估计

类条件概率密度函数的贝叶斯学习

引入标记 $\mathscr{X}^N = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N\}$

假定样本集合中各样本是独立抽取的

$$p(\mathscr{X}^N \mid \boldsymbol{\theta}) = p(\boldsymbol{X}_N \mid \boldsymbol{\theta})\, p(\boldsymbol{X}_{N-1} \mid \boldsymbol{\theta}) \Lambda\, p(\boldsymbol{X}_2 \mid \boldsymbol{\theta})\, p(\boldsymbol{X}_1 \mid \boldsymbol{\theta}) = p(\boldsymbol{X}_N \mid \boldsymbol{\theta})\, p(\mathscr{X}^{N-1} \mid \boldsymbol{\theta})$$

$$p(\boldsymbol{\theta} \mid \mathscr{X}^N) = \frac{p(\mathscr{X}^N \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int_{\Theta} p(\mathscr{X}^N \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \qquad \text{贝叶斯公式}$$

$$= \frac{p(\boldsymbol{X}_N \mid \boldsymbol{\theta})\, p(\mathscr{X}^{N-1} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int_{\Theta} p(\boldsymbol{X}_N \mid \boldsymbol{\theta})\, p(\mathscr{X}^{N-1} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \qquad \text{独立性}$$

$$= \frac{p(\boldsymbol{X}_N \mid \boldsymbol{\theta})\, p(\mathscr{X}^{N-1})\, p(\boldsymbol{\theta} \mid \mathscr{X}^{N-1})}{\int_{\Theta} p(\boldsymbol{X}_N \mid \boldsymbol{\theta})\, p(\mathscr{X}^{N-1})\, p(\boldsymbol{\theta} \mid \mathscr{X}^{N-1})\, d\boldsymbol{\theta}} \qquad \text{贝叶斯公式}$$

$$= \frac{p(\boldsymbol{X}_N \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathscr{X}^{N-1})}{\int_{\Theta} p(\boldsymbol{X}_N \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathscr{X}^{N-1})\, d\boldsymbol{\theta}} \qquad \text{约去 } p(\mathscr{X}^{N-1})$$

§ 3.10 类条件概率密度的参数估计

类条件概率密度函数的贝叶斯学习

实现参数θ在线学习的递推公式

$$p(\boldsymbol{\theta} \mid \mathscr{X}^N) = \frac{p(\boldsymbol{X}_N \mid \boldsymbol{\theta})\ p(\boldsymbol{\theta} \mid \mathscr{X}^{N-1})}{\int_{\Theta} p(\boldsymbol{X}_N \mid \boldsymbol{\theta})\ p(\boldsymbol{\theta} \mid \mathscr{X}^{N-1})\ d\boldsymbol{\theta}}$$

$N=1$时 $\quad p(\boldsymbol{\theta} \mid \mathscr{X}^0) = p(\boldsymbol{\theta})$

$N=2$时 $\quad p(\boldsymbol{\theta} \mid \mathscr{X}^1) = \dfrac{p(\boldsymbol{X}_1 \mid \boldsymbol{\theta})\ p(\boldsymbol{\theta} \mid \mathscr{X}^0)}{\int_{\Theta} p(\boldsymbol{X}_1 \mid \boldsymbol{\theta})\ p(\boldsymbol{\theta} \mid \mathscr{X}^0)\ d\boldsymbol{\theta}}$

....

$N=N$时 $\quad p(\boldsymbol{\theta} \mid \mathscr{X}^N) = \dfrac{p(\boldsymbol{X}_N \mid \boldsymbol{\theta})\ p(\boldsymbol{\theta} \mid \mathscr{X}^{N-1})}{\int_{\Theta} p(\boldsymbol{X}_N \mid \boldsymbol{\theta})\ p(\boldsymbol{\theta} \mid \mathscr{X}^{N-1})\ d\boldsymbol{\theta}}$

....

$$p(\boldsymbol{\theta}),\ p(\boldsymbol{\theta} \mid \mathscr{X}^1),\ p(\boldsymbol{\theta} \mid \mathscr{X}^2), \Lambda\ p(\boldsymbol{\theta} \mid \mathscr{X}^{N-1}),\ p(\boldsymbol{\theta} \mid \mathscr{X}^N), \Lambda$$

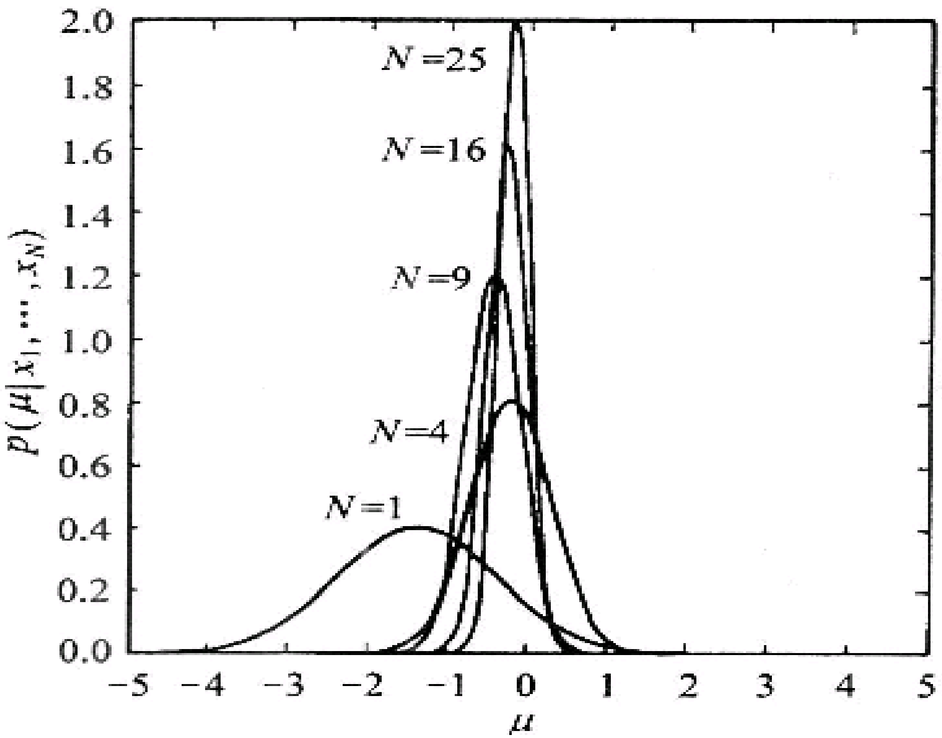
§ 3.10 类条件概率密度的参数估计

类条件概率密度函数的贝叶斯学习

$$p(\boldsymbol{\theta}), p(\boldsymbol{\theta} \mid \mathscr{X}^{1}), p(\boldsymbol{\theta} \mid \mathscr{X}^{2}), \Lambda\ \ p(\boldsymbol{\theta} \mid \mathscr{X}^{N\text{-}1}), p(\boldsymbol{\theta} \mid \mathscr{X}^{N}), \Lambda$$

随着N值的增加，相应后验概率密度一般会变得越来越尖锐。

若上述概率密度函数序列在 $N \to \infty$时收敛于以真实参数θ为中心的δ函数，则称相应的学习过程为贝叶斯学习。



$$\lim_{N \to \infty} p(\boldsymbol{X} \mid \mathscr{X}^{N})$$

$$= p(\boldsymbol{X} \mid \mathscr{X}^{N \to \infty})$$

$$= p(\boldsymbol{X} \mid \hat{\boldsymbol{\theta}} = \boldsymbol{\theta})$$

$$= p(\boldsymbol{X})$$

§ 3.10 类条件概率密度的参数估计

非监督情况下的参数估计

条件：类别数已知（S），训练样本的标签未知。

输出：分类别的估计参数。

求解方法：

分而治之，各个击破？ → 不分彼此，混合求治

$\mathscr{X}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_N\}$ 混合样本集合

混合概率密度

$$p(\boldsymbol{X})=\sum_{i=1}^{S}\underbrace{p(\boldsymbol{X}\mid\omega_i,\boldsymbol{\theta}_i)}_{\text{分量概率密度}}P(\omega_i)$$

混合参数：$P(\omega_i)$ → 数值已知

分量概率密度：$p(\boldsymbol{X}\mid\omega_i,\boldsymbol{\theta}_i)$ → 函数形式已知

§ 3.10 类条件概率密度的参数估计

非监督情况下的参数估计

$$\begin{cases}\text{最大似然估计}\\ \text{贝叶斯估计}\end{cases}$$

混合概率密度的最大似然估计

$$p(\mathscr{X}) = p(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n) = \prod_{k=1}^{N} p(\boldsymbol{X}_k)$$

$$\boldsymbol{\mathcal{L}}(\boldsymbol{\theta}) = \log p(\mathscr{X}) = \sum_{k=1}^{n} \log p(\boldsymbol{X}_k) = \sum_{k=1}^{N} \log\left(\sum_{i=1}^{S} p(\boldsymbol{X}_k \mid \omega_i, \boldsymbol{\theta}_i) P(\omega_i)\right)$$

求混合概率密度的最大似然估计

$$\nabla_{\theta_i} \boldsymbol{\mathcal{L}}(\boldsymbol{\theta}) = 0$$

§ 3.10 类条件概率密度的参数估计

非监督情况下的参数估计

$$\boldsymbol{\mathcal{L}}(\boldsymbol{\theta}) = \log p(\mathscr{X}) = \sum_{k=1}^{N} \log p(\boldsymbol{X}_k) = \sum_{k=1}^{N} \log\left(\sum_{i=1}^{S} p(\boldsymbol{X}_k \mid \omega_i, \boldsymbol{\theta}_i) P(\omega_i)\right)$$

$$\nabla_{\boldsymbol{\theta}_i} \boldsymbol{\mathcal{L}}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}_i} \sum_{k=1}^{N} \log p(\boldsymbol{X}_k) = \sum_{k=1}^{N} \frac{1}{p(\boldsymbol{X}_k)} \nabla_{\boldsymbol{\theta}_i} p(\boldsymbol{X}_k)$$

$$= \sum_{k=1}^{N} \frac{1}{p(\boldsymbol{X}_k)} \nabla_{\boldsymbol{\theta}_i} \left(\sum_{j=1}^{S} p(\boldsymbol{X}_k \mid \omega_j, \boldsymbol{\theta}_j) P(\omega_j)\right)$$

$$\because \quad \nabla_{\boldsymbol{\theta}_i} \log p(\boldsymbol{X}_k \mid \omega_i, \boldsymbol{\theta}_i) = \frac{1}{p(\boldsymbol{X}_k \mid \omega_i, \boldsymbol{\theta}_i)} \nabla_{\boldsymbol{\theta}_i} p(\boldsymbol{X}_k \mid \omega_i, \boldsymbol{\theta}_i)$$

$$\therefore \quad \nabla_{\boldsymbol{\theta}_i} p(\boldsymbol{X}_k \mid \omega_i, \boldsymbol{\theta}_i) = p(\boldsymbol{X}_k \mid \omega_i, \boldsymbol{\theta}_i) \nabla_{\boldsymbol{\theta}_i} \log p(\boldsymbol{X}_k \mid \omega_i, \boldsymbol{\theta}_i)$$

$$\therefore \quad \nabla_{\boldsymbol{\theta}_i} \boldsymbol{\mathcal{L}}(\boldsymbol{\theta}) = \sum_{k=1}^{N} \frac{p(\boldsymbol{X}_k \mid \omega_i, \boldsymbol{\theta}_i) P(\omega_i)}{p(\boldsymbol{X}_k)} \nabla_{\boldsymbol{\theta}_i} \log p(\boldsymbol{X}_k \mid \omega_i, \boldsymbol{\theta}_i)$$

$$= \sum_{k=1}^{N} P(\omega_i \mid \boldsymbol{X}_k, \boldsymbol{\theta}_i) \nabla_{\boldsymbol{\theta}_i} \log p(\boldsymbol{X}_k \mid \omega_i, \boldsymbol{\theta}_i)$$

§ 3.10 类条件概率密度的参数估计

非监督情况下的参数估计

1. $P(\omega_i) \geq 0, i = 1,2,\ldots S$ 已知

$$\sum_{k=1}^{N} P(\omega_i \mid \boldsymbol{X}_k, \hat{\boldsymbol{\theta}}_i) \nabla_{\boldsymbol{\theta}_i} \log p(\boldsymbol{X}_k \mid \omega_i, \hat{\boldsymbol{\theta}}_i) = 0, \quad i = 1,2,\ldots S$$

解之，可得到参数集的最大似然估计：$\boldsymbol{\theta}_i, i = 1,2,\ldots S$

$$p(\boldsymbol{X}) = \sum_{i=1}^{S} p(\boldsymbol{X} \mid \omega_i, \boldsymbol{\theta}_i) P(\omega_i)$$

2. $P(\omega_i) \geq 0, i = 1,2,\ldots S$ 未知

$$\begin{cases} P(\omega_i) \geq 0, i = 1,2,\ldots S \\ \sum_{i=1}^{S} P(\omega_i) = 1 \end{cases}$$

约束条件

$$J = \sum_{k=1}^{N} \log\left(\sum_{i=1}^{S} p(\boldsymbol{X}_k \mid \omega_i, \boldsymbol{\theta}_i) P(\omega_i) \right) + \lambda\left(\sum_{i=1}^{S} P(\omega_i) - 1\right)$$

目标函数　　极小化

§ 3.10 类条件概率密度的参数估计

非监督情况下的参数估计

$$J=\sum_{k=1}^{N}\log\left(\sum_{i=1}^{S}p(\boldsymbol{X}_k\mid\omega_i,\boldsymbol{\theta}_i)P(\omega_i)\right)+\lambda\left(\sum_{i=1}^{S}P(\omega_i)-1\right)$$

求条件极值（一）

$$\nabla_{P(\omega_i)}J=\frac{\partial}{\partial P(\omega_i)}\left(\sum_{k=1}^{N}\log\left(\sum_{i=1}^{S}p(\boldsymbol{X}_k\mid\omega_i,\boldsymbol{\theta}_i)P(\omega_i)\right)\right)+\lambda$$

$$=\sum_{k=1}^{N}\frac{p(\boldsymbol{X}_k\mid\omega_i,\boldsymbol{\theta}_i)}{\sum_{i=1}^{S}p(\boldsymbol{X}_k\mid\omega_i,\boldsymbol{\theta}_i)P(\omega_i)}+\lambda\quad=0$$

$$\sum_{k=1}^{N}\frac{p(\boldsymbol{X}_k\mid\omega_i,\hat{\boldsymbol{\theta}}_i)}{\sum_{i=1}^{S}p(\boldsymbol{X}_k\mid\omega_i,\hat{\boldsymbol{\theta}}_i)\hat{P}(\omega_i)}=-\lambda\qquad i=1,2,\ldots S$$

$$\sum_{k=1}^{N}\frac{p(\boldsymbol{X}_k\mid\omega_i,\hat{\boldsymbol{\theta}}_i)\hat{P}(\omega_i)}{\sum_{i=1}^{S}p(\boldsymbol{X}_k\mid\omega_i,\hat{\boldsymbol{\theta}}_i)\hat{P}(\omega_i)}=-\lambda\hat{P}(\omega_i)\qquad i=1,2,\ldots S$$

§ 3.10 类条件概率密度的参数估计

非监督情况下的参数估计

$$\sum_{k=1}^{N} \frac{p(\boldsymbol{X}_k \mid \omega_i, \hat{\boldsymbol{\theta}}_i)\hat{P}(\omega_i)}{\sum_{i=1}^{S} p(\boldsymbol{X}_k \mid \omega_i, \hat{\boldsymbol{\theta}}_i)\hat{P}(\omega_i)} = -\lambda \hat{P}(\omega_i) \qquad i = 1,2,\ldots S$$

$$\sum_{k=1}^{N} \hat{P}(\omega_i \mid \boldsymbol{X}_k, \hat{\boldsymbol{\theta}}_i) = -\lambda \hat{P}(\omega_i) \qquad i = 1,2,\ldots S$$

$$\sum_{i=1}^{S}\sum_{k=1}^{N} \hat{P}(\omega_i \mid \boldsymbol{X}_k, \hat{\boldsymbol{\theta}}_i) = -\lambda \sum_{i=1}^{S} \hat{P}(\omega_i) = -\lambda$$

$$= \sum_{k=1}^{N}\sum_{i=1}^{S} \hat{P}(\omega_i \mid \boldsymbol{X}_k, \hat{\boldsymbol{\theta}}_i) = N \Rightarrow \lambda = -N$$

$$\hat{P}(\omega_i) = \frac{1}{N}\sum_{k=1}^{N} \hat{P}(\omega_i \mid \boldsymbol{X}_k, \hat{\boldsymbol{\theta}}_i) \qquad i = 1,2,\ldots S$$

§ 3.10 类条件概率密度的参数估计

非监督情况下的参数估计

$$J=\sum_{k=1}^{N}\log\left(\sum_{i=1}^{S}p(\boldsymbol{X}_k\mid\omega_i,\boldsymbol{\theta}_i)P(\omega_i)\right)+\lambda\left(\sum_{i=1}^{S}P(\omega_i)-1\right)$$

求条件极值（二）

$$\nabla_{\boldsymbol{\theta}_i}J=0,\qquad i=1,2,\ldots S$$

$$\sum_{k=1}^{N}\hat{P}(\omega_i\mid\boldsymbol{X}_k,\hat{\boldsymbol{\theta}}_i)\nabla_{\boldsymbol{\theta}_i}\log p(\boldsymbol{X}_k\mid\omega_i,\hat{\boldsymbol{\theta}}_i)=0,\qquad i=1,2,\ldots,S$$

$$\hat{P}(\omega_i\mid\boldsymbol{X}_k,\boldsymbol{\theta}_i)=\frac{p(\boldsymbol{X}_k\mid\omega_i,\hat{\boldsymbol{\theta}}_i)\hat{P}(\omega_i)}{p(\boldsymbol{X}_k)}$$

$$=\frac{p(\boldsymbol{X}_k\mid\omega_i,\hat{\boldsymbol{\theta}}_i)\hat{P}(\omega_i)}{\sum_{j=1}^{S}p(\boldsymbol{X}_k\mid\omega_j,\hat{\boldsymbol{\theta}}_j)\hat{P}(\omega_j)},\qquad i=1,2,\ldots,S$$

§ 3.10 类条件概率密度的参数估计

多维正态分布情况下的非监督参数估计

$$\mathscr{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N\} \quad S$$

$$p(\boldsymbol{X} \mid \omega_i, \boldsymbol{\theta}_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), i = 1,2,\ldots S$$

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{k=1}^{N} \ln\left(\sum_{i=1}^{S} \frac{P(\omega_i)}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\boldsymbol{X}_k - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{X}_k - \boldsymbol{\mu}_i) \right\} \right)$$

设置目标函数

$$J = \sum_{k=1}^{N} \ln\left(\sum_{i=1}^{S} \frac{P(\omega_i)}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\boldsymbol{X}_k - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{X}_k - \boldsymbol{\mu}_i) \right\} \right) + \lambda\left(\sum_{i=1}^{S} P(\omega_i) - 1\right)$$

令 $\nabla_{\boldsymbol{\mu}_i} J = 0,\ \nabla_{\boldsymbol{\Sigma}_i} J = 0,\ \nabla_{P(\omega_i)} J = 0, \qquad i = 1,2,\ldots S$

求解联立方程，可解出待求参数应满足的关系式。

§ 3.10 类条件概率密度的参数估计

多维正态分布情况下的非监督参数估计

待求参数应满足的关系式：

$$\hat{P}(\omega_i)=\frac{1}{N}\sum_{k=1}^{N}\hat{P}(\omega_i\mid \boldsymbol{X}_k,\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\Sigma}}_i)$$

$$\hat{\boldsymbol{\mu}}_i=\frac{\sum_{k=1}^{N}\hat{P}(\omega_i\mid \boldsymbol{X}_k,\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\Sigma}}_i)\boldsymbol{X}_k}{\sum_{k=1}^{N}\hat{P}(\omega_i\mid \boldsymbol{X}_k,\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\Sigma}}_i)}\qquad \hat{\boldsymbol{\Sigma}}_i=\frac{\sum_{k=1}^{N}\hat{P}(\omega_i\mid \boldsymbol{X}_k,\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\Sigma}}_i)(\boldsymbol{X}_k-\hat{\boldsymbol{\mu}}_i)(\boldsymbol{X}_k-\hat{\boldsymbol{\mu}}_i)^T}{\sum_{k=1}^{N}\hat{P}(\omega_i\mid \boldsymbol{X}_k,\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\Sigma}}_i)}$$

$$\hat{P}(\omega_i\mid \boldsymbol{X}_k,\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\Sigma}}_i)=\frac{p(\boldsymbol{X}_k\mid\omega_i,\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\Sigma}}_i)\hat{P}(\omega_i)}{\sum_{j=1}^{S}p(\boldsymbol{X}_k\mid\omega_j,\hat{\boldsymbol{\mu}}_j,\hat{\boldsymbol{\Sigma}}_j)\hat{P}(\omega_j)}$$

$$=\frac{\left|\hat{\boldsymbol{\Sigma}}_i\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{X}_k-\hat{\boldsymbol{\mu}}_i)^T\hat{\boldsymbol{\Sigma}}_i^{-1}(\boldsymbol{X}_k-\hat{\boldsymbol{\mu}}_i)\right\}\hat{P}(\omega_i)}{\sum_{j=1}^{S}\left|\hat{\boldsymbol{\Sigma}}_i\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{X}_k-\hat{\boldsymbol{\mu}}_j)^T\hat{\boldsymbol{\Sigma}}_j^{-1}(\boldsymbol{X}_k-\hat{\boldsymbol{\mu}}_j)\right\}\hat{P}(\omega_j)}$$

§ 3.10 类条件概率密度的参数估计

多维正态分布下求解待求参数的迭代算法

Step1. 输入混合训练样本集 $\mathscr{X}$、类别数S和最大迭代次数M;

Step2. 设置程序参数：迭代计数器$m=0$, 以及

$$\boldsymbol{\mu}_i(0), \boldsymbol{\Sigma}_i(0), P(\omega_i, 0), i = 1,2,\ldots,S$$

Step3. 根据待定参数的迭代值计算后验概率

$$\hat{P}(\omega_i \mid \boldsymbol{X}_k, \hat{\boldsymbol{\mu}}_i(m), \hat{\boldsymbol{\Sigma}}_i(m))$$

$$= \frac{\left|\hat{\boldsymbol{\Sigma}}_i(m)\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{X}_k - \hat{\boldsymbol{\mu}}_i(m))^T \hat{\boldsymbol{\Sigma}}_i(m)^{-1}(\boldsymbol{X}_k - \hat{\boldsymbol{\mu}}_i(m))\right\}\hat{P}(\omega_i, m)}{\sum_{j=1}^{S}\left|\hat{\boldsymbol{\Sigma}}_i(m)\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{X}_k - \hat{\boldsymbol{\mu}}_j(m))^T \hat{\boldsymbol{\Sigma}}_j(m)^{-1}(\boldsymbol{X}_k - \hat{\boldsymbol{\mu}}_j(m))\right\}\hat{P}(\omega_j, m)}, \quad i = 1,2,\ldots,S$$

Step4. 更新先验概率

$$\hat{P}(\omega_i, m+1) = \frac{1}{N}\sum_{k=1}^{N}\hat{P}(\omega_i \mid \boldsymbol{X}_k, \hat{\boldsymbol{\mu}}_i(m), \hat{\boldsymbol{\Sigma}}_i(m)), \quad i = 1,2,\ldots,S$$

Step5. 更新均值向量 $\hat{\boldsymbol{\mu}}_i(m+1),\ i = 1,2,\ldots,S$。

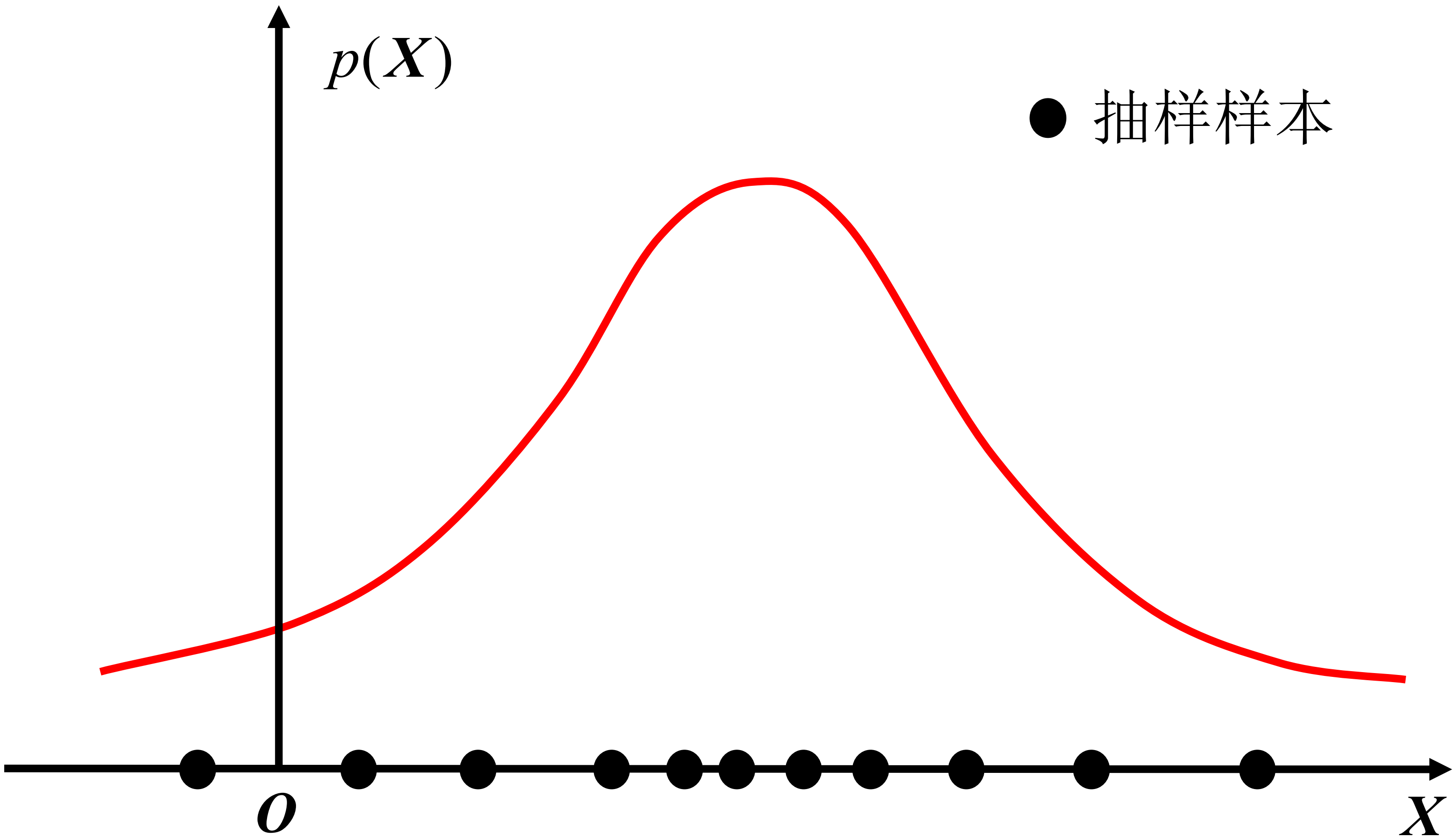
Step6. 更新协方差矩阵 $\hat{\boldsymbol{\Sigma}}_i(m+1),\ i = 1,2,\ldots,S$ 。

Step7. 若 $m \le M$, 则 $m = m+1$, 转**Step3**;否则, 算法结束。

中国科学技术大学 University of Science and Technology of China

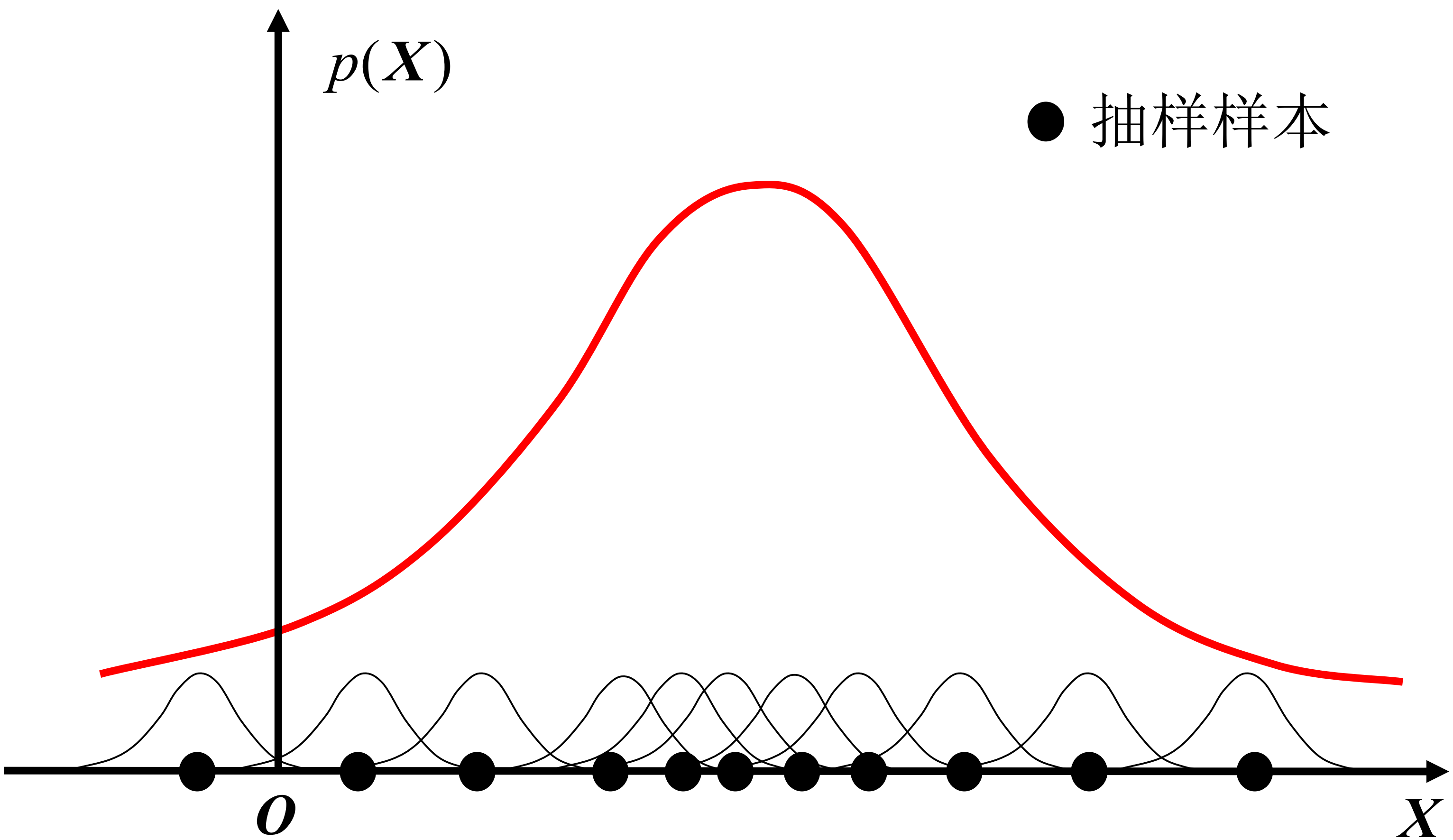
§ 3.11 类条件概率密度的非参数估计

非参数估计的基本概念和方法



§ 3.11 类条件概率密度的非参数估计

非参数估计的基本概念和方法

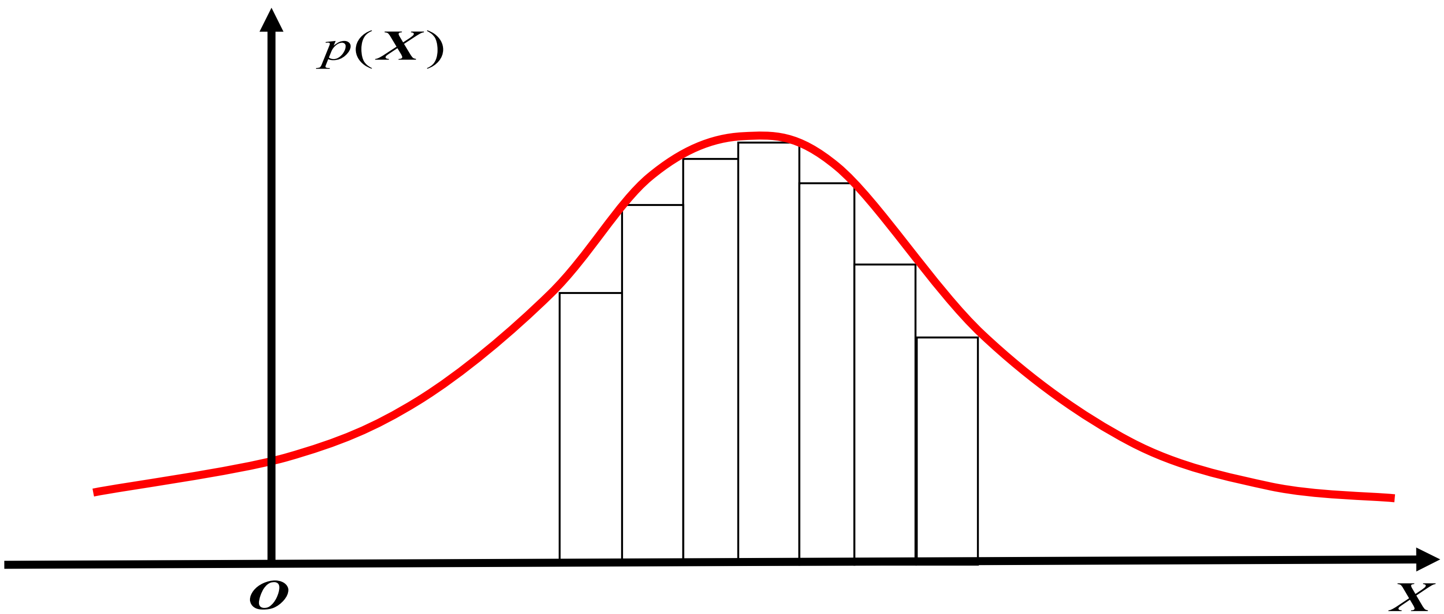


§ 3.11 类条件概率密度的非参数估计

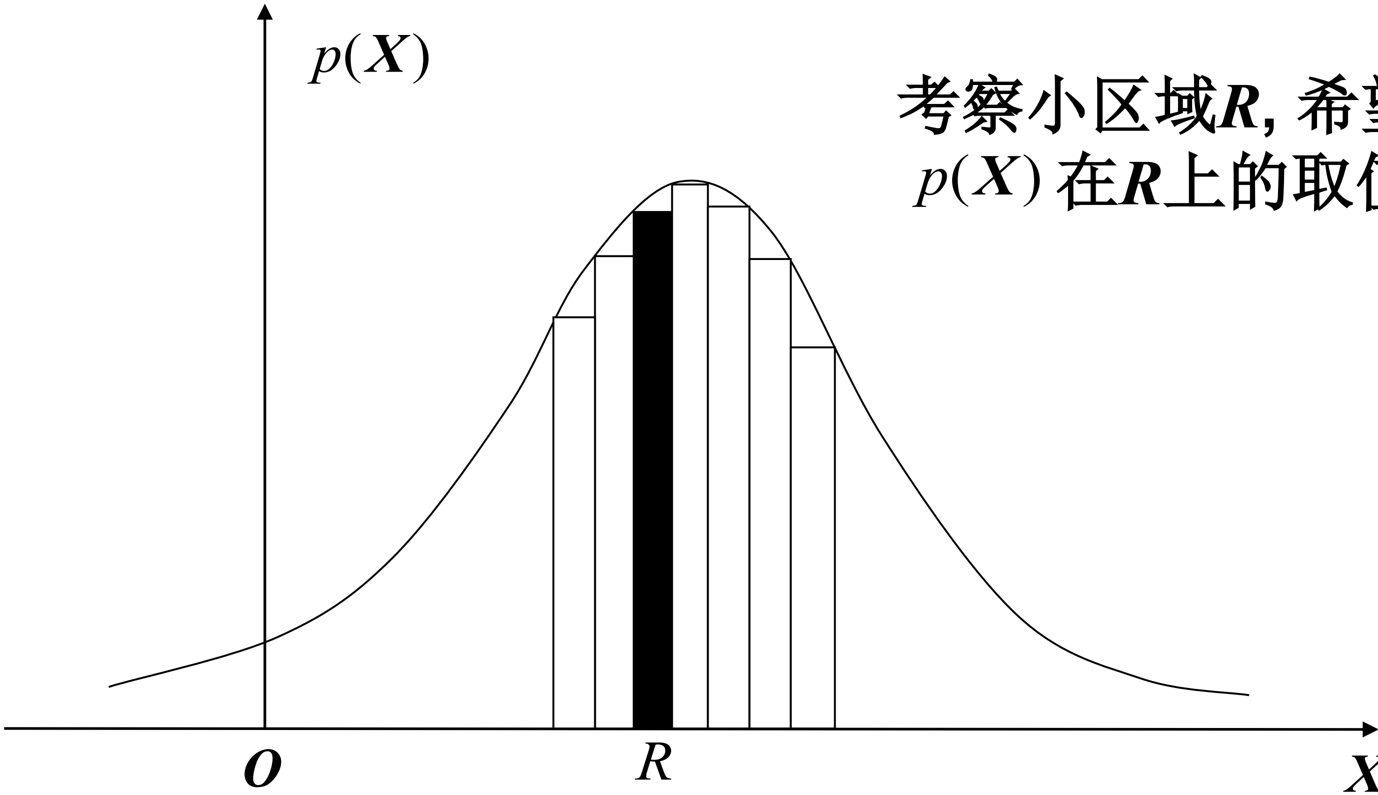
非参数估计的目标

- 当给定的样本集是样本类别属性已知的、经过预分的样本集时，能够得到相应类条件概率密度的估计；
- 当给定的样本集是样本类别属性未知的混合样本集时，能够给出相应混合概率密度的估计。

如何完成非参数估计？



§ 3.11 类条件概率密度的非参数估计



考察小区域R, 希望确定$p(X)$在R上的取值 。

样本落入区域R中的概率

R所包围的体积

$$P=\int_R p(X)dx \cong p(X)\int_R dx = p(X)V \;\Rightarrow\; \hat{p}(X)=\frac{\boxed{P}}{V}$$

R中任意一点

§ 3.11 类条件概率密度的非参数估计

如何确定P呢？ $\mathscr{X}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_n\}$

- **重复独立试验**

  将试验 E 重复进行 n 次，若各次试验的结果互不影响，即每次试验结果出现的概率都不依赖于其它各次试验的结果，则称这 n 次试验是相互独立的，或称这 n 次试验为 n 次重复独立试验。

- **n 重伯努利试验**

  若试验 E 只有两个可能结果：A或者B（非A），则称E为伯努利试验。类似地，若将E独立地重复n次，则称这 n 次试验为 n 重伯努利试验。

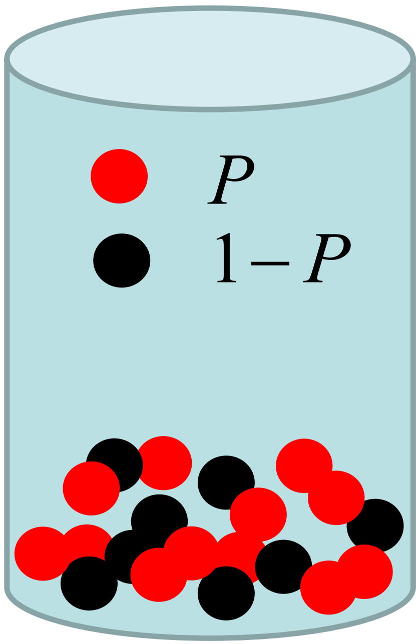
若$P(A)=\mathrm{P}\,(0<P<1)$，则$P(B)=1-P$。

§ 3.11 类条件概率密度的非参数估计

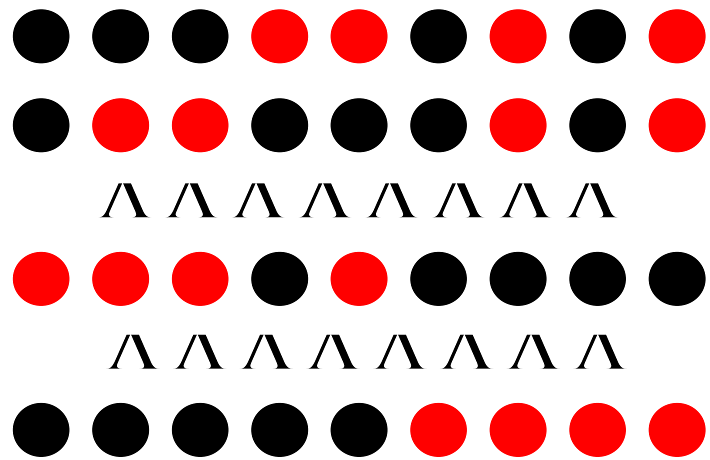
■ 二项概率公式

定义：事件A为摸到红球，$P(A)=P$。

问题：n次试验中摸到k次红球的概率P_k？



n次试验中摸到k次红球的方式如下：



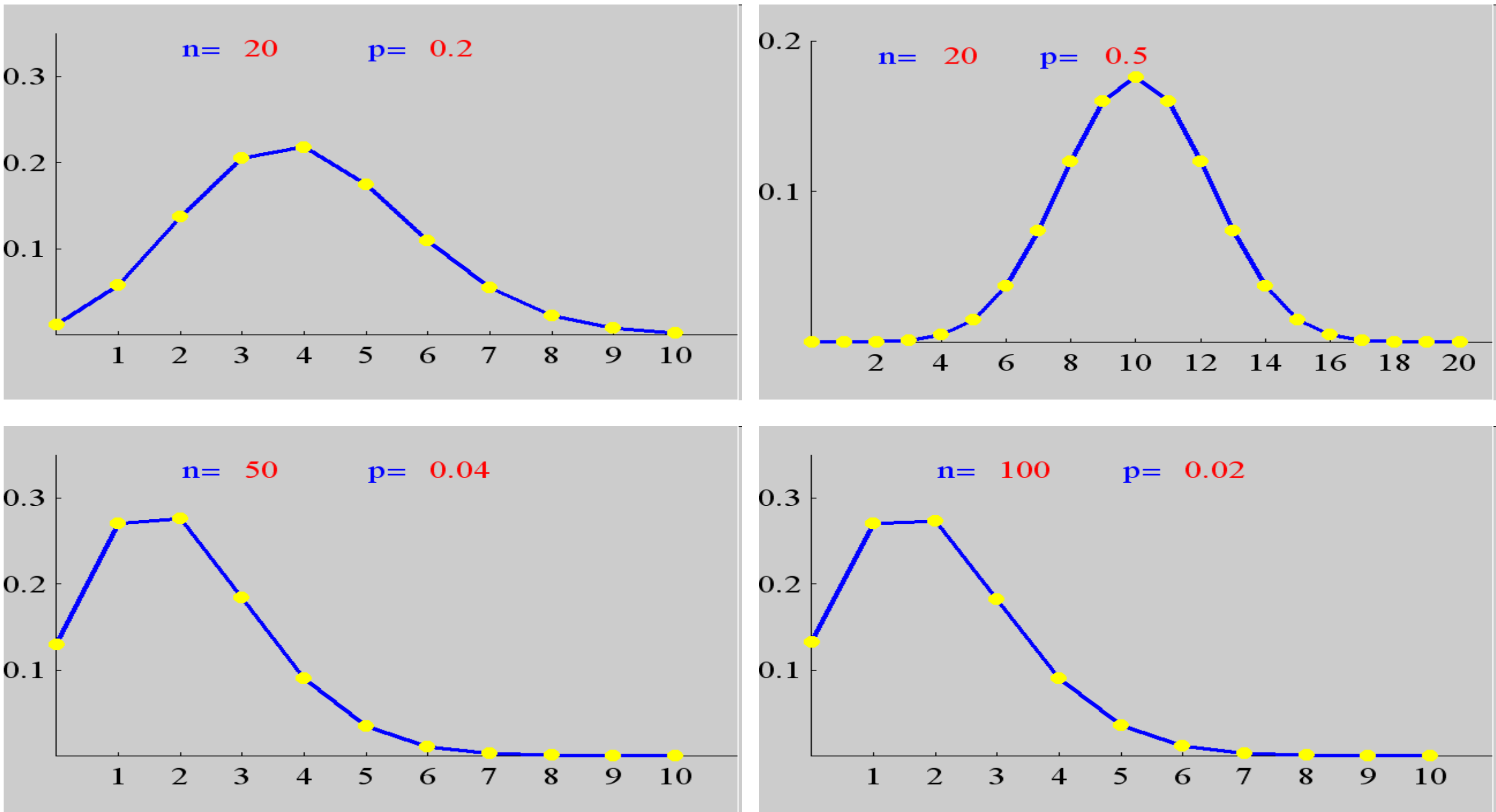
共有$\binom{n}{k}$种！

$$P_k=\binom{n}{k}P^k(1-P)^{n-k}$$

§ 3.11 类条件概率密度的非参数估计

二项分布的图形：示例

陡峭的峰



§ 3.11 类条件概率密度的非参数估计

$$E(k)=\sum_{k=1}^{n}kP_k=\sum_{k=1}^{n}k\binom{n}{k}P^k(1-P)^{n-k}=nP$$

[证明]

$$M_k(t)=E(e^{tk})=\sum_{k=0}^{n}e^{tk}P_k=\sum_{k=0}^{N}e^{tk}\binom{n}{k}P^k(1-P)^{n-k}$$ 矩母函数

$$=\sum_{k=0}^{n}\binom{n}{k}(Pe^t)^k(1-P)^{n-k}=(Pe^t+(1-P))^n$$

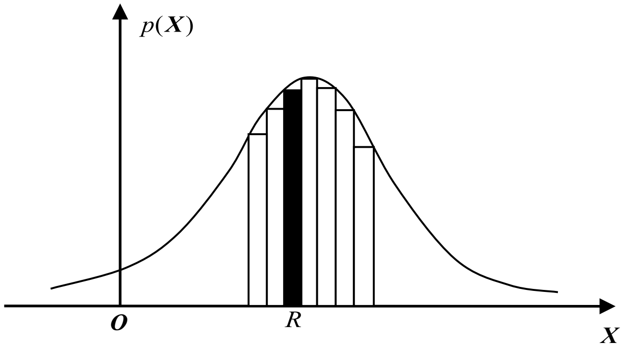
$$M_k^{'}(t)=\frac{\partial}{\partial t}\left(\sum_{k=0}^{n}e^{tk}\binom{n}{k}P^k(1-P)^{n-k}\right)=\sum_{k=0}^{n}ke^{tk}\binom{n}{k}P^k(1-P)^{n-k}$$

$$M_k^{'}(0)=\sum_{k=0}^{n}k\binom{n}{k}P^k(1-P)^{n-k}=\sum_{k=1}^{n}k\binom{n}{k}P^k(1-P)^{n-k}=E(k)$$

$$M_k^{'}(t)=\frac{\partial}{\partial t}(Pe^t+(1-P))^n=nPe^t(Pe^t+(1-P))^{n-1}$$ $$M_k^{'}(0)=nP$$

$$E(k)=nP$$ 得证。

§ 3.11 类条件概率密度的非参数估计



$$\hat{p}(\boldsymbol{X}) = \frac{P}{V}$$

P：样本落入区域***R***中的概率

V：***R***包围的体积

样本落入区域***R*** ➡ 摸到红球

样本落在区域***R***外 ➡ 摸到黑球

➡ $P_k = \binom{n}{k} P^k (1-P)^{n-k}$

$E(k) = nP$ ➡ $\hat{P} = \frac{E(k)}{n} \cong \frac{k}{n}$

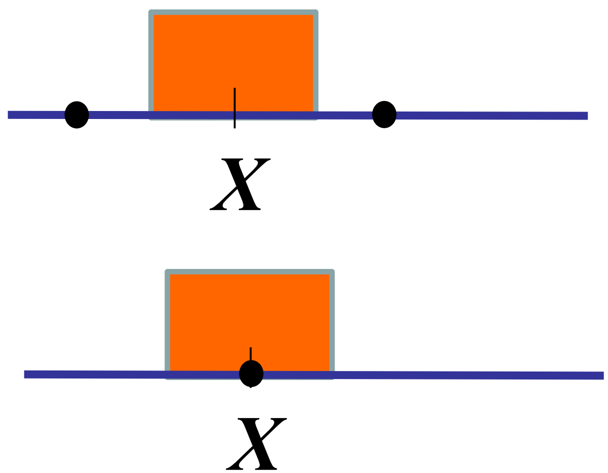
➡ $$\boxed{\hat{p}(\boldsymbol{X}) = \frac{P}{V} = \frac{k/n}{V}}$$

§ 3.11 类条件概率密度的非参数估计

分析

$$\hat{p}(\boldsymbol{X})=\frac{P}{V}=\frac{k/n}{V}$$

趋于0



$\therefore$区域***R***不允许取得太小。

- $p(\boldsymbol{X})$ 的真值和估计值之间会存在一定的误差
- 如何使误差尽可能小？

§ 3.11 类条件概率密度的非参数估计

$$\hat{p}(\boldsymbol{X}) = \frac{P}{V} = \frac{k/n}{V}$$

$k_1, k_2, \Lambda\ k_n, \Lambda$

构造一个包含*X*在内的区域序列 $R_1, R_2, \Lambda\ R_n, \Lambda$

估计值 $\hat{p}_n(\boldsymbol{X}) = \dfrac{P}{V} = \dfrac{k_n/n}{V_n}$

估计值收敛于真实概率密度的条件

1. $\lim\limits_{n\to\infty} V_n = 0$　　$\hat{p}_n(\boldsymbol{X}) \Rightarrow p(\boldsymbol{X})$

2. $\lim\limits_{n\to\infty} k_n = \infty$　　$k_n/n \Rightarrow P$

3. $\lim\limits_{n\to\infty} \dfrac{k_n}{n} = 0$　　$\hat{p}_n(\boldsymbol{X})$ 收敛的必要条件

§ 3.11 类条件概率密度的非参数估计

满足概率密度函数收敛条件的区域序列选择方法

1. $R_1(V_1)$ ➡ 选择 R_n, 使

$$V_n = \frac{V_1}{\sqrt{n}}$$

➡ **Parzen窗估计法**

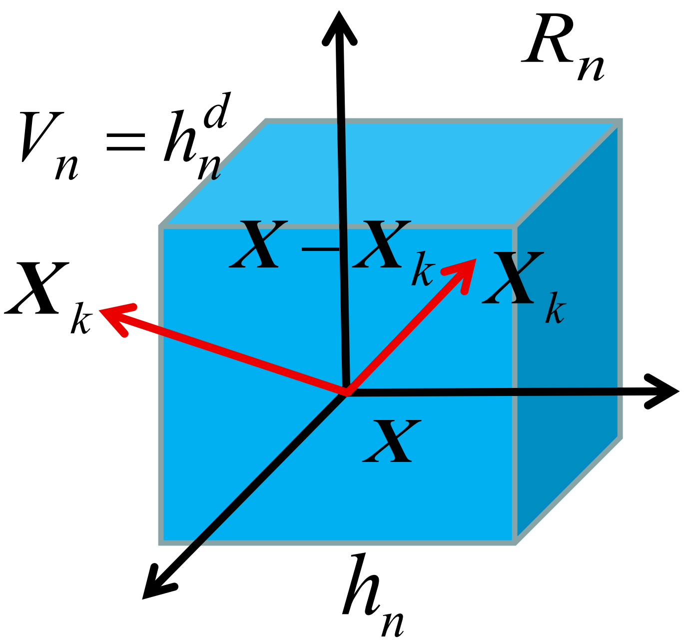
2. $R_1(k_1)$ ➡ 选择 k_n , 使

$$k_n = k_1\sqrt{n}$$

➡ ***kn*-近邻估计法**

§ 3.11 类条件概率密度的非参数估计

Parzen窗估计法



$\boldsymbol{X}_k$ 落在区域 R_n 中的条件

$\boldsymbol{X} - \boldsymbol{X}_k$ 每个分量的长度 $< h_n / 2$

计数落入区域 R_n 中的样本个数

k_n

§ 3.11 类条件概率密度的非参数估计

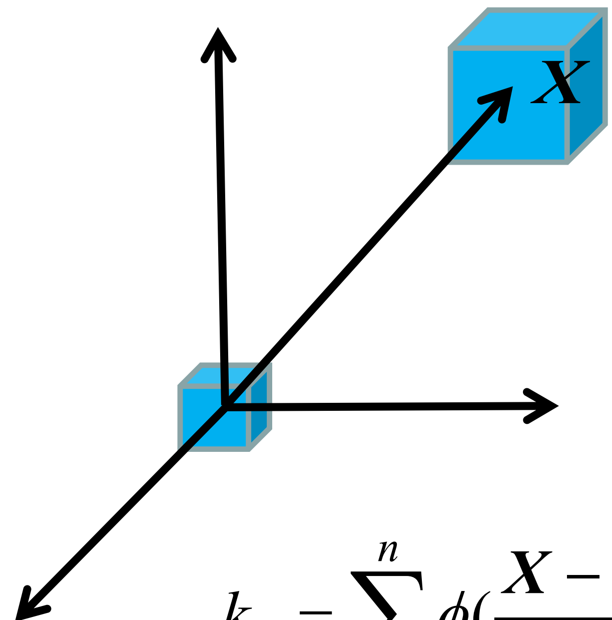
Parzen窗估计法

超方窗函数

$$\phi(\boldsymbol{u}) = \begin{cases} 1 & \left|\mathrm{u_j}\right| \le 1/2,\ \mathrm{j} = 1,2,\ldots \mathrm{d} \\ 0 & \text{否则} \end{cases}$$

$$\phi(\boldsymbol{u}) \ge 0$$

$$\int_{\boldsymbol{E}_d} \phi(\boldsymbol{u}) d\boldsymbol{u} = 1$$



$$k_n = \sum_{k=1}^{n} \phi\left(\frac{\boldsymbol{X} - \boldsymbol{X}_k}{h_n}\right)$$

$$\boldsymbol{u} = \frac{\boldsymbol{X} - \boldsymbol{X}_k}{h_n}$$

$$\Longrightarrow \quad \phi\left(\frac{\boldsymbol{X} - \boldsymbol{X}_k}{h_n}\right) = \begin{cases} 1 & \left|(\boldsymbol{X} - \boldsymbol{X}_k)_{\mathrm{j}}\right| \le \dfrac{h_n}{2},\ \mathrm{j} = 1,2,\ldots \mathrm{d} \\ 0 & \text{否则} \end{cases}$$

§ 3.11 类条件概率密度的非参数估计

Parzen窗估计法

$$\hat{p}_n(\boldsymbol{X}) = \frac{k_n / n}{V_n} = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{V_n}\phi\left(\frac{\boldsymbol{X}-\boldsymbol{X}_k}{h_n}\right)$$

$$\begin{cases} \hat{p}_n(\boldsymbol{X}) \geq 0 \\ \int_{\boldsymbol{E}_d} \hat{p}_n(\boldsymbol{X}) d\boldsymbol{X} = 1 \end{cases}$$

$$\begin{aligned}
\int_{\boldsymbol{E}_d} \hat{p}_n(\boldsymbol{X}) d\boldsymbol{X} &= \int_{\boldsymbol{E}_d} \frac{1}{n}\sum_{k=1}^{n}\frac{1}{V_n}\phi\left(\frac{\boldsymbol{X}-\boldsymbol{X}_k}{h_n}\right) d\boldsymbol{X} = \frac{1}{n}\sum_{k=1}^{n}\int_{\boldsymbol{E}_d}\frac{1}{V_n}\phi\left(\frac{\boldsymbol{X}-\boldsymbol{X}_k}{h_n}\right) d\boldsymbol{X} \\
&= \frac{1}{n}\sum_{k=1}^{n}\int_{\boldsymbol{E}_d}\frac{1}{h_n^d}\phi\left(\frac{\boldsymbol{X}-\boldsymbol{X}_k}{h_n}\right) d\boldsymbol{X} = \frac{1}{n}\sum_{k=1}^{n}\int_{\boldsymbol{E}_d}\phi\left(\frac{\boldsymbol{X}-\boldsymbol{X}_k}{h_n}\right) d\left(\frac{\boldsymbol{X}-\boldsymbol{X}_k}{h_n}\right) \\
&= \frac{1}{n}\sum_{k=1}^{n}\int_{\boldsymbol{E}_d}\phi(\boldsymbol{u}) d\boldsymbol{u} = 1
\end{aligned}$$

§ 3.11 类条件概率密度的非参数估计

Parzen窗估计法

$$\delta_n(\boldsymbol{X}) = \frac{1}{V_n}\phi\left(\frac{\boldsymbol{X}}{h_n}\right) = \frac{1}{h_n^d}\phi\left(\frac{\boldsymbol{X}}{h_n}\right)$$

$$\hat{p}_n(\boldsymbol{X}) = \frac{k_n/n}{V_n} = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{V_n}\phi\left(\frac{\boldsymbol{X}-\boldsymbol{X}_k}{h_n}\right) = \frac{1}{n}\sum_{k=1}^{n}\delta_n(\boldsymbol{X}-\boldsymbol{X}_k)$$

h_n 影响估计概率密度函数的幅度和宽度

$h_n \nearrow$ 拓展很宽、幅度很小的慢变函数的叠加 平滑估计

$h_n \searrow$ 拓展很窄、幅度很大的快变函数的叠加 空间分辨率高

$$\begin{cases} \sup\limits_{\boldsymbol{u}} \phi(\boldsymbol{u}) < \infty \\ \lim\limits_{\|\boldsymbol{u}\|\to\infty} \phi(\boldsymbol{u})\prod\limits_{j=1}^{d} u_j = 0 \\ \lim\limits_{n\to\infty} V_n = 0 \\ \lim\limits_{n\to\infty} nV_n = \infty \end{cases} \Rightarrow$$

在均方意义下, 有

$$\lim_{n\to\infty} \hat{p}_n^*(\boldsymbol{X}) = p(\boldsymbol{X}) \quad \text{均值}$$

$$\lim_{n\to\infty} \sigma_n^2(\boldsymbol{X}) = 0 \quad \text{方差}$$

§ 小结

最小错误概率判决

最小风险判决

最大似然判决

Neyman-Peason判决

最小最大判决

概率密度函数的估计

- 参数方法
- 非参数方法

若干图片材料取自网络，特此致谢。

谢谢聆听！

中国科学技术大学