模式识别与机器学习

Pattern Recognition and Machine Learning

课程内容

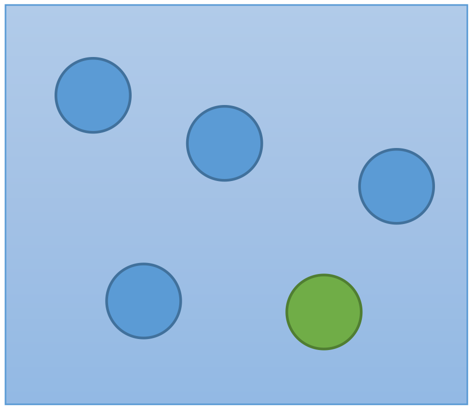
- **模式识别与机器学习概述**
- **模式识别与机器学习的基本方法**
  - 回归分析、线性判别函数、线性神经网络、核方法和支持向量机、决策树分类
  - 贝叶斯统计决策理论、概率密度函数估计
  - 无监督学习和聚类
  - 特征选择与提取

线性分类器

- 分类模型的概率视角

Box 1

$P(B_1) = 2/3$

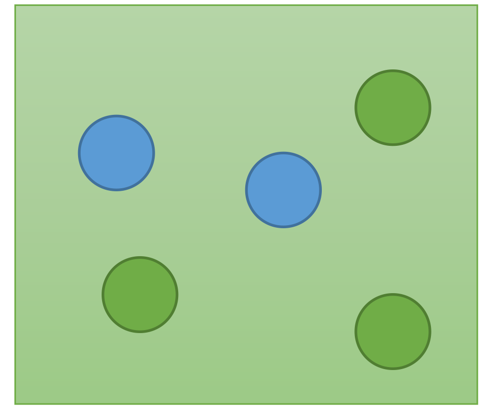


$P(\text{Blue}|B_1) = 4/5$

$P(\text{Green}|B_1) = 1/5$

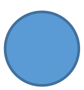
Box 2

$P(B_2) = 1/3$



$P(\text{Blue}|B_2) = 2/5$

$P(\text{Green}|B_2) = 3/5$

 from one of the boxes

Where does it come from?

$$P(B_1 \mid \text{Blue}) = \frac{P(\text{Blue}|B_1)P(B_1)}{P(\text{Blue}|B_1)P(B_1) + P(\text{Blue}|B_2)P(B_2)}$$

参考自李宏毅老师课件

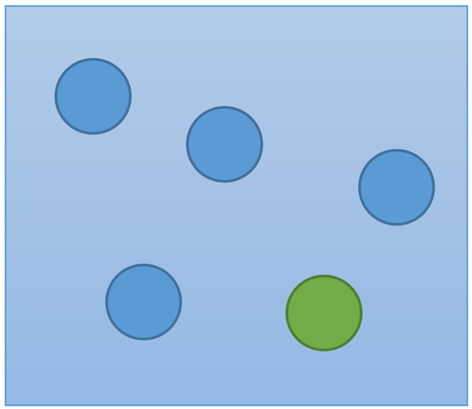
线性分类器

Estimating the Probabilities
From training data

- 分类模型的概率视角

Class 1

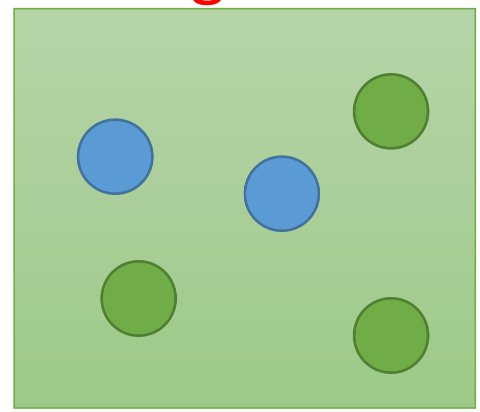
$\boxed{P(C_1)}$



$\boxed{P(x|C_1)}$

Class 2

$\boxed{P(C_2)}$



$\boxed{P(x|C_2)}$

Given an x, which class does it belong to

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

Generative Model $P(x) = P(x|C_1)P(C_1) + P(x|C_2)P(C_2)$

监督式学习--三段式解法

- 有监督学习：　找一個函数的能力

Step 0: What kind of function do you want to find?

Regression, Classification, ……

Step 1: define a set of function

Linear Classifier
Probabilistic
SVM
Deep Learning
Decision Tree

→

Step 2: goodness of function

Regression(MSE)
Classification
……

→

Step 3: pick the best function

Gradient Descent
……

参考自李宏毅老师课件

课程内容

- **线性分类**
  - 概率论基础
  - 贝叶斯统计决策
  - 参数学习
  - 线性分类器 --- 训练

概率论基础

■ 随机变量

- ➢ 随机变量：试验结果能用一个数ξ来表示，这个数ξ是随着试验的结果不同而变化的，即它是样本点的一个函数，这种量称之为随机变量。
- ➢ 离散型随机变量：试验结果ξ所可能的取值为有限个或至多可列个，这种类型的随机变量称为离散型随机变量。
- ➢ 连续型随机变量：一些随机现象所出现的试验结果不止取可列个值，这时用来描述试验结果的随机变量还是样本点的函数，但是这随机变量能取某个区间$[c, d]$或$(-\infty, +\infty)$的一切值。

概率论基础

■ 离散型概率分布

- 对于离散型随机变量，设$\{x_i\}$为离散型随机变量的所有可能取值，而$P(x_i)$是ξ取x_i的概率。
  $\{p(x_i), i=1,2,3\cdots\}$ 称为随机变量 ξ 的概率分布，
  它满足下面关系：

$$p(x_i) \geq 0, i=1,2,3\cdots$$

$$\sum_{i=1}^{\infty} p(x_i) = 1$$

- 对于离散型随机变量，通过下式求得分布函数:

$$F(x) = P\{\xi < x\} = \sum_{x_k < x} p(x_k)$$

概率论基础

■ 连续型概率分布

- 对于连续型随机变量，这种随机变量可取某个区间 $[c, d]$ 或 $(-\infty, +\infty)$ 中的一切值，其分布函数 $F(x)$ 是绝对连续函数，即存在可积函数 $p(x)$，使

$$F(x) = \int_{-\infty}^{x} p(y)\,dy$$

其中，$p(y)$称为 ξ 的概率密度函数。

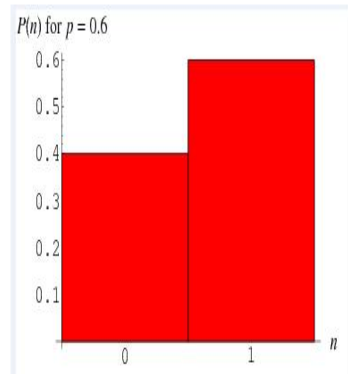
$$p(x) = F'(x)$$

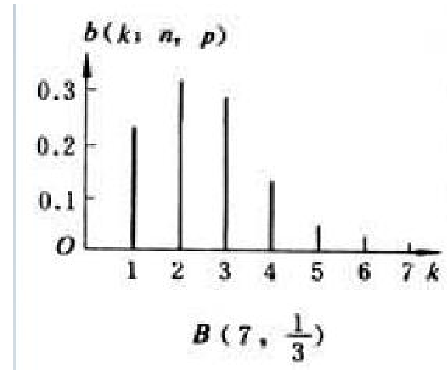
$$p(x) \geq 0$$

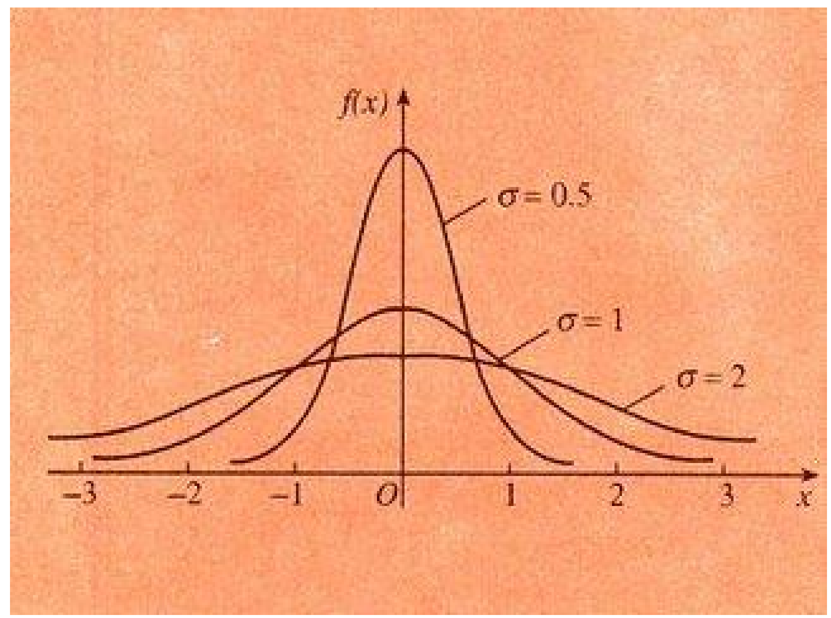
$$\int_{-\infty}^{+\infty} p(x)\,dx = 1$$

概率论基础

- **常用的随机变量分布**
  - 离散型：
    伯努利分布
    二项分布
    泊松分布
  - 连续型：
    均匀分布
    正态分布







概率论基础

■ **单变量正态分布**

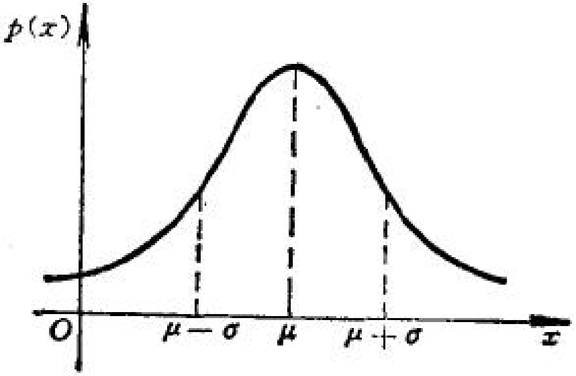
定义： $\rho(x)=\frac{1}{\sqrt{2\pi\sigma}}\exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2]$

其中： μ 为随机变量x的期望，即平均值；

σ^2 为x的方差，σ 为均方差，或标准差。

$$\mu=E(x)=\int_{-\infty}^{\infty}x\cdot\rho(x)dx$$

$$\sigma^2=\int_{-\infty}^{\infty}(x-\mu)^2\cdot\rho(x)dx$$



一维概率密度函数

概率论基础

■ 多变量正态分布

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

其中： $x=[x_1,x_2,\cdots\cdots,x_d]^T$为d维随机向量，对于d维随机向量x，它的均值向量$\mu$是$d$ 维的,Σ 是$d\times d$ 维协方差矩阵，Σ^{-1} 是Σ 的逆矩阵。$|\Sigma|$为 Σ 的行列式。

μ 和 Σ 分别是向量x和矩阵$(x-\mu)(x-\mu)^T$的期望。

- 若x_i是x的第i个分量，μ_i是 μ 的第i个分量，σ_{ij}^2是Σ的第i、j个元素

$$\mu_i=E[x_i]=\int x_i\rho(x)dx=\int_{-\infty}^{\infty}x_i\rho(x_i)dx_i$$

$$\rho(x_i)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\rho(x)dx_1dx_2\cdots dx_d$$ 为边缘分布

贝叶斯统计决策

■ 基本概念

➢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

➢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，用 $\rho(\mathbf{x}\,|\,A),\rho(\mathbf{x}\,|\,B)$ 表示

➢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 $p(A\,|\,\mathbf{x}),p(B\,|\,\mathbf{x})$ 表示

贝叶斯统计决策

- **Bayes法则**

设有 R 类样本，分别为 $\omega_1, \omega_2, ..\omega_R$，若每类的先验概率为 $p(\omega_i), i = 1, 2, .., R$。对于一随机矢量**x**，每类的条件概率为 $\rho(\mathbf{x} \mid \omega_i)$

根据Bayes公式，后验概率为：

$$p(\omega_i \mid \mathbf{x}) = \frac{\rho(\mathbf{x} \mid \omega_i)\, p(\omega_i)}{\sum_{k=1}^{R} p(\omega_k)\, p(\mathbf{x} \mid \omega_k)}$$

贝叶斯统计决策

■ 判决规则

➢目标：给定变量属于哪一类的规则(判决函数)。

- 这个规则会将输入空间划分为几个决策区域R_k，每个区域对应一个类别，如果变量**x**落入了R_k，我们就判定它属于w_k这一个类别。
- 规则的定义依据后验概率的关系给出

$$p(w_1 \mid \mathbf{x}), p(w_2 \mid \mathbf{x}), \ldots, p(w_R \mid \mathbf{x})\}$$

➢判决规则

– 最小错误率判决规则

– 最小风险判决规则

– Neyman-Pearsen判决规则

贝叶斯统计决策

■ 最小错误率

➢目标：尽可能减少错分类

以两类问题为例。

➢在两类问题中若出现错误，应该是本该是属于ω_1类别的变量却被判定为ω_2，或者是本该属于ω_2类别的变量被判定为ω_1，这个错误的概率可以表示为：

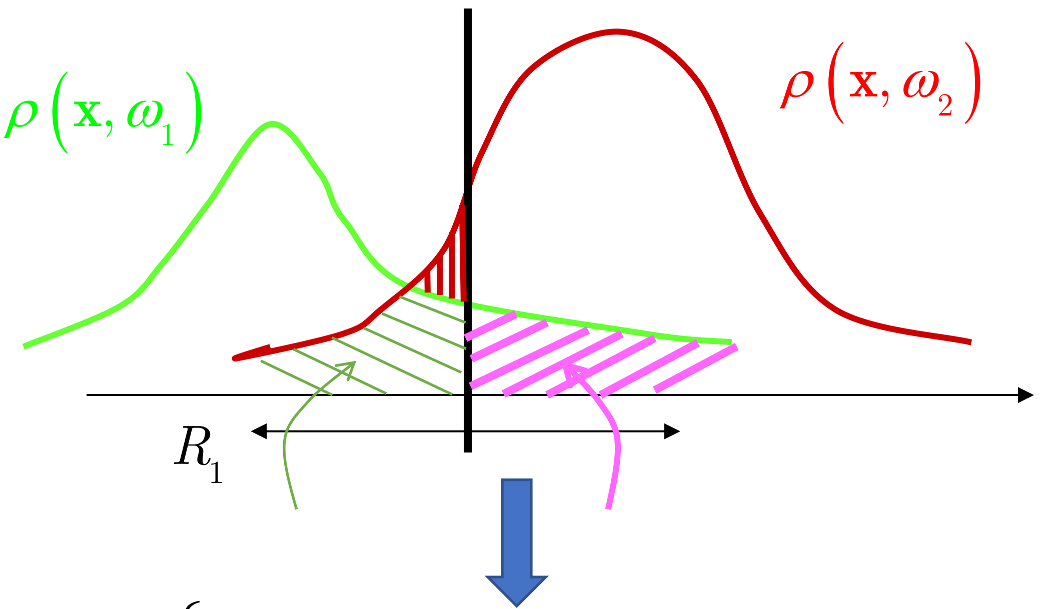
$$p(\text{mistake}) = p(\mathbf{x} \in \mathrm{R}_1, \omega_2) + p(\mathbf{x} \in \mathrm{R}_2, \omega_1)$$

$$p(\text{mistake}) = \int_{R_1} p(\mathbf{x}, \omega_2) d\mathbf{x} + \int_{R_2} p(\mathbf{x}, \omega_1) d\mathbf{x}$$

贝叶斯统计决策

■ 最小错误率

- 目标：使 $p(\text{mistake})$ 最小



$$
\begin{cases}
p(\mathbf{x}, \omega_1) > p(\mathbf{x}, \omega_2), & \mathbf{x} \in \omega_1 \\
p(\mathbf{x}, \omega_1) < p(\mathbf{x}, \omega_2), & \mathbf{x} \in \omega_2 \\
p(\mathbf{x}, \omega_1) = p(\mathbf{x}, \omega_2), & otherwise
\end{cases}
$$

贝叶斯统计决策

■ 最小错误率

- 根据贝叶斯公式，可知 $p(\mathbf{x},\omega_k)=p(\omega_k\mid\mathbf{x})p(\mathbf{x})$,

等价于

$$\boxed{\begin{cases} p(\omega_1\mid\mathbf{x})>p(\omega_2\mid\mathbf{x}), & then\ \ \mathbf{x}\in\omega_1 \\ p(\omega_1\mid\mathbf{x})<p(\omega_2\mid\mathbf{x}), & then\ \mathbf{x}\in\omega_2 \\ p(\omega_1\mid\mathbf{x})=p(\omega_2\mid\mathbf{x}), & otherwise \end{cases}}$$

最小错误率判决规则

$$\begin{cases} \rho(\mathbf{x}\mid w_1)\cdot p(w_1)>\rho(\mathbf{x}\mid w_2)\cdot p(w_2), & then\ \ \mathbf{x}\in\omega_1 \\ \rho(\mathbf{x}\mid w_1)\cdot p(w_1)<\rho(\mathbf{x}\mid w_2)\cdot p(w_2), & then\ \mathbf{x}\in\omega_2 \\ \rho(\mathbf{x}\mid w_1)\cdot p(w_1)=\rho(\mathbf{x}\mid w_2)\cdot p(w_2), & otherwise \end{cases}$$

贝叶斯统计决策

■ 最小错误率-一个例子

➢为了对癌症进行诊断，对一批人进行一次普查，各每个人打试验针，观察反应，然后进行统计，规律如下：

（1）这一批人中，每1000个人中有5个癌症病人；

（2）这一批人中，每100个正常人中有一个试验呈阳性反应；

（3）这一批人中，每100个癌症病人中有95人试验呈阳性反应。

问：若某人（甲）呈阳性反应，甲是否正常？

贝叶斯统计决策

■ 最小错误率-一个例子

➢假定x表示实验反应为阳性，

（1）人分为两类： w_1 - 正常人， w_2 - 癌症患者

$$p(w_1)+p(w_2)=1$$

（2）由已知条件计算概率值：

先验概率： $p(w_1)=0.995$ ， $p(w_2)=0.005$

类条件概率密度： $p(x|w_1)=0.01$, $p(x|w_2)=0.95$

（3）决策过程：

贝叶斯统计决策

■ **最小错误率-一个例子**

$$p(w2|x)=\frac{p(x|w_2)\cdot p(w_2)}{p(x|w_1)\cdot p(w_1)+p(x|w_2)\cdot p(w_2)}$$

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

$$=0.323$$

$$p(x|w_1)\cdot p(w_1)=0.00995,\quad p(w_1|x)=1-p(w_2|x)=1-0.323=0.677$$

$$p(x|w_2)\cdot p(w_2)=0.00475$$

$$p(w_1|x)>p(w_2|x)\qquad p(x|w_1)\cdot p(w_1)>p(x|w_2)\cdot p(w_2)$$

- 由最小错误判决规则，可知： 甲 $\in w_1$

贝叶斯统计决策

■ **最小错误率判决规则(判决函数)**

$$\begin{cases} p(\omega_1 \mid \mathbf{x}) > p(\omega_2 \mid \mathbf{x}), & then \ \mathbf{x} \in \omega_1 \\ p(\omega_1 \mid \mathbf{x}) < p(\omega_2 \mid \mathbf{x}), & then \ \mathbf{x} \in \omega_2 \\ p(\omega_1 \mid \mathbf{x}) = p(\omega_2 \mid \mathbf{x}), & otherwise \end{cases}$$

➢对于一个两类问题，有如下结论：

1.若对于某一样本**x**，有 $\rho\left(\mathbf{x} \mid \omega_1\right) = \rho\left(\mathbf{x} \mid \omega_2\right)$，则说明**x**没有提供关于类别状态的任何信息，完全取决于先验概率。

2.若 $p\left(\omega_1\right) = p\left(\omega_2\right)$，则判决完全取决于条件概率。

贝叶斯统计决策

■ 两类问题决策面方程

假定

$$g_i\left(\mathbf{x}\right)=p\left(\omega_i \mid \mathbf{x}\right)$$

$$g_i\left(\mathbf{x}\right)=p\left(\mathbf{x} \mid \omega_i\right)p\left(\omega_i\right)$$

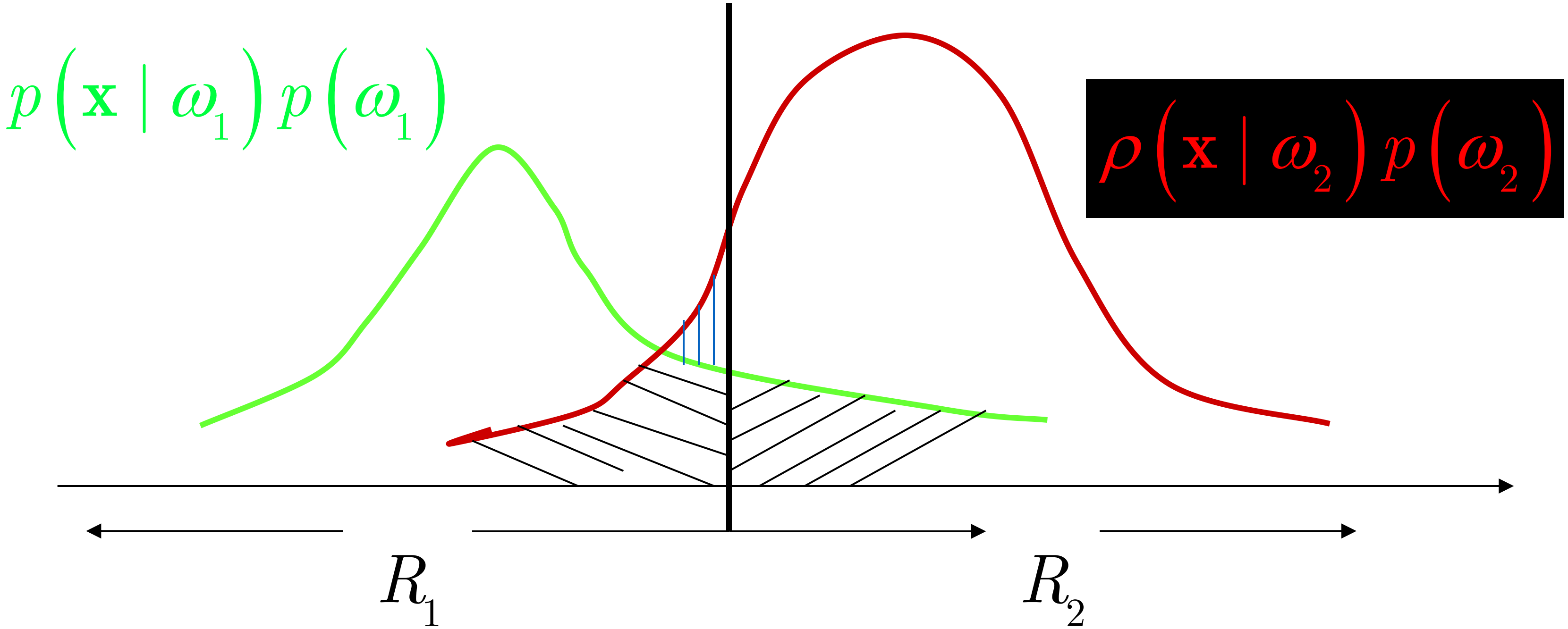
$$g_i\left(\mathbf{x}\right)=\log p\left(\mathbf{x} \mid \omega_i\right)+\log p\left(\omega_i\right)$$

贝叶斯统计决策

■ 两类问题决策面方程

二类问题：判别函数

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}) \begin{cases} > 0, & \mathbf{x} \in \omega_1 \\ < 0, & \mathbf{x} \in \omega_2 \end{cases}$$



贝叶斯统计决策

■ 两类问题决策面方程

决策面方程 $g(\mathbf{x})=0$

➢也可以表示为：

$$p(\mathbf{x}\mid w_1)\cdot p(w_1)-p(\mathbf{x}\mid w_2)\cdot p(w_2)=0$$

- 对于**上例**，用判决函数：

$$g(\mathbf{x})=p(\mathbf{x}\mid w_1)\cdot p(w_1)-p(\mathbf{x}\mid w_2)\cdot p(w_2)$$

得到对应的判决函数为：

$$g(\mathbf{x})=0.995\,p(\mathbf{x}\mid w_1)-0.005\,p(\mathbf{x}\mid w_2)$$

决策面方程为：

$$0.995\,p(\mathbf{x}\mid w_1)-0.005\,p(\mathbf{x}\mid w_2)=0$$

贝叶斯统计决策

■ 多类问题决策面方程

假定

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

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

$$g_i(\mathbf{x}) = \log p(\mathbf{x} \mid \omega_i) + \log p(\omega_i)$$

但是它们的效果是一样的，都是把特征空间分割成多个不同的决策区域。

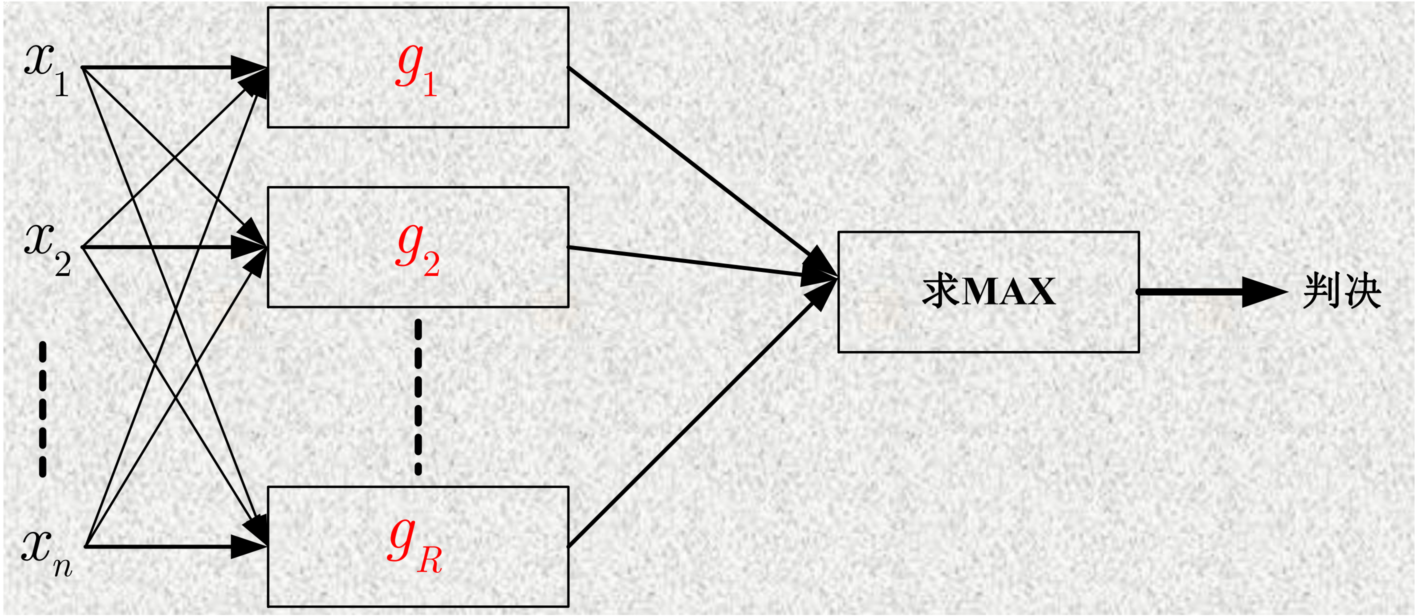
$$g_i(\mathbf{x}) > g_j(\mathbf{x}), \forall j \neq i \Rightarrow \mathbf{x} \in \omega_i$$

贝叶斯统计决策

■ **多类问题最小错误率判决**

➢ **多类分类器结构**

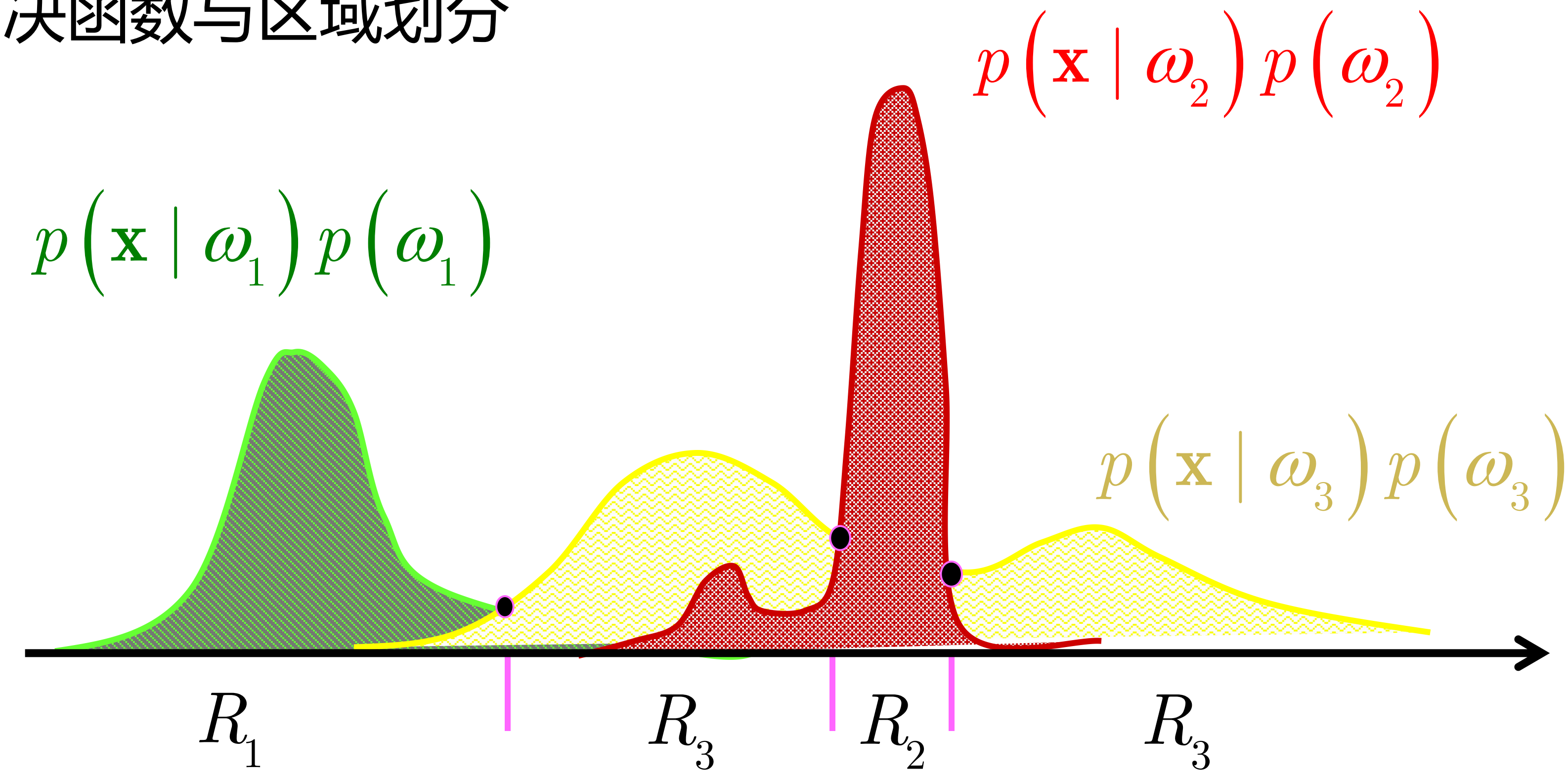
$$\mathbf{x} = (x_1, x_2, ..x_n)^T \qquad \omega_1, \omega_2, ..\omega_R$$



若有 $g_i(\mathbf{x}) > g_j(\mathbf{x}), \forall j \neq i \Rightarrow \mathbf{x} \in \omega_i$

贝叶斯统计决策

- **多类问题最小错误率判决**
  - 判决函数与区域划分



决策面： $g_i(\mathbf{x}) = g_j(\mathbf{x})$ **且** R_i **与** R_j **相邻**

贝叶斯统计决策

- **正态分布时的统计决策**

一维：

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

散布程度归一化距离

d维：

$$\rho(\mathbf{x})=\frac{1}{(2\pi)^{\frac{d}{2}}\,|\Sigma|^{\frac{1}{2}}}\exp\left[-\frac{\gamma^2}{2}\right]$$

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

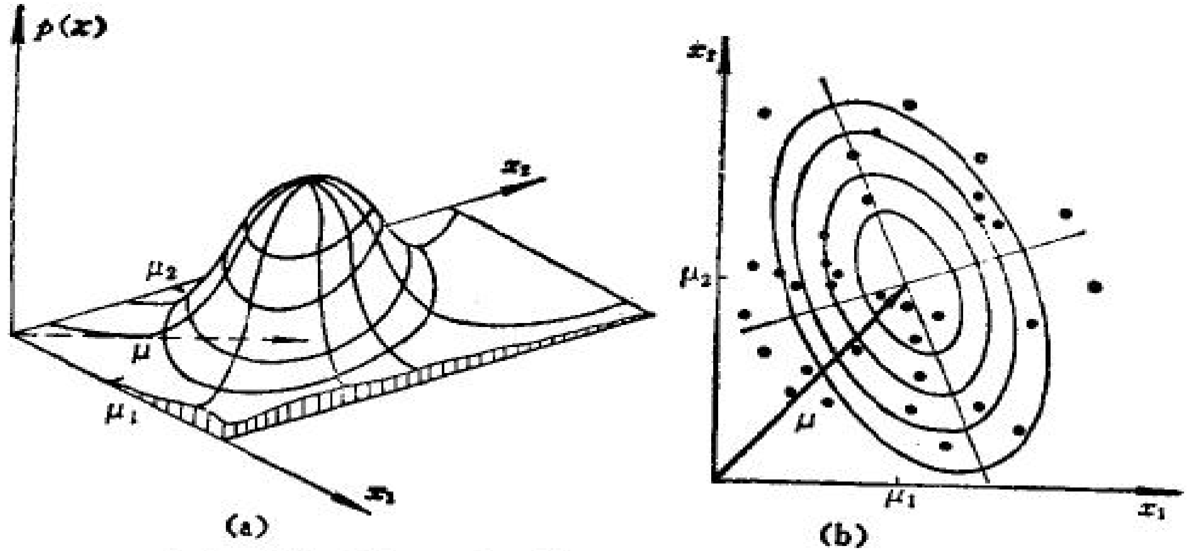
马氏距离

$$\rho(X)\sim N(\mu,\Sigma)$$

贝叶斯统计决策

■ 正态分布时的统计决策

可以证明，上式的解是一个超椭球面，其主轴方向取决于Σ的本征向量（特征向量），主轴的长度与相应的本征值成正比。如下图所示：



(a) $p(x)$ ($d=2$); (b) 等密度点轨迹

贝叶斯统计决策

■ 多元正态分布函数的性质

（1）参数μ和Σ对分布的决定性:

$\rho(x)$可由μ、Σ完全确定

（2）等密度点的轨迹为一超椭球面：

由 $\rho(x)$ 的定义公式可知，当右边指数项为常数时，密度 $\rho(x)$ 的值不变，所以等密度点满足：

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

（3）不相关性等价于独立性：

对服从正态分布的两个分量互不相关，则它们之间一定独立。

贝叶斯统计决策

■ 多元正态分布函数的性质

（4）边缘分布与条件分布的等价性：

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。

（5）线性变换的正态性：

对于多元随机向量的线性变换，仍为多元正态分布的随机向量。

（6）线性组合的正态性

若x为多元正态随机向量，则线性组合 $y = a^T x$ 是一维的正态随机变量：

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

其中，a与x同维。

贝叶斯统计决策

■ 正态分布中的Bayes分类方法

- 符合Bayes判决

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

- 若条件概率密度为

$$\rho(\mathbf{x} \mid \omega_i) \sim N(\mu_i, \Sigma_i)$$

则

$$g_i(\mathbf{x}) = -\frac{d}{2}\log 2\pi - \frac{1}{2}\log |\Sigma_i| - \frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1} (\mathbf{x}-\mu_i) + \log p(\omega_i)$$

各特征分量统计独立，且具有相同的方差
协方差矩阵相同
协方差矩阵互不相同

贝叶斯统计决策

- **正态分布中的Bayes分类方法--第一种情况**
  - **各特征分量统计独立，且具有相同的方差**

$$\Sigma_i = \sigma^2 I = \sigma^2 \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

则

$$|\Sigma_i| = \sigma^2 \qquad \Sigma_i^{-1} = \frac{I}{\sigma^2}$$

特点: 抽样落在同样大小的超球内。第i类抽样，以 μ_i为中心。

贝叶斯统计决策

■ 正态分布中的Bayes分类方法--第一种情况

判决函数

$$g_i(\mathbf{x}) = -\frac{(\mathbf{x}-\mu_i)^T(\mathbf{x}-\mu_i)}{2\sigma^2} + \log p(\omega_i)$$

若 $P(\omega_i)$ 相同，则

最小距离分类器

$$g_i(\mathbf{x}) = -\frac{\boxed{(\mathbf{x}-\mu_i)^T(\mathbf{x}-\mu_i)}}{2\sigma^2}$$

欧氏距离

展开

贝叶斯统计决策

■ 正态分布中的Bayes分类方法--第一种情况

判决函数

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}(\mathbf{x}^T\mathbf{x} - 2\mu_i^T\mathbf{x} + \mu_i^T\mu_i) + \log p(\omega_i)$$

$$g_i(\mathbf{x}) = W_i^T X + W_{i0}$$

其中， $W_i = \frac{1}{\sigma^2}\mu_i$ $\quad W_{i0} = -\frac{1}{2\sigma^2}\mu_i^T\mu_i + \log p(\omega_i)$

说明： $\mathbf{x}^T\mathbf{x}$ 对于所有的i均相同。

贝叶斯统计决策

■ 正态分布中的Bayes分类方法--第一种情况

决策面

若类 ω_i 与类 ω_j 相邻，则决策面由

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

确定。

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}(-2\mu_i^T\mathbf{x} + \mu_i^T\mu_i) + \log p(\omega_i)$$

$$g_j(\mathbf{x}) = -\frac{1}{2\sigma^2}(-2\mu_j^T\mathbf{x} + \mu_j^T\mu_j) + \log p(\omega_j)$$

贝叶斯统计决策

■ 正态分布中的Bayes分类方法--第一种情况

决策面 $$W^T\mathbf{x} - W^T\mathbf{x}_0 = 0$$

$$
\begin{aligned}
0 &= g_i(\mathbf{x}) - g_j(\mathbf{x}) \\
&= \frac{1}{\sigma^2}(\mu_i^T - \mu_j^T)\mathbf{x} - \frac{1}{\sigma^2}\left[\frac{1}{2}(\mu_i^T\mu_i - \mu_j^T\mu_j) - \sigma^2 \log\frac{p(\omega_i)}{p(\omega_j)}\right] \\
&= (\mu_i^T - \mu_j^T)\mathbf{x} \\
&\quad -(\mu_i - \mu_j)^T\left[\frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{(\mu_i - \mu_j)^T(\mu_i - \mu_j)}(\mu_i - \mu_j)\log\frac{p(\omega_i)}{p(\omega_j)}\right] \\
&= W^T\mathbf{x} - W^T\mathbf{x}_0
\end{aligned}
$$

$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{(\mu_i - \mu_j)^T(\mu_i - \mu_j)}(\mu_i - \mu_j)\log\frac{p(\omega_i)}{p(\omega_j)}$$

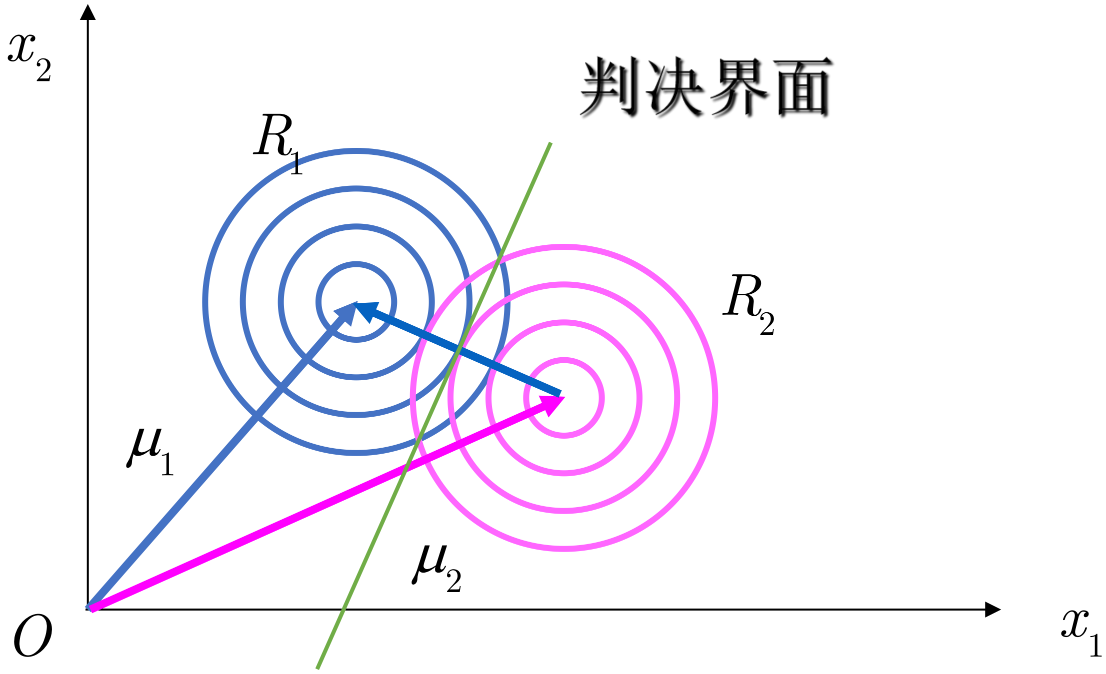
$$W = (\mu_i - \mu_j)$$

贝叶斯统计决策

- **正态分布中的Bayes分类方法--第一种情况**

结论 $$W^T\mathbf{x} - W^T\mathbf{x}_0 = 0$$

(1) 满足此方程的超平面通过x_0平行于W^T。由于$W = (\mu_i - \mu_j)$，故超平面垂直于均值之间的连线。

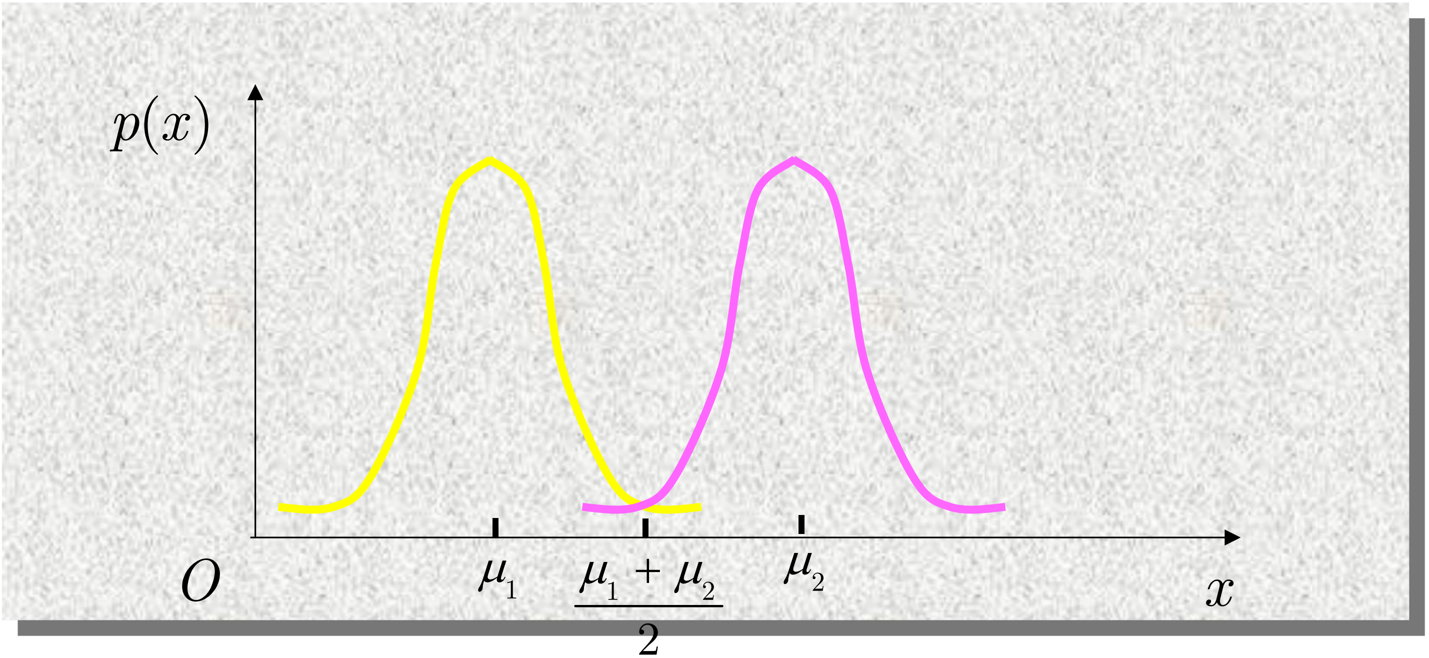


贝叶斯统计决策

■ **正态分布中的Bayes分类方法--第一种情况**

结论

(2) 如果$p(\omega_i) = p(\omega_j)$，则$\mathbf{x}_0$将通过均值连线的中点。

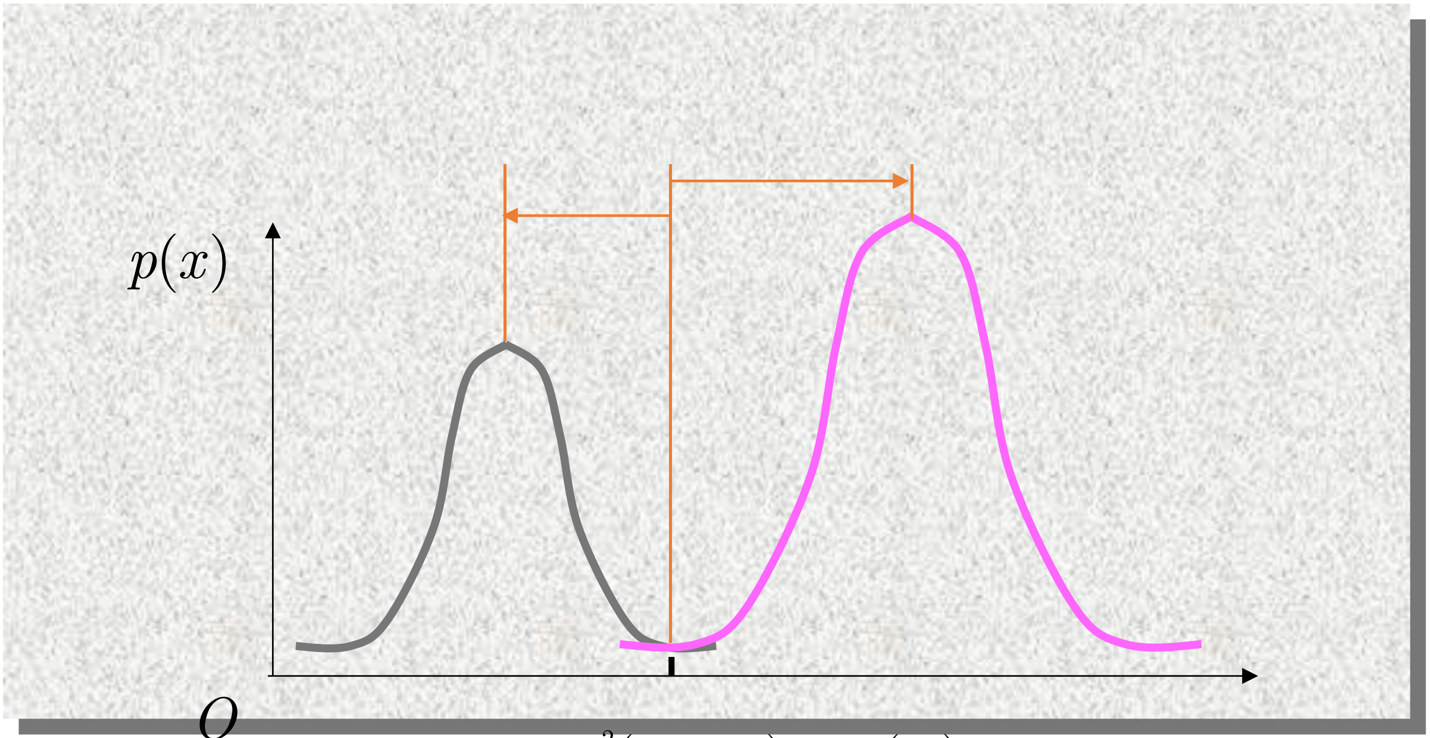


贝叶斯统计决策

■ 正态分布中的Bayes分类方法--第一种情况

结论

(3)如果 $p(\omega_i) \neq p(\omega_j)$ ，则x0离开先验概率较大的均值。



$$\frac{\mu_1+\mu_2}{2}-\frac{\sigma^2(\mu_1-\mu_2)}{|\mu_1-\mu_2|^2}\log\frac{p(\omega_1)}{p(\omega_2)}$$

贝叶斯统计决策

■ **正态分布中的Bayes分类方法--第二种情况**

协方差矩阵相同 $\Sigma_i = \Sigma \quad \rho \neq 0 \quad \sigma_1^2 \neq \sigma_2^2$

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

与类别i无关

$$g_i(\mathbf{x}) = \boxed{-\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_i|} - \frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i) + \log p(\omega_i)$$

$$= -\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma^{-1}(\mathbf{x}-\mu_i) + \log p(\omega_i)$$

展开

特点: 抽样落在超椭球簇内。第i类的簇以μ_i为中心。

■ 正态分布中的Bayes分类方法--第二种情况

➢判决函数

$$g_i(\mathbf{x}) = W_i^T \mathbf{x} + W_{i0}$$

$$W_i = \Sigma^{-1} \mu_i \qquad W_{i0} = -\frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \log p(\omega_i)$$

➢判决面

$$W^T(\mathbf{x} - \mathbf{x}_0) = 0$$

$$W = \Sigma^{-1}(\mu_i - \mu_j)$$

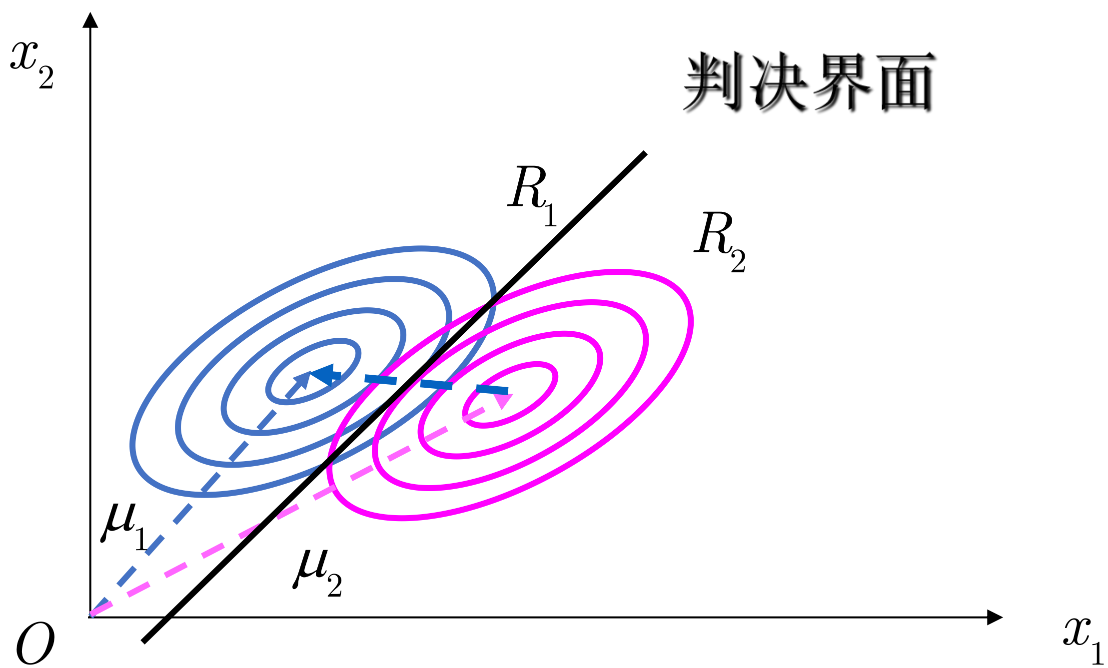
$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{1}{(\mu_i - \mu_j)^T \Sigma^{-1} (\mu_i - \mu_j)} (\mu_i - \mu_j) \log \frac{p(\omega_i)}{p(\omega_j)}$$

贝叶斯统计决策

■ **正态分布中的Bayes分类方法--第二种情况**

结论

(1) 由于 $\Sigma^{-1}(\mu_i - \mu_j)$ 通常与 $(\mu_i - \mu_j)$ 方向不一致，故分割R_1与R_2的超平面一般不垂直于均值的连线。



贝叶斯统计决策

- **正态分布中的Bayes分类方法--第二种情况**

  结论

  (2)若先验概率相等，则此超平面与均值连线相交在均值连线的中点，即

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)$$

  (3)若先验概率不相等，则判断界面就是离开先验概率较大的那个均值。

贝叶斯统计决策

- **正态分布中的Bayes分类方法--第三种情况**
  - **协方差矩阵互不相同**

    判别函数不再是线性的，而是二次型的。

$$g_i(\mathbf{x}) = \mathbf{x}^T W_i \mathbf{x} + W_{i1}^T \mathbf{x} + W_{i0}$$

$$W_i = -\frac{1}{2}\Sigma_i^{-1}$$

$$W_{i1} = \Sigma_i^{-1}\mu_i$$

$$W_{i0} = -\frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i - \frac{1}{2}\log\left|\Sigma_i\right| + \log p(\omega_i)$$

贝叶斯统计决策

■ 正态分布中的Bayes分类方法--第三种情况

➢决策面

- 超二次曲面对：超平面对、超球对、超椭球对、超抛物面对、超双曲面对
- 二维情况下，假设分量x,y是类条件独立的($\rho=0$)，故协方差矩阵是对角形的，从而不同的决策面与各自的方差有关。

参数学习

- 参数估计是知道概率密度的分布形式，但其中的部分未知或全部未知。概率密度函数估计就是通过样本来估计这些参数。
- 非参数估计是既不知道分布形式，也不知道分布里的参数，通过样本的分布把概率密度函数值数值化估计出来

➢参数估计方法
  - 最大似然估计
  - 贝叶斯估计
  - EM估计方法

非参数估计方法
  -Parzen窗法
  -Kn近邻法

参数学习

- 最大似然估计
  - 设已知样本集有样本类 $\mathscr{X}_1, \mathscr{X}_2, \cdots, \mathscr{X}_c$ ，其中 $\mathscr{X}_j$ 类有样本 $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ ，是按概率密度 $p(\mathbf{x} \mid \omega_j)$ 从总体中独立地抽取的，但是其中某一参数 μ 或参数矢量 (μ, σ) 不知道，记作参数 θ_j 。
  - 似然函数：$p(\mathscr{X} \mid \theta)$

    同一类的样本子集 $\mathscr{X} = \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$，它们具有概率密度 $p(\mathbf{x}_k \mid \theta), k = 1, 2, \cdots n$，且样本是独立抽取的

参数学习

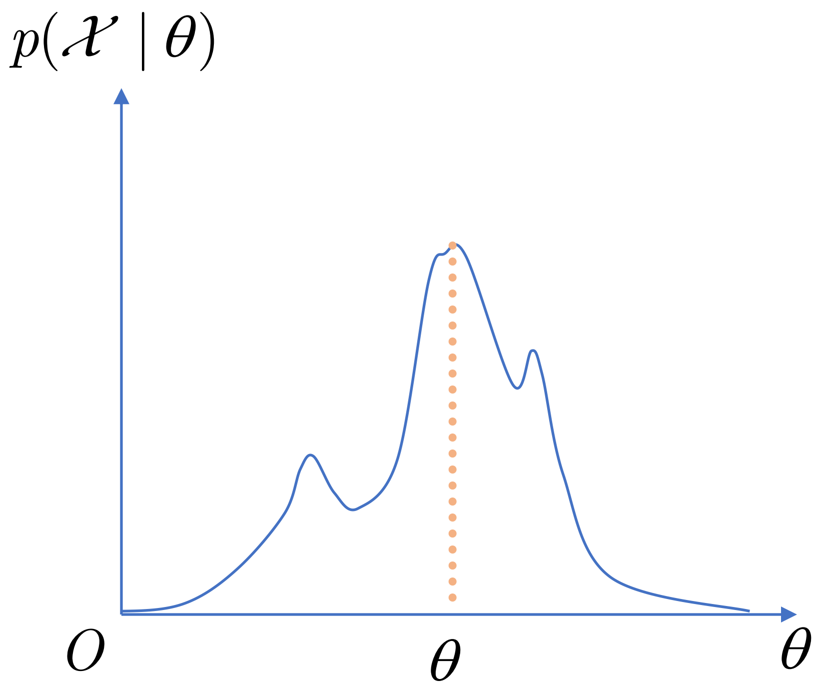
- 最大似然估计
  - 似然函数：$p(\mathscr{X} \mid \theta)$

    同一类的样本子集 $\mathscr{X} = \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$，它们具有概率密度 $p(\mathbf{x}_k \mid \theta), k = 1, 2, \cdots n$，且样本是独立抽取的

$$p(\mathscr{X} \mid \theta) = \prod_{k=1}^{n} p(\mathbf{x}_k \mid \theta),$$

$$L(\theta) = p(\mathscr{X} \mid \theta)$$

$$\hat{\theta} = \arg\max L(\theta)$$



参数学习

- 最大似然估计
  - 似然函数：$p(\mathscr{X} \mid \theta)$

$$L(\theta) = \log p(\mathscr{X} \mid \theta) = \sum_{k=1}^{n} \log p(\mathbf{x}_k \mid \theta),$$

$$\hat{\theta} = \arg\max L(\theta)$$

计算:

$$\nabla_\theta L = \frac{\partial}{\partial \theta}\left(\log p(\mathscr{X} \mid \theta)\right) = \sum_{k=1}^{n} \frac{\partial}{\partial \theta}\left[\log p(\mathbf{x}_k \mid \theta)\right] = 0$$

$$\nabla_\theta = \begin{Bmatrix} \dfrac{\partial}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial}{\partial \theta_p} \end{Bmatrix}$$

参数学习

■ 正态分布下的最大似然估计

➢ 均值、方差未知的一维正态情况

$$\theta_1 = \mu, \quad \theta_2 = \sigma^2$$

$$p(\mathbf{x}_k \mid \theta) = \frac{1}{\sqrt{2\pi\theta_2}} \exp\left[-\frac{1}{2}\frac{(\mathbf{x}_k - \theta_1)^2}{\theta_2}\right]$$

log

$$\log p(\mathbf{x}_k \mid \theta) = -\frac{1}{2}\log 2\pi\theta_2 - \frac{1}{2}\frac{(\mathbf{x}_k - \theta_1)^2}{\theta_2}$$

参数学习

- 正态分布下的最大似然估计
  - 均值、方差未知的一维正态情况

$$\log p(\mathbf{x}_k \mid \theta) = -\frac{1}{2}\log 2\pi\theta_2 - \frac{1}{2}\frac{(\mathbf{x}_k - \theta_1)^2}{\theta_2}$$

$$\nabla_{\theta_1}$$

$$\nabla_{\theta_1} \log p(\mathbf{x}_k \mid \theta) = -\left[\frac{1}{2\theta_2}\cdot 2(\mathbf{x}_k - \theta_1)\cdot(-1)\right] = \frac{\mathbf{x}_k - \theta_1}{\theta_2}$$

- 均值

$$\sum_{k=1}^{n}\nabla_{\theta_1} L = \frac{1}{\theta_2}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\theta}_1) = 0 \qquad \sum_{k=1}^{n}(\mathbf{x}_k - \hat{\theta}_1) = 0$$

$$\boxed{\hat{\theta}_1 = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k}$$

参数学习

- 正态分布下的最大似然估计
  - 均值、方差未知的一维正态情况

$$\log p(\mathbf{x}_k \mid \theta) = -\frac{1}{2}\log 2\pi\theta_2 - \frac{1}{2}\frac{(\mathbf{x}_k - \theta_1)^2}{\theta_2}$$

∇_{θ_2}

$$\nabla_{\theta_2} \log p(\mathbf{x}_k \mid \theta) = -\frac{1}{2}\left[\left(\frac{1}{2\theta_2}\cdot 2\pi\right) - \frac{(\mathbf{x}_k - \theta_1)^2}{2}\cdot(-1)\theta_2^{-2}\right]$$

$$= -\frac{1}{2\theta_2} + \frac{(\mathbf{x}_k - \theta_1)^2}{2\theta_2^2}$$

- 方差：有偏估计

$$\sum_{k=1}^{n}\nabla_{\theta_2} L = \frac{1}{2\theta_2}\left[\sum_{k=1}^{n}(-1) + \sum_{k=1}^{n}\frac{(\mathbf{x}_k - \theta_1)^2}{\theta_2}\right] = 0$$

$$\hat{\theta}_2 = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \theta_1)^2 \quad \approx \quad \sigma^2 = \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k - \mu)^2$$

参数学习

- 正态分布下的最大似然估计
  - 均值未知的d维正态情况

设$\mathscr{X}$ 中的某一样本 $\mathbf{x}_k=(x_{k1},x_{k2},\cdots,x_{kd})^T$ 具有正态形式，参数 μ 未知，

$$p(\mathbf{x}_k \mid \mu)=\frac{1}{(2\pi)^{\frac{d}{2}} \mid \Sigma \mid^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\mathbf{x}_k-\mu)^T\Sigma^{-1}(\mathbf{x}_k-\mu)\right]$$

log

$$\log p(\mathbf{x}_k \mid \mu)=-\frac{1}{2}\log\left[(2\pi)^d \mid \Sigma \mid\right]-\frac{1}{2}(\mathbf{x}_k-\mu)^T\Sigma^{-1}(\mathbf{x}_k-\mu)$$

若干基础知识：

$$\nabla_\theta \log p(\mathbf{x}_k \mid \mu)=\frac{\partial}{\partial\mu}(\mathbf{x}_k-\mu)^T\Sigma^{-1}(\mathbf{x}_k-\mu)$$

$$=\ldots$$

参数学习

- 正态分布下的最大似然估计
  - 均值未知的d维正态情况

$$\nabla_{\mu} \log p(\mathbf{x}_k \mid \mu) = \frac{\partial}{\partial \mu}(\mathbf{x}_k - \mu)^T \Sigma^{-1} \mathbf{x}_k - \mu)$$

$$= \frac{\partial}{\partial \mu}(\mathbf{x}_k - \mu)^T \left[\Sigma^{-1}(\mathbf{x}_k - \mu)\right] + (\mathbf{x}_k - \mu)^T \frac{\partial}{\partial \mu}\left[\Sigma^{-1}(\mathbf{x}_k - \mu)\right]$$

$$= [-1]^T \left[\Sigma^{-1}(\mathbf{x}_k - \mu)\right] + \left[\Sigma^{-1}(\mathbf{x}_k - \mu)\right]^T [-1]^T$$

$$= 2[-1]^T \left[\Sigma^{-1}(\mathbf{x}_k - \mu)\right]$$

$$\nabla_{\mu} L = 2[-1]^T \left[\Sigma^{-1}(\mathbf{x}_k - \hat{\mu})\right] = 0 \Rightarrow \sum_{k=1}^{n} (\mathbf{x}_k - \hat{\mu}) = 0$$

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k$$