数据挖掘与机器学习

1

潘斌
panbin@nankai.edu.cn
范孙楼227

上节回顾

- 线性分类器
  - 垂直平分分类器
  - 感知准则
- 梯度下降

本节提要

- 最小错分样本数准则
- 最小误差准则
- 贝叶斯分类器
  - 最小错误率Bayes决策
  - 最小风险Bayes决策
  - 最大最小Bayes决策
  - 贝叶斯分类器的设计

5　最小错分样本数准则

- 5.1　问题与思路
- 5.2　最小错分样本数准则一
- 5.3　最小错分样本数准则二
- 5.4　特点

5.1　问题与思路

- **问题的提出**
  - 感知准则只适用线性可分样本集——无错分
  - 实际情况未必线性可分——有错分
  - 另外线性可分的判断也很困难

  - 既然存在错分样本——求错分样本数最少

5.1 问题与思路

- **数学描述**
  - 仿照线性可分样本集的规范化（ω_2类样本的增广向量乘以-1）
    - $a^T y_i > 0$——正确分类
    - $a^T y_i < 0$——错误分类
  - 设样本数为N，N个不等式联立
    - $a^T y_i > 0 \qquad (i = 1,\ldots,N)$
  - 求满足不等式最多的解（权向量）

5.1 问题与思路

- **数学描述**
  - 写成矩阵形式

用矩阵形式重写式(4-44)所表示的不等式组：

$$Ya > 0$$

$$Y = \begin{bmatrix} \boldsymbol{y}_1^T \\ \boldsymbol{y}_2^T \\ \vdots \\ \boldsymbol{y}_N^T \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1\hat{d}} \\ y_{21} & y_{22} & \cdots & y_{2\hat{d}} \\ \cdots & \cdots & \cdots & \cdots \\ y_{N1} & y_{N2} & \cdots & y_{N\hat{d}} \end{bmatrix}$$

为使解更可靠，引入余量 $\boldsymbol{b}>0$

$$Ya \geqslant \boldsymbol{b} > 0$$

5.2 最小错分样本数准则一

- **最小错分样本数准则一**
  - 准则函数

$$\min J_{q1}(\boldsymbol{a}) = \| (Y\boldsymbol{a} - \boldsymbol{b}) - |Y\boldsymbol{a} - \boldsymbol{b}| \|^2$$

  - 求极值解
    - 共轭梯度下降法

共轭梯度下降法

在数值线性代数中，共轭梯度法是一种求解对称正定线性方程组Ax=b的迭代方法。

事实上，求解Ax=b等价于求解：$min||Ax-b||_2^2$,将其展开后可以得到：$min \quad x^TA^TAx-b^TAx+b^Tb$,也就是等价于求解 $min \quad \frac{1}{2}x^TA^TAx-b^TAx$ 。于是解方程问题就转化为了求解二次规划问题(QP)。

共轭梯度法是介于梯度下降法与牛顿法之间的一个方法，是一个**一阶方法**。它克服了梯度下降法收敛慢的缺点，又避免了存储和计算牛顿法所需要的二阶导数信息。

在n维的优化问题中，共轭梯度法最多n次迭代就能找到最优解（是找到，不是接近），但是只针对二次规划问题。

共轭梯度法的思想就是找到n个两两共轭的共轭方向，每次沿着一个方向优化得到该方向上的极小值，后面再沿其它方向求极小值的时候，不会影响前面已经得到的沿哪些方向上的极小值，所以理论上对n个方向都求出极小值就得到了n维问题的极小值。

5.3　最小错分样本数准则二

- **最小错分样本数准则二**
  - 准则函数

$$\max J_{q2}(\boldsymbol{a}) = \sum_{i=1}^{N} \frac{1+\mathrm{sgn}(\boldsymbol{y}_i \boldsymbol{a})}{2}$$

$$\mathrm{sgn}(\boldsymbol{y}_i \boldsymbol{a}) = \begin{cases} +1, & \text{对于 } \boldsymbol{y}_i \boldsymbol{a} \geq 0 \\ -1, & \text{对于 } \boldsymbol{y}_i \boldsymbol{a} < 0 \end{cases}$$

  - 求极值解
    - 搜索算法

5.4 特点

- **最小错分样本数准则（分类器）的特点**
  - 解决两类问题的线性分类器
  - 样本集不限，可以是线性不可分的
  - 求满足不等式个数最多的权向量（最优）
  - 分类器设计过程复杂

6　最小平方误差准则

6.1 问题与思路

- **问题的提出**
  - 对于线性不可分问题
  - 最小错分样本数准则——求错分样本数最少
  - 工程上往往是求误差平方和最小

6.1 问题与思路

- **数学描述**
  - 引入余量b_i，将不等式组改造为等式组
    - $a^T y_i = b_i > 0 \quad (i = 1,\ldots,N)$
  - 求满足等式组的最小平方误差解（权向量）

6.1　问题与思路

- **数学描述**
  - 写成矩阵形式

$$Ya = b$$

$$Y = \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_N^T \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1d} \\ y_{21} & y_{22} & \cdots & y_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ y_{N1} & y_{N2} & \cdots & y_{Nd} \end{bmatrix}$$

$$b = [b_1, b_2, \cdots, b_N]^T$$

6.2 最小平方误差准则

- **最小平方误差准则——工程上常用准则**
  - 定义优化准则函数

$$e = Ya - b$$

$$J_s(a) = \|e\|^2 = \|Ya - b\|^2 = \sum_{n=1}^{N} (a^T y_n - b_n)^2$$

6.2 最小平方误差准则

- **最小平方误差准则优化结果**
  - 直接求极值解

首先对式(4-63)中的 $J_s(\boldsymbol{a})$ 求梯度，

$$\nabla J_s(\boldsymbol{a}) = \sum_{n=1}^{N} 2(\boldsymbol{a}^T \boldsymbol{y}_n - b_n)\boldsymbol{y}_n = 2Y^T(Y\boldsymbol{a} - \boldsymbol{b})$$

令 $\nabla J_s(\boldsymbol{a}) = 0$，得

$$Y^T Y \boldsymbol{a}^* = Y^T \boldsymbol{b} \tag{4-65}$$

这样，求解 $Y\boldsymbol{a}=\boldsymbol{b}$ 的问题转化为求解 $Y^T Y\boldsymbol{a}^* = Y^T\boldsymbol{b}$ 的问题了。这一方程的最大优点是，矩阵 $Y^T Y$ 是 $\hat{d}\times\hat{d}$ 方阵，而且一般是非奇异的，因此可唯一地解得

$$\boldsymbol{a}^* = (Y^T Y)^{-1} Y^T \boldsymbol{b} = Y^+ \boldsymbol{b} \tag{4-66}$$

式中 $(\hat{d}\times N)$ 矩阵

$$Y^+ = (Y^T Y)^{-1} Y^T \tag{4-67}$$

是 Y 的左逆矩阵，$\boldsymbol{a}^*$ 就是式(4-62)的 MSE 解。

6.3 特点

- **最小平方误差准则（分类器）的特点**
  - 解决两类问题的线性分类器
  - 样本集不限，可以是线性不可分的
  - 求最小平方误差的权向量（最优）
  - 分类器设计过程相对简单

Bayes分类器

- 4.1 基本概念
- 4.2 最小错误率Bayes决策
- 4.3 最小风险Bayes决策
- 4.4 最小最大Bayes决策
- 4.5 Bayes分类器设计

4.1 基本概念

- [错误率] 几乎所有的分类器在识别时都有可能出现错误分类（简称错分 / 误判）的情况，这种错误分类的可能性称为分类器识别结果的错误概率，简称错误率 / 误判率。

- [正确率] （通常意义的）正确率 = 1 – 错误率

4.1 基本概念

- **线性分类器**
  - 垂直平分分类器
    - 未经优化，错误率通常较大
  - 感知器
    - 优化（求线性可分样本集的解），最终错误率未知
  - 最小平方误差
    - 优化（样本集MSE的解），最终错误率未知

4.1 基本概念

- **Bayes分类器设计思路**
  - 寻求概率意义上的最小错误率的分类器
  - 即具有最小错分概率的分类器——分类器设计的最优解

4.1 基本概念

- **数学基础回顾**
  - 概率论与数理统计
    - 随机事件
    - 概率
    - 条件概率
    - Bayes公式
    - 随机变量
    - 概率密度函数

4.1 基本概念

- **数学基础回顾**
  - Bayes分类相关
    - 随机事件——样本的状态/类别
    - 概率——状态 / 类别的概率
    - 随机变量——随机向量
    - 概率密度函数

贝叶斯公式

$$P(\theta|X) = \frac{P(X|\theta) \times P(\theta)}{P(X)}$$

$$P\left(B_i|A\right) = \frac{P\left(B_i\right)P\left(A|B_i\right)}{\sum_{j=1}^{n} P\left(B_j\right)P\left(A|B_j\right)}$$

- 先验和后验：**P(θ)和P(θ|x)**

机器学习两大流派——贝叶斯派和频率派

- **频率派旨在求最大似然估计**
  - 认为待求参数θ是唯一存在的
  - θ可以是模型参数，也可以是分类标签或预测结果
  - 利用已知的样本结果信息，反推最具有可能（最大概率）导致这些样本结果出现的模型参数值

$$
\begin{aligned}
\hat{\theta}_{\mathrm{MLE}} &= \arg\max \boxed{P(X;\theta)} \\
&= \arg\max P(x_1;\theta)P(x_2;\theta)\cdots\cdot P(x_n;\theta) \\
&= \arg\max \log \prod_{i=1}^{n} P(x_i;\theta) \\
&= \arg\max \sum_{i=1}^{n} \log P(x_i;\theta) \\
&\boxed{= \arg\min -\sum_{i=1}^{n} \log P(x_i;\theta)}
\end{aligned}
$$

– 负对数似然函数

贝叶斯派和频率派

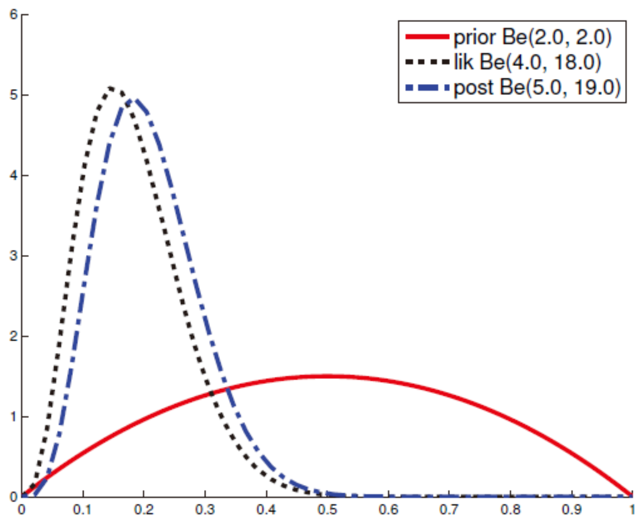
- 贝叶斯派旨在求最大后验估计
  - 认为待求参数 θ 是一个随机变量，符合一定的概率分布
  - 预设一个参数 θ 的概率分布，再用已有样本去修正这个预设（先验概率），得到最有利于样本出现的分布参数（后验概率）

$$
\begin{aligned}
\hat{\theta}_{\mathrm{MAP}} &= \arg\max \boxed{P(\theta|X)} \\
&= \arg\min -\log P(\theta|X) \\
&= \arg\min -\log P(X|\theta) - \log P(\theta) + \log P(X) \\
&= \arg\min \boxed{-\log P(X|\theta)} - \log P(\theta)
\end{aligned}
$$

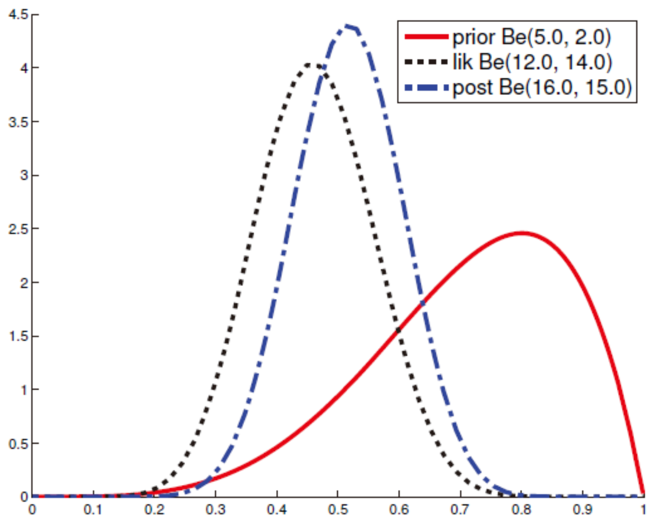
*分号表示确定性，竖条表示不确定性

贝叶斯派和频率派

- **频率派的优势**
  - 样本足够大的情况下较容易得到接近无偏的估计
  - 样本少的情况下，偏差较大（例：投5次硬币）
- **贝叶斯派的优势**
  - 实际上是基于先验的校正，由于先验的存在，样本少时效果也不会太差
  - 先验非常重要

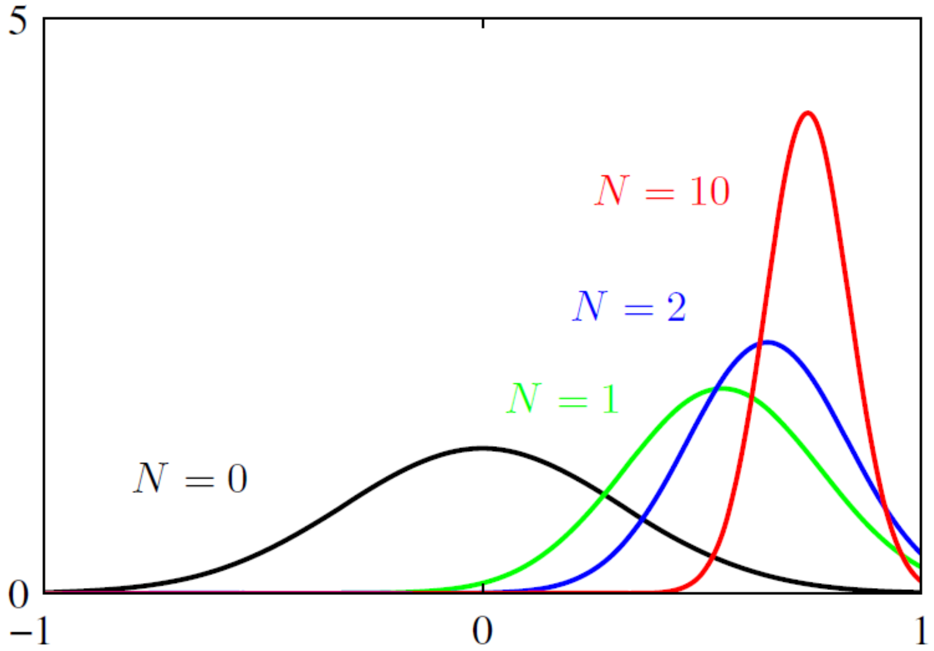


(a)



(b)

- 不同先验的影响



- 训练样本的影响–稀释先验，X和x的区别

- 频率派和贝叶斯派的等价关系

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

4.2 最小错误率Bayes决策

- 4.2.1 问题
- 4.2.2 后验概率
- 4.2.3 决策规则
- 4.2.4 实例
- 4.2.5 其它问题
- 4.2.6 特点

4.2.1 问题（P代表概率，p代表概率密度函数）

- **Bayes决策已知条件和问题（先考虑C = 2分类问题）**
  - 两类问题：ω_1和ω_2
  - 先验概率：$P(\omega_1)$ 和$P(\omega_2)$
  - 类条件概率密度函数：$p(x|\omega_1)$ 和$p(x|\omega_2)$
  - 发生了一个随机事件，其观察值为：特征向量x

  - **求最小错误率分类器**

4.2.2 后验概率

- **Bayes（条件概率）公式**
  - $p(x|\omega_1)\,P(\omega_1) = p(x)\,P(\omega_1|x)$
  - $p(x|\omega_2)\,P(\omega_2) = p(x)\,P(\omega_2|x)$

- 后验概率
  - $P(\omega_1|x) = p(x|\omega_1)\,P(\omega_1) / p(x)$
  - $P(\omega_2|x) = p(x|\omega_2)\,P(\omega_2) / p(x)$
  - $p(X) = p(x|\omega_1)\,P(\omega_1) + p(x|\omega_2)\,P(\omega_2)$

4.2.3 决策规则

- **决策规则**
  - 比较后验概率，取最大值进行类别判断
  - 对于未知样本x，若$P(\omega_1|x) > P(\omega_2|x)$，则$x \in \omega_1$
  - 若$P(\omega_1|x) < P(\omega_2|x)$，则$x \in \omega_2$

4.2.3 决策规则

- **等价规则一 后验概率分子**
  - 比较分子$p(x|\omega_1)P(\omega_1)$ 和 $p(x|\omega_2)P(\omega_2)$，取最大
  - 对于未知样本x，若$p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$，则$x \in \omega_1$
  - 若$p(x|\omega_1)P(\omega_1) < p(x|\omega_2)P(\omega_2)$，则$x \in \omega_2$

4.2.3 决策规则

- **等价规则二　似然比**
  - 定义似然比函数$l(x) = p(x|\omega_1) / p(x|\omega_2)$
  - 对于未知样本x，若$l(x) > P(\omega_2) / P(\omega_1)$，则$x \in \omega_1$
  - 若$l(x) < P(\omega_2) / P(\omega_1)$，则$x \in \omega_2$

4.2.3 决策规则

- **等价规则三 负对数似然比**
  - 定义负对数似然比函数$h(x) = -\ln l(x) = -\ln p(x|\omega_1) + \ln p(x|\omega_2)$
  - 对于未知样本x，若$h(x) < \ln [P(\omega_1) / P(\omega_2)]$，则$x \in \omega_1$
  - 若$h(x) > \ln [P(\omega_1) / P(\omega_2)]$，则$x \in \omega_2$

4.2.4 实例

- 已知
  - 癌细胞图像识别：正常和异常两类（即C = 2）
  - 已知未知样本特征观察值：x = 0.5
  - 又已知$P(\omega_1) = 0.9$和$P(\omega_2) = 0.1$
  - 查函数曲线得$p(0.5|\omega_1) = 0.2$和$p(0.5|\omega_2) = 0.4$（可以是先验或学习得到）

  - 试对未知样本x = 0.5进行分类

4.2.4 实例

- 已知
  - 癌细胞图像识别：正常和异常两类（即C = 2）
  - 已知未知样本特征观察值：x = 0.5
  - 又已知$P(\omega_1) = 0.9$和$P(\omega_2) = 0.1$
  - 查函数曲线得$p(0.5|\omega_1) = 0.2$和$p(0.5|\omega_2) = 0.4$
  - 试对未知样本x = 0.5进行分类

解：利用贝叶斯公式，分别计算出 ω_1 及 ω_2 的后验概率。

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

$$P(\omega_2|x)=1-p(\omega_1|x)=0.182$$

根据贝叶斯决策规则式(2-2)，有

$$P(\omega_1|x)=0.818>P(\omega_2|x)=0.182$$

所以合理的决策是把 x 归类于正常状态。

4.2.5 最小错误率的说明

- **最小错误率的说明（设C=2，D=1）**
  - 错误率P(e)的定义

首先应指出所谓错误率是指平均错误率，以 $P(e)$ 来表示，其定义为

$$P(e)=\int_{-\infty}^{\infty}P(e,\mathbf{x})\mathrm{d}\mathbf{x}=\int_{-\infty}^{\infty}P(e|\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x} \tag{2-6}$$

其中 $\int_{-\infty}^{\infty}(\quad)\mathrm{d}\mathbf{x}$ 表示在整个 d 维特征空间上的积分。

对两类别问题，从式(2-2)的决策规则可知，如果 $P(\omega_2|\mathbf{x})>P(\omega_1|\mathbf{x})$，则决策应为 ω_2，显然在作出决策 ω_2 时，$\mathbf{x}$ 的条件错误概率为 $P(\omega_1|\mathbf{x})$；反之，则应为 $P(\omega_2|\mathbf{x})$。可表示为

$$P(e|\mathbf{x})=\begin{cases}P(\omega_1|\mathbf{x}),\text{当 } P(\omega_2|\mathbf{x})>P(\omega_1|\mathbf{x})\\ P(\omega_2|\mathbf{x}),\text{当 } P(\omega_1|\mathbf{x})>P(\omega_2|\mathbf{x})\end{cases} \tag{2-7}$$

4.2.5 最小错误率的说明

- **最小错误率的说明（设C=2，D=1）**
  - 错误率P(e)的推导
    - 做ω_1判别时的错误概率

$$P(e_{12}) = \int_{\Re_1} P(\omega_2 \mid x) p(x) dx = \int_{\Re_1} P(\omega_2) p(x \mid \omega_2) dx$$

    - 做ω_2判别时的错误概率

$$P(e_{21}) = \int_{\Re_2} P(\omega_1 \mid x) p(x) dx = \int_{\Re_2} P(\omega_1) p(x \mid \omega_1) dx$$

    - 总错误概率

$$P(e) = \int_{\Re_1} P(\omega_2) p(x \mid \omega_2) dx + \int_{\Re_2} P(\omega_1) p(x \mid \omega_1) dx$$

4.2.5 最小错误率的说明

- **最小错误率的说明（设C=2，D=1）**
  - 再假设t为唯一分界点（**C=2，D=1**）
    - 做ω_1判别时的错误概率

      $$P(e_{12}) = \int_{-\infty}^{t} P(\omega_2)\, p(x \mid \omega_2) dx$$

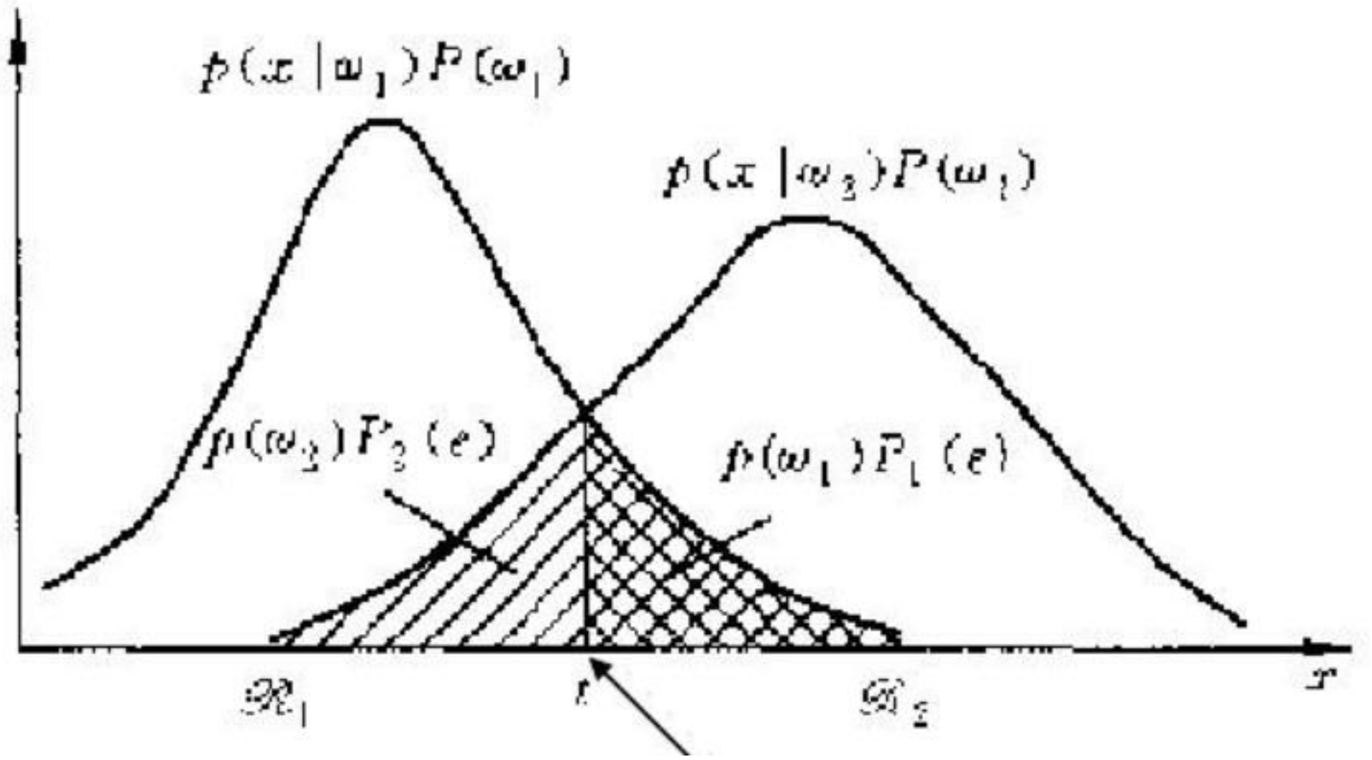
    - 做ω_2判别时的错误概率

      $$P(e_{21}) = \int_{t}^{\infty} P(\omega_1)\, p(x \mid \omega_1) dx$$

    - 总错误概率

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

4.2.5 最小错误率的说明



$$P_1(e)=\int_{R_2} p(\mathbf{x}|\omega_1)d\mathbf{x}$$

$$P_2(e)=\int_{R_1} p(\mathbf{x}|\omega_2)d\mathbf{x}$$

4.2.5 最小错误率的说明

- **推广到多类（任意C类）**
  - 已知条件和问题

  - C类D维问题：ω_1， ω_2， …， ω_C
  - 先验概率：$P(\omega_1)$， $P(\omega_2)$， …， $P(\omega_C)$
  - 类条件概率密度函数：$p(x \mid \omega_1)$， $p(x \mid \omega_2)$， …， $p(x \mid \omega_C)$
  - 发生了一个随机事件，其观察值为：特征向量x

  - **求最小错误率分类器**

4.2.5 最小错误率的说明

- **推广到多类（任意C类）**
  - 判别函数
    - $P(\omega_i|x) = p(x|\omega_i)\, P(\omega_i) / p(x) \qquad i = 1, \ldots, C$
  - 决策规则
    - 对于未知样本x，若$P(\omega_j|x) = \max P(\omega_i|x)$，则$x \in \omega_j$

4.2.6 特点

- **最小错误率Bayes决策的特点**
  - 已知条件多——各类概率分布
  - 最小错误率——概率意义上最优
  - 非线性分类器
  - 设计过程复杂

4.3 最小风险Bayes决策

- **4.3.1 问题的提出**
- **4.3.2 决策规则**
- **4.3.3 其它说明**
- **4.3.4 特点**

4.3.1 问题的提出

- **最小错误率Bayes决策**
  - 最小错误率——概率意义上最优
  - 工程上是否最优?
- 错误分类的结果、代价或风险会是怎样的?
  - 考虑癌细胞图像识别的例子
- 出错的可能情况
  - 正常细胞ω_1错分为异常ω_2
  - 异常细胞ω_2错分为正常ω_1

4.3.1 问题的提出

- **区别状态和决策概念**
  - 状态：识别的目的是分类，把样本归类于其可能的自然状态（即类别）之一，将这种自然状态简称为状态，记为ω
  - 状态空间：所有可能的状态的集合构成状态空间，记为$\boldsymbol{\Omega}$

4.3.1 问题的提出

- **区别状态和决策概念**
  - 决策：把样本归类于某个状态，或不能进行归类，都是决策，记为α
  - 决策空间：所有可能的决策（包括拒绝决策）的集合构成决策空间，记为A

4.3.1 问题的提出

- **已知条件（类别C = 2，决策A = 2）**
  - ω_1、$P(\omega_1)$、 $p(x \mid \omega_1)$和ω_2 、$P(\omega_2)$、 $p(x \mid \omega_2)$
  - $\Omega = \{\omega_1，\omega_2\}$
  - $A = \{\alpha_1，\alpha_2\}$
  - 定义损失函数$\lambda(\alpha_i, \omega_j)$，简记$\lambda_{ij}$
  - 发生了一个随机事件，其观察值为特征向量x

  - **求最小风险分类器**

4.3.2 决策规则

- **判别函数**
  - $P(\omega_j \mid x) = p(x \mid \omega_j)\, P(\omega_j) / p(x)$　　　$j = 1, 2$
  - $R(\alpha_i \mid x) = E[\lambda(\alpha_i, \omega_j) \mid x] = \lambda_{i1} P(\omega_1 \mid x) + \lambda_{i2} P(\omega_2 \mid x)$　　　$i = 1, 2$

- **决策规则**
  - 对于未知样本x，若$R(\alpha_k \mid x) = \min R(\alpha_i \mid x)$，则$x \in \omega_k$，**即决策$\alpha_k$**

4.3.2 决策规则

- **推广到任意情况（C个类别，A个决策）**
  - ω_j、$P(\omega_j)$、 $p(x|\omega_j)$ $j = 1, \ldots, C$
  - $\Omega = \{\omega_1, \omega_2, \ldots, \omega_C\}$
  - $A = \{\alpha_1, \alpha_2, \ldots, \alpha_A\}$
  - 定义损失函数$\lambda(\alpha_i, \omega_j)$，简记$\lambda_{ij}$
  - 发生了一个随机事件，其观察值为特征向量x

  - **求最小风险分类器**

4.3.2 决策规则

- **判别函数**
  - $P(\omega_j \mid x) = p(x \mid \omega_j)\, P(\omega_j) / p(x)$ $\quad j = 1, 2, \ldots, C$
  - $R(\alpha_i \mid x) = E[\lambda(\alpha_i, \omega_j) \mid x] = \sum_{j=1} \lambda(\alpha_i, \omega_j) P(\omega_j \mid x)$ $\quad i = 1, 2, \ldots, A$

- **决策规则**
  - 对于未知样本x，若$R(\alpha_k \mid x) = \min R(\alpha_i \mid x)$，则$x \in \omega_k$，**即决策**$\alpha_k$

4.3.3 其它说明

- **最小风险与最小错误率的关系**
  - 0-1损失函数

$$\lambda(\alpha_i, \omega_j) = 0 \qquad i = j$$
$$\lambda(\alpha_i, \omega_j) = 1 \qquad i \neq j$$

$$R(\alpha_i, x) = E\left[\lambda(\alpha_i, \omega_j)\right] = \sum_{j=1} \lambda(\alpha_i, \omega_j) P(\omega_j \mid x)$$

$$R(\alpha_i, x) = \sum_{\substack{j=1 \\ j \neq i}} P(\omega_j \mid x) = 1 - P(\omega_i \mid x)$$

4.3.3 其它说明

- **最小风险与最小错误率的关系**

$$R(\alpha_i, x) = \sum_{\substack{j=1 \\ j \neq i}}^{L} P(\omega_j \mid x) = 1 - P(\omega_i \mid x)$$

  - 对于未知样本x，若$R(\alpha_k \mid x) = \min R(\alpha_i \mid x)$，则$x \in \omega_k$
  - 对于未知样本x，若$P(\omega_k|x) = \max P(\omega_i|x)$，则$x \in \omega_k$

  - **结论：最小错误率Bayes决策，等价于0-1损失函数的最小风险Bayes决策**

4.3.4 特点

- **最小风险Bayes决策的特点**
  - 已知条件多——各类概率分布及风险系数
  - 最小错误风险——概率意义上最优
  - 非线性分类器
  - 设计过程复杂

例题

- **1、已知**
  - 甲类：$P(\omega_1) = 0.7$和类条件概率密度函数$p(x|\omega_1)$
  - 乙类：$P(\omega_2) = 0.3$和类条件概率密度函数$p(x|\omega_2)$
  - 今有待分类样本特征观察值$x = 10$，且由函数曲线查得$p(10|\omega_1) = 0.2$，$p(10|\omega_2) = 0.5$
  - (1)试用最小错误率Bayes决策对样本$x = 10$进行分类
  - (2)试用最小风险Bayes决策对该样本进行分类，设$\lambda_{11}=\lambda_{22}=0$，$\lambda_{12}=2$，$\lambda_{21}=1$

4.4 最小最大决策

- 4.4.1 问题的提出
- 4.4.2 期望损失
- 4.4.3 最小最大风险
- 4.4.4 特点

4.4.1 问题的提出

- **已知条件和问题（C = 2情况）**
  - 先验概率：考虑$P(\omega_1)$ 和$P(\omega_2)$未知或不确定的情况
  - 此时绝对意义的最小风险不存在
  - 如何求Bayes分类器

  - 思路
    - 假设$P(\omega_1)$ 和$P(\omega_2)$确定
    - 设计一系列最小风险Bayes分类器
    - 取其中最大风险为最小的一个来用
    - 目的是控制最大风险

4.4.1 问题的提出

- **问题（类别C = 2，决策A = 2）**
  - ω_1、$p(x \mid \omega_1)$和ω_2 、$p(x \mid \omega_2)$
  - $\Omega = \{\omega_1，\omega_2\}$
  - $A = \{\alpha_1，\alpha_2\}$
  - 损失函数$\lambda(\alpha_i, \omega_j)$，简记$\lambda_{ij}$
  - 发生了一个随机事件，其观察值为特征向量x

  - **求最小最大风险分类器**

4.4.2 期望损失

- 期望损失的推导（回顾最小错误率证明）

$$P(e) = \int_{\Re_1} P(\omega_2)\, p(x \mid \omega_2) dx + \int_{\Re_2} P(\omega_1) p(x \mid \omega_1) dx$$

$$P(e_{12}) = \int_{\Re_1} P(\omega_2)\, p(x \mid \omega_2) dx$$

$$P(e_{21}) = \int_{\Re_2} P(\omega_1) p(x \mid \omega_1) dx$$

4.4.2 期望损失

- **期望损失的推导**

联合或条件的形式均可

$$R=\int R(\alpha(\mathbf{x})|\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x}=\int_{\mathcal{R}_1}\boxed{R(\alpha_1|\mathbf{x})p(\mathbf{x})}\mathrm{d}\mathbf{x}+\int_{\mathcal{R}_2}R(\alpha_2|\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x}$$

$$=\int_{\mathcal{R}_1}[\lambda_{11}P(\omega_1)p(\mathbf{x}|\omega_1)+\lambda_{12}P(\omega_2)p(\mathbf{x}|\omega_2)]\mathrm{d}\mathbf{x}+\int_{\mathcal{R}_2}[\lambda_{21}P(\omega_1)p(\mathbf{x}|\omega_1)$$

$$+\lambda_{22}P(\omega_2)p(\mathbf{x}|\omega_2)]\mathrm{d}\mathbf{x}$$

$P(\omega_2)=1-P(\omega_1)$，代入

4.4.2 期望损失

- 期望损失的推导

$$
\begin{aligned}
R = \lambda_{22} + (\lambda_{12} - \lambda_{22})\int_{\mathscr{R}_1} p(\boldsymbol{x}|\omega_2)\mathrm{d}\boldsymbol{x} + P(\omega_1)[(\lambda_{11} - \lambda_{22}) \\
+ (\lambda_{21} - \lambda_{11})\int_{\mathscr{R}_2} p(\boldsymbol{x}|\omega_1)\mathrm{d}\boldsymbol{x} - (\lambda_{12} - \lambda_{22})\int_{\mathscr{R}_1} p(\boldsymbol{x}|\omega_2)\mathrm{d}\boldsymbol{x}]
\end{aligned}
$$

$$
a = \lambda_{22} + (\lambda_{12} - \lambda_{22})\int_{\mathscr{R}_1} p(\boldsymbol{x}|\omega_2)\mathrm{d}\boldsymbol{x}
$$

$$
b = (\lambda_{11} - \lambda_{22}) + (\lambda_{21} - \lambda_{11})\int_{\mathscr{R}_2} p(\boldsymbol{x}|\omega_1)\mathrm{d}\boldsymbol{x} - (\lambda_{12} - \lambda_{22})\int_{\mathscr{R}_1} p(\boldsymbol{x}|\omega_2)\mathrm{d}\boldsymbol{x}
$$

推导思路：替换后改成只有P(w1)有关的形式

4.4.3 最小最大风险

- **最小最大风险的解**
  - 任给$P(\omega_1)^*$的值，$P(\omega_2)^* = 1 - P(\omega_1)^*$
  - ↓
  - 求对应的最小风险Bayes分类器，设决策域为R_1和R_2
  - ↓
  - 计算对应的a^*和b^*，得到$R^* = a^* – b^* P(\omega_1)^*$（注意！**R1R2**随**P(w)**变，即决策会根据先验改变。故**R**和**P(w)**不是线性）
  - ↓
  - 重复上述计算步骤，得到一系列最小风险Bayes分类器，并 可得$P(\omega_1)^*$——R^*关系曲线
  - ↓
  - 比较所有最大风险，取其中最小的一个，作为最终的分类器

4.4.4 最小最大决策

- **最小最大决策的特点**
  - 已知条件多——各类概率分布及风险系数
  - 最小最大风险——概率意义上最优
  - 非线性分类器
  - 设计过程很复杂

4. 5 Bayes分类器的线性化设计

- 4.5.1 原理设计
- 4.5.2 Bayes决策面
- 4.5.3 错误率估计
- 4.5.4 其它

4.5.1 原理设计

- **两类情况（C=2）设计方法一**
  - 定义判别函数g(x)

① $g(\boldsymbol{x})=P(\omega_1|\boldsymbol{x})-P(\omega_2|\boldsymbol{x})$

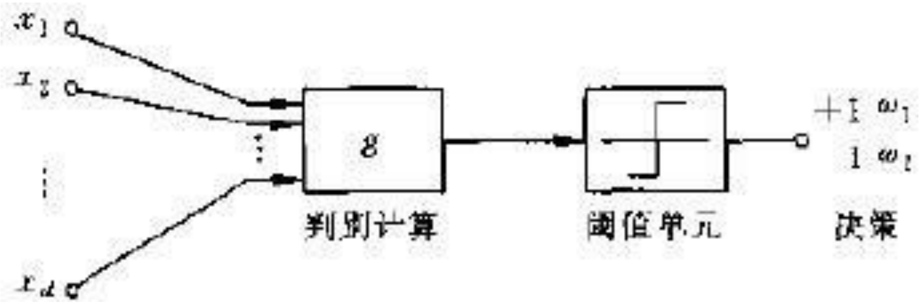
② $g(\boldsymbol{x})=p(\boldsymbol{x}|\omega_1)P(\omega_1)-p(\boldsymbol{x}|\omega_2)P(\omega_2)$

③ $g(\boldsymbol{x})=\ln\dfrac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)}+\ln\dfrac{P(\omega_1)}{P(\omega_2)}$

  - 决策面H方程
    - g(x) = 0

4.5.1 原理设计

- **两类情况（C=2）设计方法一**
  - 决策规则
    - 对于未知样本x，若g(x) > 0，则x决策为ω_1类
    - 若g(x) < 0，则x决策为ω_2类
  - 原理图



4.5.1 原理设计

- **两类情况（C=2）设计方法二**
  - 定义判别函数$g_1(x)$和$g_2(x)$

    ① $g_i(\boldsymbol{x}) = P(\omega_i|\boldsymbol{x})$

    ② $g_i(\boldsymbol{x}) = p(\boldsymbol{x}|\omega_i)P(\omega_i)$

    ③ $g_i(\boldsymbol{x}) = \ln p(\boldsymbol{x}|\omega_i) + \ln P(\omega_i)$

  - 决策面H方程
    - $g_1(x) = g_2(x)$

4.5.1 原理设计

- **两类情况（C=2）设计方法二**
  - 决策规则
    - 对于未知样本x，若$g_1(x) > g_2(x)$，则x决策为ω_1类
    - 若$g_1(x) < g_2(x)$，则x决策为ω_2类
  - 原理图

4.5.1 原理设计

- **多类情况（C任意）设计方法**
  - 定义判别函数$g_i(x)$　　$i = 1, 2, \ldots, C$

    ① $g_i(\boldsymbol{x}) = P(\omega_i \mid \boldsymbol{x})$

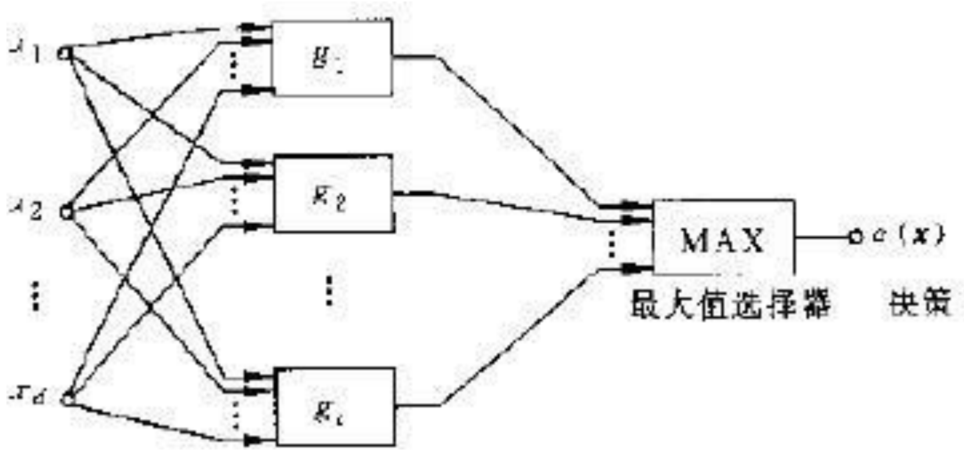
    ② $g_i(\boldsymbol{x}) = p(\boldsymbol{x} \mid \omega_i) P(\omega_i)$

    ③ $g_i(\boldsymbol{x}) = \ln p(\boldsymbol{x} \mid \omega_i) + \ln P(\omega_i)$

  - 决策面H方程
    - $g_i(x) = g_j(x)$　　$i, j = 1, 2, \ldots, C$

4.5.1 原理设计

- **多类情况（C任意）设计方法**
  - 决策规则
    - 对于未知样本x，若$g_j(x) = MAX\ g_i(x)$，则x决策为ω_j类
  - 原理图



4.5.2 正态分布决策面

- 单变量正态分布 / 一元正态分布

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

4.5.2 正态分布决策面

- 多变量正态分布 / 多元正态分布

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

$$\mu = E\{\boldsymbol{x}\}$$

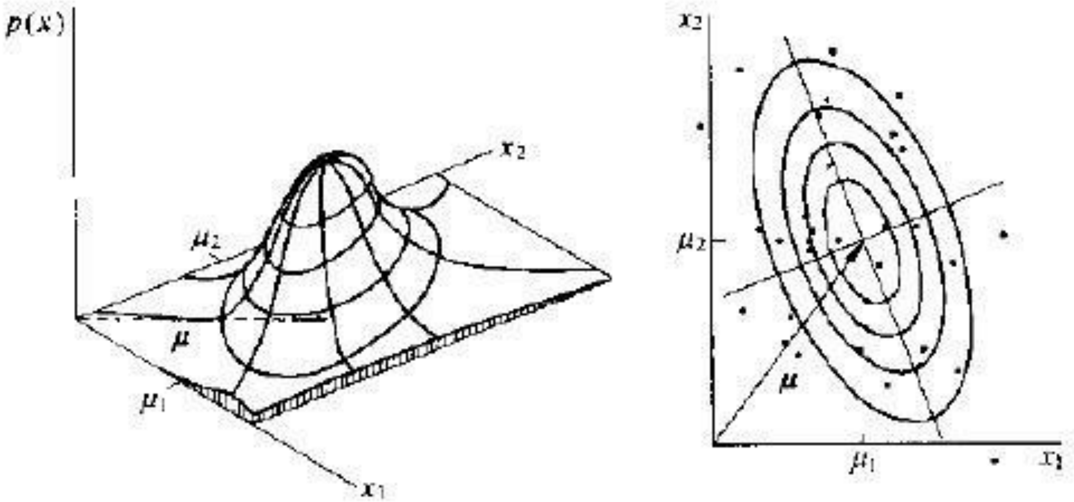
$$\Sigma = E\{(\boldsymbol{x}-\mu)(\boldsymbol{x}-\mu)^T\}$$

4.5.2 正态分布决策面

- **多元正态分布的性质**
  - 均值与方差
  - 等概率密度点的轨迹

正态分布的等密度点的轨迹为超椭球面

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



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

4.5.2 正态分布决策面

- 多元正态分布的性质
  - 不相关性与独立性：　　不相关性 = 独立性

$$\sigma_{ij}^2 = E[(x_i-\mu_i)(x_j-\mu_j)], \quad i,j=1,2,\cdots,d; i\neq j$$
$$= E(x_i-\mu_i)\cdot E(x_j-\mu_j)$$
$$=0$$

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \cdots & 0 \\ \vdots & \ddots & \\ 0 & \cdots & \sigma_{dd}^2 \end{bmatrix}$$

$$\Sigma^{-1} = \begin{bmatrix} 1/\sigma_{11}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_{dd}^2 \end{bmatrix}$$

4.5.2 正态分布决策面

- **多元正态分布的性质**
  - 边缘分布：　　正态分布
  - 条件分布：　　正态分布
  - 线性组合：　　正态分布

4.5.2 正态分布决策面

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

- 正态分布时分类器的决策面方程
  - 判别函数

③ $g_i(\boldsymbol{x})=\ln p(\boldsymbol{x}|\omega_i)+\ln P(\omega_i)$

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

  - 决策面方程

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

$$-\frac{1}{2}\left[(\boldsymbol{x}-\mu_i)^T\Sigma_i^{-1}(\boldsymbol{x}-\mu_i)-(\boldsymbol{x}-\mu_j)^T\Sigma_j^{-1}(\boldsymbol{x}-\mu_j)\right]-\frac{1}{2}\ln\frac{|\Sigma_i|}{|\Sigma_j|}+\ln\frac{P(\omega_i)}{P(\omega_j)}=0$$

4.5.2 正态分布决策面

- **正态分布时分类器的决策面方程**
  - 二次型判别函数

展开

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

$$= \boldsymbol{x}^T W_i \boldsymbol{x} + \boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}$$

$$W_i = -\frac{1}{2}\Sigma_i^{-1} \qquad (d \times d \text{ 矩阵})$$

$$\boldsymbol{w}_i = \Sigma_i^{-1}\boldsymbol{\mu}_i \qquad (d \text{ 维列向量})$$

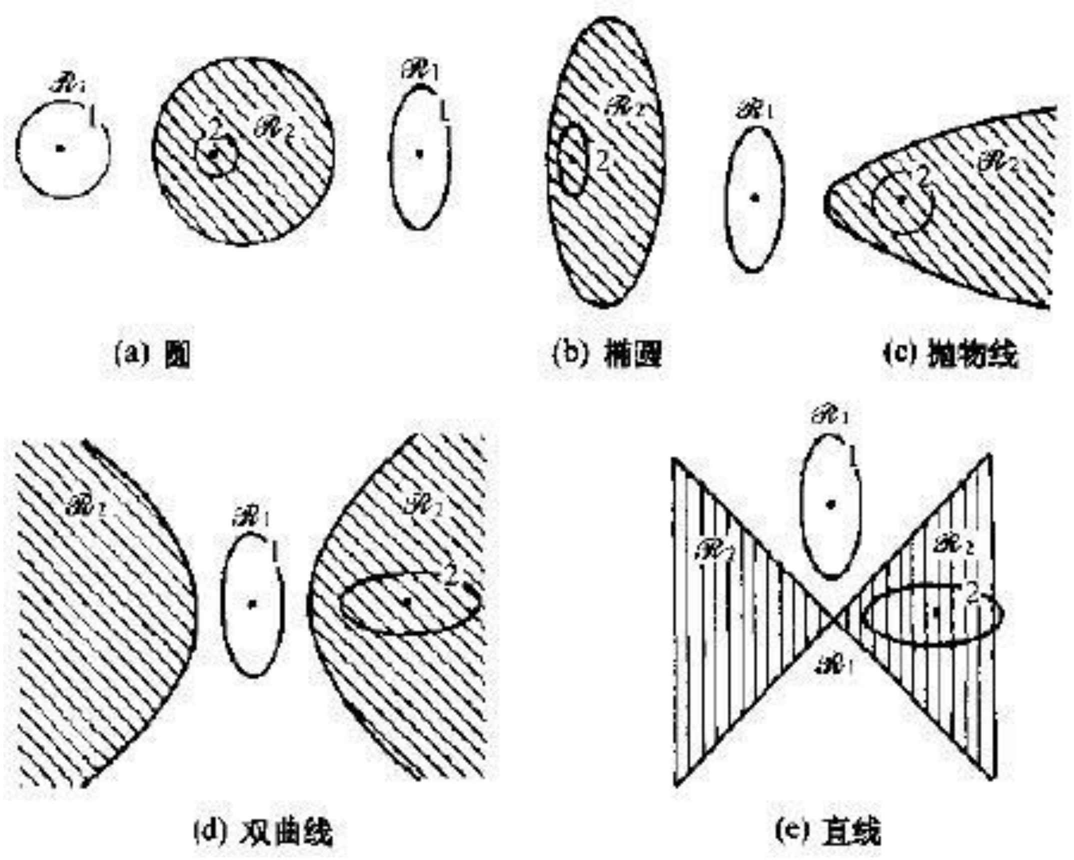
$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T\Sigma_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

  - 决策面方程

相减

4.5.2 正态分布决策面

- 决策面示例



THE END !