第三章　统计模式识别中的概率分类法

中科大 自动化系　郑志刚

2018.10

- 3.1 引言
- 3.2 最小错误率判决规则
- 3.3 最大似然比判决规则
- 3.4 最小风险判决规则
- 3.5 最小最大判决规则
- 3.6 Neyman-Pearsen判决规则
- 3.7 分类器设计
- 3.8 正态分布时的统计决策
- 3.9 参数估计与非参数估计

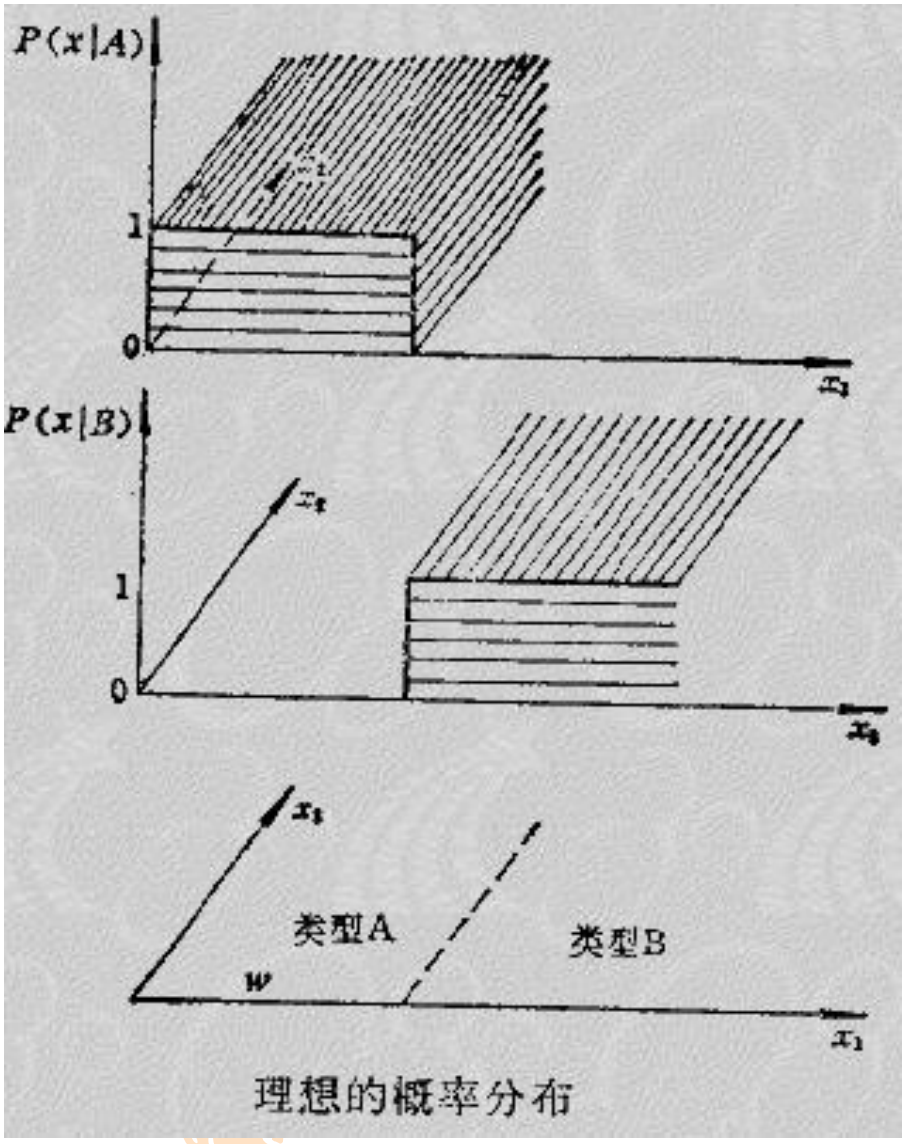
3.1 引言

随机模式：在可以觉察到的客观世界中，存在着大量的物体和事件，他们在基本条件不变时，具有某种不确定性，每一次观测的结果没有重复性，这种模式就是随机模式。

虽然随机模式样本测量值具有不确定性，但同类抽样实验的大量样本的观测值具有某种统计特性，这个统计特性是建立各种分类方法的基本依据。

先看一下确定性模式判决函数的问题。

如下图所示：



理想的概率分布

通过判决函数，特征空间被区分界面划分成两种类型的区域A和B。由于模式样本的观测值是确定性的，经常被正确分配到类型区域A、B之中。假如我们用概率的形式来表达，就是：在类型A的条件下观测模式样本x，则x位于区域A的概率为1，而位于区域B的概率为0。同样，在类型B的条件下观测模式样本x，情况正好相反，x位于区域A的概率为0，而位于区域B的概率为1。这实际上是将概率的方法引入到确定模式，对于大多数实际情况，这是非常理想的概率分布。

- 许多实际情况，即使在类型A的条件下，模式样本x位于区域A的概率也往往小于1，而位于区域B的概率也不为0。对于类型B的条件也一样。这种交错分布的样本使分类发生错误，是模式随机性的一种表现。此时，分类方法就从确定性模式转到随机模式。
- “如何使分类错误率尽可能小，是研究各种分类方法的中心议题。”
- Bayes决策理论是随机模式分类方法最重要的基础。下面是几个重要的概念：

1. 先验概率

先验概率是预先已知的或者可以估计的模式识别系统位于某种类型的概率。

若仍然用两个类型 A 和 B 为例，可用 $P(A)$ 和 $P(B)$ 表示各自的先验概率，此时满足 $P(A)+P(B)=1$ 。

推广到一般的 c 类问题中，用 $w_1, w_2, \cdots\cdots w_c$ 表示类型，则各自的先验概率用 $P(w_1), P(w_2), \cdots\cdots, P(w_c)$ 表示，且满足：

$$P(w_1)+P(w_2)+\cdots\cdots+P(w_c)=1$$

其实，在处理实际问题时，有时不得不以先验概率的大小作为判决的依据。如：有一批木材，其中桦木占 70%，松木占 30%，A——桦木，B——松木，则 $P(A)=0.7$ ， $P(B)=0.3$ ，如果从中任取一块木材，而又要用先验概率作出判决，那就判为桦木。

先验概率不能作为判决的唯一依据，但当先验概率相当大时，它也能成为主要因素。

2．类（条件）概率密度

它是系统位于某种类型条件下，模式样本 x 出现的概率密度分布函数，常用 $\rho(x \mid A), \rho(x \mid B)$，以及 $\rho(x \mid w_i)(i \in 1,2,\cdots\cdots,c)$ 来表示。

类概率密度在分类方法中起至关重要的作用，它的函数形式及主要参数或者是已知的，或者是可通过大量抽样实验估计出来。

3．后验概率

它是系统在某个具体的模式样本 x 条件下，位于某种类型的概率，常以 $P(A \mid x), P(B \mid x)$，以及 $P(w_i \mid x)(i \in 1,2,\cdots\cdots,c)$ 表示。

后验概率可以根据贝叶斯公式计算出来，可直接用作分类判决的依据。

Bayes法则—最大后验概率准则

对于两类ω_1， ω_2问题，直观地，可以根据后验概率做判决：

$$若\quad p(\omega_1|\vec{x}) > p(\omega_2|\vec{x}) \quad 则\quad \vec{x} \in \omega_1$$

$$若\quad p(\omega_1|\vec{x}) < p(\omega_2|\vec{x}) \quad 则\quad \vec{x} \in \omega_2$$

根据Bayes公式，后验概率$p(\omega_i / \vec{x})$可由类ω_i的先验概率$P(\omega_i)$和条件概率密度$p(\vec{x} / \omega_i)$来表示，即

$$p(\omega_i | \vec{x}) = \frac{p(\vec{x} | \omega_i)P(\omega_i)}{p(\vec{x})} = \frac{p(\vec{x} | \omega_i)P(\omega_i)}{\sum_{i=1}^{2} p(\vec{x} | \omega_i)P(\omega_i)}$$

式中，$p(x|\omega_i)$又称似然函数(likelihood function of class ω_i)，可由已知样本求得。

将$P(\omega_i|x)$代入判别式，判别规则可表示为

$$若\quad p(\vec{x}|\omega_1)P(\omega_1) > p(\vec{x}|\omega_2)P(\omega_2) \quad 则\quad \vec{x}\in\omega_1$$

$$若\quad p(\vec{x}|\omega_1)P(\omega_1) < p(\vec{x}|\omega_2)P(\omega_2) \quad 则\quad \vec{x}\in\omega_2$$

或改写为

$$l_{12}=\frac{p(\vec{x}\,|\,\omega_1)}{p(\vec{x}\,|\,\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}=\theta_{12} \quad 则\quad \vec{x}\in\omega_1$$

$$l_{12}=\frac{p(\vec{x}\,|\,\omega_1)}{p(\vec{x}\,|\,\omega_2)} < \frac{P(\omega_2)}{P(\omega_1)}=\theta_{12} \quad 则\quad \vec{x}\in\omega_2$$

l_{12}称为似然比（likelihood ratio），θ_{12}称为似然比的判决阈值。

原则：要确定x是属于ω1类还是ω2类，要看x是来自于ω1类的概率大还是来自ω2类的概率大。

例题1： 鱼类加工厂对鱼进行自动分类，ω_1:鲈鱼；ω_2:鲑鱼。模式特征x=*x*(长度)。

已知：（统计结果）

先验概率：　$P(\omega_1)=1/3$（鲈鱼出现的概率）

$P(\omega_2)=1-P(\omega_1)=2/3$　（鲑鱼出现的概率）

条件概率：　$p(x|\omega_1)$ 见图示（鲈鱼的长度特征分布概率）

$p(x|\omega_2)$见图示（鲑鱼的长度特征分布概率）

求：后验概率：$P(\omega|x=10)=?$

（如果一条鱼x＝10，是什么类别？）

解法1：

利用Bayes公式

$$P(\omega_1 \mid x=10)=\frac{p(x=10\mid\omega_1)P(\omega_1)}{p(x=10)}$$

$$=\frac{p(x=10\mid\omega_1)P(\omega_1)}{p(x=10\mid\omega_1)P(\omega_1)+p(x=10\mid\omega_2)P(\omega_2)}$$

$$=\frac{0.05\times1/3}{0.05\times1/3+0.50\times2/3}=0.048$$

因为，$P(\omega_2|x=10)=1-P(\omega_1|x=10)=1-0.048=0.952$

$$P(\omega_1|x=10)<P(\omega_2|x=10)$$

故判决：$(x=10)\in\omega_2$，即是鲑鱼。

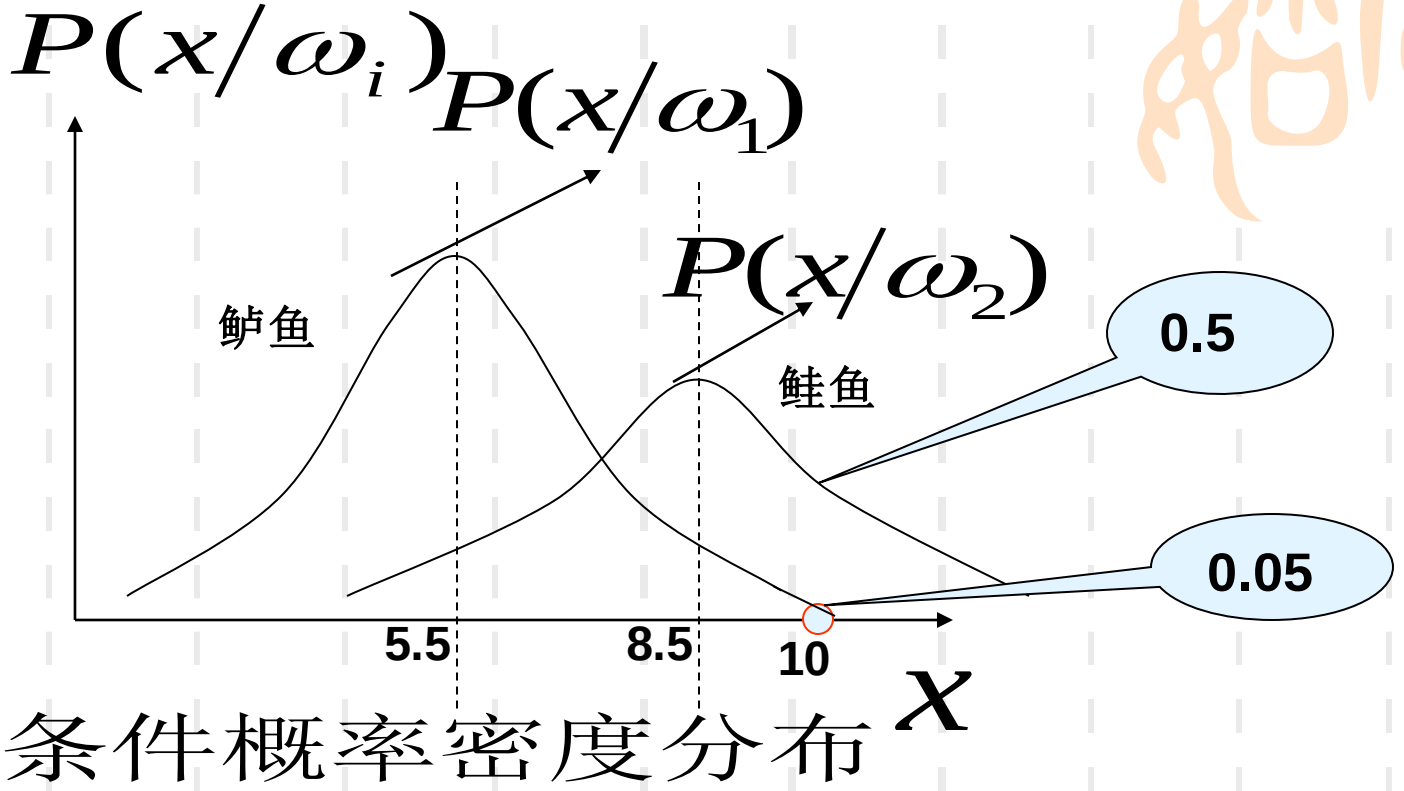
解法2：

写成似然比形式

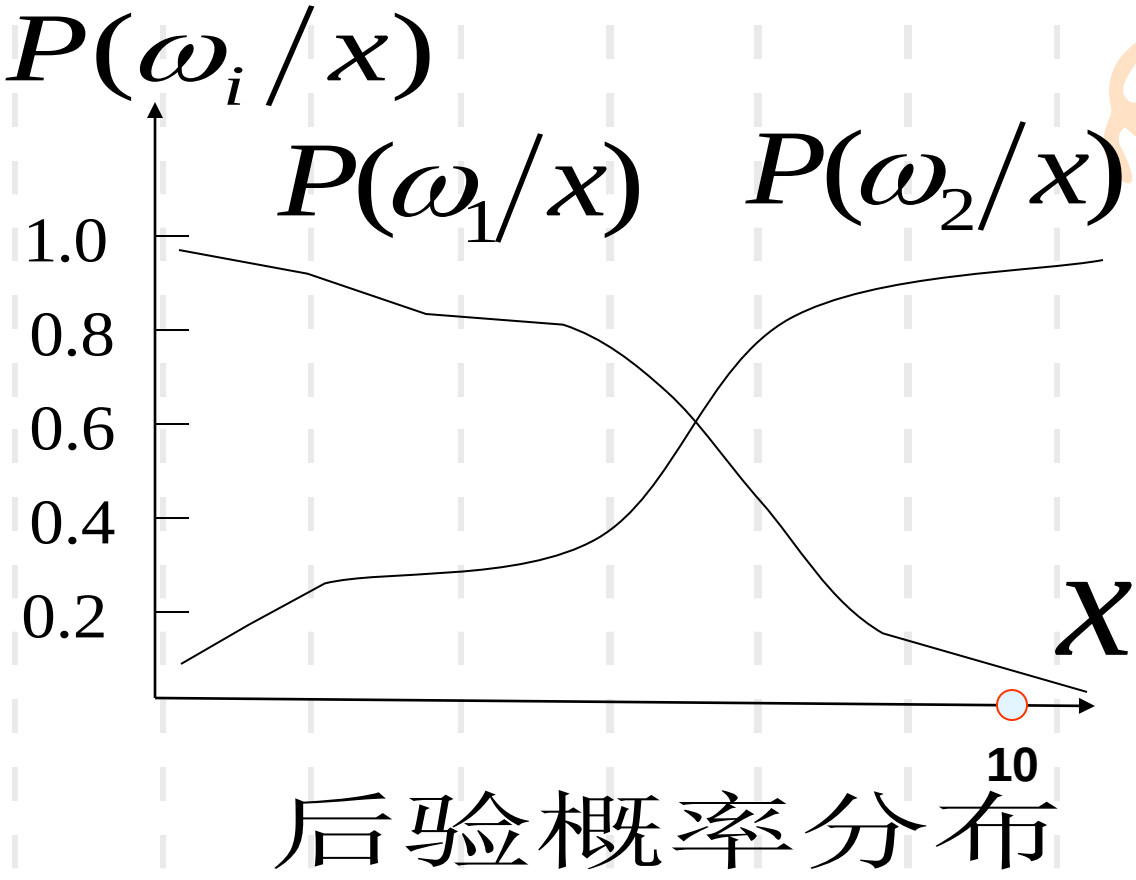
$$l_{12}(x=10)=\frac{p(x=10\mid\omega_1)}{p(x=10\mid\omega_2)}=\frac{0.05}{0.50}=0.1$$

$$\text{判决阈值}\theta_{12}=\frac{P(\omega_2)}{P(\omega_1)}=\frac{2/3}{1/3}=2$$

$$\because\quad l_{12}(x=10)<\theta_{12}\ ,\quad \therefore\quad x\in\omega_2\ ,\ \text{即是鲑鱼。}$$



[例题1图示]



例题1图示

- 最小错误率判决准则
- 最小损失判决准则
- 最小最大损失准则
- N-P(Neyman—Pearson)判决

3.2 最小错误率判决规则（最简单的 Bayes 分类方法）

分析一个“两类问题”。

以上一个例子为例，用 w1 和 w2 表示两种不同的类型，如 w1 表示诊断正常，w2 表示诊断出患有癌症。

用 $P(w_1)$ 和 $P(w_2)$ 分别表示先验概率。如：$P(w_1)$ 诊断正常的概率，$P(w_2)$ 表示某地人患癌症的概率，可通过大量的统计得到。

用 $\rho(x \mid w_1)$ 和 $\rho(x \mid w_2)$ 表示两个类概率密度。

样本 x 表示“试验反应阳性”，则 $\rho(x \mid w_1)$ 诊断为无癌症且试验反应为阳性，$P(w_1 \mid x)$ 试验为阳性且没有癌症。

根据全概率公式，模式样本 x 出现的全概率密度为：

$$\rho(x) = \rho(x \mid w_1) \cdot P(w_1) + \rho(x \mid w_2) \cdot P(w_2) \qquad (3.2-1)$$

根据 Bayes 公式，在模式样本 x 出现的条件下，两个类型的后验概率为：

$$P(w_1 \mid x)=\frac{\rho(x \mid w_1)\cdot P(w_1)}{\rho(x)}, \qquad P(w_2 \mid x)=\frac{\rho(x \mid w_2)\cdot P(w_2)}{\rho(x)} \qquad (3.2-2)$$

此时，样本归属于“后验概率较高”的那种类型。

也就是：

$$\left.\begin{array}{l} P(w_1 \mid x) > P(w_2 \mid x)，\ 则\ x \in w_1 \\ P(w_2 \mid x) > P(w_1 \mid x)，\ 则\ x \in w_2 \end{array}\right\} \qquad (3.2-3)$$

$P(w_1 \mid x) = P(w_2 \mid x)$，则偶然决定 $x \in w_1$，或 $x \in w_2$

根据（3.2－2）式，上述判决规则等价于：

$$\left.\begin{array}{l} \rho(x \mid w_1)\cdot P(w_1) > \rho(x \mid w_2)\cdot P(w_2)，\ 则\ x \in w_1 \\ \rho(x \mid w_2)\cdot P(w_2) > \rho(x \mid w_1)\cdot P(w_1)，\ 则\ x \in w_2 \end{array}\right\} \qquad (3.2-4)$$

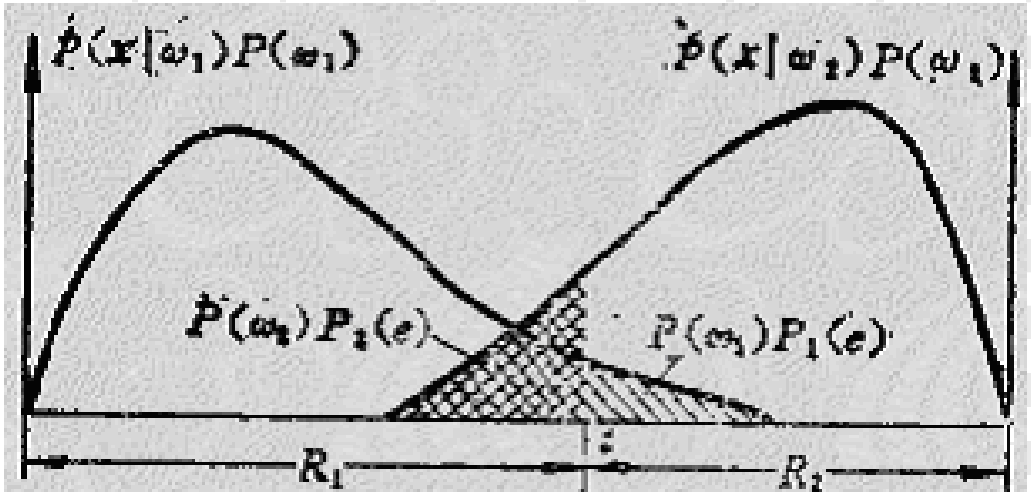
$\rho(x \mid w_2)\cdot P(w_2) = \rho(x \mid w_1)\cdot P(w_1)$，则偶然决定 $x \in w_1$，或 $x \in w_2$

上面只是给出了最小错误率贝叶斯决策规则，但没有证明按这种规则进行分类确实使错误率最小。下面用一维情况来证明最小错误率贝叶斯决策规则，其结果不难推广到多维。

如下图所示，在一维特征空间里，判决门限 t 把空间划分为两个类型区域 $\mathbf{R_1}$，$\mathbf{R_2}$。

在 $\mathbf{R_1}$ 中，$\rho(x \mid w_1) \cdot P(w_1) > \rho(x \mid w_2) \cdot P(w_2)$，则 $x \in w_1$；

在 $\mathbf{R_2}$ 中，$\rho(x \mid w_2) \cdot P(w_2) > \rho(x \mid w_1) \cdot P(w_1)$，则 $x \in w_2$；



阴影区域是两类样本的交错分配区域，阴影面积就是这种分类方法的错误概率。总错误率有两种情况：

$x \in w_1$，而判为 $x \in w_2$，斜线区域。

$x \in w_2$，而判为 $x \in w_1$，纹线区域。

所以，总错误率：

$$P(e) = \int_{-\infty}^{\infty} P(e \mid x)\rho(x)dx$$

其中，$\int_{-\infty}^{\infty}()dx$ 表示在整个 **d** 维特征空间上的积分。

对上述两类问题：当 $P(w_2 \mid x) > P(w_1 \mid x)$ 时，则 $x \in w_2$。显然作出决策 w_2 时，**x** 的条件错误概率为 $P(w_1 \mid x)$，反之为 $P(w_2 \mid x)$。

也就是：

$$P(e \mid x) = \begin{cases} P(w_1 \mid x) & \text{当 } P(w_2 \mid x) > P(w_1 \mid x) \\ P(w_2 \mid x) & \text{当 } P(w_1 \mid x) > P(w_2 \mid x) \end{cases}$$

若令 t 为两类分界面，特征向量 x 为一维时，t 为 x 轴上的一个点，如上图所示：

$$P(e)=\int_{-\infty}^{t}P(w_2\mid x)\cdot\rho(x)dx+\int_{t}^{\infty}P(w_1\mid x)\rho(x)dx$$

$$=\int_{-\infty}^{t}\rho(x\mid w_2)\cdot P(w_2)dx+\int_{t}^{\infty}\rho(x\mid w_1)\cdot P(w_1)dx$$

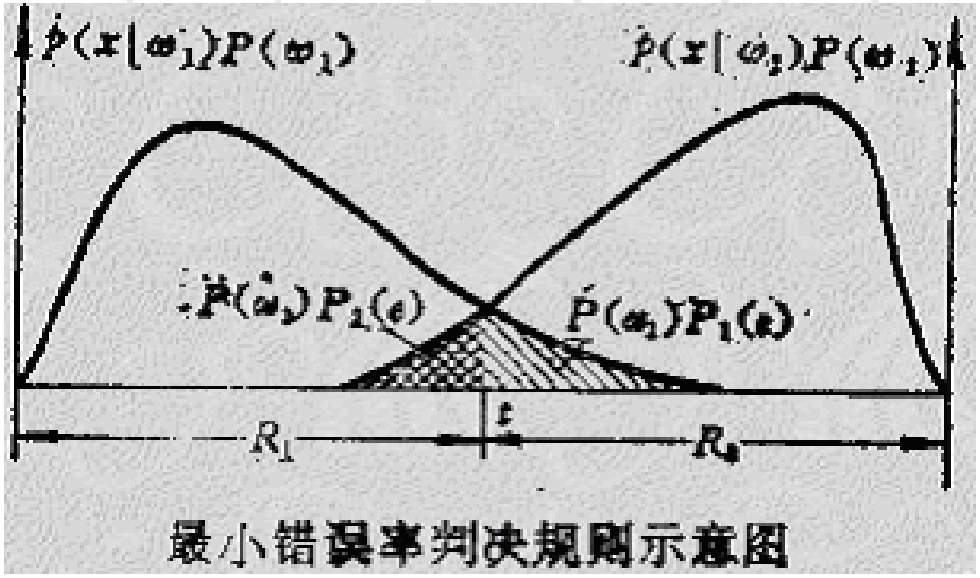
也可写为：

$$P(e)=P(x\in R_1,w_2)+P(x\in R_2,w_1)$$

$$=P(w_2)\cdot P(x\in R_1\mid w_2)+P(w_1)\cdot P(x\in R_2\mid w_1)$$

$$=P(w_2)\cdot\int_{R_1}\rho(x\mid w2)dx+\rho(w_1)\cdot\int_{R_2}\rho(x\mid w1)dx$$

$$=P(w_2)\cdot P_2(e)+P(w_1)\cdot P_1(e)$$



最小错误率判决规则示意图

所以要使$P(e)$最小，判决门限应如上图所示，否则就会有多余的阴影面。而（3.2-3）、（3.2-4）表达的判决规则，判决门限正好如上图所示，所以称之为“最小错误概率判决规则”。

可以把上述两类问题导出的最小错误率判决规则一般化，推广到c类问题中，表达为：

若：$P(w_i \mid x) = \max_{j=1,\cdots,c}\{P(w_j \mid x)\}$，则$x \in w_i$，

等价于：$\rho(x \mid w_i)\cdot P(w_i) = \max_{j=1,\cdots,c}\{\rho(x \mid w_j)\cdot P(w_j)\}$，则$x \in w_i$

最小误判概率准则下的判决规则：

如果，$P(\omega_1)p(\vec{x}|\omega_1) \gtrless P(\omega_2)p(\vec{x}|\omega_2)$

则判 $\vec{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$

或等价地，如果，$l_{12}(\vec{x}) = \dfrac{p(\vec{x}|\omega_1)}{p(\vec{x}|\omega_2)} \gtrless \dfrac{P(\omega_2)}{P(\omega_1)}$

则判 $\vec{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$

称 $l_{12}(\vec{x})$为似然比(Likelihood Ratio)，

称 θ_{12} 为似然比阈值记为 $P(\omega_2)/P(\omega_1)$

由贝叶斯定理 $p(\vec{x})P(\omega_i|\vec{x}) = P(\omega_i)p(\vec{x}|\omega_i)$

另一个等价形式是：

如果 $P(\omega_2|\vec{x}) \gtrless P(\omega_1|\vec{x})$

则判 $\vec{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$

对于多类问题，最小误判概率准则有如下几种等价的判决规则：

(1) 若 $P(\omega_i|\vec{x}) > P(\omega_j|\vec{x})$，$\forall j \neq i$ ，则判 $\vec{x} \in \omega_i$

(2) 若 $P(\omega_i|\vec{x}) = \max_j [P(\omega_j|\vec{x})]$，则判 $\vec{x} \in \omega_i$

（后验概率形式）

(3) 若 $p(\vec{x}|\omega_i)P(\omega_i) > p(\vec{x}|\omega_j)P(\omega_j)$，$\forall j \neq i$，则判 $\vec{x} \in \omega_i$

(4) 若 $p(\vec{x}|\omega_i)P(\omega_i) = \max_j [p(\vec{x}|\omega_j)P(\omega_j)]$，则判 $\vec{x} \in \omega_i$

（条件概率形式）

(5) 若 $l_{ij}(\vec{x}) \triangleq \dfrac{p(\vec{x}|\omega_i)}{p(\vec{x}|\omega_j)} > \dfrac{P(\omega_j)}{P(\omega_i)} \triangleq \theta_{ij}$，$\forall j \neq i$ ，则判 $\vec{x} \in \omega_i$

（似然比形式）

(6) 如果 $\ln p(\vec{x}|\omega_i) + \ln P(\omega_i) > \ln p(\vec{x}|\omega_j) + \ln P(\omega_j)$ ，$\forall j \neq i$ ，则判 $\vec{x} \in \omega_i$

（条件概率的对数形式）

例：对一批人进行癌症普查，患癌症者定为属ω_1类，正常者定为属ω_2类。统计资料表明人们患癌的概率 $P(\omega_1)=0.005$ ，从而 $P(\omega_2)=0.995$ 。设有一种诊断此病的试验，其结果有阳性反应和阴性反应之分，依其作诊断。化验结果是一维离散模式特征。统计资料表明：癌症者有阳性反映的概率为**0.95**即 $P(x=阳|\omega_1)=0.95$ ，从而可知 $P(x=阴|\omega_1)=0.05$ ，正常人阳性反映的概率为**0.01**即 $P(x=阳|\omega_2)=0.01$ ，可知 $P(x=阴|\omega_2)=0.99$ 。

问有阳性反映的人患癌症的概率有多大？

解：

$$P(\omega_1|x=\text{阳})=\frac{P(x=\text{阳}|\omega_1)P(\omega_1)}{P(x=\text{阳})}$$

$$=\frac{P(x=\text{阳}|\omega_1)P(\omega_1)}{P(x=\text{阳}|\omega_1)P(\omega_1)+P(x=\text{阳}|\omega_2)P(\omega_2)}$$

$$=\frac{0.95\times 0.005}{0.95\times 0.005+0.01\times 0.995}$$

$$=0.323$$

说明有阳性反应的人其患癌的概率有32. 3%

写成似然比形式：

$$l_{12}(x)=\frac{P(x=\text{阳}|\omega_1)}{P(x=\text{阳}|\omega_2)}=\frac{0.95}{0.01}=95$$

$$\theta_{12}=\frac{P(\omega_2)}{P(\omega_1)}=\frac{0.995}{0.005}=197$$

$$\because l_{12}(x)<\theta_{12} \qquad \therefore x\in\omega_2$$

3.4 最小损失准则判决

3.4.1 损失概念、损失函数与平均损失

设模式空间中存在c个类别：$\omega_1, \omega_2, \cdots, \omega_c$，

决策空间由a个决策：$\alpha_1, \alpha_2, \cdots, \alpha_a$，

决策α_j常指将模式x指判为某一类w_j或者是拒判。

对一个实属ω_i类的模式采用了决策α_j所造成的损失记为：

$$\lambda(\alpha_j / \omega_i) \triangleq \lambda_{ij}$$

于是就有$\{\omega_1, \omega_2, \cdots, \omega_c\} \times \{\alpha_1, \alpha_2, \cdots, \alpha_a\}$空间中的二元函数，称其为**损失函数**。

决策-损失表

λ_{ij}

	ω_1	ω_2	…	ω_c
α_1	$\lambda(\alpha_1/\omega_1)$	$\lambda(\alpha_1/\omega_2)$	…	$\lambda(\alpha_1/\omega_c)$
α_2	$\lambda(\alpha_2/\omega_1)$	$\lambda(\alpha_2/\omega_2)$	…	$\lambda(\alpha_2/\omega_c)$
…				…
α_c	$\lambda(\alpha_c/\omega_1)$	$\lambda(\alpha_c/\omega_2)$	…	$\lambda(\alpha_c/\omega_c)$
α_{c+1}	$\lambda(\alpha_{c+1}/\omega_1)$	$\lambda(\alpha_{c+1}/\omega_2)$	…	$\lambda(\alpha_{c+1}/\omega_c)$

- 决策α_j指将模式x指判为w_j或者是拒判。

$$\lambda_{ij} = \begin{cases} = 0 & i = j \\ = 1 & i \neq j \end{cases}$$

0-1损失函数

条件平均风险

令决策的数目*a*等于类数*c*，如果决策α_j定义为判$\vec{x}$属于ω_j类，那么对于给定的模式$\vec{x}$在采取决策α_j的条件下损失的期望为

$$R_j(\vec{x}) = R(\alpha_j|\vec{x}) = \sum_{i=1}^{c} \lambda_{ij} P(\omega_i|\vec{x}) \triangleq E_i\left[\lambda_{ij}|\vec{x}\right] \quad (j = 1,2,\cdots,c)$$

条件期望损失$R_j(\vec{x})$刻划了在模式为$\vec{x}$、决策为α_j条件下的平均损失，故也称$R_j(\vec{x})$为**条件平均损失或条件平均风险（Risk）**。由贝叶斯公式，上式可以写为

$$R_j(\vec{x}) = \sum_{i=1}^{c} \lambda_{ij}\, p(\vec{x}|\omega_i)\, P(\omega_i) / p(\vec{x})$$

$$= \sum_{i=1}^{c} \lambda_{ij}\, p(\vec{x}|\omega_i) P(\omega_i) \Big/ \sum_{i=1}^{c} p(\vec{x}|\omega_i) P(\omega_i)$$

3.4.2 最小损失准则判决

- 可以将最小条件平均损失判决规则表示为

如果 $R_j(\vec{x}) = \min_i [R_i(\vec{x})]$

则判 $\vec{x} \in \omega_j$

定理：使条件平均损失最小的判决也必然使总的平均损失最小。

所以最小条件平均损失准则也称为最小平均损失准则或最小平均风险准则，简称为**最小损失准则**。

$$R_j(\vec{x}) = \sum_{i=1}^{c} \lambda_{ij} p(\vec{x}|\omega_i) P(\omega_i) / p(\vec{x}) = \sum_{i=1}^{c} \lambda_{ij} p(\vec{x}|\omega_i) P(\omega_i) \Big/ \sum_{i=1}^{c} p(\vec{x}|\omega_i) P(\omega_i)$$

对于两类问题，

$$R_1(\vec{x}) = \left[\lambda_{11} p(\vec{x}|\omega_1) P(\omega_1) + \lambda_{21} p(\vec{x}|\omega_2) P(\omega_2)\right] / p(\vec{x})$$

$$R_2(\vec{x}) = \left[\lambda_{12} p(\vec{x}|\omega_1) P(\omega_1) + \lambda_{22} p(\vec{x}|\omega_2) P(\omega_2)\right] / p(\vec{x})$$

如果 $R_1(\vec{x}) \lessgtr R_2(\vec{x})$ **则：** $\vec{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$

这时最小损失判决规则可以表为：

$$\lambda_{11} p(\vec{x}|\omega_1) P(\omega_1) + \lambda_{21} p(\vec{x}|\omega_2) P(\omega_2) \lessgtr \lambda_{12} p(\vec{x}|\omega_1) P(\omega_1) + \lambda_{22} p(\vec{x}|\omega_2) P(\omega_2)$$

$$\lambda_{11}p(\vec{x}|\omega_1)P(\omega_1)+\lambda_{21}p(\vec{x}|\omega_2)P(\omega_2) \lessgtr \lambda_{12}p(\vec{x}|\omega_1)P(\omega_1)+\lambda_{22}p(\vec{x}|\omega_2)P(\omega_2)$$

经整理可得：

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

两类问题的最小损失准则的似然比形式的判决规则为：

如果 $\dfrac{p(\vec{x}|\omega_1)}{p(\vec{x}|\omega_2)} \gtrless \dfrac{P(\omega_2)(\lambda_{21}-\lambda_{22})}{P(\omega_1)(\lambda_{12}-\lambda_{11})}$ **则判** $\vec{x}\in\begin{cases}\omega_1\\ \omega_2\end{cases}$

若记似然比阈值 $\theta_{12} = \dfrac{P(\omega_2)}{P(\omega_1)} \dfrac{(\lambda_{21} - \lambda_{22})}{(\lambda_{12} - \lambda_{11})}$

则两类问题的判决规则为：

如果 $l_{12}(\vec{x}) \gtrless \theta_{12}$ 则判： $\vec{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$

注意，若 $l_{12}(\vec{x}) = \theta_{12}$

我们规定任判或拒判。

$$\theta_{12}=\frac{P(\omega_2)}{P(\omega_1)}\frac{(\lambda_{21}-\lambda_{22})}{(\lambda_{12}-\lambda_{11})}$$

如果 $l_{12}(\vec{x})\gtrless\theta_{12}$ 则判：$\vec{x}\in\begin{cases}\omega_1\\ \omega_2\end{cases}$

损失函数如何确定依赖于实际问题和经验，有时为了方便，对于一般的 C 类问题，令

$$\lambda_{ij}=\begin{cases}0 & ,\ i=j\\ 1 & ,\ i\neq j\end{cases}$$ （0-1损失函数）

此时：$\theta_{12}=\frac{P(\omega_2)}{P(\omega_1)}$ 此即为最小误判概率准则的判决规则

3.4.2 最小损失准则判决

取0-1损失函数时，最小损失准则等价于最小误判概率准则，此时的平均损失就是误判概率，使平均损失最小即使误判概率最小。这也表明，最小误判概率准则是最小损失准则的特例。

例： 设信源信号为$\{0,1\}$,其通过一受加性噪声干扰的信道，噪声为正态分布，均值为零，方差为 σ^2 信道输出为 x, 试求出最优判决规则，以区分输入是0还是1。

解： 设信号0为ω_0类，信号1为ω_1类。因输入信号受正态分布 $N(0,\sigma$的加性噪声干扰，所以输出信号的加性噪声干扰，所以输出信号为信源信号与噪声的迭加，输出信号的概型由噪声确定，方差仍为 σ^2 均值分别为0和1。

输入信号值为0和1而输出幅值为x 的概密函数为：

$$\begin{cases} p(x|\omega_0)=\dfrac{1}{\sqrt{2\pi}\sigma}\exp\left[-\dfrac{x^2}{2\sigma^2}\right] \\ p(x|\omega_1)=\dfrac{1}{\sqrt{2\pi}\sigma}\exp\left[-\dfrac{(x-1)^2}{2\sigma^2}\right] \end{cases}$$

似然比为：

$$l_{01}(x)=\frac{p(x|\omega_0)}{p(x|\omega_1)}=\exp\left[\frac{1-2x}{2\sigma^2}\right]$$

$$l_{01}(x)=\frac{p(x|\omega_0)}{p(x|\omega_1)}=\exp\left[\frac{1-2x}{2\sigma^2}\right]$$

运用最小损失准则，判决规则为：

$$l_{01}(x)=\exp\left[\frac{1-2x}{2\sigma^2}\right]\lessgtr\frac{(\lambda_{10}-\lambda_{11})}{(\lambda_{01}-\lambda_{00})}\frac{P(\omega_1)}{P(\omega_0)}\text{ 时 }\vec{x}\in\begin{cases}\omega_1\\ \omega_2\end{cases}$$

两边取对数：当 $x<\frac{1}{2}-\sigma^2\ln\left[\frac{(\lambda_{10}-\lambda_{11})}{(\lambda_{01}-\lambda_{00})}\frac{P(\omega_1)}{P(\omega_0)}\right]$ 时

判 $x\in\omega_0$ 即信号为“0”。

若取 $\lambda_{01}=\lambda_{10}=1$，$\lambda_{00}=\lambda_{11}=0$，$P(\omega_0)=P(\omega_1)=\frac{1}{2}$

则 $x<\frac{1}{2}$ 时，应判信号为“0”；$x>\frac{1}{2}$ 时应判信号为“1”。

这和我们通常的习惯处理方法是一致的。

例： 设正常细胞属于ω_1类，异常细胞属于ω_2类，它们的先验概率分别为$P(\omega_1)=0.9$，$P(\omega_2)=0.1$现有一个待识细胞，其观测矢量为$\vec{x}$，从类概密曲线上查得$p(\vec{x}|\omega_1)=0.2$，$p(\vec{x}|\omega_2)=0,4$如果损失系数取为$\lambda_{11}=0,\ \lambda_{12}=1\ \ \lambda_{21}=6\ \lambda_{22}=0$试用最小误判概率准则和最小损失准则判断该细胞是正常的还是异常的。

解： 由贝叶斯定理可以分别算出ω_1和ω_2的后验概率。

$$P(\omega_1|\vec{x})=\frac{p(\vec{x}|\omega_1)P(\omega_1)}{\sum_{i=1}^{2}p(\vec{x}|\omega_i)P(\omega_i)}=\frac{0.2\times0.9}{0.2\times0.9+0.4\times0.1}=0.818$$

$$P(\omega_1|\vec{x})=0.818 \qquad P(\omega_2|\vec{x})=1-P(\omega_1|\vec{x})=0.182$$

因 $P(\omega_1|\vec{x})=0.818>P(\omega_2|\vec{x})=0.182$，所以按最小误判概率准则，应判 $\vec{x}$ 归于正常细胞。

当依据损失进行判决时，需计算条件平均损失。

$$R(\alpha_1|\vec{x})=\sum_{i=1}^{2}\lambda_{i1}P(\omega_i|\vec{x})=\lambda_{21}P(\omega_2|\vec{x})=1.092$$

$$R(\alpha_2|\vec{x})=\sum_{i=1}^{2}\lambda_{i2}P(\omega_i|\vec{x})=\lambda_{12}P(\omega_1|\vec{x})=0.818$$

取损失最小的判决，由于 $R(\alpha_1|\vec{x})>R(\alpha_2|\vec{x})$，故 $\vec{x}\in\omega_2$

之所以这两个判决结果相反，是因为λ_{21}取得较大的缘故。

3.4.3 含拒绝判决的最小损失判决

拒绝判决可以作为最小损失判决中的一个可能判决，

α_{c+1} =“拒绝判决”。

令$\lambda(\alpha_{c+1}/\omega_i)$表示模式$\vec{x}$实属$\omega_i$类但拒绝作出判决所造成的损失，于是在模式$\vec{x}$条件下拒绝判决的平均损失为

$$R(\alpha_{c+1}|\vec{x}) = \sum_{i=1}^{c} \lambda(\alpha_{c+1}/\omega_i) P(\omega_i|\vec{x})$$

如果 $R(\alpha_{c+1}|\vec{x}) < R(\alpha_j|\vec{x})$ j=1, 2, …, c

则作出拒绝判决。

3.5 最小最大损失准则

最小最大损失准则的基本思想：

实际中，类先验概率 $P(\omega_i)$ 往往不能精确知道或在分析过程中是变动的，从而导致判决域不是最佳的。所以应考虑如何解决在 $P(\omega_i)$ 不确知或变动的情况下使平均损失变大的问题。

应该立足最差的情况争取最好的结果。

对于两类问题，设一种分类识别决策将特征空间Ω分划为两个子空间Ω_1和Ω_2，记λ_{ij}为将实属ω_i类的模式判为ω_j的损失函数，各种判决的平均损失为

$$R = \int_{\Omega} R(\alpha(\vec{x})|\vec{x})\, p(\vec{x})d\vec{x}$$

$$= \int_{\Omega_1} R(\alpha_1(\vec{x})|\vec{x})\, p(\vec{x})d\vec{x} + \int_{\Omega_2} R(\alpha_2(\vec{x})|\vec{x})\, p(\vec{x})d\vec{x}$$

$$= \int_{\Omega_1}\sum_{i=1}^{2}\lambda_{i1}p(\vec{x}|\omega_i)P(\omega_i)d\vec{x} + \int_{\Omega_2}\sum_{i=1}^{2}\lambda_{i2}p(\vec{x}|\omega_i)P(\omega_i)d\vec{x}$$

$$= \lambda_{11}P(\omega_1)\int_{\Omega_1} p(\vec{x}|\omega_1)d\vec{x} + \lambda_{21}P(\omega_2)\int_{\Omega_1} p(\vec{x}|\omega_2)d\vec{x}$$

$$+ \lambda_{12}P(\omega_1)\int_{\Omega_2} p(\vec{x}|\omega_1)d\vec{x} + \lambda_{22}P(\omega_2)\int_{\Omega_2} p(\vec{x}|\omega_2)d\vec{x}$$

利用 $P(\omega_2)=1-P(\omega_1)$ 和

$$\int_{\Omega_1} p(\vec{x}|\omega_i)d\vec{x} = 1 - \int_{\Omega_2} p(\vec{x}|\omega_i)d\vec{x}$$

则平均损失可写成

$$R = \lambda_{22} + (\lambda_{21} - \lambda_{22})\int_{\Omega_1} p(\vec{x}|\omega_2)d\vec{x}$$

$$+ P(\omega_1)\left[(\lambda_{11} - \lambda_{22}) + (\lambda_{12} - \lambda_{11})\int_{\Omega_2} p(\vec{x}|\omega_1)d\vec{x} + (\lambda_{22} - \lambda_{21})\int_{\Omega_1} p(\vec{x}|\omega_2)d\vec{x}\right]$$

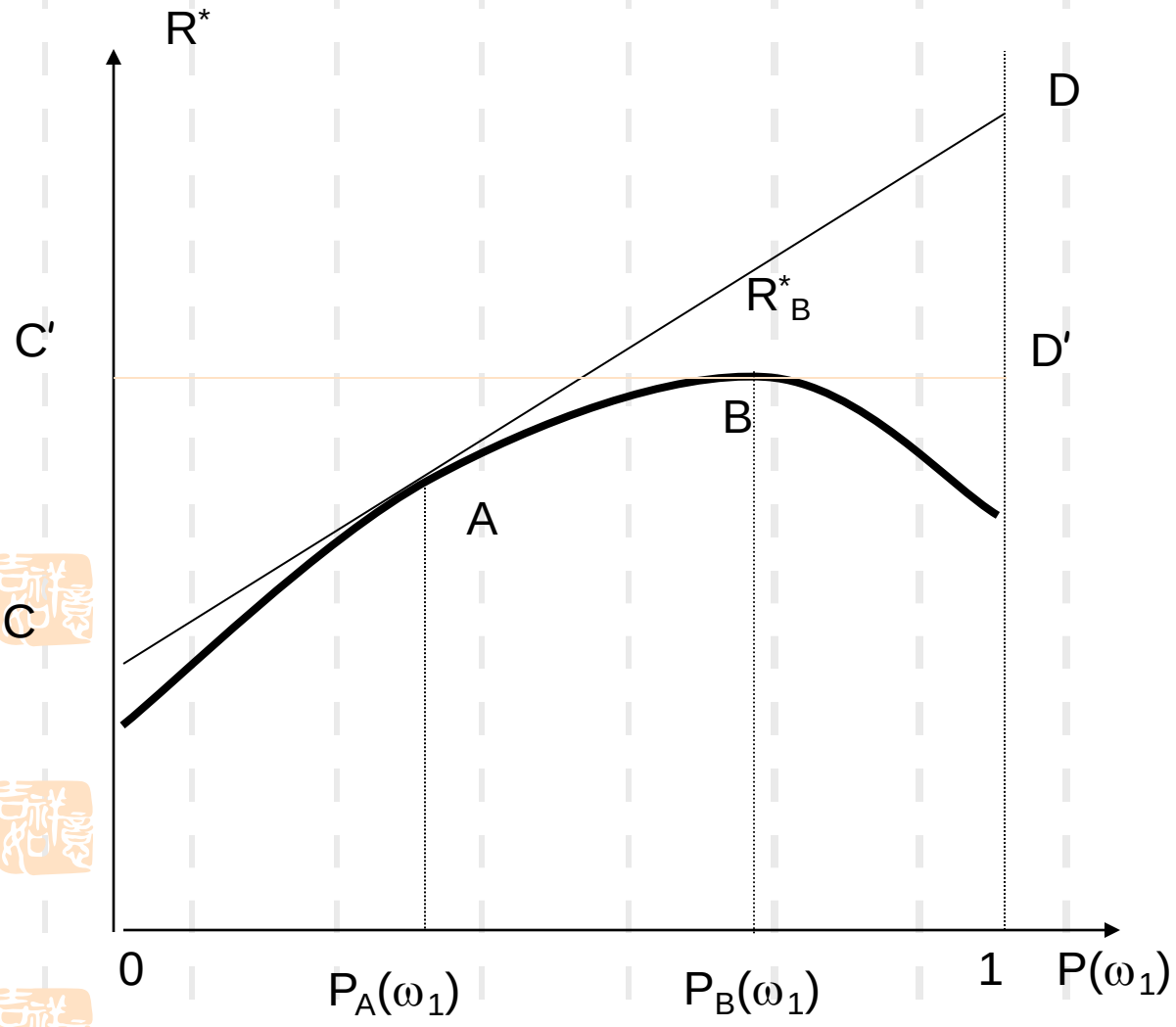
$$\underline{\Delta} \; a + bP(\omega_1)$$

由于 $P(\omega_1)$ 在 0 和 1 之间取值，所以平均损失值有

$$a \le R \le a + b$$

$$R = a + bP(\omega_1)$$

- 由上式可见，当类概密、损失函数λ_{ij}、类域Ω_i取定后，R是$P(\omega_1)$的线性函数。
- 考虑$P(\omega_1)$的各种可能取值情况，为此在区间$(0, 1)$中取若干个不同的$P(\omega_1)$值，并分别按最小损失准则确定相应的最佳决策类域Ω_1、Ω_2，然后计算出其相应的最小平均损失R^*，从而可得最小平均损失R^*与先验概率$P(\omega_1)$的关系曲线。


R*
D
R*B
C′
D′
B
A
C
0
PA(ω1)
PB(ω1)
1
P(ω1)

如果能求出某个 $P(\omega_1) \triangleq P_B(\omega_1)$，相对于 $P_B(\omega_1)$ 的最佳判决类域 Ω_1 和 Ω_2 能使该式中的 $b=0$，即

$$b=(\lambda_{11}-\lambda_{22})+(\lambda_{12}-\lambda_{11})\int_{\Omega_2} p(\vec{x}|\omega_1)d\vec{x}+(\lambda_{22}-\lambda_{21})\int_{\Omega_1} p(\vec{x}|\omega_2)d\vec{x}=0$$

在此决策类域下，无论 $P(\omega_1)$ 如何变化，因 $b=0$ 而使 R 与 $P(\omega_1)$ 无关，从而使得平均损失 R 恒等于常数 a，即

$$R=\lambda_{22}+(\lambda_{21}-\lambda_{22})\int_{\Omega_1} p(\vec{x}|\omega_2)d\vec{x} \triangleq R^*=a$$

求使 $b=0$ 的 $P(\omega_1)$ 等价于在最小平均损失 $R^* \sim P(\omega_1)$ 关系中求使

$$dR^* \Big/ dP(\omega_1)=0$$

的 $P(\omega_1)$，显然，此时的 $P(\omega_1)$ 使 R^* 取所有最小损失的最大值 R_m^*。所以 R_m^* 是最大的最小损失。

具体的设计过程是:

(1) 按最小损失准则找出对应于 $(0,1)$ 中的各个不同值的 $P(\omega_1)$ 的最佳决策类域 Ω_1、Ω_2,

(2) 计算相应各个 $P(\omega_1)$ 及最佳决策类域的最小平均损失，得 $R^* \sim P(\omega_1)$ 曲线，找出使 R^* 取最大值的 $P^*(\omega_1)$,

(3) 运用 $P^*(\omega_1)$、$1-P^*(\omega_1)$ 及 λ_{ij} 构造似然比阈值并运用最小损失准则下的决策规则对具体的模式分类识别。

最小最大损失判决规则 为：

如果 $\dfrac{p(\vec{x}|\omega_1)}{p(\vec{x}|\omega_2)} \gtrless \dfrac{(\lambda_{21}-\lambda_{22})(1-P^*(\omega_1))}{(\lambda_{12}-\lambda_{11})P^*(\omega_1)}$ 则判 $\vec{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$

当采用 0-1 损失函数时，由 $b=0$ 可得

$$\int_{\Omega_1} p(\vec{x}|\omega_2)d\vec{x} = \int_{\Omega_2} p(\vec{x}|\omega_1)d\vec{x}$$

上式表明，最小最大损失判决所导出的最佳分界面应使两类错误概率相等 ，可知此时的平均损失

$$R = \int_{\Omega_1} p(\vec{x}|\omega_2)d\vec{x}$$

3.6 N-P(Neyman—Pearson)判决

在某些实际问题中，可能存在以下几种情况：

(1) 不知道各类的先验概率 $P(\omega_i)$；

(2) 难于确定误判的代价 λ_{ij}；

(3) 某一种错误较另一种错误更为重要。

针对(1)，可以采用最小最大损失准则或令各类概率相等的办法克服；

针对(2)，如果允许的话，可以避开使用损失函数而采用最小误判概率准则；

针对(3)，可以采用最小损失准则判决。针对上面的三个问题，更主要的是针对(3)，我们采用N-P准则。

所谓N-P准则，是严格限制较重要的一类错误概率令其等于某常数而使另一类误判概率最小。

对两类问题，$d(\vec{x})=0$将特征空间Ω分成两个子空间Ω_1和Ω_2，其中$\Omega_1 \cup \Omega_2 = \Omega$，$\Omega_1 \cap \Omega_2 = \Phi$。当一模式特征点$\vec{x} \in \Omega_1$时，指判$\vec{x} \in \omega_1$；当$\vec{x} \in \Omega_2$时，指判$\vec{x} \in \omega_2$。

将实属 ω_1 类的模式 $\vec{x}$ 判属 ω_2 类的误判概率为

$$\varepsilon_{12} = \int_{\Omega_2} p(\vec{x}|\omega_1)d\vec{x}$$

将实属 ω_2 类的模式判属 ω_1 类的误判概率为

$$\varepsilon_{21} = \int_{\Omega_1} p(\vec{x}|\omega_2)d\vec{x}$$

N-P准则是在使某一类误判概率等于常数的约束下使另一类误判概率最小。

令 $\varepsilon_{21}=\varepsilon_0=$常数，求使 ε_{12} 最小。运用拉格朗日乘数法求条件极值，为此作辅助函数:

$$y=\varepsilon_{12}+\lambda(\varepsilon_{21}-\varepsilon_0)$$

$$=\int_{\Omega_2} p(\vec{x}|\omega_1)d\vec{x}+\lambda\left[\int_{\Omega_1} p(\vec{x}|\omega_2)d\vec{x}-\varepsilon_0\right]$$

$$=(1-\lambda\varepsilon_0)-\int_{\Omega_1}(p(\vec{x}|\omega_1)-\lambda p(\vec{x}|\omega_2))d\vec{x}$$

$$y=(1-\lambda\varepsilon_0)-\int_{\Omega_1}(p(\vec{x}|\omega_1)-\lambda p(\vec{x}|\omega_2))d\vec{x}$$

求 Ω_1 使 y 取极小值。

一般地讲，Ω_1 无法直接用解析的办法求得。

但注意到 λ 在式子中是确定的，$p(\vec{x}|\omega_1)$、$p(\vec{x}|\omega_2)$ 在 Ω 空间中也是确定的，如果选择满足条件 $p(\vec{x}|\omega_1)-\lambda p(\vec{x}|\omega_2)>0$ 的 $\vec{x}$ 的全体作为 Ω_1^* 就能保证这时所求得的 y 值 y^* 比 Ω_1 的其它取法时的 y 值要小。

因为这种取法下，Ω_1^* 是使被积函数取正数的最大的域。

因此选择满足条件 $p(\vec{x}|\omega_1)-\lambda p(\vec{x}|\omega_2)>0$ 的全体 $\vec{x}$ 作为 Ω_1^*

同理，　由

$$y=(\lambda-\lambda\varepsilon_0)+\int_{\Omega_2}(\,p(\vec{x}|\omega_1)-\lambda p(\vec{x}|\omega_2))\,d\vec{x}$$

选择满足条件 $p(\vec{x}|\omega_1)-\lambda p(\vec{x}|\omega_2)<0$ 的全体 $\vec{x}$ 作为 Ω_2^*

综上，即：

$$\begin{cases} \text{在}\,\Omega_1\,\text{中} & p(\vec{x}|\omega_1)-\lambda p(\vec{x}|\omega_2)>0 \\ \text{在}\,\Omega_2\,\text{中} & p(\vec{x}|\omega_1)-\lambda p(\vec{x}|\omega_2)<0 \end{cases}$$

于是将其中一类错误概率作为控制量而使另一类错误概率最小的 N-P 判决规则为：

如果 $\dfrac{p(\vec{x}|\omega_1)}{p(\vec{x}|\omega_2)} \gtrless \lambda$ ，则判 $\vec{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$

上式中，λ 是判决阈值。

可以看出，N-P判决规则的形式和最小误判概率准则及最小损失准则的形式相同，只是似然比阈值不同。

这里 λ 是由下列关系式确定：

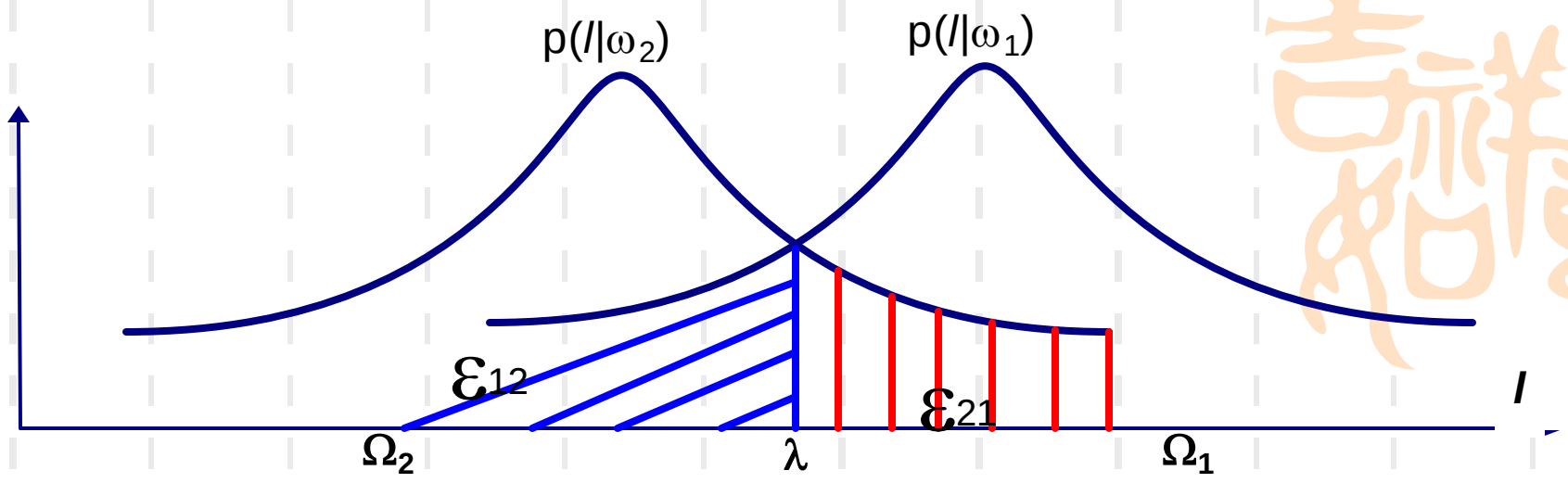
$$\varepsilon_{21}=\int_{\Omega_1} p(\vec{x}|\omega_2)d\vec{x}=\varepsilon_0$$

即适当地选取 λ 以保证使 $\varepsilon_{21}=\varepsilon_0$，因此 λ 的值决定着类域 Ω_1、Ω_2。

为求 λ，令 $l(\vec{x})=\dfrac{p(\vec{x}|\omega_1)}{p(\vec{x}|\omega_2)}$

因当 $l>\lambda$ 时就判 $\vec{x}\in\omega_1$，所以当 ε_0 给定后，拉格朗日乘子 λ 可由式 $\varepsilon_{21}=\int_{\mu(\lambda)}^{+\infty} p(l|\omega_2)dl=\varepsilon_0$ 确定。

这里 $p(l|\omega_2)$ 为似然比 l 的条件概密。



λ的值决定着类域Ω_1、Ω_2，这里的λ是由ε_0所确定的，即适当地选取λ使$\varepsilon_{21}=\varepsilon_0$。为求$\lambda$，令$p(l|\omega_2)$为似然比$l(\vec{x})$在$\vec{x}\in\omega_2$的条件下的概密，因当$l>\lambda$时就判$\vec{x}\in\omega_1$，所以当$\varepsilon_0$给定后，拉格朗日乘子$\lambda$可由式

$$\varepsilon_{21}=\int_{\lambda}^{+\infty}p(l|\omega_2)dl=\varepsilon_0$$

确定。

实际上难于由 $\varepsilon_{21}=\int_{\mu(\lambda)}^{+\infty}p(l|\omega_2)dl=\varepsilon_0$ 求得 λ 的解析显式，通常采用近似的办法。

上式的积分是λ的单调减函数，给出一系列λ值，由上式可以相应的算得一系列 ε_{21} 值，总存在一个λ使 ε_{21} 最接近 ε_0，那个λ值便是我们所求的λ的近似值。当作得足够精细时，λ的近似值的精度就很高。

在具体运用N-P准则时，首先根据给定的控制量 ε_0 计算门限λ，然后运用判决规则进行判决分类。

例：设两类问题中，二维模式均为正态分布，其均值矢量和协方差阵分别为： $\vec{\mu}_1=(-1,0)'$

$\vec{\mu}_2=(1,0)'$ $\Sigma_1=\Sigma_2=\mathrm{I}$，取定 $\varepsilon_{21}=0.04$

试求N-P判决阈值。

解：由公式和给定的条件可算得两类的概密分别为：

$$p(\vec{x}|\omega_1)=\frac{1}{2\pi}\exp\left[-((x_1+1)^2+x_2{}^2)/2\right]$$

$$p(\vec{x}|\omega_2)=\frac{1}{2\pi}\exp\left[-((x_1-1)^2+x_2{}^2)/2\right]$$

由上面二式可以算得：

$$\frac{p(\vec{x}|\omega_1)}{p(\vec{x}|\omega_2)}=\exp\left[-2x_1\right]=\lambda$$

$$\frac{p(\vec{x}|\omega_1)}{p(\vec{x}|\omega_2)} = \exp[-2x_1] = \lambda$$

其为判决界面，上式两边取对数，于是可得判决规则：

$$x_1 \lessgtr -\frac{1}{2}\ln\lambda \quad \Rightarrow \vec{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

由于界面只是x_1的函数，需求$p(\vec{x}|\omega_2)$的边缘密度$p(x_1|\omega_2)$

$$p(x_1|\omega_2) = \int_{-\infty}^{\infty} p(\vec{x}|\omega_2)dx_2 = \int_{-\infty}^{\infty} \frac{1}{2\pi}\exp\left[(-(x_1-1)^2 + x_2^2)/2\right]dx_2$$

$$= \frac{1}{\sqrt{2\pi}}\exp\left[-(x_1-1)^2/2\right]$$

$$p(x_1|\omega_2)=\frac{1}{\sqrt{2\pi}}\exp\left[-(x_1-1)^2/2\right]$$

由上面的判决规则，有：

$$\varepsilon_{21}=\int_{-\infty}^{-\frac{1}{2}\ln\lambda}\frac{1}{\sqrt{2\pi}}\exp\left[\frac{-(x_1-1)^2}{2}\right]dx_1$$

令 $y=x_1-1$，$dy=dx_1$

$$=\int_{-\infty}^{-\frac{1}{2}\ln\lambda-1}\frac{1}{\sqrt{2\pi}}\exp\left[\frac{-y^2}{2}\right]dy$$

$$=\frac{1}{2}-\int_{0}^{\frac{1}{2}\ln\lambda+1}\frac{1}{\sqrt{2\pi}}\exp\left[\frac{-y^2}{2}\right]dy$$

$$=\frac{1}{2}-\int_{0}^{\frac{\sqrt{2}}{4}\ln\lambda+\frac{\sqrt{2}}{2}}\frac{1}{\sqrt{\pi}}\exp\left[-z^2\right]dz$$

有数学手册可查得:

$$\varphi(x)=\int_0^x e^{-t^2}\,dt=\frac{x}{1}-\frac{1}{1!}\cdot\frac{x^3}{3}+\frac{1}{2!}\cdot\frac{x^5}{5}-\cdots$$

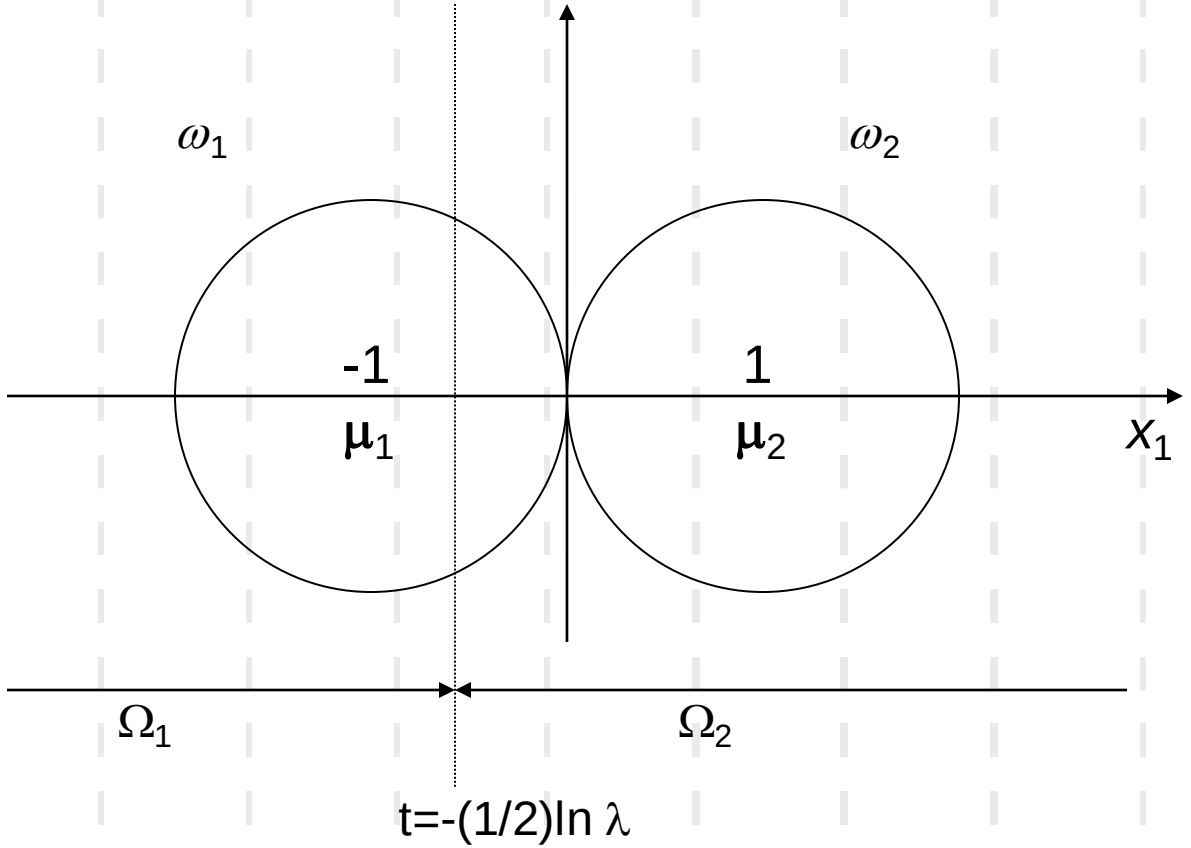
$$\varepsilon_{21}=\frac{1}{2}-\varphi\left(\frac{\sqrt{2}}{4}\ln\lambda+\frac{\sqrt{2}}{2}\right)$$

可算得ε_{21}与 λ 的关系如下表所示:

λ	4	2	1	1/2	1/4
ε_{21}	0.046	0.089	0.0159	0.258	0.378

由设定的$\varepsilon_{21}=0.04$，查上表可得$\lambda=4$，对应的$-(\ln\lambda)/2=-0.693$，从而得此问题的判决规则为:

若 $x_1 \lessgtr -0.693$，则判 $\vec{x}\in\begin{cases}\omega_1\\ \omega_2\end{cases}$



本章主要介绍了贝叶斯统计决策理论为基础的贝叶斯分类方法，其中包括了**最小误判概率、最小损失准则**等，依据这些准则设计的分类器，从理论上讲是最优的性能，即分类的错误率或风险在所有可能的分类器中为最小，因此经常被用来作为衡量其他分类器设计方法优劣的标准。

由于正态分布在**物理上**的合理性和**数学上**的计算简便性，我们详细介绍了贝叶斯分类方法在正态分布下的几种特殊情形，导出了其对应的判决函数、决策面方程及相应的几何描述。

下面我们简单回顾一下本章所学的几种贝叶斯决策准则：

1、最小误判概率准则

两类时： 如果 $l_{12}(\vec{x})=\dfrac{p(\vec{x}|\omega_1)}{p(\vec{x}|\omega_2)}\gtrless\dfrac{P(\omega_2)}{P(\omega_1)}$ 则判 $\vec{x}\in\begin{cases}\omega_1\\ \omega_2\end{cases}$

多类时： 如果 $P(\omega_i|\vec{x})=\max_j\left[P(\omega_j|\vec{x})\right]$ 判 $\vec{x}\in\omega_i$

2、最小损失准则

两类时： 如果 $\dfrac{p(\vec{x}|\omega_1)}{p(\vec{x}|\omega_2)} \gtrless \dfrac{P(\omega_2)(\lambda_{21}-\lambda_{22})}{P(\omega_1)(\lambda_{12}-\lambda_{11})}$ 则判 $\vec{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$

多类时： 计算 $R(\alpha_j|\vec{x}) = \sum_{i=1}^{c} \lambda_{ij} P(\omega_i|\vec{x}) \quad j = 1,2,\cdots c$

若 $R(\alpha_l|\vec{x}) = \min_j R(\alpha_j|\vec{x})$ 则判 $\vec{x} \in \omega_l$

3、含拒绝判决的最小损失准则

两类时：

$$\frac{p(\vec{x}|\omega_1)}{p(\vec{x}|\omega_2)} \geq \frac{P(\omega_2)(1-t)}{P(\omega_1)t}$$

$$\frac{p(\vec{x}|\omega_1)}{p(\vec{x}|\omega_2)} \leq \frac{P(\omega_2)t}{P(\omega_1)(1-t)}$$

$$\frac{P(\omega_2)t}{P(\omega_1)(1-t)} < \frac{p(\vec{x}|\omega_1)}{p(\vec{x}|\omega_2)} < \frac{P(\omega_2)(1-t)}{P(\omega_1)t}$$

$\vec{x} \in \omega_1$

$\vec{x} \in \omega_2$

拒判

其中：

$$t = \frac{\lambda_r - \lambda_c}{\lambda_e - \lambda_c}$$

- λ_r —— **拒判损失**
- λ_e —— **误判损失**
- λ_c —— **正确判决损失**

3、含拒绝判决的最小损失准则

多类时：计算 $R(\alpha_j|\vec{x})=\sum_{i=1}^{c}\lambda_{ij}P(\omega_i|\vec{x})\quad j=1,2,\cdots c,c+1$

若 $R(\alpha_l|\vec{x})=\min_j R(\alpha_j|\vec{x})$ **则判** $\vec{x}\in\omega_l$

其中 α_{c+1} **为拒绝判决。**

4、最小最大损失准则

如果 $\frac{p(\vec{x}|\omega_1)}{p(\vec{x}|\omega_2)} \gtrless \frac{(\lambda_{21}-\lambda_{22})(1-P^*(\omega_1))}{(\lambda_{12}-\lambda_{11})P^*(\omega_1)}$ 则判 $\vec{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$

$P^*(\omega_1)$ 是如何获得的？

$P^*(\omega_2)$

让$P(\omega_1)$从0 逐渐变化到1，按最小损失准则算出最小平均损失，即取各类条件平均损失的最小者。

由此可得出R—P（ω_1）曲线，最大的R对应的P（ω_1）就是 $P^*(\omega_1)$

5、N-P准则

如果 $\frac{p(\vec{x}|\omega_1)}{p(\vec{x}|\omega_2)} \gtrless \lambda$ ，则判 $\vec{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$

λ是如何获得的？

由 $\varepsilon_{21} = \int_{\mu(\lambda)}^{+\infty} p(l|\omega_2)dl = \varepsilon_0$ 固定ε_0反求λ

例：在军事目标识别中，假定有灌木丛和坦克两种类型，它们的先验概率分别是0.7和0.3，损失函数如下表所示，其中，类型ω_1和ω_2分别表示灌木和坦克，判决$\alpha_1=\omega_1$，$\alpha_2=\omega_2$，α_3表示拒绝判决。现在做了四次试验，获得四个样本的类概率密度如下：

$P(x|\omega_1)$:0.1, 0.15, 0.3, 0.6,　$P(x|\omega_2)$:0.8, 0.7, 0.55, 0.3

问：

（1）用最小误判概率准则，判断四个样本各属哪一个类型。

（2）假定只考虑前两种情况，试用最小损失准则判断四个样本各属于哪一个类型。

损失 \ 类型 / 判决	ω_1	ω_2
α_1	2.5	2.0
α_2	4.0	1.0
α_3	1.5	1.5

（3）把拒绝判决考虑在内，重新考核四次试验的结果。

答　求出四个样本两类的似然比。

：

$$l_{12}=\frac{P(x\mid\omega_1)}{P(x\mid\omega_2)}=(\frac{0.1}{0.8},\frac{0.15}{0.7},\frac{0.3}{0.55},\frac{0.6}{0.3})=(0.125,0.214,0.545,2)$$

最小误判概率准则时的阈值：

$$\theta_{12}=\frac{P(\omega_2)}{P(\omega_1)}=\frac{0.3}{0.7}=0.429$$

（1）因此按最小误判概率准则判决时，第一、第二样本属于第二类即坦克，第三、第四属于第一类即灌木丛。

（2） 按最小损失准则判决

$$l_{12}=\frac{P(x|\omega_1)}{P(x|\omega_2)}=(\frac{0.1}{0.8},\frac{0.15}{0.7},\frac{0.3}{0.55},\frac{0.6}{0.3})=(0.125,0.214,0.545,2)$$

最小损失准则时的阈值：

$$\theta_{12}=\frac{P(\omega_2)(\lambda_{21}-\lambda_{22})}{P(\omega_1)(\lambda_{12}-\lambda_{11})}=\frac{0.3(2-1)}{0.7(4-2.5)}=0.286$$

因此按最小损失准则判决时，第一、第二样本属于第二类即坦克，第三、第四属于第一类即灌木丛。

（3） 带拒绝的最小损失准则判决

$$R_1(\vec{x}) = \left[\lambda_{11} p(\vec{x}|\omega_1)P(\omega_1) + \lambda_{21} p(\vec{x}|\omega_2)P(\omega_2)\right] / p(\vec{x})$$

$$R_2(\vec{x}) = \left[\lambda_{12} p(\vec{x}|\omega_1)P(\omega_1) + \lambda_{22} p(\vec{x}|\omega_2)P(\omega_2)\right] / p(\vec{x})$$

$$R_3(\vec{x}) = \left[\lambda_{13} p(\vec{x}|\omega_1)P(\omega_1) + \lambda_{23} p(\vec{x}|\omega_2)P(\omega_2)\right] / p(\vec{x})$$

由于是比较大小，可忽略p(x)，即只需计算

$$R_1(\vec{x}) = \left[\lambda_{11} p(\vec{x}|\omega_1)P(\omega_1) + \lambda_{21} p(\vec{x}|\omega_2)P(\omega_2)\right]$$

$$R_2(\vec{x}) = \left[\lambda_{12} p(\vec{x}|\omega_1)P(\omega_1) + \lambda_{22} p(\vec{x}|\omega_2)P(\omega_2)\right]$$

$$R_3(\vec{x}) = \left[\lambda_{13} p(\vec{x}|\omega_1)P(\omega_1) + \lambda_{23} p(\vec{x}|\omega_2)P(\omega_2)\right]$$

（3） 带拒绝的最小损失准则判决

$$R_1(\vec{x}) = \left[\lambda_{11} p(\vec{x}|\omega_1)P(\omega_1) + \lambda_{21} p(\vec{x}|\omega_2)P(\omega_2)\right]$$

=2.5*0.7*(0.1,0.15,0.3,0.6)+2.0*0.3*(0.8,0.7,0.55,0.3)

=(0.655, 0.683, 0.855, 1.23)

$$R_2(\vec{x}) = \left[\lambda_{12} p(\vec{x}|\omega_1)P(\omega_1) + \lambda_{22} p(\vec{x}|\omega_2)P(\omega_2)\right]$$

=4.0*0.7*(0.1,0.15,0.3,0.6)+1.0*0.3*(0.8,0.7,0.55,0.3)

=(0.52, 0.63, 1.005, 1.77)

$$R_3(\vec{x}) = \left[\lambda_{13} p(\vec{x}|\omega_1)P(\omega_1) + \lambda_{23} p(\vec{x}|\omega_2)P(\omega_2)\right]$$

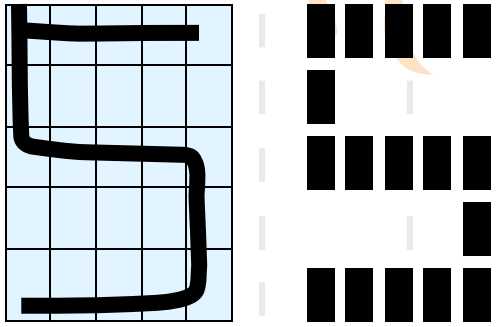
=1.5*0.7*(0.1,0.15,0.3,0.6)+1.5*0.3*(0.8,0.7,0.55,0.3)

=(0.465, 0.473, 0.563, 0.765)

因此第一、 第二、第三、第四样本均拒判。

利用贝叶斯分类器实现手写数字识别的例子

对每个手写的数字样品，按NxN方式划分，共有25份，如图所示。



1、理论基础

对每一份内的象素个数进行累加统计,除以每一份内的象素总数,设定阈值T=0.05,若每一份内的象素占有率大于T则对应的特征值为1,否则为0.

2、实现步骤

（1）先计算先验概率 $P(\omega_i)$

$$P(\omega_i) \approx N_i / N$$

$P(\omega_i)$类别为数字i的先验概率

N_i数字i的样品数

N为样品总数

利用贝叶斯分类器实现手写数字识别的例子

2、实现步骤

（2）计算 $P_j(\omega_i)$ ，再计算类条件概率$P(X \mid \omega_i)$

$$P_j(\omega_i) = \left(\sum_{\substack{k=0 \\ X \in \omega_i}}^{N_i} x_{kj} + 1\right) / (N_i + 2) \qquad \begin{array}{l} i = 0,1,\cdots,9 \\ j = 0,1,\cdots,24 \end{array}$$

$P_j(\omega_i)$ 表示样品X属于ω_i类条件下，X的第j个分量为1的概率估计值。

$$P(x_j = 1 \mid X \in \omega_i) = P_j(\omega_i)$$

$$P(x_j = 0 \mid X \in \omega_i) = 1 - P_j(\omega_i)$$

利用贝叶斯分类器实现手写数字识别的例子

2、实现步骤

$$P(X \mid \omega_i) = P[X = (x_0, x_1, x_2, \cdots, x_{24}) \mid X \in \omega_i]$$

$$= \prod_{j=0}^{24} P(x_j = \alpha \mid X \in \omega_i) \qquad i = 0,1,\cdots,9$$

其中α=0或1

（3）利用贝叶斯公式求后验概率

$$P(\omega_i \mid X) = \frac{P(\omega_i)P(X \mid \omega_i)}{P(\omega_0)P(X \mid \omega_0) + P(\omega_1)P(X \mid \omega_1) + \cdots + P(\omega_9)P(X \mid \omega_9)}$$

$$i = 0,1,\cdots,9$$

利用贝叶斯分类器实现手写数字识别的例子

2、实现步骤

（4）后验概率的最大值的类别（0—9）就是手写数字的所属类别。

利用贝叶斯分类器实现手写数字识别的例子

2、实现步骤

（1）先计算先验概率 $P(\omega_i)$

（2）计算 $P_j(\omega_i)$ ，再计算类条件概率 $P(X \mid \omega_i)$

（3）利用贝叶斯公式求后验概率

（4）后验概率的最大值的类别（0—9）就是手写数字的所属类别。

3.7 分类器设计

1. 判别函数和决策面

定义：用于表达决策规则的函数称为判别函数。

决策面：将划分决策域的边界面称为决策面。可用数学表达式表达为决策面方程。

对两类最小错误率 Bayes 决策规则，有 4 种表达方式：

（1） $P(w_1 \mid x) \underset{<}{\overset{>}{}} P(w_2 \mid x)$，对应 $x \in \begin{cases} w_1 \\ w_2 \end{cases}$

（2） $\rho(x \mid w_1) \cdot P(w_1) \underset{<}{\overset{>}{}} \rho(x \mid w_2) \cdot P(w_2)$，对应样本 $x \in \begin{cases} w_1 \\ w_2 \end{cases}$

（3） $l(x) = \dfrac{\rho(x \mid w_1)}{\rho(x \mid w_2)} \underset{<}{\overset{>}{}} \dfrac{P(w_2)}{P(w_1)}$，对应 $x \in \begin{cases} w_1 \\ w_2 \end{cases}$

（4） $\ln \rho(x \mid w_1) + \ln P(w_1) \underset{<}{\overset{>}{}} \ln \rho(x \mid w_2) + \ln P(w_2)$，对应 $x \in \begin{cases} w_1 \\ w_2 \end{cases}$

对多类别情况：

$\Omega = \{w_1, w_2, \cdots\cdots, w_c\}$，c 类

$x = \{x_1, x_2, \cdots\cdots, x_d\}^T$

同样存在 4 个决策规则：

（1） $P(w_i \mid x) > P(w_j \mid x), j = 1,2,\cdots\cdots,c,$ 且$j \neq i$， 对应样本 $x \in w_i$

（2） $\rho(x \mid w_i) \cdot P(w_i) > \rho(x \mid w_j) \cdot P(w_j), j = 1,2,\cdots\cdots,c,$ 且$j \neq i$， 对应样本 $x \in w_i$

（3） $l(x) = \dfrac{\rho(x \mid w_i)}{\rho(x \mid w_j)} > \dfrac{P(w_i)}{P(w_j)}, j = 1,2,\cdots\cdots,c,$ 且$j \neq i$， 对应 $x \in w_i$

（4） $\ln \rho(x \mid w_i) + \ln P(w_i) > \ln \rho(x \mid w_j) + \ln P(w_j), j = 1,2,\cdots\cdots,c,$ 且$j \neq i$，对应 $x \in w_i$

对于最小风险 Bayes 决策，同样有：

$$R(\alpha_1 \mid x) \underset{<}{\overset{>}{}} R(\alpha_2 \mid x)\text{，对应样本 } x \in \begin{cases} w_2 \\ w_1 \end{cases}$$

推广到多维情况：

$$R(\alpha_i \mid x) \underset{<}{\overset{>}{}} R(\alpha_j \mid x), i, j = 1,2,\cdots\cdots,c,\text{且} i \neq j\text{，} \quad x \in \begin{cases} w_j \\ w_i \end{cases}$$

2. 多类判别函数和分类器

（1）判别函数

一般定义，一组函数 $g_i(x), i=1,2,\cdots\cdots,c$，表示多类决策规则：

$$g_i(x) > g_j(x) \Rightarrow \text{对于} i \neq j, \text{样本} x \in w_i$$

对于多类情况，$g_i(x)$ 可以定义为：

① $g_i(x) = P(w_i \mid x)$

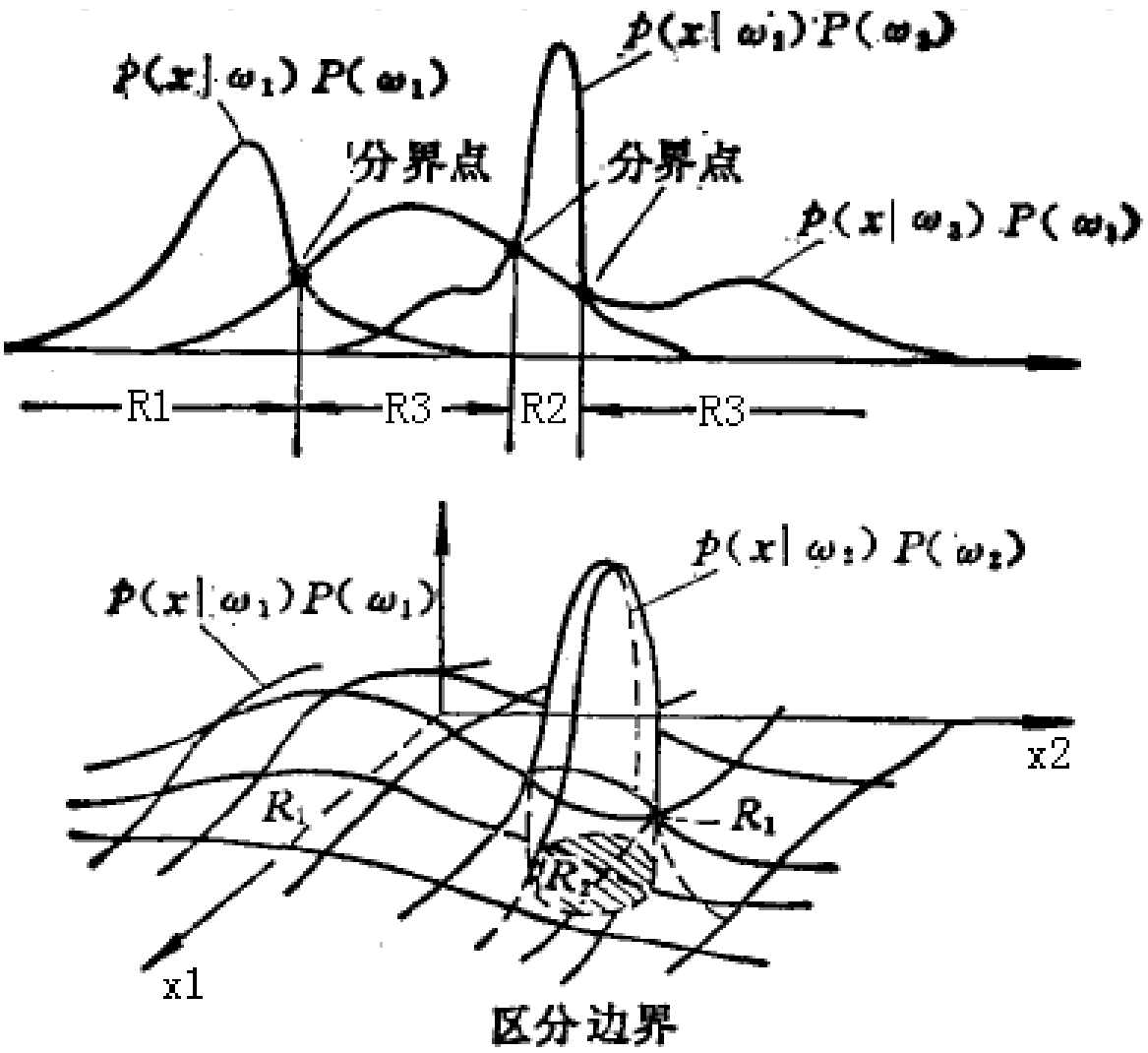
② $g_i(x) = \rho(x \mid w_i) \cdot P(w_i)$

③ $g_i(x) = \ln \rho(x \mid w_i) + \ln P(w_i)$

（2） 决策面方程

各决策域 R_i 被决策面所分割，这些决策面是特征空间种的超曲面，对于相邻的两个决策域 R_i 和 R_j，分割它们的决策面方程应满足：

$g_i(x) = g_j(x)$　　（显然它们在决策面上相邻决策函数相等）

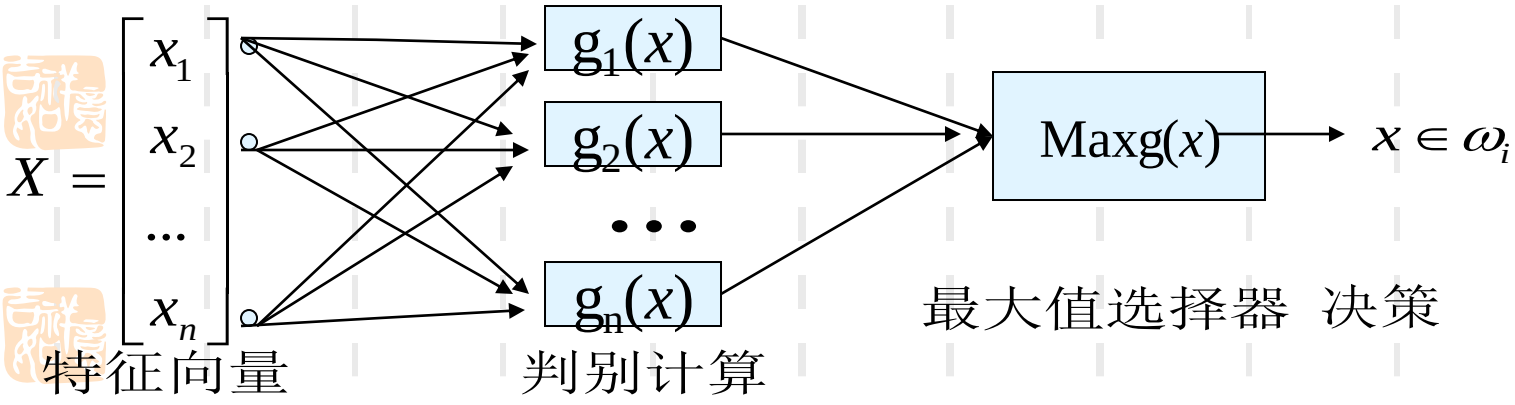


随机模式的判决函数及决策面
（a）一维特征空间三个类型区域
（b）二维特征空间两个类型区域

指x的维数：
在一维空间，对应的是点
在二维空间，对应的是曲线
在三维空间，对应的是曲面
在四维空间对应的是超曲面

3）分类器设计

功能：先设计出c个判别函数 $g_i(x)$，再从中选出对应于判决函数为最大值的类作为决策结果。



3．两类情况

（1）判决函数 $g(x)=g_1(x)-g_2(x)$

决策规则：

$$\begin{cases} g(x)>0, \text{决策} x\in w_1 \\ g(x)<0, \text{决策} x\in w_2 \end{cases}$$

具体来说，可定义：

① $g(x)=\rho(w_1\mid x)-\rho(w_2\mid x)$

② $g(x)=\rho(x\mid w_1)\cdot P(w_1)-\rho(x\mid w_2)\cdot P(w_2)$

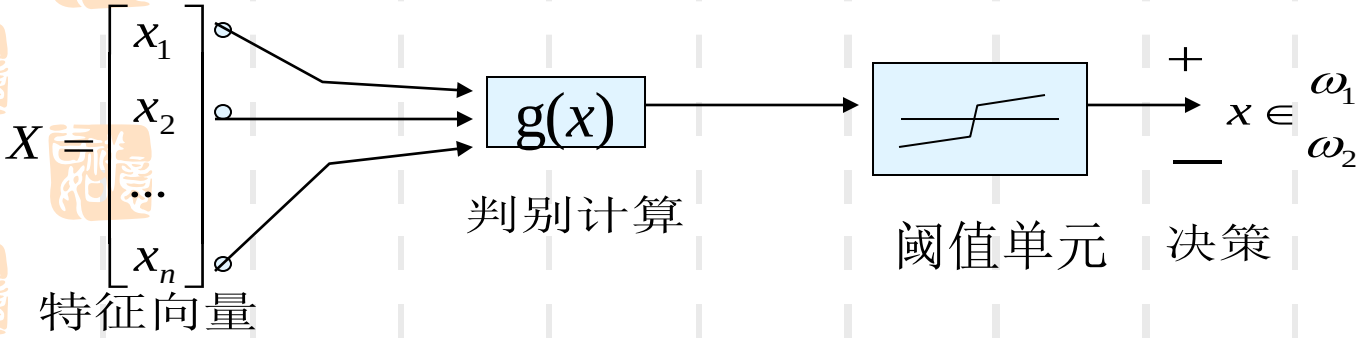
③ $g(x)=\ln\dfrac{\rho(x\mid w_1)}{\rho(x\mid w_2)}-\ln\dfrac{P(w_1)}{P(w_2)}$

（2）决策面方程 $g(x)=0$

也可以表示为：$\rho(x\mid w_1)\cdot P(w_1)-\rho(x\mid w_2)\cdot P(w_2)=0$

（3）分类器设计

通过计算，根据计算结果的符号将x分类。



例 3：对例 1，例 2 分别写出其判决函数和决策面方程。

解：对于例 1，用判决函数：$g(x)=\rho(x|w_1)\cdot P(w_1)-\rho(x|w_2)\cdot P(w_2)$ 得到对应的判决函数为：

$$g(x)=0.995\rho(x|w_1)-0.005\rho(x|w_2)\,;$$

决策面方程为：$0.995\rho(x|w_1)-0.005\rho(x|w_2)=0$ 。

对例 2，判决函数定义为：

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

其中：$g_1(x)=1-R(a_1=w_1|x)$，$g_2(x)=1-R(a_2=w_2|x)$，带入上式：

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

$$=R(a_2=w_2|x)-R(a_1=w_1|x)$$

$$R(\alpha_1 = w_1 \mid x) = \sum_{j=1}^{2} L(\alpha_1 \mid w_j) \cdot P(w_j \mid x) = 0.5 * P(w_1 \mid x) + 6 * P(w_2 \mid x)$$

$$R(\alpha_2 = w_2 \mid x) = \sum_{j=1}^{2} L(\alpha_2 \mid w_j) \cdot P(w_j \mid x) = 2 * P(w_1 \mid x) + 0.5 * P(w_2 \mid x)$$

$$g(x) = 1.5 * P(w_1 \mid x) - 5.5P(w_2 \mid x)$$

决策面方程为：$1.5 * P(w_1 \mid x) - 5.5P(w_2 \mid x) = 0$ 。

3.8 正态分布决策理论

一、正态分布判别函数

1、为什么采用正态分布：

a、正态分布在物理上是合理的、广泛的。

b、正态分布数学上简单，$N(\mu,\sigma^2)$ 只有均值和方差两个参数。

2、单变量正态分布：

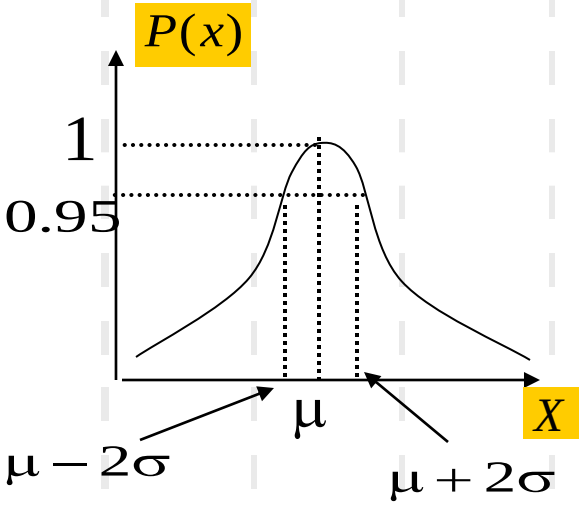
$$P(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]=N(\mu,\sigma^2)$$

$$\text{其中：}\mu=E(x)=\int_{-\infty}^{\infty}xP(x)dx,\text{(均值或数学期望)}$$

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

概率密度函数应满足下列关系：

$$\begin{cases}P(x)\geq 0,(-\infty<x<\infty)\\ \int_{-\infty}^{\infty}P(x)dx=1\end{cases}$$



3、（多变量）多维正态分布

（1）函数形式：

$$P(x)=\frac{1}{(2\pi)^{n/2}\left|\sum\right|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)^T\sum{}^{-1}(x-\mu)\right]$$

其中：$x=(x_1,x_2,...,x_n)^T$，n维特征向量

$\mu=(\mu_1,\mu_2,...,\mu_n)^T$，$n$维均值向量

$\sum$为$n\times n$维协方差矩阵，$\sum^{-1}$为$\sum$的逆阵，$\left|\sum\right|$为$\sum$的行列式

$$\mu_i=E(x_i)=\int_{-\infty}^{\infty}x_iP(x_i)dx_i$$

$$\sum=E\left[(x-\mu)(x-\mu)^T\right]$$

$$=E\left\{\begin{bmatrix}(x_1-\mu_1)\\ \cdots\cdots\\ (x_n-\mu_n)\end{bmatrix}[(x_1-\mu_1),\ldots,(x_n-\mu_n)]\right\}$$

$$=E\left\{\begin{bmatrix}(x_1-\mu_1)(x_1-\mu_1)\ldots(x_1-\mu_1)(x_n-\mu_n)\\ \cdots\cdots\\ (x_n-\mu_n)(x_1-\mu_1)\ldots(x_n-\mu_n)(x_n-\mu_n)\end{bmatrix}\right\}$$

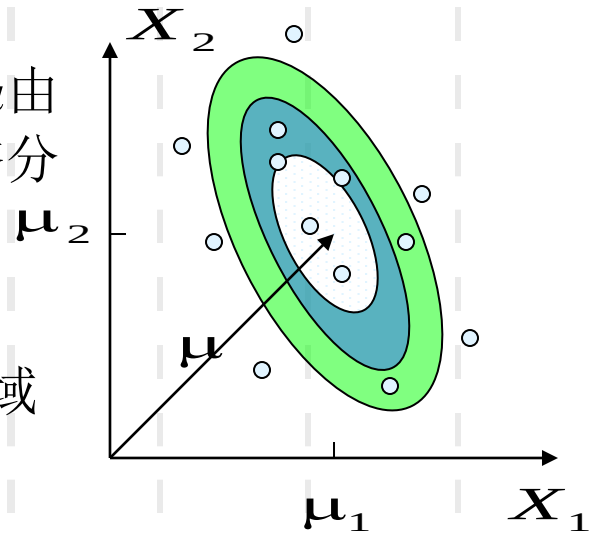
$$= \begin{bmatrix} E[(x_1-\mu_1)(x_1-\mu_1)] \cdots E[(x_1-\mu_1)(x_n-\mu_n)] \\ \cdots\cdots \\ E[(x_n-\mu_n)(x_1-\mu_1)] \cdots E[(x_n-\mu_n)(x_n-\mu_n)] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1n}^2 \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{n1}^2 & \sigma_{n2}^2 & \cdots & \sigma_{nn}^2 \end{bmatrix}, \begin{pmatrix} \text{对角线}\sigma_{ij}^2, i=j\text{是方差} \\ \text{非对角线}\sigma_{ij}^2,\ i\neq j\text{是协方差} \end{pmatrix}$$

(2)、性质：

①、μ与$\sum$对分布起决定作用P(χ)=N(μ, $\sum$), μ由n个分量组成，$\sum$由n(n+1)/2元素组成。∴多维正态分布由n+n(n+1)/2个参数组成。 。

常记为： $\rho(x) \sim N(\mu,\Sigma)$



②、等密度点的轨迹是一个超椭球面。区域中心由μ决定，区域形状由$\sum$决定。

$$(x-\mu)^T \Sigma^{-1}(x-\mu) = \text{常数}$$

$$\gamma^2 = (x-\mu)^T \Sigma^{-1}(x-\mu)$$

③、不相关性等价于独立性。若xi与xj互不相关,则xi与xj一定独立。

不相关：$E(x_i x_j) = E(x_i) \cdot E(x_j)$　独立：$P(x_i, x_j) = P(x_i) \cdot P(x_j)$

证明：根据定义，x_i 和 x_j 的协方差 $\sigma_{ij}^2 = E[(x_i - \mu_i)(x_j - \mu_j)]$

又根据不相关定义 $E(x_i, x_j) = E(x_i) \cdot E(x_j)$ 有：

$$\sigma_{ij}^2 = E[(x_i - \mu_i)(x_j - \mu_j)] = E(x_i - \mu_i) \cdot E(x_j - \mu_j)$$

又：$\mu_i = E(x_i)$，$E[(x_i - \mu_i) = E(x_i) - E(\mu_i) = E(x_i) - \mu_i = 0$

所以：有 $\sigma_{ij}^2 = 0$

协方差矩阵 $\Sigma = \begin{bmatrix} \sigma_{11}^2 & & \\ & \ddots & \\ & & \sigma_{dd}^2 \end{bmatrix}$ 成为对角阵。

可以计算出：$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_{11}^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_{dd}^2} \end{bmatrix}$

$$|\Sigma| = \prod_{i=1}^{d} \sigma_{ii}^2, \quad |\Sigma|^{\frac{1}{2}} = \prod_{i=1}^{d} \sigma_{ii}$$

$$(x-\mu)T \cdot \Sigma^{-1}(x-\mu) = [x_1-\mu_1, \cdots\cdots, x_d-\mu_d] \begin{bmatrix} \frac{1}{\sigma_{11}^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_{dd}^2} \end{bmatrix} \begin{bmatrix} x_1-\mu_1 \\ \vdots \\ x_d-\mu_d \end{bmatrix}$$

$$= \sum_{i=1}^{d} \left(\frac{x_i-\mu_i}{\sigma_{ii}}\right)^2$$

因此，$$\rho(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu)\right]$$

$$= \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_{ii}} \cdot \exp\left\{-\frac{1}{2}\left(\frac{x_i-\mu_i}{\sigma_{ii}}\right)^2\right\} = \prod_{i=1}^{d} \rho(x_i)$$

根据独立性的定义：正态分布随机向量的各分量间互不相关性与相互独立等价。

④边缘分布与条件分布的正态性

不难证明正态随机向量的边缘分布与条件分布仍服从正态分布。

从③证明得出的结论 $\rho(x)$ 表达式，如果 x 用 x_1 表示，有：

$$\rho(x_1)=\frac{1}{\sqrt{2\pi}\sigma_{ii}}\cdot\exp(-\frac{1}{2}(\frac{x_1-\mu_1}{\sigma_{11}})^2)$$

也就是说，边缘分布 $\rho(x_1)$ 服从均值为 μ_1，方差为 σ_{11}^2 的正态分布：

$$\rho(x_1)\sim N(\mu_1,\sigma_{11}^2)$$

同理，$\rho(x_2)\sim N(\mu_2,\sigma_{22}^2)$

另外，条件分布，给定 x_1 的条件下 x_2 的分布：$\rho(x_2\mid x_1)=\rho(x_1,x_2)\Big/\rho(x_1)$

$$\rho(x_1,x_2)=\frac{1}{2\pi\mid\Sigma\mid^{1/2}}\exp\left\{-\frac{1}{2\mid\Sigma\mid}\left[\sigma_{22}^2(x_1-\mu_1)^2+\sigma_{11}^2(x_2-\mu_2)^2-\sigma_{12}^2(x_1-\mu_1)(x_2-\mu_2)\right]\right\}$$

代入上式，$\rho(x_2\mid x_1)$ 服从正态分布，同理 $\rho(x_1\mid x_2)$ 也服从正态分布。

⑤、线性变换的正态性Y=AX，A为线性变换矩阵。若X为正态分布，则Y也是正态分布。

$$\rho(x) \sim N(\mu, \Sigma) \xrightarrow{y = Ax \quad |A| \neq 0} \rho(y) \sim N(A\mu, A\Sigma A^T)$$

⑥、线性组合的正态性

$$y = a^T x \longrightarrow \rho(y) \sim N(a^T \mu, a^T \Sigma a)$$

其中，a与x同维。

- 判别函数：类条件概率密度用正态来表示：

$$g(x) = P(x|\omega_i)P(\omega_i)$$

$$= \frac{1}{(2\pi)^{n/2}\left|\sum_i\right|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^T\sum_i^{-1}(x-\mu_i)\right]P(\omega_i)$$

$$= \ln\left\{\frac{1}{(2\pi)^{n/2}\left|\sum_i\right|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^T\sum_i^{-1}(x-\mu_i)\right]\right\} + \ln P(\omega_i)$$

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

- 决策面方程$g(x_i) - g(x_j) = 0$

$$g(x_i) - g(x_j) = -\frac{1}{2}((x-\mu_i)\Sigma_i^{-1}(x-\mu_i) - (x-\mu_j)\Sigma_j^{-1}(x-\mu_j) + \ln|\Sigma_i| - \ln|\Sigma_j|) + \ln\frac{P(\omega_i)}{P(\omega_j)}$$

$$= 0$$

- 二、最小错误率(Bayes)分类器：从最小错误率这个角度来分析Bayes 分类器

1.第一种情况：各个特征统计独立，且同方差情况。(最简单情况)

即：$\sum_i = \sigma^2 I = \begin{bmatrix} \sigma_{11}^2 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \sigma_{nn}^2 \end{bmatrix}$，只有方差，协方差为零。

- 判别函数：

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \sum_i^{-1}(x-\mu_i) - \frac{n}{2}\ln 2\pi - \frac{1}{2}\ln\left|\sum_i\right| + \ln P(\omega_i)$$

因为$\sum_i = \sigma^2 I, \sum_i^{-1} = \left(1/\sigma^2\right)I, \left|\sum_i\right| = \sigma^2 I, \frac{n}{2}\ln 2\pi$都与$i$无关。

对分类无影响。

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \sum_i^{-1}(x-\mu_i) + \ln P(\omega_i)$$

$$= -\frac{\|x-\mu_i\|^2}{2\sigma^2} + \ln P(\omega_i), \text{其中} \|x-\mu_i\|^2 = (x-\mu_i)^T(x-\mu_i)$$

- 如果M类先验概率相等：

$$P(\omega_1) = P(\omega_2) = \ldots = P(\omega_m)$$

$$\therefore g(x) = -\|x-\mu_i\|^2 / 2\sigma^2, \text{(欧氏距离)}$$

- 最小距离分类器：未知x与μ_i相减，找最近的μ_i把x归类

$(x-\mu_i)^T(x-\mu_i)=x^Tx-2\mu_i x+\mu_i^T\mu_i$，因为二次项$x^Tx$与$i$无关

$\therefore$简化可得：$g_i(x)=w_i^Tx+w_{i0}$,(线性判别函数)

其中：$w_i=\frac{1}{\sigma^2}\mu_i, w_{i0}=-\frac{1}{2\sigma^2}\mu_i^T\mu_i+\ln P(\omega_i)$

判别规则：$g_i(x)=w_i^Tx+w_{i0}=\max\limits_{1\le w\le M}\{w_j^Tx+w_{j0}\}\Rightarrow x\in\omega_i$

对于二类情况 $g(x)=g_2(x)-g_1(x)$

$$=\frac{1}{\sigma^2}(\mu_2-\mu_1)^T x+\frac{1}{2\sigma^2}(\mu_1{}^T\mu_1-\mu_2{}^T\mu_2)\begin{matrix}<\\>\end{matrix}\ln\frac{P(\omega_1)}{P(\omega_2)}\Rightarrow x\in\begin{matrix}\omega_1\\\omega_2\end{matrix}$$

决策面方程：$g_i(x)-g_j(x)=0$

$$W(x-x_0)=0$$

其中 $W=\mu_i-\mu_j$

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

- 讨论：

二类情况下 $\omega_i = \omega_1\omega_2$

(a)：因为 $\sum_i = \delta^2 I$，协方差为零。所以等概率面是一个圆形。

(b)：因 W 与 $(x - x_0)$ 点积为0，因此分界面 H 与 W 垂直

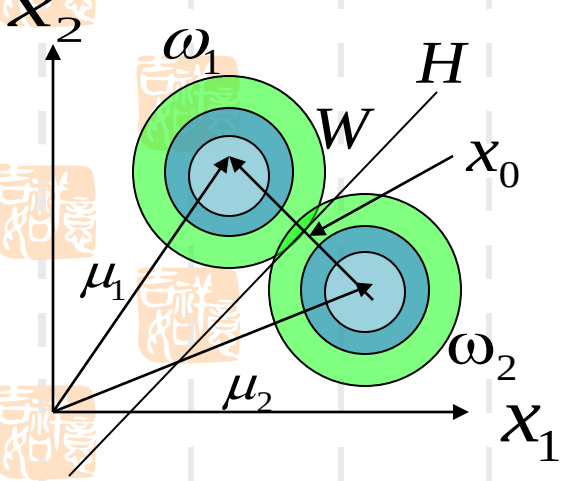
又因为 $W = \mu_i - \mu_j = \mu_1 - \mu_2$，所以 W 与 $\mu_1 - \mu_2$ 同相(同方向)

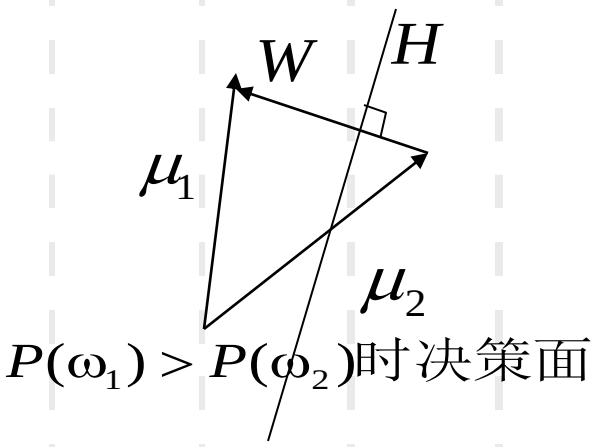
$\therefore$ 决策面 H 垂直于 μ 的联线。

(c)：如果先验概率相等 $P(\omega_1) = P(\omega_2)$，$H$ 通过 μ 联线的中点。

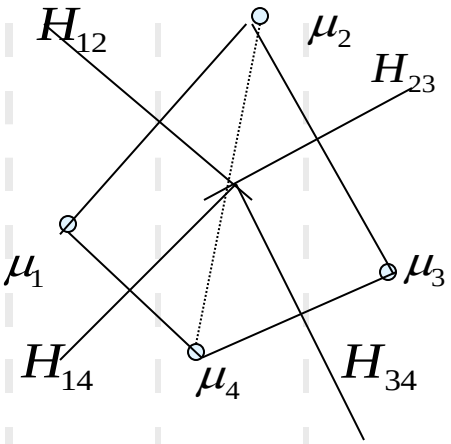
否则就是 $P(\omega_1) \neq P(\omega_2)$，$H$ 离开先验概率大的一类。

(d)：对多类情况，用各类的均值联线的垂直线作为界面。





$P(\omega_1) > P(\omega_2)$ 时决策面



- **2、第二种情况：** $\Sigma_i = \Sigma$相等，即各类协方差相等。

  因为$\sum_1 = \sum_2 = \ldots = \sum_M = \sum$与$i$无关

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

  若先验概率相等$P(\omega_1) = P(\omega_2) = P(\omega_3) = \ldots = P(\omega_i)$

  $$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \sum^{-1}(x-\mu_i) = r^2,(\text{马氏距离})$$

- 未知x，把x与各类均值相减，把x归于最近一类。最小距离分类器。

  把$(x-\mu_i)^T \sum^{-1}(x-\mu_i)$展开；$x^T \sum^{-1} x$与$i$无关。

  $$\therefore g_i(x) = W_i^T x + w_{i0},(\text{线性函数})$$

  其中$W_i = \sum^{-1}\mu_i$

  $$w_{i0} = -\frac{1}{2}\mu_i^T \sum^{-1}\mu_i + \ln P(\omega_i)$$

- 决策规则：$g_i(x)=W_i^T x+w_{i0}=\max\limits_{1\le j\le M}\{W_j^T x+w_{j0}\}\Rightarrow x\in\omega_i$

  对于二类情况$g(x)=g_2(x)-g_1(x)=(\mu_2-\mu_1)\sum{}^{-1}x$

  $$+\frac{1}{2}\sum{}^{-1}(\mu_1^T\mu_1-\mu_2^T\mu_2)\underset{>}{\overset{<}{}}\ln\frac{P(\omega_1)}{P(\omega_2)}\Rightarrow x\in\begin{matrix}\omega_1\\ \omega_2\end{matrix}$$

- 决策界面：若ω_i与ω_j相邻$\therefore g_i(x)-g_j(x)=0$

  $$\therefore W^T(x-x_0)=0,\text{其中}W=\sum{}^{-1}(\mu_i-\mu_j)。$$

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

讨论：针对ω_1, ω_2二类情况，如图：

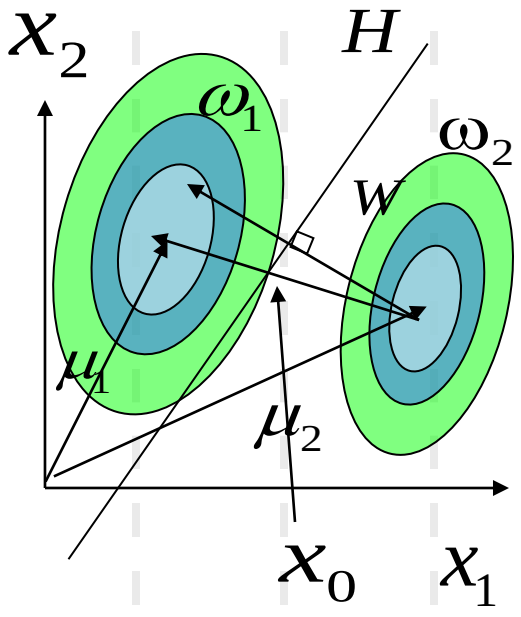
(a):因为$\sum_i = \sum \neq \delta I$,所以等概率面是椭圆，长轴由$\sum_i$本征值决定

(b):因为W与$(x-x_0)$点积为0,所以W与$(x-x_0)$正交,$\therefore H$通过x_0点。

(c):因为$W=\sum^{-1}(\mu_i-\mu_j)$;所以W与$(\mu_i-\mu_j)$不同相;$\therefore H$不垂直于μ值联线。

(d):若各类先验概率相等，则$x_0=\frac{1}{2}(\mu_i+\mu_j)$,则$H$通过均值联线中点;

否则H离开先验概率大的一类。



- **3、第三种情况**(一般情况)：Σ_i为任意，各类协方差矩阵不等，二次项$x^T \Sigma_i x$与i有关。所以判别函数为二次型函数。

  - 判别函数：$g_i(x) = x^T \overline{W}_i x + W_i^T x + w_{i0}$，其中$\overline{W}_i = -\frac{1}{2}\sum_i^{-1}$，$(n \times n$矩阵$)$

    $W_i = \sum_i^{-1} \mu_i$（n维列向量），$w_{i0} = -\frac{1}{2}\mu_i^T \sum_i^{-1} \mu_i - \frac{1}{2}\ln\left|\sum_i\right| + \ln P(\omega_i)$

  - 决策规则：$g_i(x) = x^T \overline{W}_i x + W_i^T x + w_{i0}$

    $$= \max_{1 \le j \le M} \left\{ x^T \overline{W}_j x + W_j^T x + w_{j0} \right\} \Rightarrow x \in \omega_i$$

对于二类情况$g(x) = g_2(x) - g_1(x) = \frac{1}{2}(x-\mu_1)^T \sum_1^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_2)^T \sum_2^{-1}(x-\mu_2)$

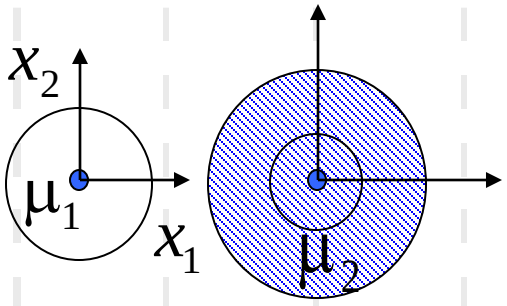
$$+\frac{1}{2}\ln\frac{\left|\sum_1\right|}{\left|\sum_2\right|} \begin{matrix} < \\ > \end{matrix} \ln\frac{P(\omega_1)}{P(\omega_2)} \Rightarrow x \in \begin{matrix} \omega_1 \\ \omega_2 \end{matrix}$$

决策面方程：$g_i(x)-g_j(x)=0$

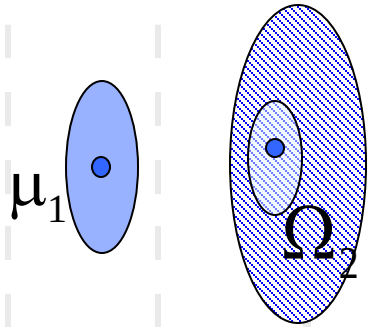
下面看一下决策界面的 各种图形：

对于二类问题，条件： a：二类情况$\omega_1\omega_2$； b：x_1x_2为条件独立；

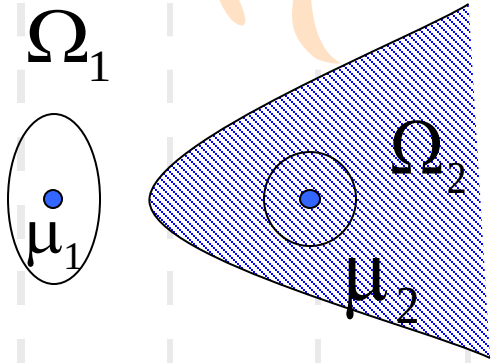
c：先验概率相等。



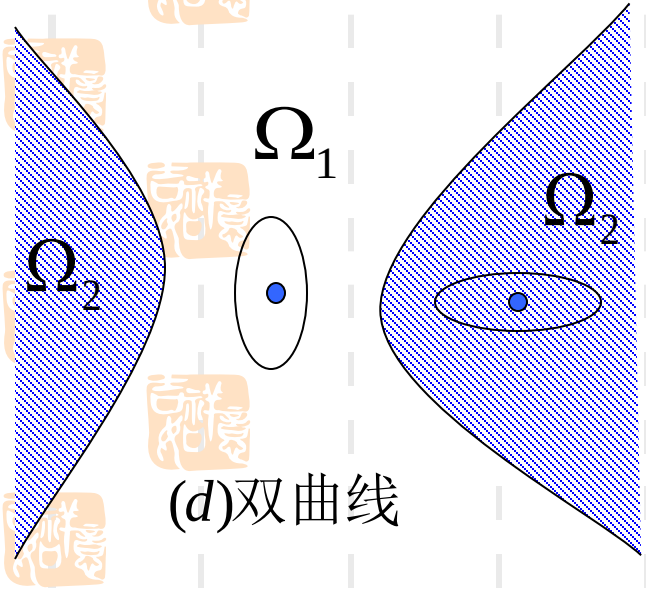
(*a*)圆



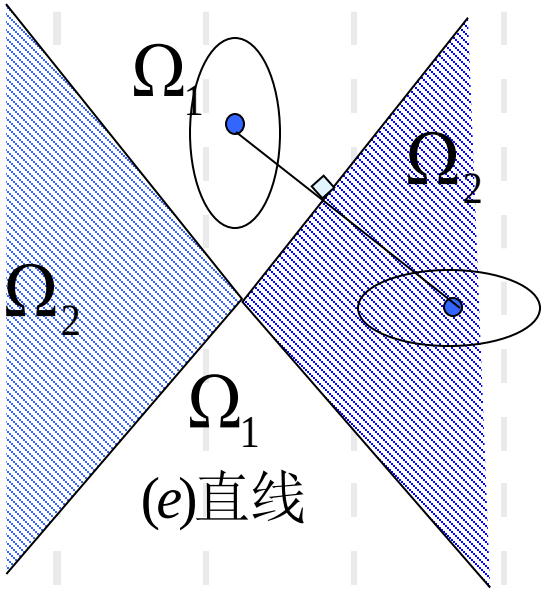
(*b*)椭圆



(*c*)抛物线

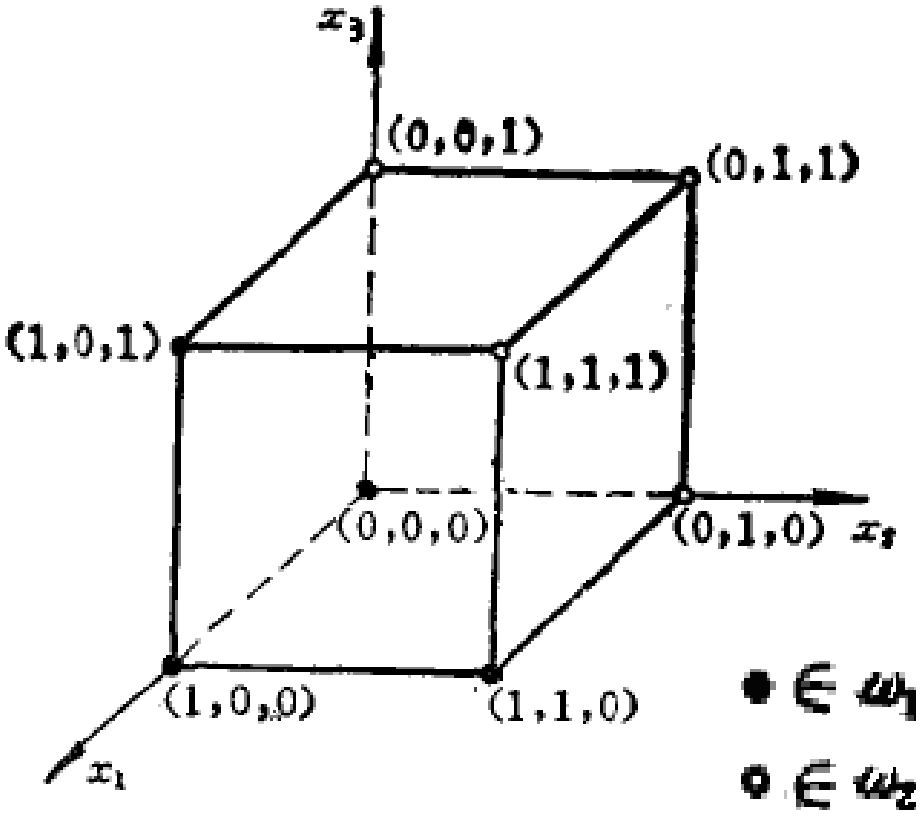


(*d*)双曲线



(*e*)直线

例 1：设在三维特征空间里，两类的类概率密度是正态分布的，分别在两个类型中获得 4 个样本，位于一个单位立方体的顶点上，如下图。两类的先验概率相等，试确定两类之间的决策面及相应的类型区域 R_1 和 R_2 。



解：w_1和w_2表示两个类型，由图可知，两个类型的样本：

w_1：$(0,0,0)^T$，$(1,0,0)^T$，$(1,1,0)^T$，$(1,0,1)^T$

w_2：$(0,1,0)^T$，$(0,0,1)^T$，$(0,1,1)^T$，$(1,1,1)^T$

用各类样本的算术平均值近似代替各类均值向量，也就是：

$$\mu_i \approx \frac{1}{N_i}\sum_{k=1} x_{ik}$$

N_i为w_i中的样本数，x_{ik}表示w_i的第k个样本。

协方差矩阵由其定义求得：

$$\Sigma_i = R_i - \mu_i \mu_j^T = \frac{1}{N_i}\sum_{k=1}^{N_i} x_{ik} \cdot x_{ik}^T - \mu_i \mu_i^T$$

式中R_i为类w_i的自相关函数。

由题中所给条件：$i=1,2$，$N_1 = N_2 = 4$

有：$\mu_1 = \frac{1}{4}(3,1,1)^T$，$\mu_2 = \frac{1}{4}(1,3,3)^T$

$$\mu_1\mu_1{}^T = (\frac{3}{4},\frac{1}{4},\frac{1}{4})^T \cdot (\frac{3}{4},\frac{1}{4},\frac{1}{4}) = \begin{pmatrix}\frac{3}{4}\\ \frac{1}{4}\\ \frac{1}{4}\end{pmatrix} \cdot (\frac{3}{4},\frac{1}{4},\frac{1}{4}) = \frac{1}{16}\begin{bmatrix}9 & 3 & 3\\ 3 & 1 & 1\\ 3 & 1 & 1\end{bmatrix}$$

$$\mu_2\mu_2{}^T = \begin{pmatrix}\frac{1}{4}\\ \frac{3}{4}\\ \frac{3}{4}\end{pmatrix} \cdot (\frac{1}{4},\frac{3}{4},\frac{3}{4}) = \frac{1}{16}\begin{bmatrix}1 & 3 & 3\\ 3 & 9 & 9\\ 3 & 9 & 9\end{bmatrix}$$

$$R_1 = \frac{1}{4}\left[\begin{pmatrix}0\\0\\0\end{pmatrix} \cdot (0,0,0) + \begin{pmatrix}1\\0\\0\end{pmatrix} \cdot (1,0,0) + \begin{pmatrix}1\\1\\0\end{pmatrix} \cdot (1,1,0) + \begin{pmatrix}1\\0\\1\end{pmatrix} \cdot (1,0,1)\right] = \frac{1}{4}\begin{bmatrix}3 & 1 & 1\\ 1 & 1 & 0\\ 1 & 0 & 1\end{bmatrix}$$

同理： $R_2 = \frac{1}{4}\begin{bmatrix}1 & 1 & 1\\ 1 & 3 & 2\\ 1 & 2 & 2\end{bmatrix}$

$$\Sigma_1 = R_1 - \mu_1\mu_1^T = \frac{1}{4}\begin{bmatrix}3 & 1 & 1\\ 1 & 1 & 0\\ 1 & 0 & 1\end{bmatrix} - \frac{1}{16}\begin{bmatrix}9 & 3 & 3\\ 3 & 1 & 1\\ 3 & 1 & 1\end{bmatrix} = \frac{1}{16}\begin{bmatrix}3 & 1 & 1\\ 1 & 3 & -1\\ 1 & -1 & 3\end{bmatrix}$$

$$\Sigma_2 = \frac{1}{16}\begin{bmatrix}3 & 1 & 1\\ 1 & 3 & -1\\ 1 & -1 & 3\end{bmatrix}$$

因此，$\Sigma_1 = \Sigma_2 = \Sigma$ 符合情况二。用情况二的公式确定决策面。

$$\Sigma^{-1} = 4\begin{bmatrix}2 & -1 & -1\\ -1 & 2 & 1\\ -1 & 1 & 2\end{bmatrix}$$

决策面为 $g_1(x) - g_2(x) = 0 \Rightarrow w^T(x - x_0) = 0$ ，$w = \Sigma^{-1}(\mu_1 - \mu_2)$ ，$x_0 = \frac{1}{2}(\mu_1 + \mu_2)$，先验概率相等 $P(w_1) = P(w_2)$

$$w = \Sigma^{-1}(\mu_1 - \mu_2) = 4\begin{bmatrix}2 & -1 & -1\\ -1 & 2 & 1\\ -1 & 1 & 2\end{bmatrix} \cdot \frac{1}{4}\begin{bmatrix}2\\ -2\\ -2\end{bmatrix} = \begin{bmatrix}8\\ -8\\ -8\end{bmatrix}$$

$$x_0 = \frac{1}{2}(\mu_1 + \mu_2) = \frac{1}{2}(1,1,1)^T$$

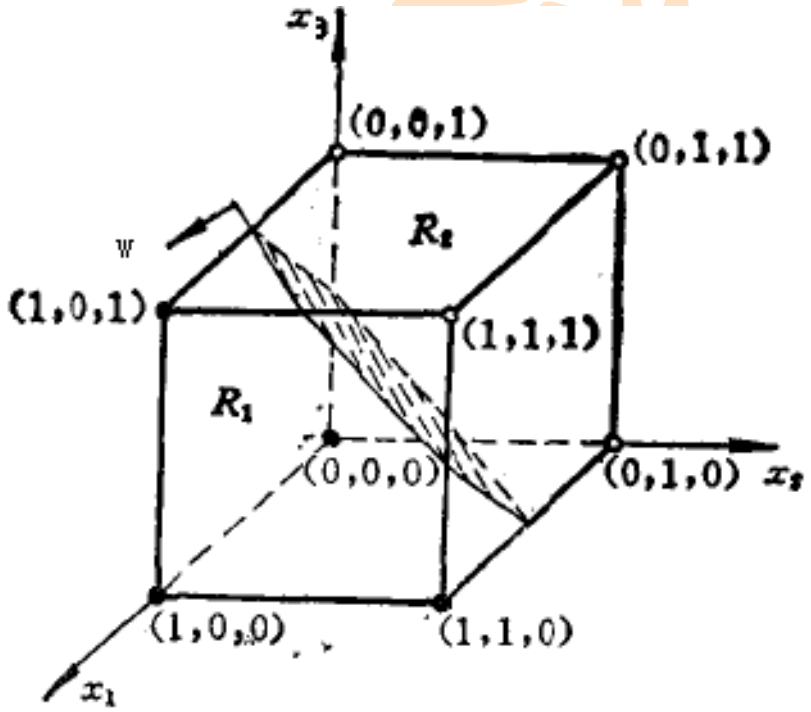
决策方程： $w^T(x-x_0)=0$

$$(8,-8,-8)\begin{bmatrix} x_1-\frac{1}{2} \\ x_2-\frac{1}{2} \\ x_3-\frac{1}{2} \end{bmatrix}=0$$

也就是： $8(x_1-\frac{1}{2})-8(x_2-\frac{1}{2})-8(x_3-\frac{1}{2})=0$

$$8x_1-8x_2-8x_3+4=0$$

$$2x_1-2x_2-2x_3+1=0$$ 如右图所示。



w指向的一侧为正，是w1的区域R1，负向的一侧为w2。

3.9 参数估计与非参数估计

- 参数估计与监督训练
- 参数估计理论
- 非参数估计理论

3.9.1 参数估计与监督训练

贝叶斯分类器中只要知道先验概率$P(\omega_i)$，条件概率$P(x/\omega_i)$就可以设计分类器了。现在来研究如何用已知训练样本的信息去估计$P(\omega_i)$,$P(x/\omega_i)$, $P(\omega_i/x)$

一．参数估计与非参数估计

参数估计：先假定研究的问题具有某种数学模型，如正态分布，二项分布，再用已知类别的训练样本估计里面的参数。

非参数估计：不假定数学模型，直接用已知类别的训练样本的先验知识直接估计数学模型。

二．监督训练与无监督训练

监督训练：在已知类别的样本指导下的训练。

无监督训练：不知道样本类别，只知道样本的某些信息去估计。

- 参数估计
  - 监督的
    - 最大似然估计
      （参数未知，但确定）
    - 贝叶斯估计
      （参数本身也是随机量）
  - 非监督的
- 非参数估计

3.9.2 参数估计的基本概念

- 统计量：样本集$K=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$的某种函数$f(K)$
- 参数空间：总体分布的未知参数 θ 所有可能取值组成的集合（Θ）

- 点估计、估计量和估计值：

  θ的估计量$\hat{\theta} = d(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = d(K)$是样本集的函数，它对样本集的一次实现称为估计值

- 区间估计：

3.9.2.1 最大似然估计

- Maximum Likelihood (ML)估计
  - 估计的参数 θ 是确定而未知的，而Bayes估计方法则视 θ 为随机变量。
  - 样本集可按类别分开，不同类别的密度函数的参数分别用各类的样本集来训练。
  - 概率密度函数的形式已知，参数未知，为了描述概率密度函数 $p(x/\omega_i)$ 与参数 θ 的依赖关系，用 $p(x/\omega_i, \theta)$ 表示。
  - 对应于不同类别的参数在函数上是独立的
- 独立地按概率密度 $p(x/\theta)$ 抽取样本集 $K=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$，用 K 估计未知参数 θ

- 独立地按概率密度$p(x/\theta)$抽取样本集$K=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$，则有：

$$p(K|\theta)=p(\mathrm{x}_1,\mathrm{x}_2,\ldots,\mathrm{x}_\mathrm{N}|\theta)$$
$$=\prod_{\mathrm{k}=1}^{\mathrm{N}} p(\mathrm{x}_\mathrm{k}|\theta)$$

似然函数：

N个随机变量$\mathrm{x}_1,\mathrm{x}_2,\ldots,\mathrm{x}_\mathrm{N}$的似然函数是N个随机变量的联合密度$l(\theta)=p(K|\theta)=p(\mathrm{x}_1,\mathrm{x}_2,\ldots,\mathrm{x}_\mathrm{N}|\theta)$

使似然函数最大的$\hat{\theta}=d(\mathrm{x}_1,\mathrm{x}_2,\ldots,\mathrm{x}_\mathrm{N})$是样本的函数，称为$\theta$的最大似然估计量

似然函数

- 似然函数：

$$l(\boldsymbol{\theta}) = p(K \mid \boldsymbol{\theta}) = p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \mid \boldsymbol{\theta})$$

$$= \prod_{k=1}^{N} p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

  - 对数(loglarized)似然函数：

$$H(\boldsymbol{\theta}) = \sum_{k=1}^{N} \ln p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

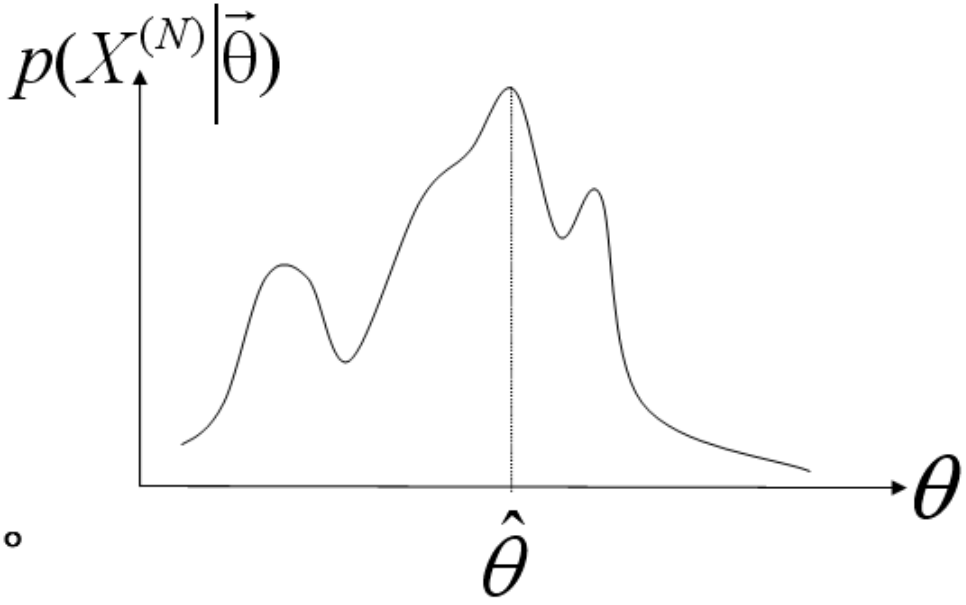
最大似然估计：

最大似然估计的思想是：如果对总体的独立观测中得样本 $X^{(N)}=\{\vec{x}_1,\vec{x}_2,\cdots,\vec{x}_N\}$，则认为 $p(X^{(N)}|\vec{\theta})$ 在点 $X^{(N)}$ 处最可能取得最大值，不同的 $\vec{\theta}$ 使 $p(X^{(N)}|\vec{\theta})$ 是同一概型而不是同一概密，从而它的最大值点 $X^{(N)}$ 也不同，于是有理由选取满足 $p(X^{(N)}|\hat{\vec{\theta}})=\max_{\vec{\theta}}\left[p(X^{(N)}|\vec{\theta})\right]$ 的 $\hat{\vec{\theta}}$ 作为未知参数集的估计值。

最大似然估计(MLE)

(Maximum Likelihood Estimate)

按此方法所得到的 $\vec{\theta}$ 的估值 $\hat{\vec{\theta}}$ 称为最大似然估计(值)，显然 $\vec{\theta}$ 的最大似然估计 $\hat{\vec{\theta}}$ 是 $\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_N$ 的函数。



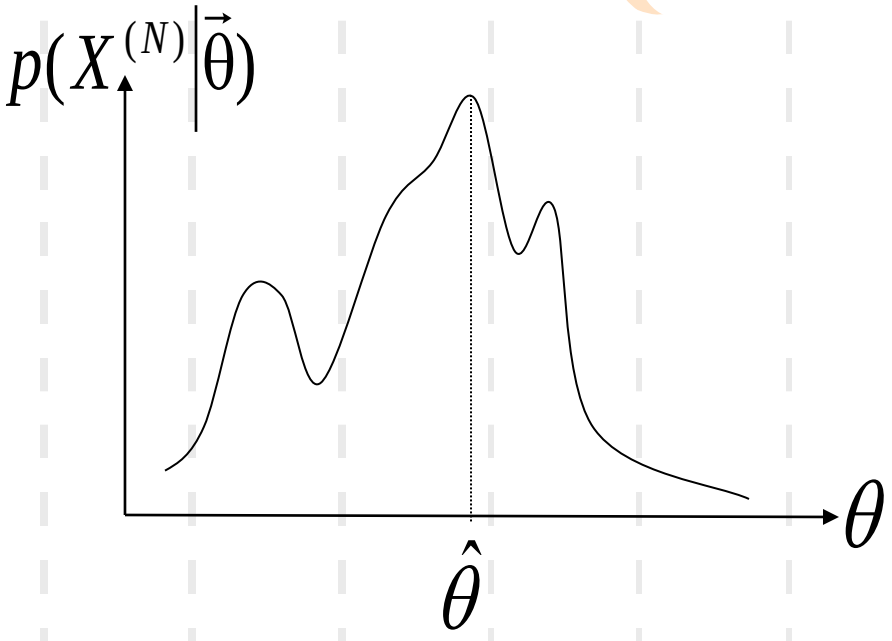
在实际中多是独立取样和经常处理正态变量，而且对数函数是单值单调函数，对数似然函数与似然函数在相同的 $\vec{\theta}$ 处取得最大值。

最大似然估计(MLE)

(Maximum Likelihood Estimate)

在似然函数可微的条件下，

求下面微分方程组的解：



$$\frac{\partial p(X^{(N)}|\vec{\theta})}{\partial \vec{\theta}} = 0$$

或等价地求

$$\frac{\partial \left[\ln p(X^{(N)}|\vec{\theta})\right]}{\partial \vec{\theta}} = \sum_{j=1}^{N} \frac{\partial}{\partial \vec{\theta}} \ln p(\vec{x}_j|\vec{\theta}) = 0$$

作为极值的必要条件。

对数似然方程组

最大似然估计(MLE)

(Maximum Likelihood Estimate)

下面我们以多维正态分布为例进行说明。

（1）假设 Σ 是已知的，未知的只是均值 μ，则：

$$\ln p(x_k \mid \theta) = \frac{-1}{2}\ln((2\pi)^d \bullet |\Sigma|) - \frac{1}{2}(x_k - \mu)^T \Sigma^{-1}(x_k - \mu)$$

$$\frac{\partial \ln p(x_k \mid \theta)}{\partial \mu} = \Sigma^{-1}(x_k - \mu)$$

$$\sum_{k=1}^{N} \Sigma^{-1}(x_k - \mu) = 0 \qquad \hat{\mu} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

最大似然估计(MLE)

(Maximum Likelihood Estimate)

这说明，样本总体的未知均值的最大似然估计就是训练样本的平均值。它的几何解释就是：若把N个样本看成是一群质点，则样本均值便是它们的质心。

（2）一维正态的情况，设 $\theta_1 = \mu, \theta_2 = \sigma^2$ 则：

$$\ln p(x_k \mid \theta) = -\frac{1}{2}\ln((2\pi)\theta_2) - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\frac{\partial \ln p(x_k \mid \theta)}{\partial \theta} = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ \frac{-1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

$$\sum_{k=1}^{N} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 \qquad -\sum_{k=1}^{N} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{N} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0$$

（2）一维正态的情况，设 $\theta_1 = \mu, \theta_2 = \sigma^2$ 则：

$$\sum_{k=1}^{N} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 \qquad -\sum_{k=1}^{N} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{N} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0$$

$$\hat{\mu} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

$$\sigma^2 = \frac{1}{N}\sum_{k=1}^{N} (x_k - \hat{\mu})^2$$

可见，正态分布中的协方差阵 Σ 的最大似然估计量等于N个矩阵的算术平均值。

（3）对于一般的多维正态密度的情况，计算方法完全是类似的。最后的结果是：

$$\hat{\mu} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{k=1}^{N}(x_k - \hat{\mu})(x_k - \hat{\mu})^T$$

可以证明上式的均值是无偏估计，但协方差阵并不是无偏估计，无偏估计是：

$$\hat{\Sigma} = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \hat{\mu})(x_k - \hat{\mu})^T$$

3.9.2.2 贝叶斯估计

- 用一组样本集$K=\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$估计未知参数 θ
- 未知参数 θ视为随机变量，先验分布为$p(\theta)$，而在已知样本集K出现的条件下的后验概率为$p(\theta \mid K)$
- 在贝叶斯风险最小的意义下估计未知参数 θ

贝叶斯决策问题与贝叶斯估计问题

- 贝叶斯决策问题:
  样本$\mathbf{x}$
  决策a_i
  真实状态w_j
  状态空间A是离散空间
  先验概率$P(w_j)$

- 贝叶斯参数估计问题：
  样本集K
  估计量^ θ
  真实参数θ
  参数空间S是连续空间
  参数的先验分布p(θ)

$$R(\alpha_i \mid x) = E[L(\alpha_i \mid w_j)] = \sum_{j=1}^{c} L(\alpha_i \mid w_j) \cdot P(w_j \mid x)$$

贝叶斯(最小风险)估计

- 参数估计的条件风险：给定x条件下，估计量的期望损失

$$R(\hat{\theta}\,|\,\mathbf{x}) = \int_{\Theta} \lambda(\hat{\theta},\theta)\, p(\theta\,|\,\mathbf{x})\, d\theta$$

- 参数估计的风险：估计量的条件风险的期望

$$R = \int_{E^d} R(\hat{\theta}\,|\,\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}$$

- 贝叶斯估计：使风险最小的估计

$$\hat{\theta}_{\mathrm{BE}} = \underset{\hat{\theta}}{\operatorname{argmax}}\, R$$

贝叶斯估计(II)

- 损失函数定义为误差平方： $\lambda(\hat{\theta},\theta)=(\theta-\hat{\theta})^2$

$$
\begin{aligned}
&R(\hat{\theta}\,|\,\mathbf{x})\\
&=\int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta\,|\,\mathbf{x})d\theta\\
&=\int_{\Theta}[\theta-E(\theta\,|\,\mathbf{x})]^2 p(\theta\,|\,\mathbf{x})d\theta+\\
&\qquad\int_{\Theta}[E(\theta\,|\,\mathbf{x})-\hat{\theta}]^2 p(\theta\,|\,\mathbf{x})d\theta
\end{aligned}
$$

定理：如果定义损失函数为误差平方函数，则有：

$$\hat{\theta}_{\mathrm{BE}}=E[\theta\,|\,\mathbf{x}]=\int_{\Theta}\theta p(\theta\,|\,\mathbf{x})d\theta$$

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

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

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

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

$$\int_{\Theta}[\theta-E(\theta\mid x)][E(\theta\mid x)-\hat{\theta}]p(\theta\mid x)d\theta=[E(\theta\mid x)-\hat{\theta}]\int_{\Theta}[\theta-E(\theta\mid x)]p(\theta\mid x)d\theta$$

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

$$=[E(\theta\mid x)-\hat{\theta}][E(\theta\mid x)-E(\theta\mid x)]=0$$

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

$$\hat{\theta}-E(\theta\mid x)=0 \quad\Rightarrow\quad R(\hat{\theta}\mid x)\text{最小}$$

$$\hat{\theta}=E(\theta\mid x)=\int\theta\,p(\theta\mid x)d\theta$$

贝叶斯估计的步骤

1. 确定 θ 的先验分布 $p(\theta)$
2. 由样本集$K=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$求出样本联合分布：$p(K|\theta)$
3. 计算 θ 的后验分布

$$p(\theta|K)=\frac{p(K|\theta)p(\theta)}{\int_{\Theta}p(K|\theta)p(\theta)d\theta}$$

4. 计算贝叶斯估计

$$\hat{\theta}_{\mathrm{BE}}=\int_{\Theta}\theta p(\theta|K)d\theta$$

正态分布参数的贝叶斯估计

单变量情况：

μ 为未知参数

(1) $P(x|\mu) \sim N(\mu, \sigma^2)$

$P(\mu) \sim N(\mu_0, \sigma_0^2)$。

(μ_0 和 σ_0 已知!)

关于均值μ的先验知识

(2)样本联合分布（样本独立抽取）：

$$p(\vec{x}\,|\,\theta=\mu)=\prod_{k=1}^{n}p(\vec{x}_k\,|\,\mu)=(\frac{1}{\sqrt{2\pi}\sigma})^n\exp\{-\frac{1}{2}\sum_{k=1}^{n}(\frac{x_k-\mu}{\sigma})^2\}$$

(3)计算后验分布

$$p(\mu\,|\,D)=\frac{p(\vec{x}\,|\,\mu)p(\mu)}{p(\vec{x})}=\alpha p(\vec{x}\,|\,\mu)p(\mu)$$

$$=\alpha(\frac{1}{\sqrt{2\pi}\sigma})^n\exp\{-\frac{1}{2}\sum_{k=1}^{n}(\frac{x_k-\mu}{\sigma})^2\}\cdot\frac{1}{\sqrt{2\pi}\sigma_0}\exp\{-\frac{1}{2}(\frac{\mu-\mu_0}{\sigma_0})^2\}$$

$$p(\mu\,|\,D)=\alpha'\exp\{-\frac{1}{2}[\sum_{k=1}^{n}(\frac{x_k-\mu}{\sigma})^2+(\frac{\mu-\mu_0}{\sigma_0})^2]\}$$

$$=\alpha''\exp\{-\frac{1}{2}[(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2})\mu^2-2(\frac{1}{\sigma^2}\sum_{k=1}^{n}x_k+\frac{\mu_0}{\sigma_0^2})\mu]\}$$

- 再生概率密度函数

$$P(\mu \mid \mathrm{D}) \sim N(\mu_n, \sigma_n^2)$$

$$p(\mu \mid D) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp[-\frac{1}{2}(\frac{\mu - \mu_n}{\sigma_n})^2]$$

$$= \alpha' \exp\{-\frac{1}{2}(\frac{1}{\sigma_n^2}\mu^2 - 2\frac{\mu_n}{\sigma_n^2}\mu + \frac{\mu_n^2}{\sigma_n^2})\}$$

$$= \alpha'' \exp\{-\frac{1}{2}(\frac{1}{\sigma_n^2}\mu^2 - 2\frac{\mu_n}{\sigma_n^2}\mu)\}$$

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}$$

$$\frac{\mu_n}{\sigma_n^2} = \frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}$$

设 $\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k$

则有 $\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0$

$\sigma_n^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2}$

(4)计算贝叶斯估计

$$\hat{\mu} = E[\mu \mid D] = \int_{\Theta} \mu p(\mu \mid D) d\mu = \mu_n$$

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

$$\sigma_n^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2}$$

结果讨论：

1）再生密度均值μ_n 是样本均值 $\hat{\mu}_n$ 和先验均值的线性组合。

2）若先验方差不为0，当N→∞，μ_n 主要由样本均值确定；当先验方差等于0，不论样本数多大，都由先验均值确定。

3）再生方差随样本数增加而减小。

最大似然估计与贝叶斯估计的比较

- 参数的解释不同：
  - ML：待估参数具有确定值，它的估计量才是随机的
  - Bayes：待估参数服从某种分布的随机变量。
- 利用的信息不同
  - ML：只利用样本信息
  - Bayes：要求事先提供一个参数的先验分布，即人们对有关参数的主观认识，是非样本信息。在参数估计中它们与样本信息一起被利用。
- 选择参数估计量的准则不同：
  - ML：最大似然为准则求解参数估计量
  - Bayes：要构造一个损失函数并以损失函数最小化为准则

贝叶斯学习

- 贝叶斯学习：利用θ的先验分布 $p(\theta)$及样本提供的信息求出θ的后验分布$p(\theta|K)$，然后直接求类条件概率密度$p(\mathbf{x}|K)$

方法

$$p(\mathbf{x}\,|\,K)=\int p(\mathbf{x},\theta\,|\,K)d\theta=\int p(\mathbf{x}\,|\,\theta)\,p(\theta\,|\,K)d\theta$$

$$p(K^N\,|\,\theta)=p(K_N\,|\,\theta)*p(K^{N-1}\,|\,\theta)$$

$$p(\theta\,|\,K^N)=\frac{p(K_N\,|\,\theta)*p(\theta\,|\,K^{N-1})}{\int p(K_N\,|\,\theta)*p(\theta\,|\,K^{N-1})d\theta}$$

参数的贝叶斯学习性质

$$p(\theta\,|\,K^{N\to\infty})=\delta(\hat{\theta}-\theta)$$

以高斯分布为例：

当观察一个样本时，N=1就会有一个 μ 的估计值的修正值

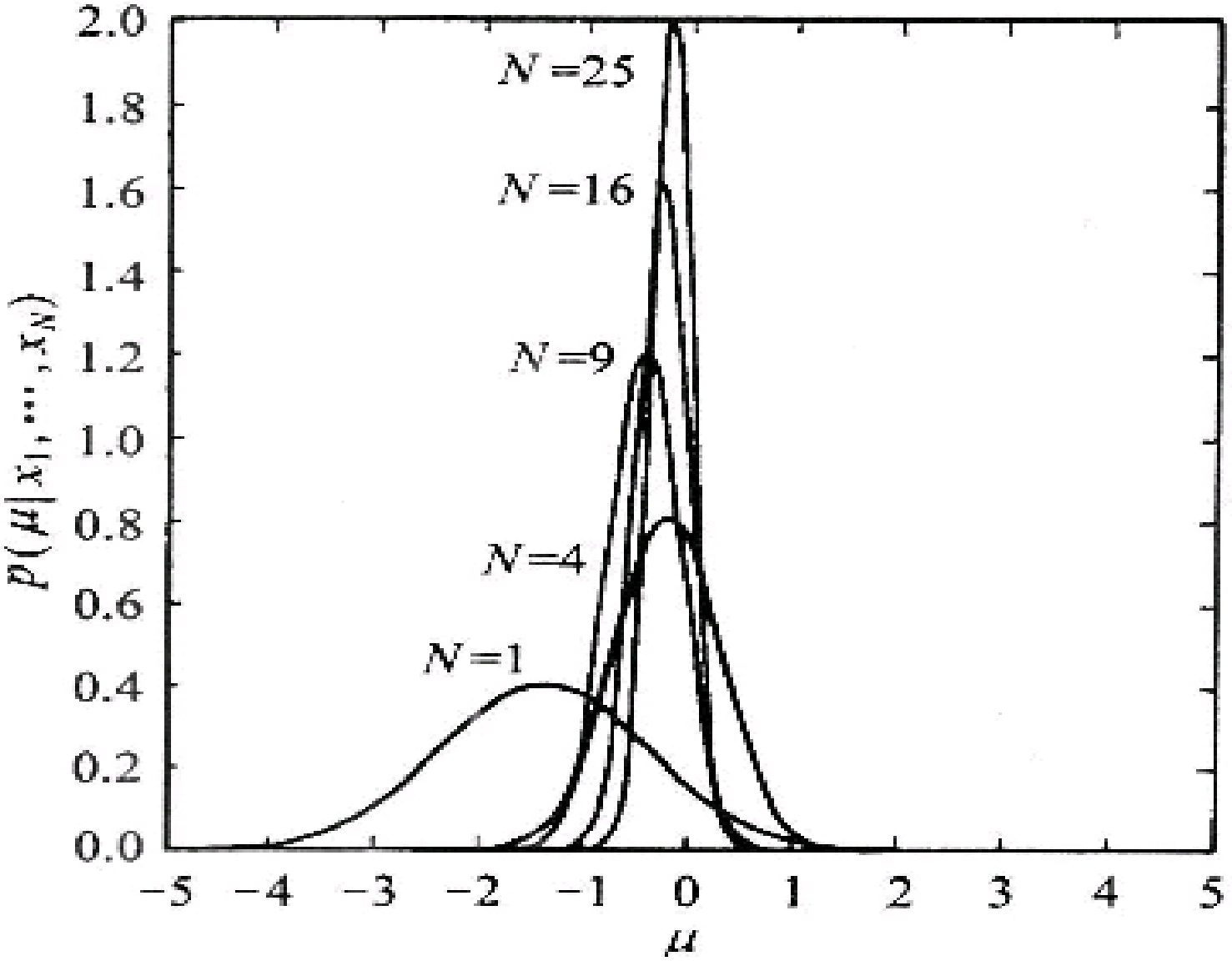
当观察N=4时，对 μ 进行修正，向真正的 μ 靠近

当观察N=9时，对 μ 进行修正，向真正的 μ 靠的更近

当N↑, μ_N就反映了观察到N个样本后对 μ 的最好推测，而σ_N^2反映了这种推测的不确定性, N↑, $\sigma_N^2 \downarrow$, σ_N^2 随观察样本增加而单调减小，且当$N \to \infty$， $\sigma_N^2 \to 0$

当N↑，$P(\mu|x^i)$越来越尖峰突起

$N \to \infty$, $P(\mu|x^i) \to \sigma$函数，这个过程成为贝叶斯学习。


2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
$p(\mu|x_1,\cdots,x_N)$
$N=25$
$N=16$
$N=9$
$N=4$
$N=1$
−5
−4
−3
−2
−1
0
1
2
3
4
5
μ

贝叶斯学习步骤

（1）确定未知参数 $\vec{\theta}$ 的先验概率分布 $p(\vec{\theta})$

（2）由训练样本集 $D=\{\vec{x}_1,\vec{x}_2,...,\vec{x}_n\}$ 及公式

$$p(\vec{x}\mid\vec{\theta})=\prod_{k=1}^{n}p(\vec{x}_k\mid\vec{\theta})$$

求得 $p(\vec{x}\mid\vec{\theta})$

（3）由 $$p(\vec{\theta}\mid D)=\frac{p(\vec{x}\mid\vec{\theta})p(\vec{\theta})}{\int p(\vec{x}\mid\vec{\theta})p(\vec{\theta})d\vec{\theta}}=\frac{p(\vec{x}\mid\vec{\theta})p(\vec{\theta})}{p(\vec{x})}$$

计算后验概率密度 $p(\vec{\theta}\mid D)$

（4）由 $$p(\vec{x}\mid D)=\int_{\Theta}p(\vec{x},\vec{\theta}\mid D)d\vec{\theta}=\int_{\Theta}p(\vec{x}\mid\vec{\theta},D)p(\vec{\theta}\mid D)d\vec{\theta}$$

求得类条件概率密度 $p(\vec{x}\mid\omega_i)$

样本对应于某一类

类概率密度的估计

在求出u的后验概率P(μ |x^i)后，可以直接利用式

$$P(x|x^i)=\int P(x|\theta)\cdot P(\theta|x^i)d\theta$$

推断类条件概率密度。

即P(x|x^i)＝ P(x|ω$_i$，x^i)

(1)一维正态：已知σ^2， μ 未知

∵ μ 的后验概率为

$$\left.\begin{aligned} P(\theta|x^i)=P(\mu|x^i)&=\frac{1}{\sqrt{2\pi}\,\sigma_N}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2\right] \\ P(x|\mu)&=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \end{aligned}\right\}$$ 服从正态分布

代入$P(x \mid x^i) = \int P(x \mid \theta) \cdot P(\theta \mid x^i) d\theta = \int P(x \mid \mu) \cdot P(\mu \mid x^i) d\mu$

$$= \int \frac{1}{\sqrt{2\pi}\,\sigma} \exp[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2] \frac{1}{\sqrt{2\pi}\,\sigma_N} \exp[-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2] d\mu$$

$$= \frac{1}{2\pi\sigma\ \sigma_N} \exp[-\frac{1}{2}\left(\frac{x-\mu_N}{\sigma_N^2+\sigma^2}\right)^2] \int \exp[-\frac{1}{2}\frac{\sigma_N^2+\sigma^2}{\sigma_N^2 \cdot \sigma^2}\left(\mu - \frac{\sigma_N^2 x + \sigma^2 \mu_N}{\sigma_N^2+\sigma^2}\right)^2] d\mu$$

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

$= N(\mu_N, \sigma_N^2 + \sigma^2)$为正态函数

- 结论：

①把第i类的先验概率$P(\omega_i)$与第i类概率密度$P(x|x^i)$相乘可以得到第i类的后验概率$P(\omega_i/x)$，根据后验概率可以分类。

②对于正态分布$P(x|x^i)$，用样本估计出来的μ_N代替原来的μ用$\sigma_N^2+\sigma^2$代替原来的方差σ^2即可。

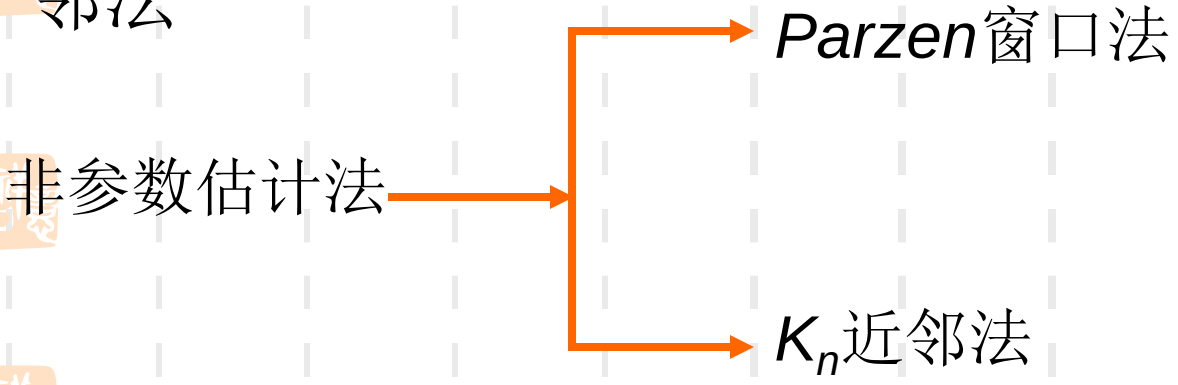
③把估计值μ_N作为μ的实际值，那么使方差由原来的σ^2变为$\sigma_N^2+\sigma^2$,使方差增大

3.9.3 非参数估计

- 原理
  - 在实际应用中，类概率密度函数形式已知的条件并不一定成立，特别是多峰的概率分布，用普通函数难以拟合，这就需要用非参数估计技术。
  - 非参数估计的原理是：不需获取类类概率密度的函数形式，而是直接利用学习样本估计特征空间任意点的类概率密度的值。
  - 即直接由学习样本来直接设计分类器。

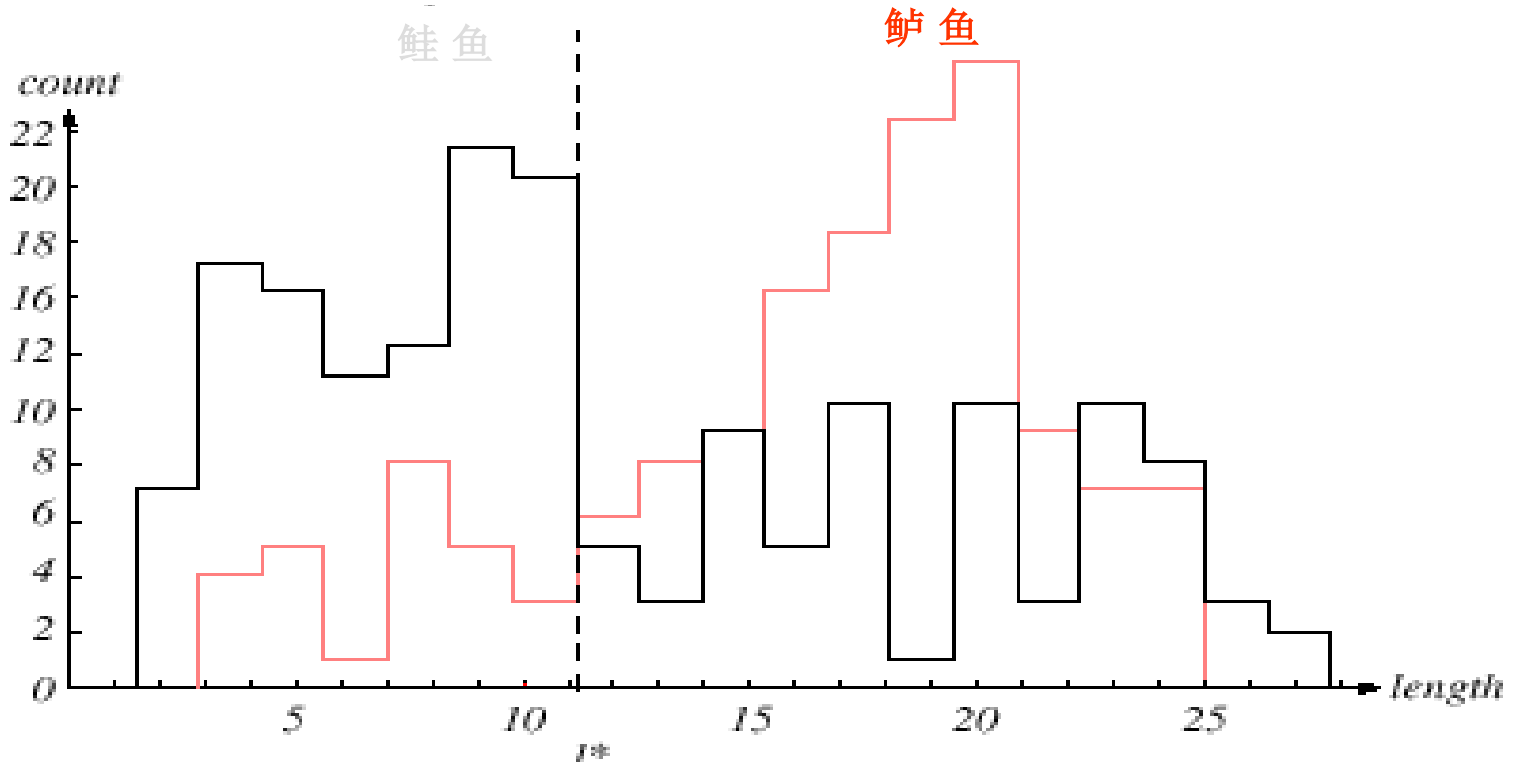
非参数估计

- 方法
  - 直接由学习样本估计类概率密度$P(X/\omega_i)$，*Parzen*窗口法
  - 直接由学习样本估计后验概率$P(\omega_i/X)$，K_n近邻法



非参数估计

- 思路



非参数估计

- 思路
  - 用已知类别的学习样本在x处出现的频度来近似$P(X/\omega_i)$，即：

$$p(X) \approx \frac{k/n}{v}$$

其中：v为包含X点的区域

一维	v为一直线
二维	圆
三维	球
四维	超球体

非参数估计

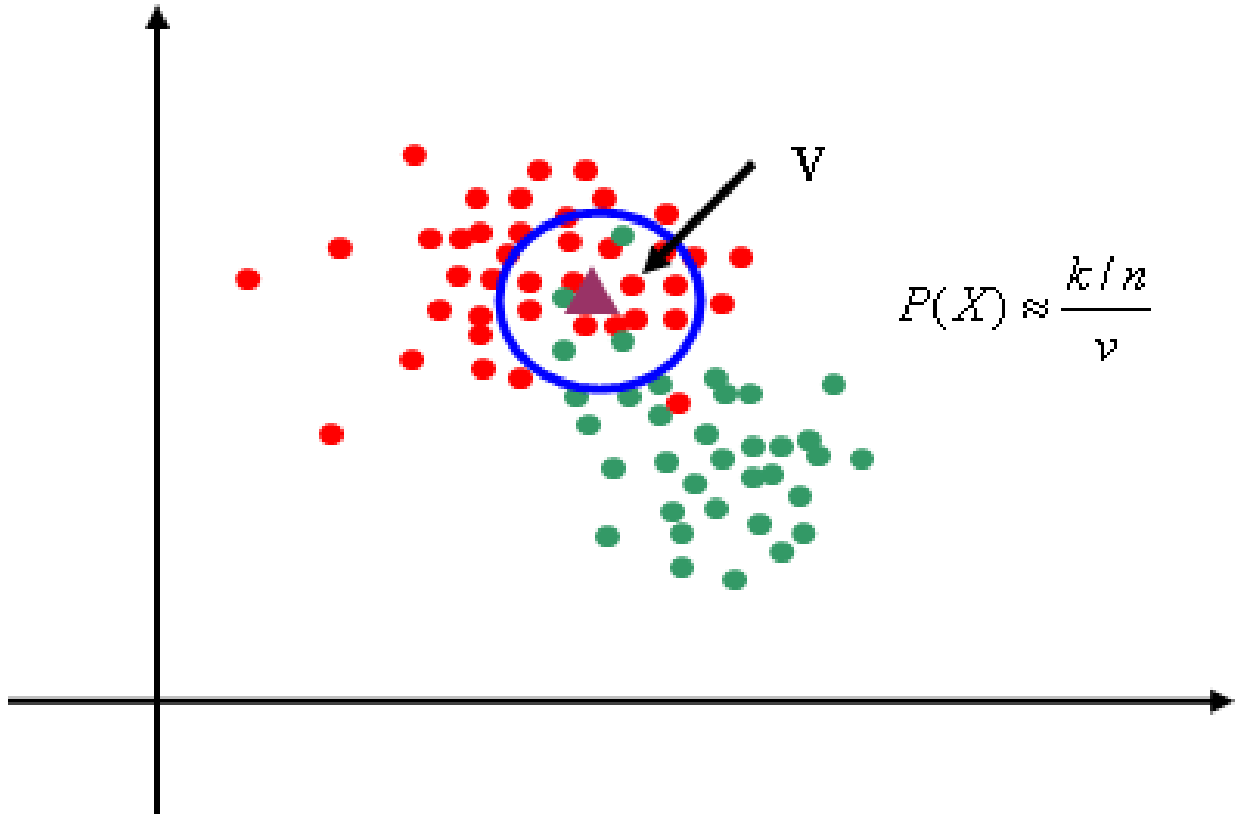
K为n个样本中落入体积v的样本数。

故：

$$p(X) \approx \frac{k/n}{v}$$

表示单位体积内落入x点邻域的样本在总样本中的比例，可以此来近似样本在X点处的类概率密度值。

非参数估计



假设有N个样本$X=(X_1, X_2,\dots X_N)^T$都是按照P(X)从总体中独立抽取的

若N个样本中有k个落入在R内的概率符合二项分布

$$P_k = C_N^k p^k (1-P)^{N-k}$$

其中P是样本X落入R内的概率

P_k是k个样本落入R内的概率

数学期望:E(k)=k=NP

∴对概率P的估计:$P = \frac{k}{N}$ 。 $\frac{k}{N}$ 是P的一个比较好的估计

$$\therefore P = \int_R P(x')dx' = \frac{k}{N}$$

设P(x')在R内连续变化,当R逐渐减小的时候,小到使P(x)在其上几乎没有变化时，则

$$P = \int_R P(x')dx' \approx P(x)\cdot V = \frac{k}{N}$$

其中 $V = \int_R dx'$ 是R包围的体积

$$\therefore P(x)\cdot V \approx P = \frac{k}{N}$$

$\therefore$ 条件密度的估计：$P(x) = \dfrac{k/N}{V}$ (V足够小)

讨论:① 当V固定的时候N增加, k也增加,当$N \to \infty$ 时$k \to \infty$

$$\therefore P = \frac{k}{N} \to 1 \quad P(x) = \frac{k/N}{V} \to \frac{1}{V}$$ 反映了P(x)的空间平均估计

而反映不出空间的变化

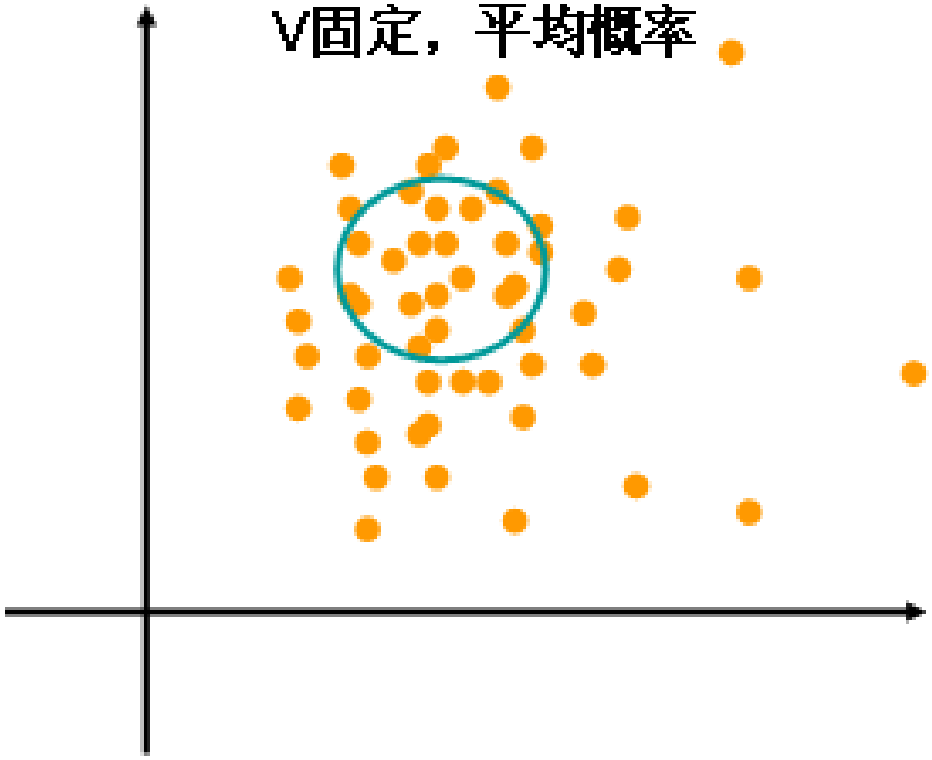
② N固定,体积变小

当 $V \to 0$ 时,k=0时 $P(x) = \dfrac{k/N}{V} \to 0$

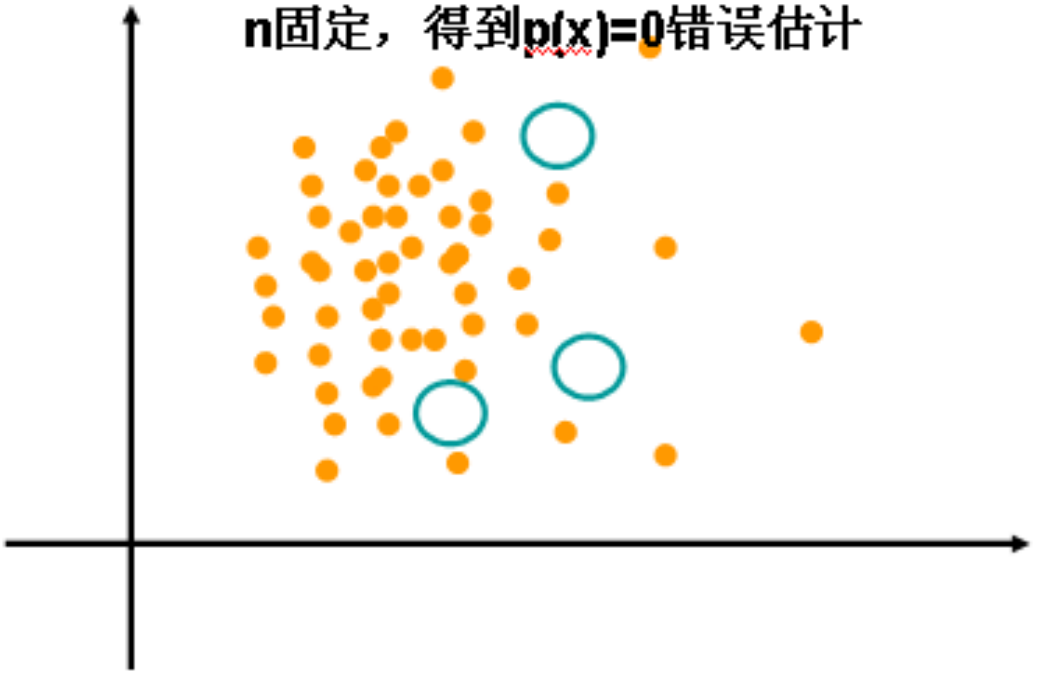
$k \neq 0$ 时$P(x) = \dfrac{k/N}{V} \to \infty$

所以起伏比较大,噪声比较大,需要对V进行改进.

非参数估计



非参数估计



非参数估计

- 解决方案
  - 考虑让v和k都随n的变化进行调整，即：

$$P(x)=\frac{k/n}{v}\Rightarrow P_n(x)=\frac{k_n/n}{v_n}$$

若$P_N(x)$收敛于$P(x)$应满足三个条件：

① $\lim_{N\to\infty} V_N = 0$ ，当N↑时，V_N↓，N→∞，V_N→0

这时虽然样本数多，但由于V_N↓，落入V_N内的样本K_N

也减小，所以空间变化才反映出来

② $\lim_{N\to\infty} K_N = \infty$ ，N↑，k_N↑，N与K_N同相变化

③ $\lim_{N\to\infty} \frac{K_N}{N} = 0$ ，K_N的变化远小于N的变化。因此尽管在R内落入了很多的样本，但同总数N比较，仍然是很小的一部分。

非参数估计

- 基本方法
  - Parzen窗口法：主动选择v_n与n的关系，k_n被动确定，指n个样本中落入区域v的样本数
  - k_n近邻法：主动选择k_n与n的关系， v_n被动确定，指包含k_n个样本的x邻域

2.Parzen窗口估计

假设R_N为一个d维的超立方体，h_N为超立方体的长度

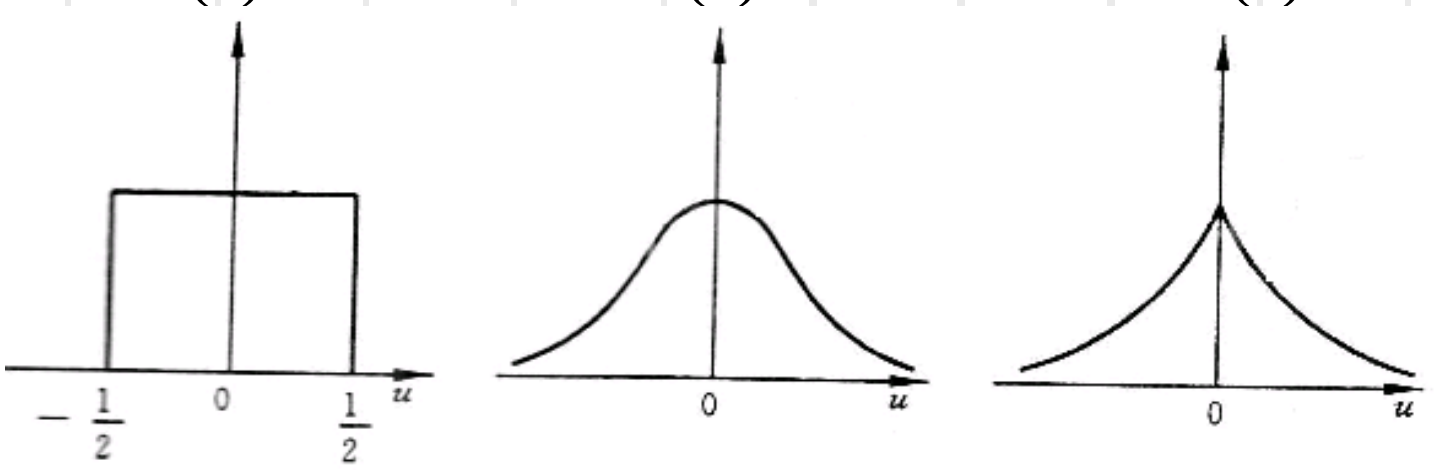
∴超立方体体积为：$V_N = h_N^d$

d=1，窗口为一线段　　d=2，窗口为一平面

d=3，窗口为一立方体　　d>3，窗口为一超立方体

窗口的选择：

方窗函数　　正态窗函数　　指数窗函数



$$\varphi(u)=\begin{cases}1, |u|\le\frac{1}{2}\\ 0.\text{其他}\end{cases}$$

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}u^2\}$$

$$\varphi(u)=\exp\{-|u|\}$$

∵ φ(u) 是以原点x为中心的超立方体。

∴在x_i落入方窗时，则有

$$\begin{cases} x - x_i \le \dfrac{h_N}{2} & \text{在}V_N\text{内为1} \\ x - x_i > \dfrac{h_N}{2} & \text{不在}V_N\text{内为0} \end{cases}$$

$$\therefore \varphi[\frac{|x - x_i|}{h_N}] = \varphi[\frac{h_N/2}{h_N}] = \varphi[\frac{1}{2}] = 1$$

落入V_N的样本数为所有为1者之和

$$\therefore K_N = \sum_{i=1}^{N} \varphi(\frac{|x - x_i|}{h_N})$$

∴ 密度估计 $P_N(x) = \dfrac{K_N/N}{V_N} = \dfrac{1}{N}\sum_{i=1}^{N}\dfrac{1}{V_N}\varphi(\dfrac{|x - x_i|}{h_N})$

讨论：

① 每个样本对估计所起的作用依赖于它到x的距离，即

$|x-x_i| \le h_N/2$时，x_i在V_N内为1，否则为0。

② $\varphi(\frac{|x-x_i|}{h_N})$ 称为 $\frac{|x-x_i|}{h_N}$ 的窗函数，取0，1两种值，但有

时可以取0，0.1，0.2……多种数值，例如随x_i离x接近的程度，

$\varphi(\frac{|x-x_i|}{h_N})$ 取值由0, 0.1, 0.2……到1。

③ 要求估计的$P_N(x)$应满足：$\begin{cases} P_N(x) \geq 0 \\ \int P_N(x)dx = 1 \end{cases}$

为满足这两个条件，要求窗函数满足：

$$\begin{cases} \varphi(\dfrac{|x - x_i|}{h_N}) > 0 \\ \int_{\frac{|x-x_i|}{h_N}}^{\alpha} \varphi(\dfrac{|x - x_i|}{h_N}) d(\dfrac{|x - x_i|}{h_N}) > 0 \end{cases}$$

④ 窗长度h_N对$P_N(x)$的影响

若h_N太大, $P_N(x)$是$P(x)$的一个平坦, 分辨率低的估计, 有平均误差

若h_N太小, $P_N(x)$是$P(x)$的一个不稳定的起伏大的估计,有噪声误差

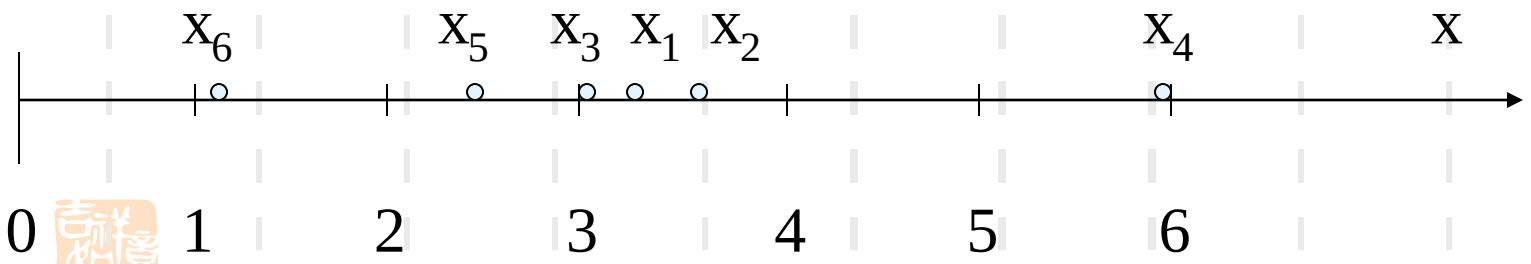
为了使这些误差不严重， h_N应很好选择

例1：对于一个二类（ ω_1 ， ω_2 ）识别问题，随机抽取ω_1类的6个样本X=(x_1，x_2，…. x_6)

ω_1=(x_1，x_2，…. x_6)

=(x_1=3.2，x_2=3.6，x_3=3，x_4=6，x_5=2.5，x_6=1.1)

估计P(x|ω_1)即$P_N(x)$



解：选正态窗函数

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)$$

$$\therefore \varphi(u)=\varphi(\frac{|x-x_i|}{h_N})=\frac{1}{\sqrt{2\pi}}\exp[-\frac{1}{2}(\frac{|x-x_i|}{h_N})^2]$$

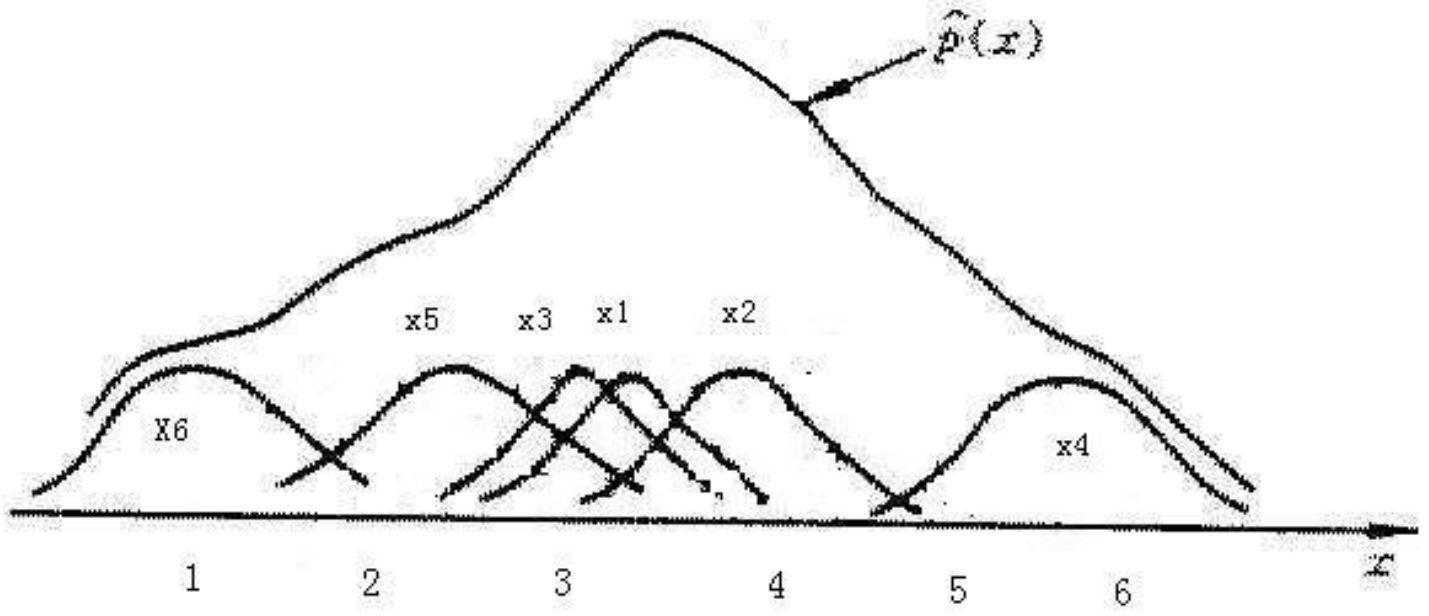
$\because$x是一维的　　$\therefore V_N = h_N = \frac{h_1}{\sqrt{N}}$,其中选$h_1 = 0.5\sqrt{6}$，$N = 6$

$$\therefore V_N = \frac{0.5\sqrt{6}}{\sqrt{6}} = 0.5$$

$$\therefore P_N(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi\left(\frac{|x - x_i|}{h_N}\right) = 0.134\exp\left[-\frac{1}{2}\left(\frac{|x-3.2|}{05}\right)^2\right] + \ldots 0.134\exp\left[-\frac{1}{2}\left(\frac{|x-1.1|}{05}\right)^2\right]$$

上式用图形表示是6个分别以3.2，3.6，3，6，2.5，1.1为中心的丘形曲线(正态曲线)，而$P_N(x)$则是这些曲线之和。

由图看出，每个样本对估计的贡献与样本间的距离有关，样本越多， $P_N(x)$越准确。



例2：设待估计的P(x)是个均值为0，方差为1的正态密度函数。若随机地抽取X样本中的1个、 16个、 256个作为学习样本x_i,试用窗口法估计$P_N(x)$。

解：设窗口函数为正态的， σ＝1， μ ＝0

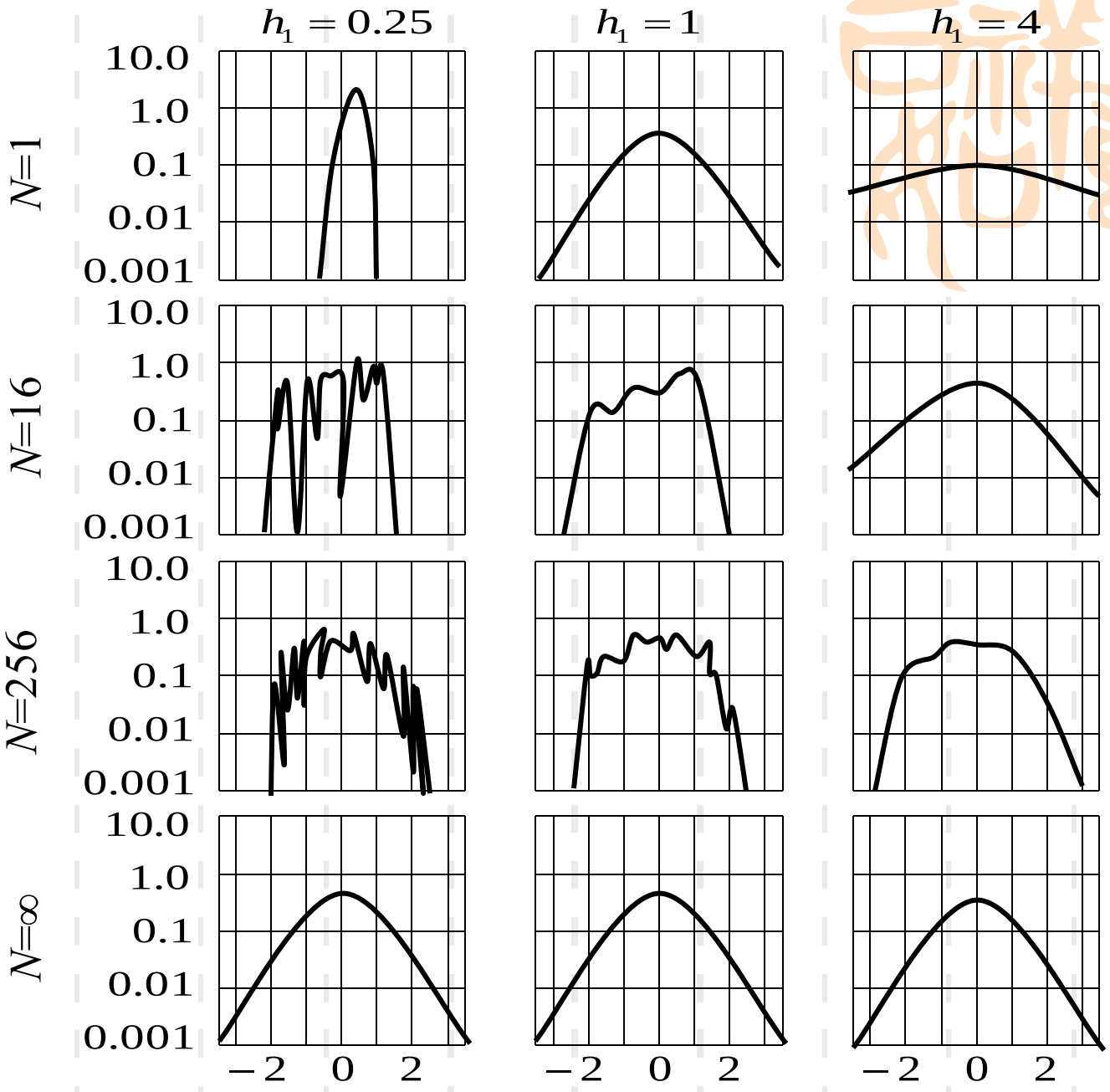
$$\varphi(\frac{|x-x_i|}{h_N})=\frac{1}{\sqrt{2\pi}}\exp[-\frac{1}{2}(\frac{|x-x_i|}{h_N})^2]$$

$$设h_N=\frac{h_1}{\sqrt{N}}$$

h_N:窗长度，N为样本数，h1为选定可调节的参数。

$$P_N(x)=\frac{1}{N}\sum_{i=1}^{N}\frac{\sqrt{N}}{h_1}\varphi(\frac{|x-x_i|}{h_N})=\frac{1}{h_1\sqrt{N}}\sum_{i=1}^{N}\frac{1}{\sqrt{2\pi}}\exp[-\frac{1}{2}\left(\frac{|x-x_i|\sqrt{N}}{h_1}\right)^2]$$

❖ 用 *Parzen* 窗法估计单一正态分布的实验



讨论：由图看出，$P_N(x)$随N，h_1的变化情况

①当N＝1时，$P_N(x)$是一个以第一个样本为中心的正态形状的小丘，与窗函数差不多。

②当N＝16及N=256时

h_1＝0.25　曲线起伏很大，噪声大

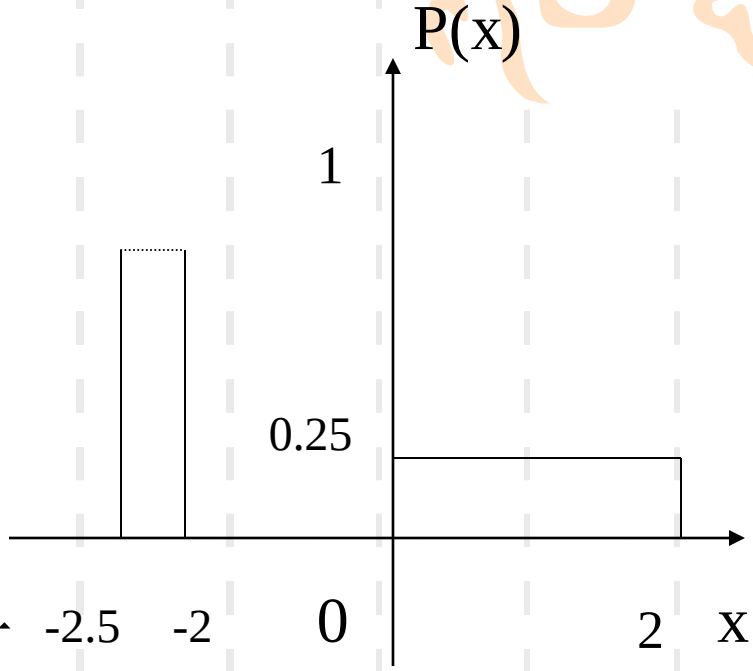
h_1＝1　起伏减小

h_1＝4　曲线平坦，平均误差

③当N→∞时，$P_N(x)$收敛于一平滑的正态曲线，估计曲线较好。

例3。待估的密度函数为二项分布

$$P(x)=\begin{cases}1 & -2.5<x<-2\\ 0.25 & 0<x<2\\ 0 & x\text{为其它}\end{cases}$$

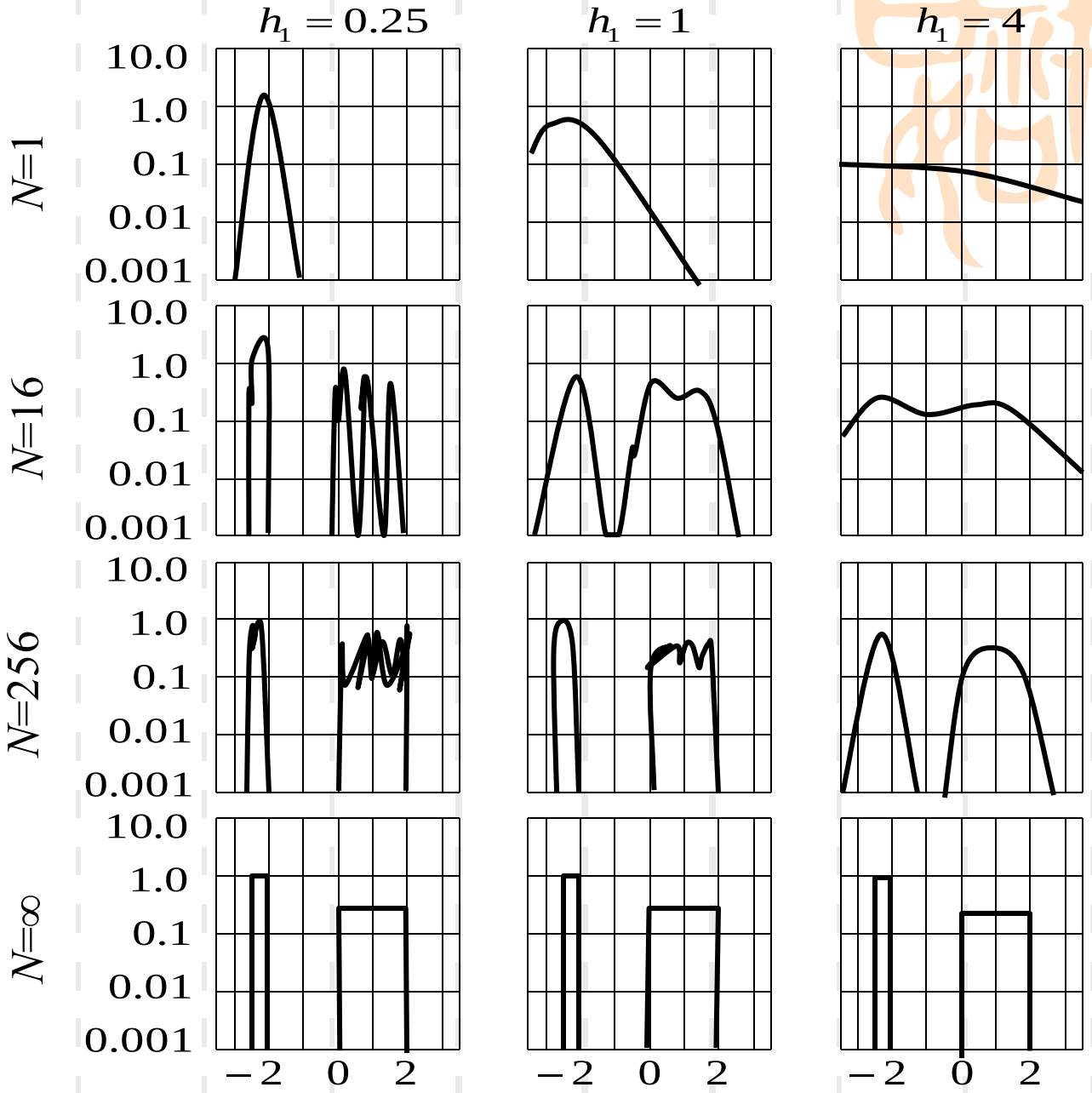


解：此为多峰情况的估计

设窗函数为正态

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}\exp[-\frac{1}{2}u^2],h_N=\frac{h_1}{\sqrt{N}}$$

❖ 用 *Parzen* 窗法估计两个均匀分布的实验



当N=1、16、256、∞时的$P_N(x)$估计如图所示

①当N＝1时，$P_N(x)$ 实际是窗函数。

②当N＝16及N=256时

h_1＝0.25　曲线起伏大

h_1＝1　曲线起伏减小

h_1＝4　曲线平坦

③当N→∞时，曲线较好。

Parzen窗口法

- 可以看出：
  - 当时 $n \to \infty$ ， $p_n(X)$ 可以收敛于任何复杂形式的 $p(X)$
  - 当n=1时， $p_n(X)$ 即是窗函数的形式
  - 当n较小时， $p_n(X)$ 对 h_1 的大小较为敏感， h_1 过小则产生噪声性误差，过大则又产生平均性误差

Parzen窗口法

- 所需样本数较多，计算量大，不易求得 $p(X)$ 的解析表达式
- 当特征空间的维数较大时，实用性差

Parzen窗口法

- 如何用Parzen窗口法进行分类器设计？
  - 获取n个学习样本 $x_1, x_2, x_3, \ldots\ldots, x_n$
  - 令 $v_n = \frac{h_1}{\sqrt{n}}$ 或 $v_n = \frac{h_1}{\log n}$
  - 令 $h_n = \sqrt[d]{v_n}$

Parzen窗口法

- 当待识别样本到来时，分别计算每一类样本的 $p_n(X)$ ，

  即计算

$$P_n(X)=\frac{1}{nv_n}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{X-X_i}{h_n}\right)^2\right)$$

- 对每一类样本重复上述过程，得各类的类概率密度 $p_n(X)$
- 将样本归类到 $p_n(X)P(\omega_j)$ 最大的类别中去

K_n近邻法

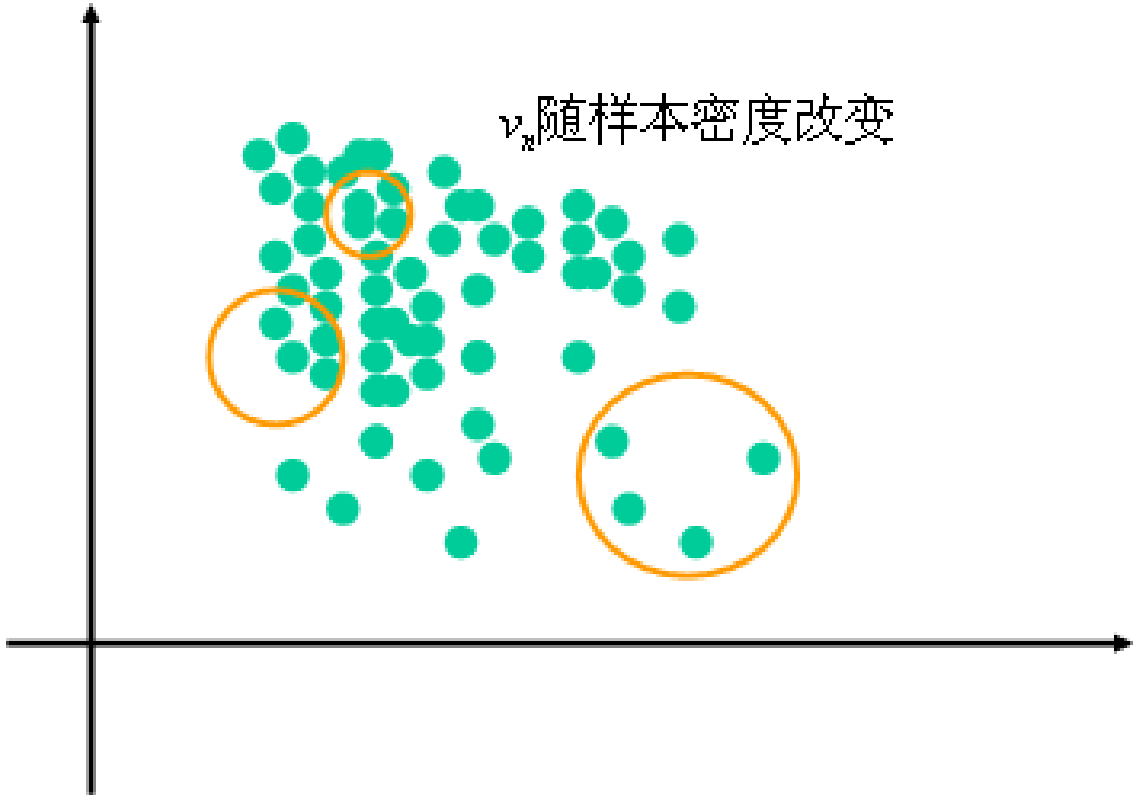
- Parzen窗口法的估计效果取决于样本总数n及h_1，当n较小时，对h_1较为敏感，即：

$$\begin{cases} h_1\text{较大容易产生平均性误差} \\ h_1\text{较小则容易产生噪声性误差} \end{cases}$$

K_n近邻法

- 其原因是由于 $v_n = \frac{h_1}{\sqrt{n}}$ 只与总样本数有关，即进行概率密度 $P_n(x)$ 估计时，任何x点处的 v_n 都是相同的
- 一种合理的选择是对样本出现密度大的x处，v_n 可较小，而对样本密度较小的x处，v_n 则相对大一些，这就是近邻法。

K_n近邻法



K_n近邻法

- 基本原理
  - 主动选择 k_n 与n的关系，v_n 被动确定，即使得体积 v_n 为样本密度的函数，而不是样本总数的函数。
  - 可选择 $k_n = \sqrt{n}$ ，该条件可满足：

$$\lim_{n\to\infty} k_n = \infty \qquad \lim_{n\to\infty} k_n / n = 0 \qquad \lim_{n\to\infty} v_n = \frac{k_n}{n \cdot P_n(x)} = 0$$

K_n近邻法

- K_n近邻法，有效地解决了Parzen存在的问题，对平均误差和噪声性误差均有较好的改善

- 选择 $k_n = \sqrt{n}$ 后，$P_n(x) = \dfrac{k_n/n}{v_n} = \dfrac{1}{\sqrt{n}v_n}$

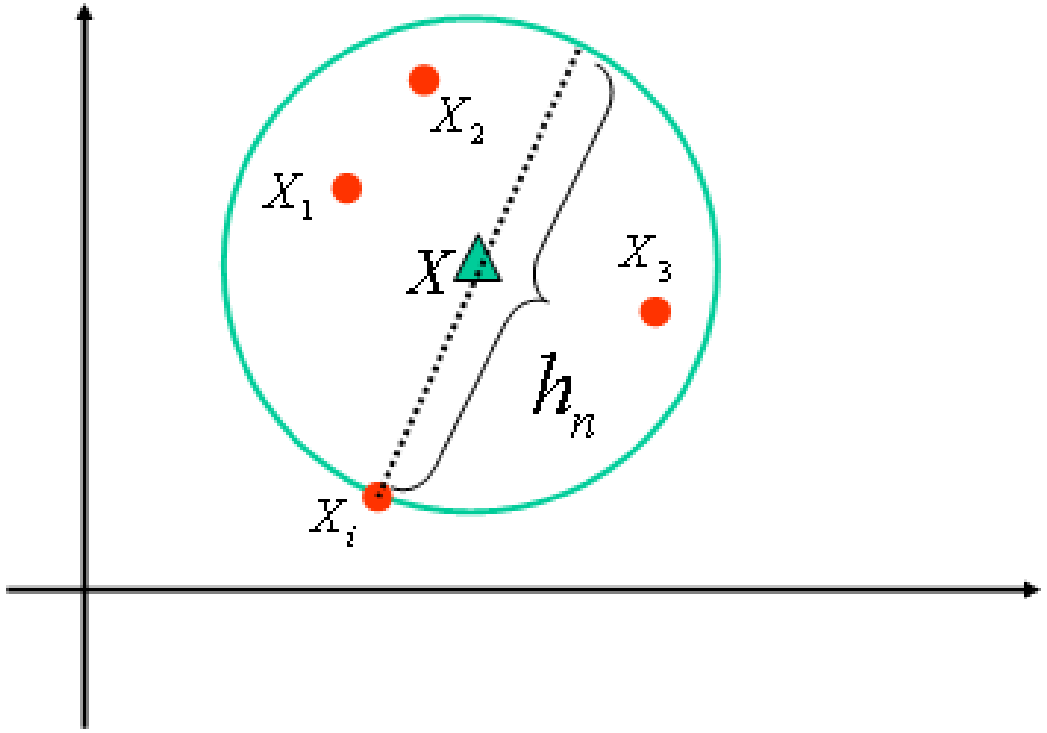
- 如何计算 v_n ？

K_n近邻法

- v_n 为与x点相邻的 k_n 个近邻样本中，与x距离最远的样本所构成的区域，即

$$v_n = h_n^d$$

K_n近邻法



K_n近邻法

- 用K_n近邻法设计分类器的过程：
  - 获取n个学习样本 $X_1, X_2, X_3, \ldots\ldots, X_n$
  - 令 $k_n = \sqrt{n}$
  - 找到待识样本X处的K_n个近邻
  - 计算K_n个邻近到X的距离，找到最远距离的样本
  - 计算邻域的直径 h_n ，计算邻域的体积 $v_n = h_n^{\ d}$

K_n近邻法

- 则

$$P_n(X)=\frac{k_n/n}{v_n}=\frac{1}{v_n\sqrt{n}}$$

- 对每一类样本重复上述过程，得各类的类概率密度 $p_n(X)$

- 将样本X归类到 $p_n(X)P(\omega_j)$ 最大的类别中去

用K_n近邻法估计后验概率

- 非参数估计法的基本思想是：

$$P_n(X)=\frac{k_n/n}{v_n}$$

- 上式即可以用来估计各类样本的类概率密度，也可以用来估计所有类别样本的概率密度分布。

用K_n近邻法估计后验概率

- 设共有C个待识类别，各个类别的学习样本数分别为$n_1,n_2,\ldots\ldots,n_C$

  总的学习样本数为$N= n_1+n_2+\ldots\ldots+n_C$

则

$$P_N(X)=\frac{k_N/N}{v_N}$$

表示所有类别样本在特征空间X处的概率密度

其中 k_N 为落入体积 v_N 中的样本数

用K_n近邻法估计后验概率

- 而联合概率密度

$$P_N(\omega_i, X) == \frac{k_{i,N} / N}{v_N}$$

$k_{i,N}$ 为N个落入 v_N 中的样本中属于第i类的样本数

用K_n近邻法估计后验概率

又由于

$$P_N(\omega_i, X) = P_N(\omega_i / X) P_N(X)$$

则后验概率

$$P_N(\omega_i / X) = \frac{P_N(X, \omega_i)}{P_N(X)} = \frac{k_{i,N}}{k_N}$$

上式表明，待识样本在x点处属于第i类的后验概率即是落入其近邻体积内第i类样本与落入总样本数之比

用K_n近邻法估计后验概率

- K_n近邻准则：
  - 设各类总的学习样本为N,令 $k_N = \sqrt{N}$
  - 当待识样本x到来时，找出x的 k_N 个近邻，其中属于第i类的样本为 $k_{i,N}$ ，则：

$$P_N(\omega_i / X) = \frac{k_{i,N}}{k_N} \qquad i = 1, 2, ..., c$$

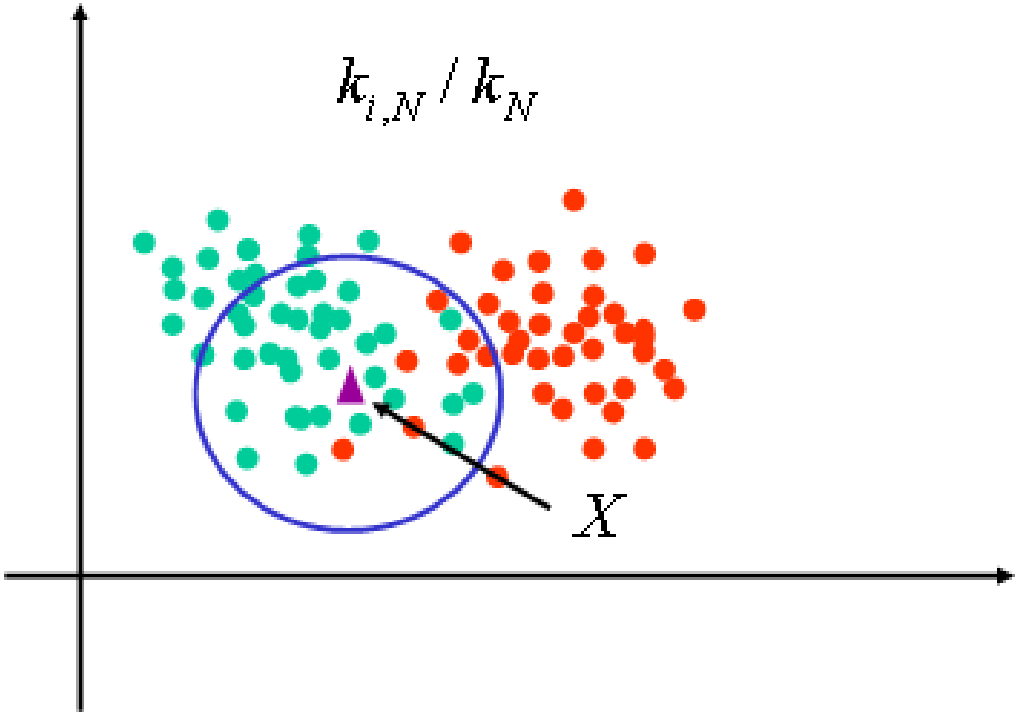
  - 取 $P_N(\omega_i / X)$ 最大的一类为识别结果

用K_n近邻法估计后验概率

- 近邻法的特点
  - 简单，容易实现
  - 计算量和计算机存储容量较大
  - 多特征，高维空间效率低
  - 需要较多的学习样本

用K_n近邻法估计后验概率

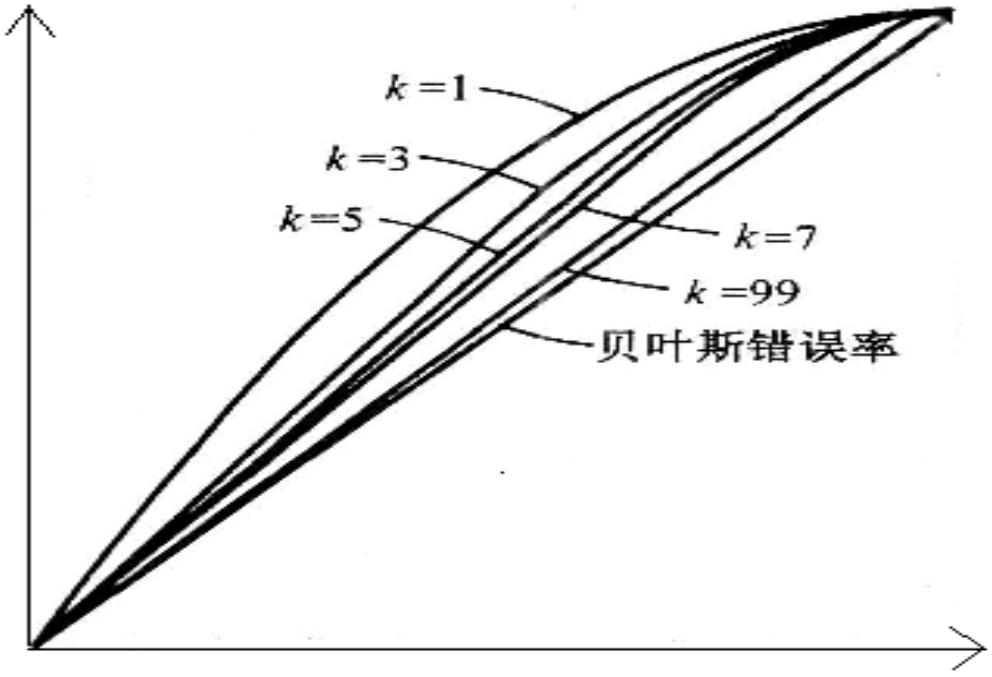
- 例：



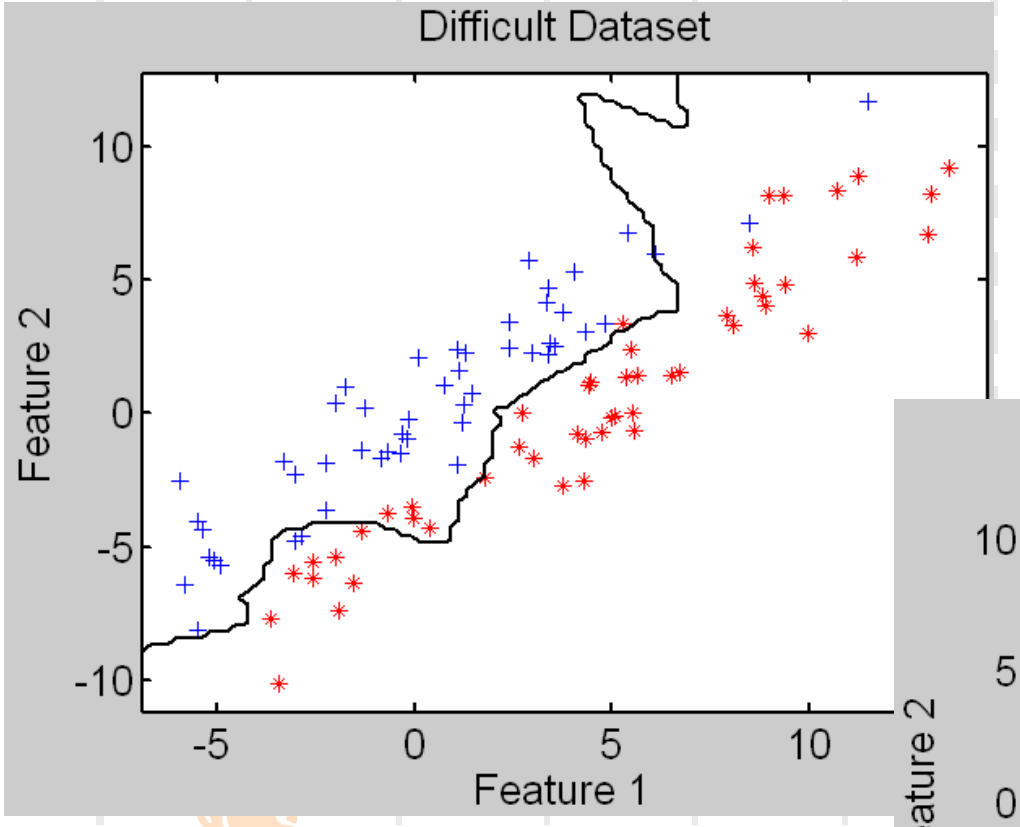
用K_n近邻法估计后验概率

- 最近邻准则：
- 取 $k_N = 1$，则K_n近邻准则变为：当待识别样本X到来时，找出其最近邻的样本，并将X判为最近邻样本类

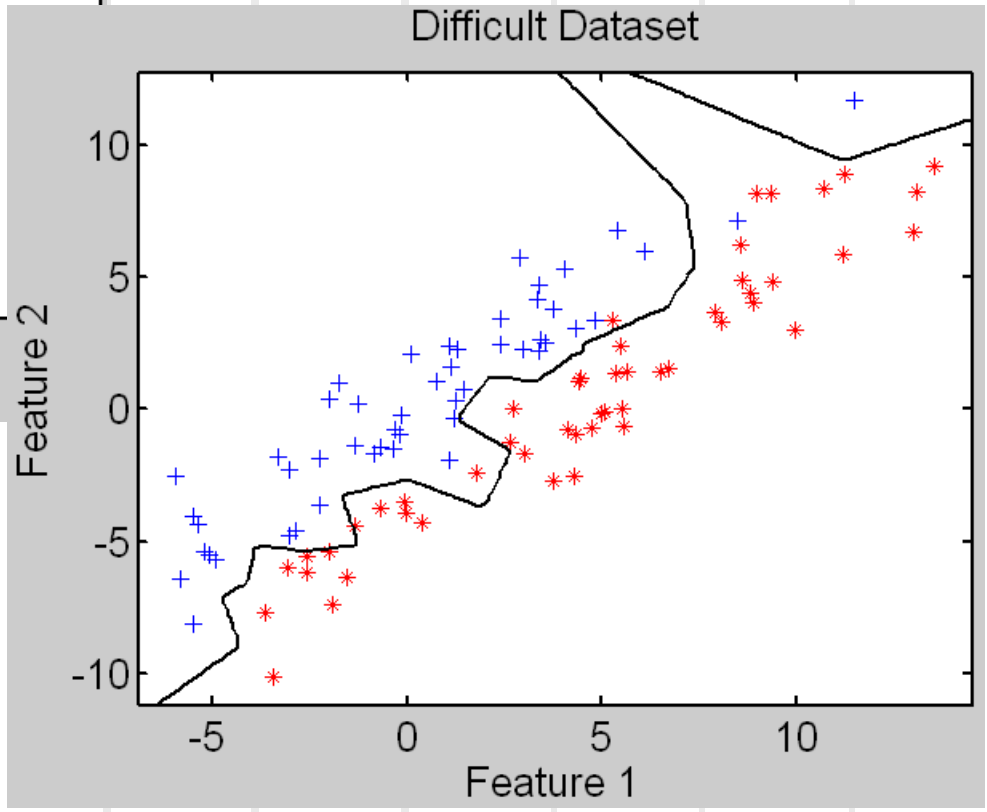
K近邻分类的错误率随K↑，P_k↓,最低的错误率为Bayes分类。



最近邻与k-近邻法分类演示



最近邻



3-近邻