模式识别与机器学习

Pattern Recognition and Machine Learning

课程内容

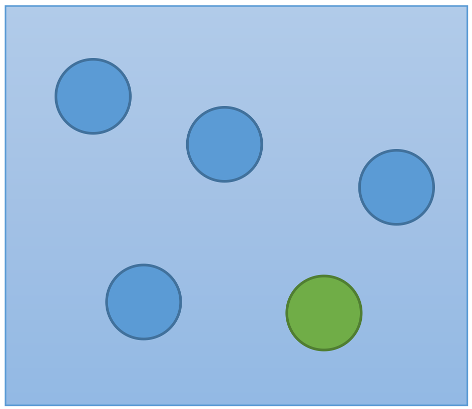
- **模式识别与机器学习概述**
- **模式识别与机器学习的基本方法**
  - 回归分析、线性判别函数、线性神经网络、核方法和支持向量机、决策树分类
  - 贝叶斯统计决策理论、概率密度函数估计
  - 无监督学习和聚类
  - 特征选择与提取

线性分类器

- 分类模型的概率视角

Box 1

$P(B_1) = 2/3$

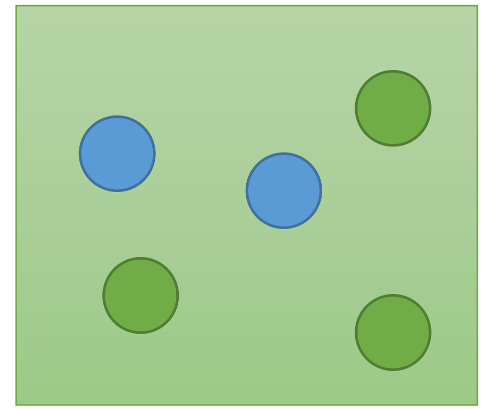


$P(\text{Blue}|B_1) = 4/5$

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

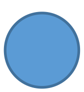
Box 2

$P(B_2) = 1/3$



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

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

 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

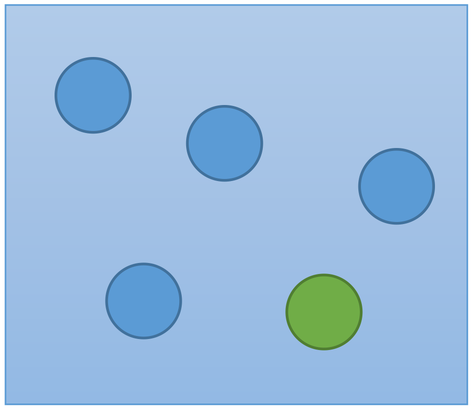
Estimating the Probabilities

From training data

■ 分类模型的概率视角

Class 1

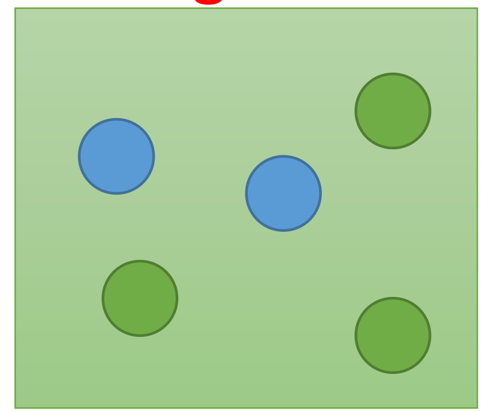
$P(C_1)$



$P(x|C_1)$

Class 2

$P(C_2)$



$P(x|C_2)$

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

参考自李宏毅老师课件

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:

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

■ 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 $F(\mathrm{x})$ 是绝对连续函数，即存在可积函数 $p(\mathrm{x})$，使

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

其中，$p(\mathrm{y})$称为 ξ 的概率密度函数。

$$p(\mathrm{x}) = \mathrm{F}'(\mathrm{x})$$

$$p(\mathrm{x}) \geq 0$$

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

概率论基础

- **常用的随机变量分布**
  - 离散型：

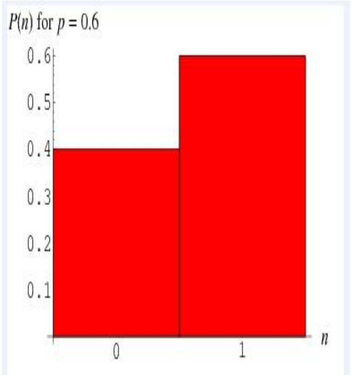
    伯努利分布

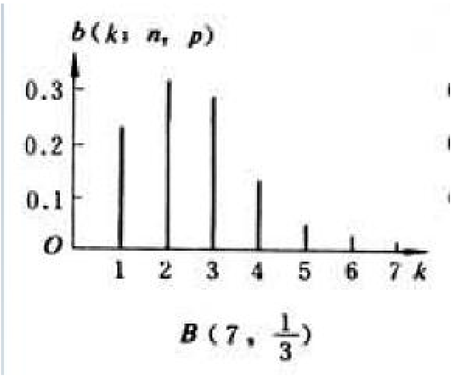
    二项分布

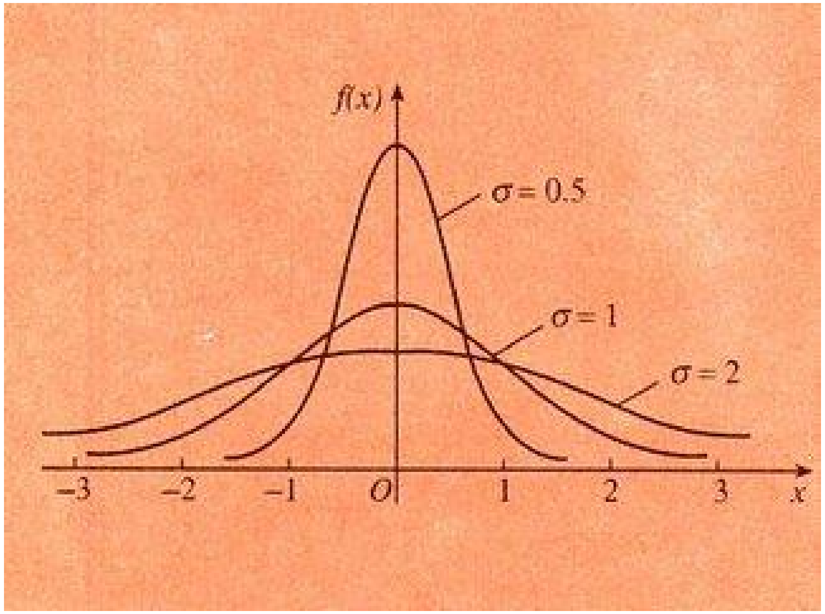
    泊松分布
  - 连续型：

    均匀分布

    正态分布







概率论基础

■ 单变量正态分布

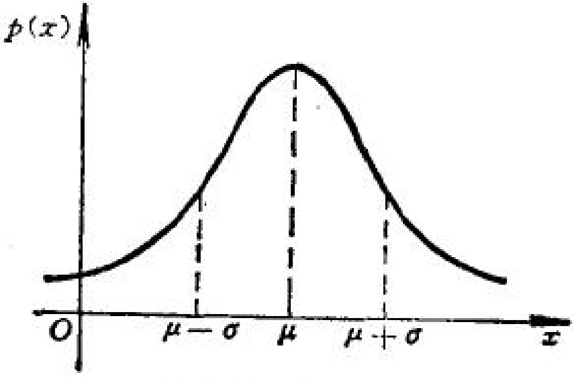
定义： $\rho(x)=\frac{1}{\sqrt{2\pi\sigma}}\exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2]$

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 $\rho(\mathbf{x} \mid A), \rho(\mathbf{x} \mid B)$ 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 $p(A \mid \mathbf{x}), p(B \mid \mathbf{x})$ 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 $p\left(\omega_i\right), i = 1, 2, .., R$。对于一随机矢量**x**，每类的条件概率为 $\rho\left(\mathbf{x} \mid \omega_i\right)$

根据Bayes公式，后验概率为：

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

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：

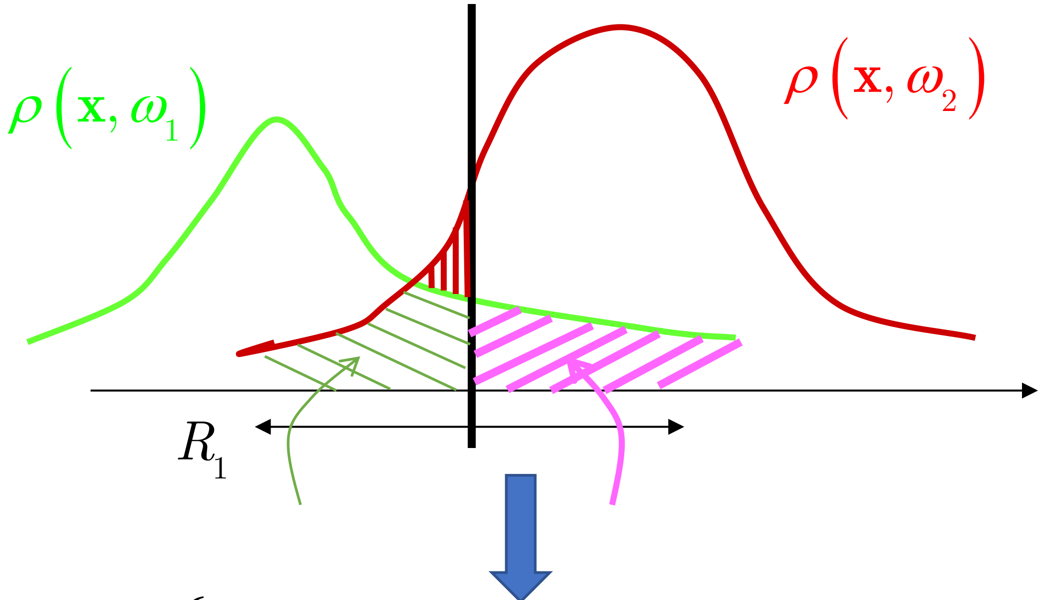
$$p(\text{mistake}) = p(\mathbf{x} \in \mathrm{R}_1, \omega_2) + p(\mathbf{x} \in \mathrm{R}_2, \omega_1)$$

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

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

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

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

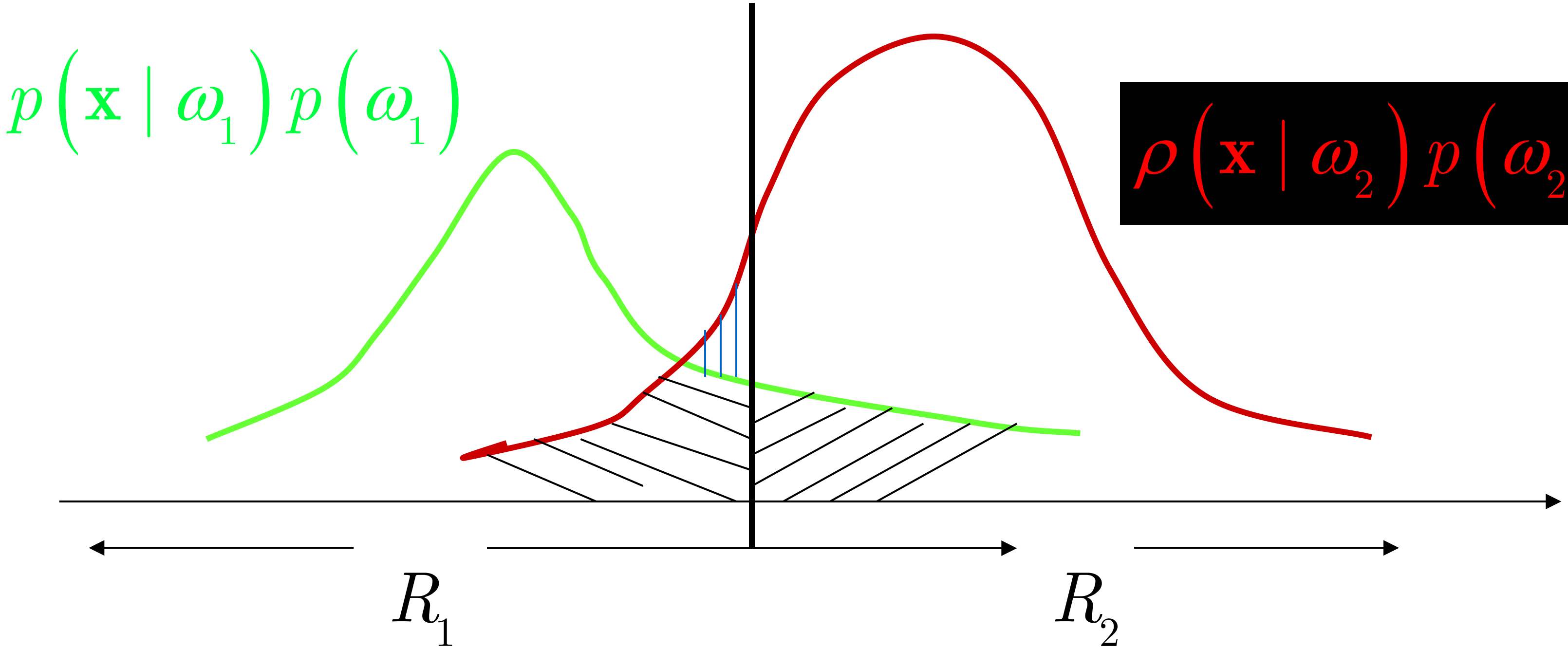
$$g_i(\mathbf{x}) = \log p(\mathbf{x} \mid \omega_i) + \log p(\omega_i)$$

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：

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

- 对于**上例**，用判决函数：

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

得到对应的判决函数为：

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

决策面方程为：

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

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。

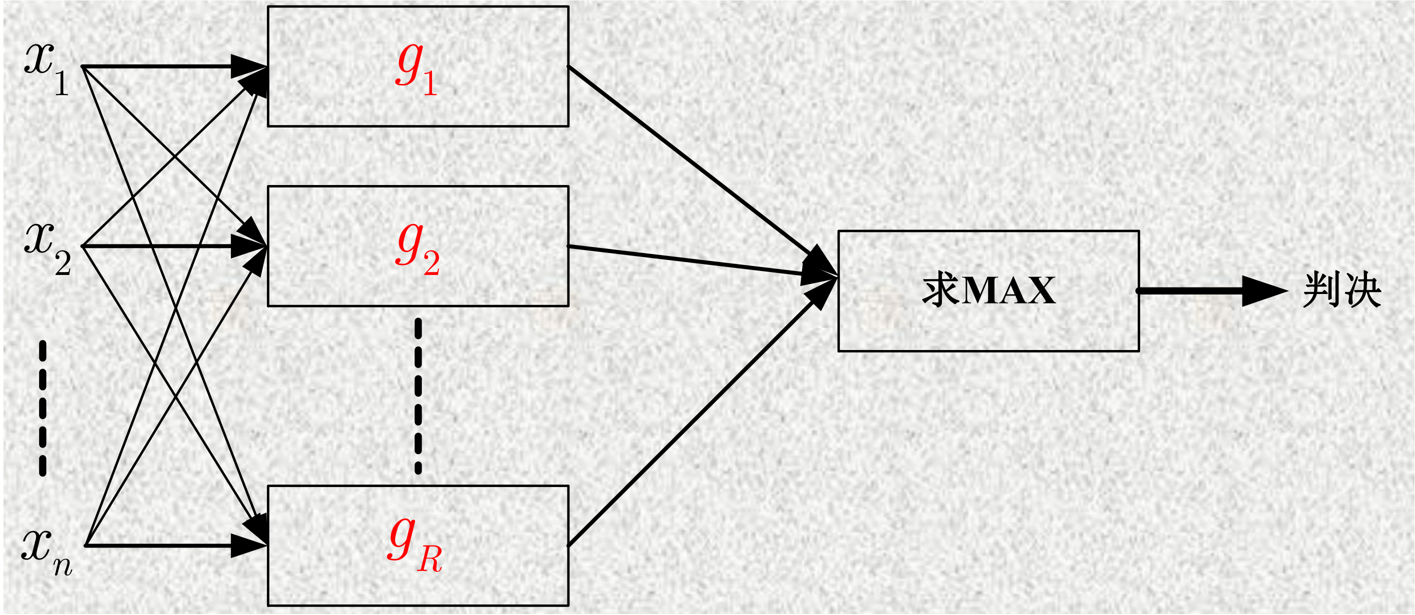
$$g_i(\mathbf{x}) > g_j(\mathbf{x}), \forall j \neq i \Rightarrow \mathbf{x} \in \omega_i$$

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**

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

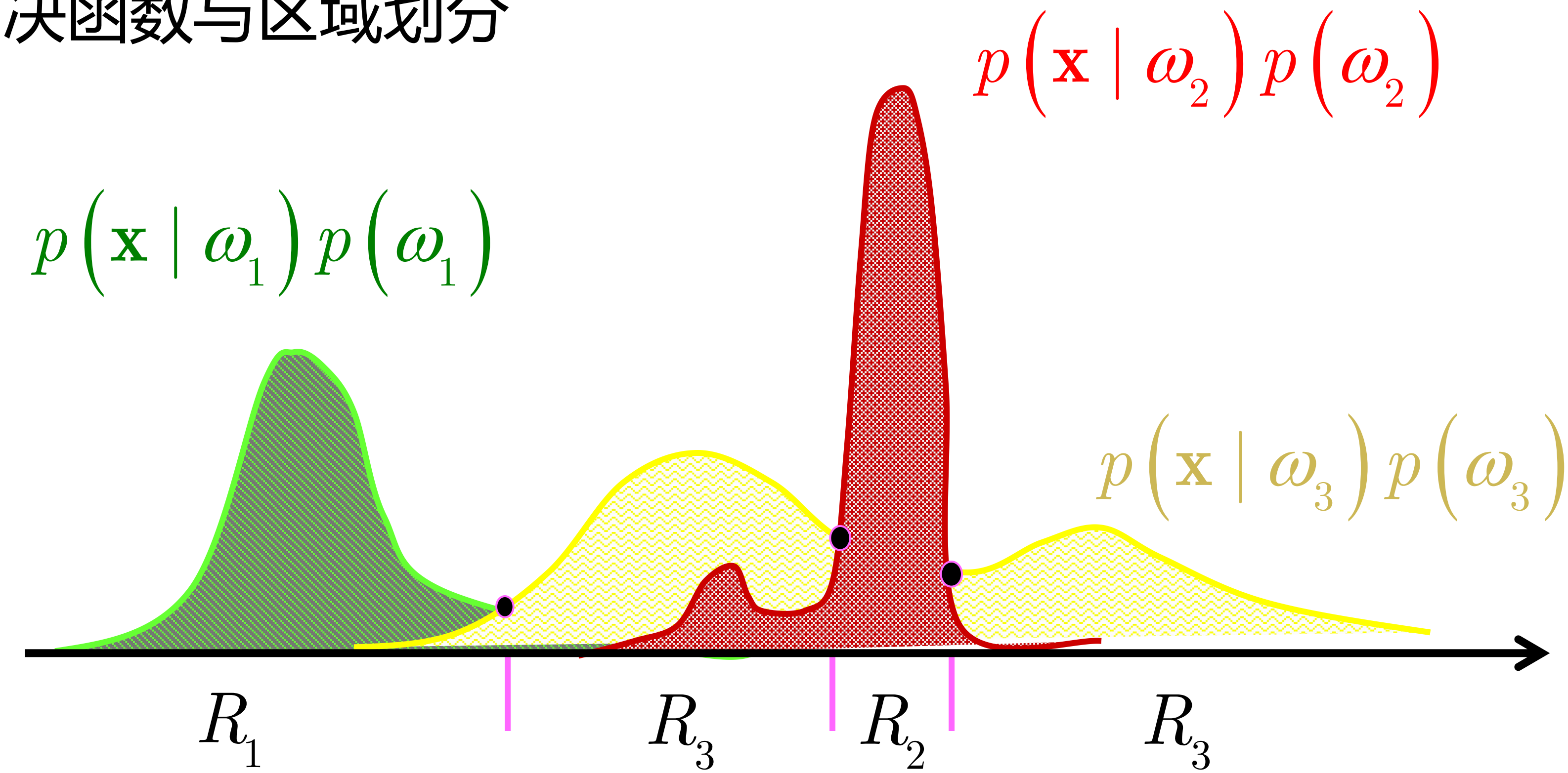


若有 $g_i(\mathbf{x}) > g_j(\mathbf{x}), \forall j \neq i \Rightarrow \mathbf{x} \in \omega_i$

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：

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

马氏距离

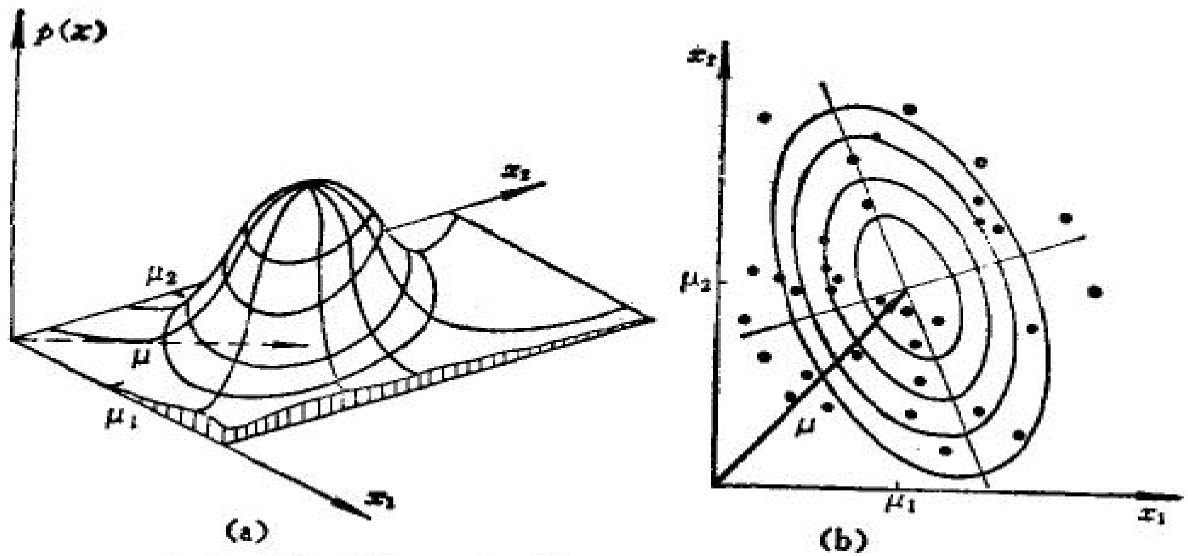
$$\gamma^2 = (\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)$$

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

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。

■ 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：

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

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

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

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

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

最小距离分类器

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

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

确定。

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

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

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

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

$||$

$\mathbf{x}_0$

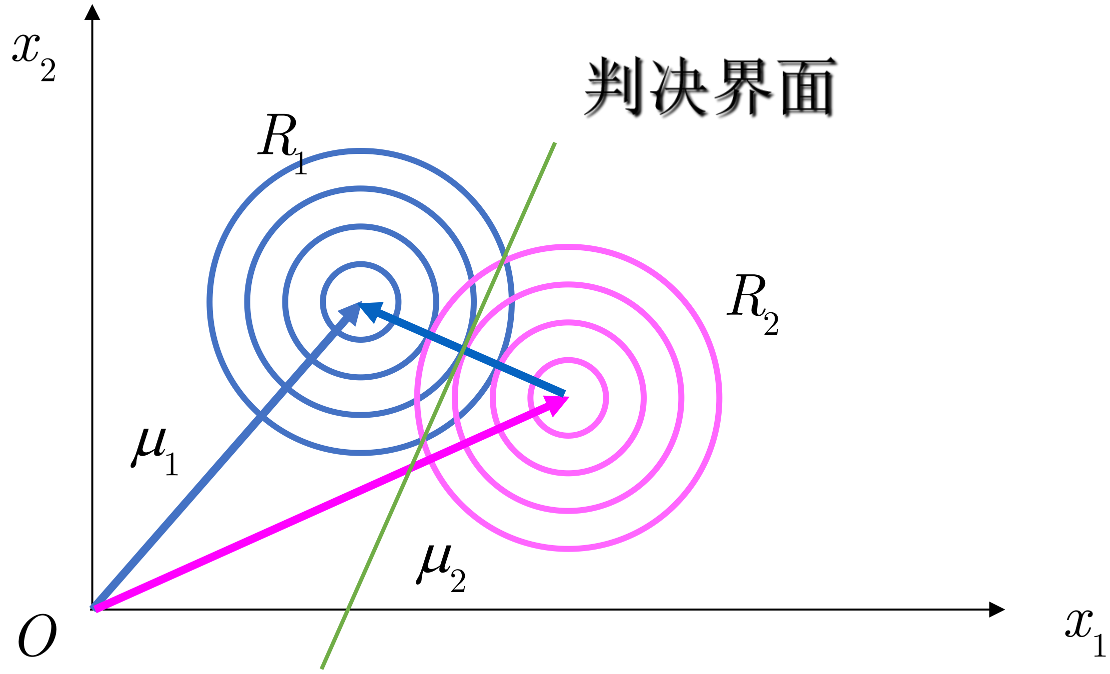
$$W = (\mu_i - \mu_j)$$

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。

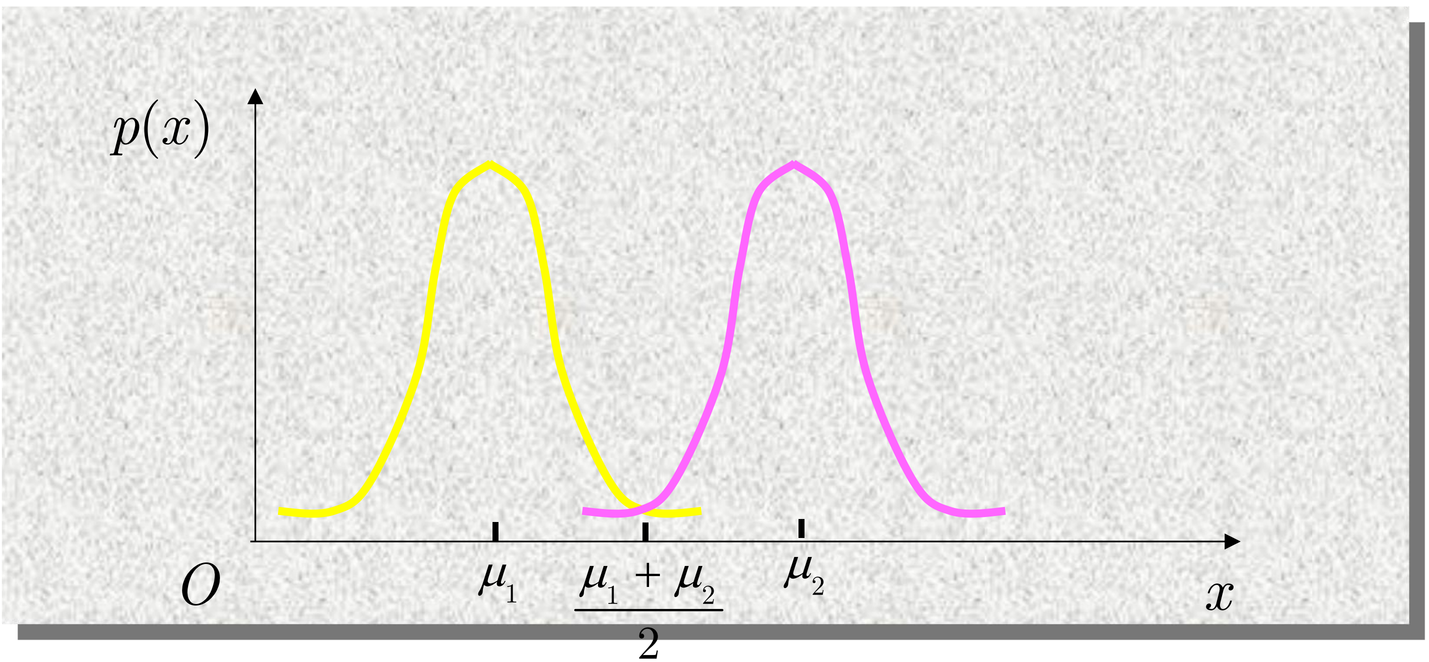


贝叶斯统计决策

■ 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。

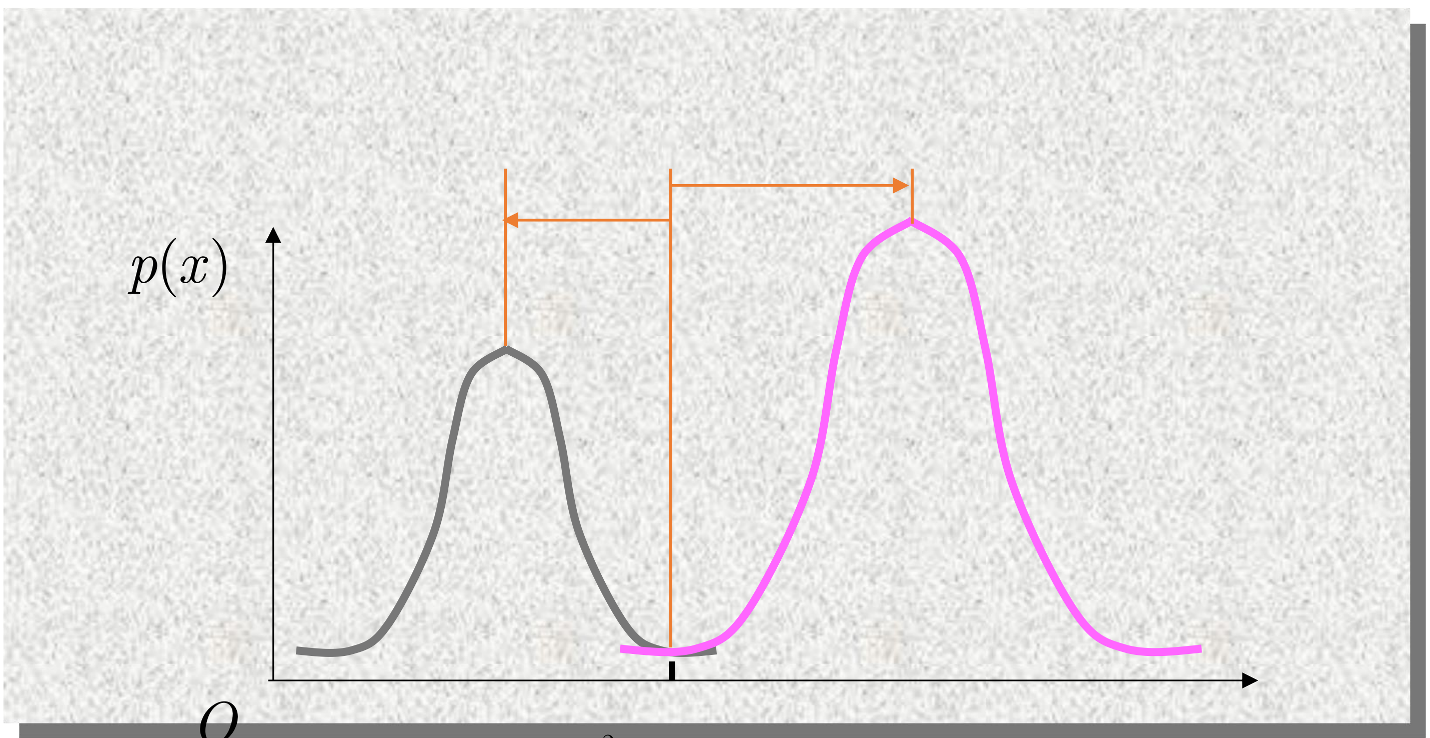


贝叶斯统计决策

■ 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。

贝叶斯统计决策

■ 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

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

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

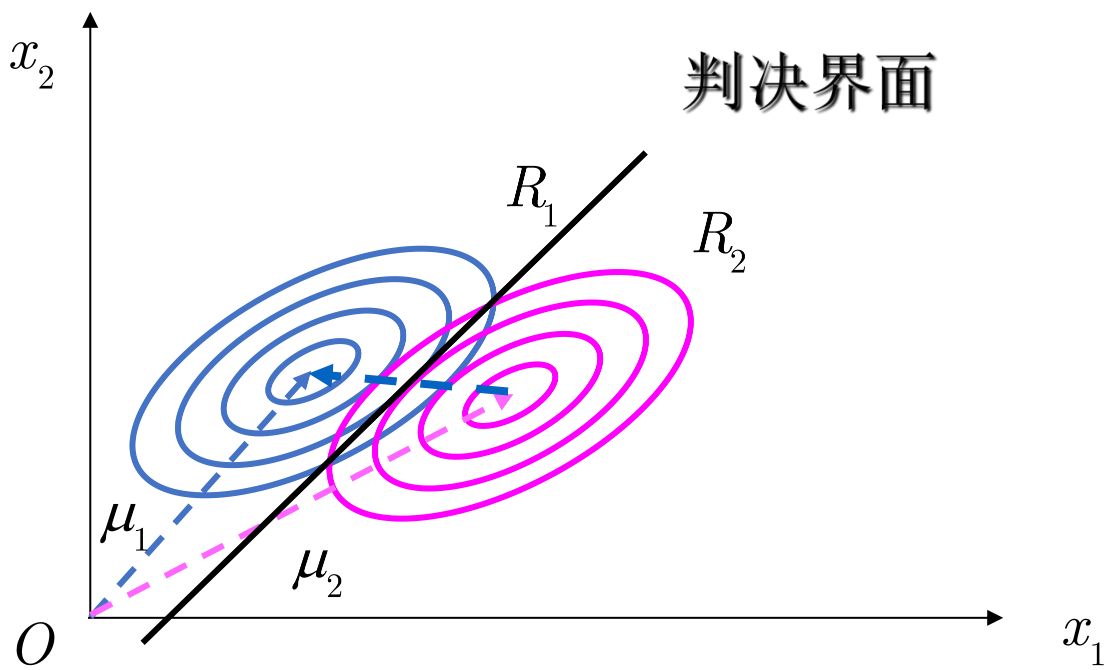
$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{1}{(\mu_i - \mu_j)^T \Sigma^{-1} (\mu_i - \mu_j)}(\mu_i - \mu_j)\log\frac{p(\omega_i)}{p(\omega_j)}$$

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

$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j)$$

(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(ρ=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

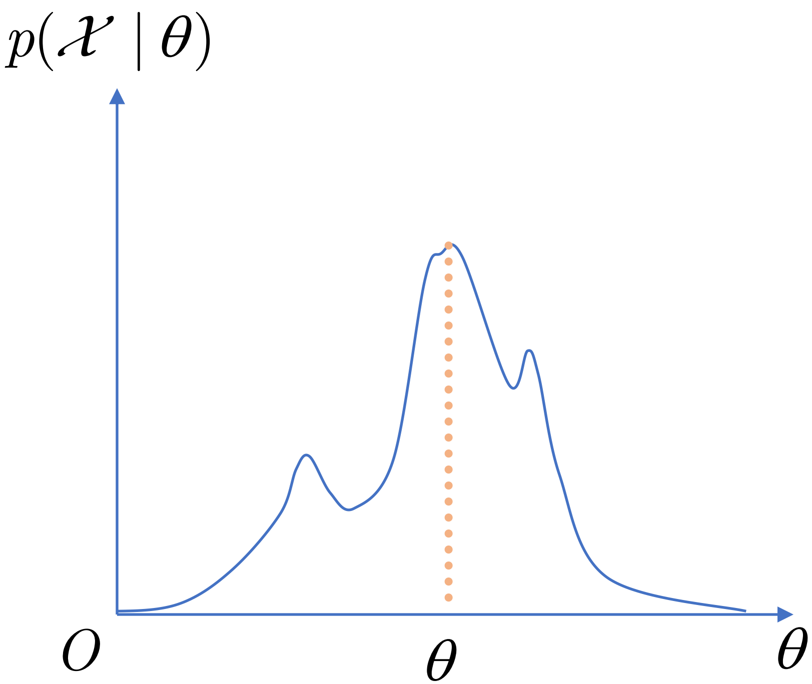
- 最大似然估计
  - 似然函数：$p(\mathcal{X} \mid \theta)$

    同一类的样本子集 $\mathcal{X} = \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$，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

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

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

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



参数学习

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

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

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

计算:

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

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

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

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

∇_{θ_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

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

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

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

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

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

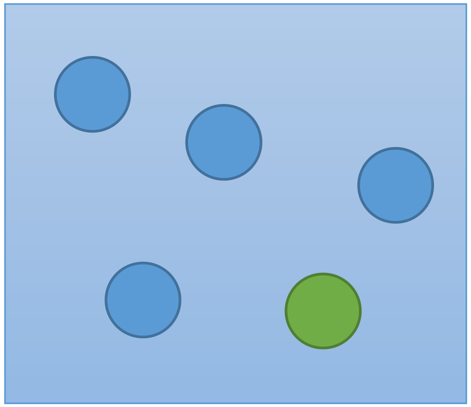
$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

概率分类

- Prior

Class 1

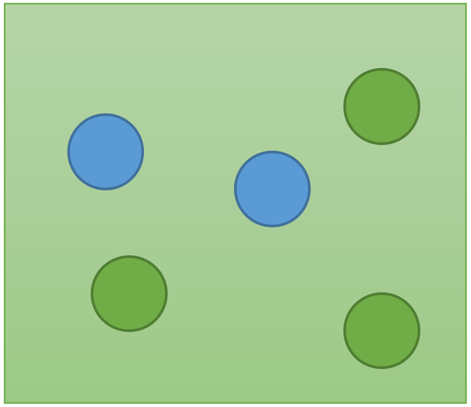
$\boxed{P(C_1)}$



Water

Class 2

$\boxed{P(C_2)}$



Normal

Water and Normal type with ID < 400 for training, rest for testing

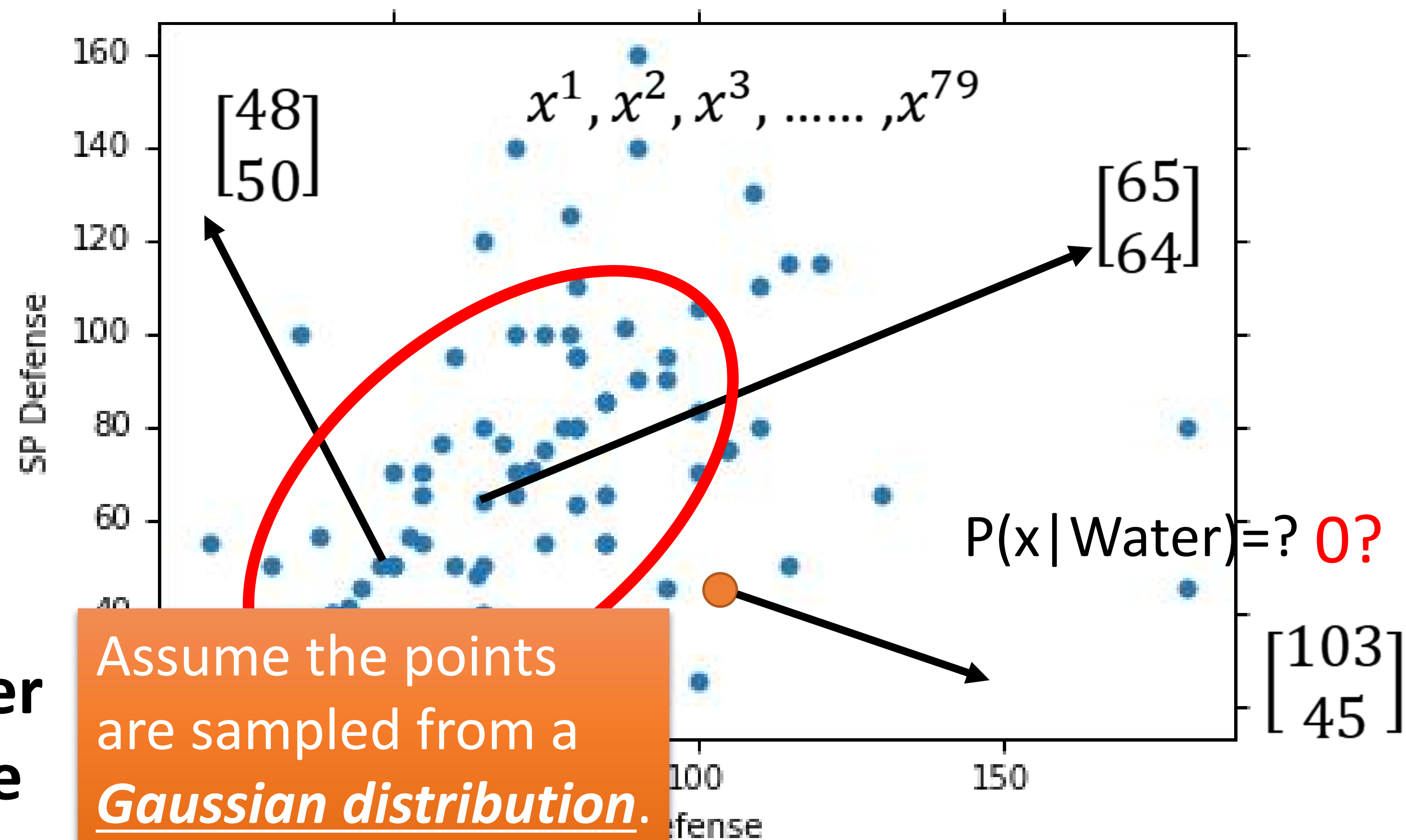
Training: 79 Water, 61 Normal

$$P(C_1) = 79 / (79 + 61) = 0.56$$

$$P(C_2) = 61 / (79 + 61) = 0.44$$

概率分类

- Probability from Class - Feature
  - Considering **Defense** and **SP Defense**



Water Type

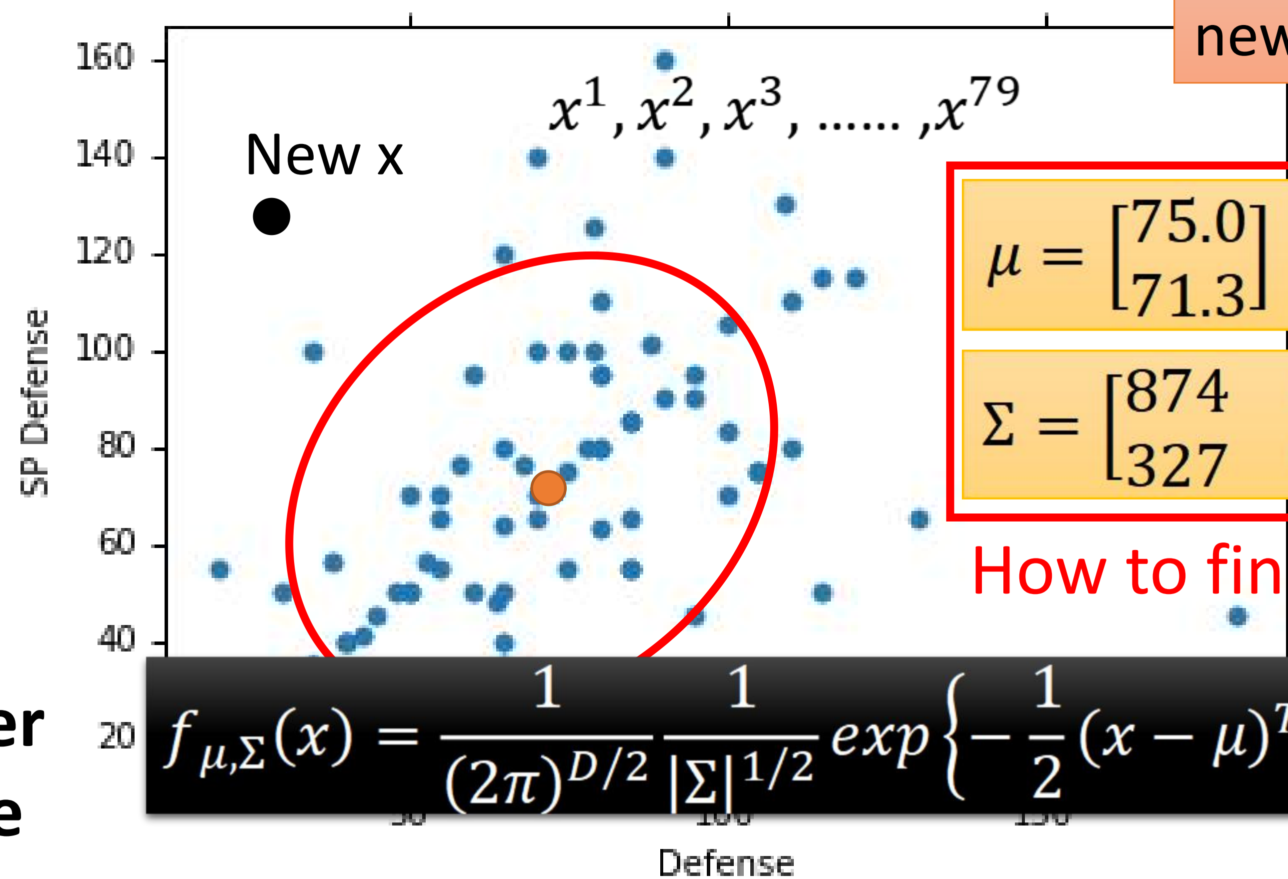
Assume the points are sampled from a ***Gaussian distribution***.

概率分类

- Probability from Class

Assume the points are sampled from a Gaussian distribution

Find the Gaussian distribution behind them → Probability for new points



$$\mu = \begin{bmatrix} 75.0 \\ 71.3 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 874 & 327 \\ 327 & 929 \end{bmatrix}$$

How to find them?

Water Type

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right\}$$

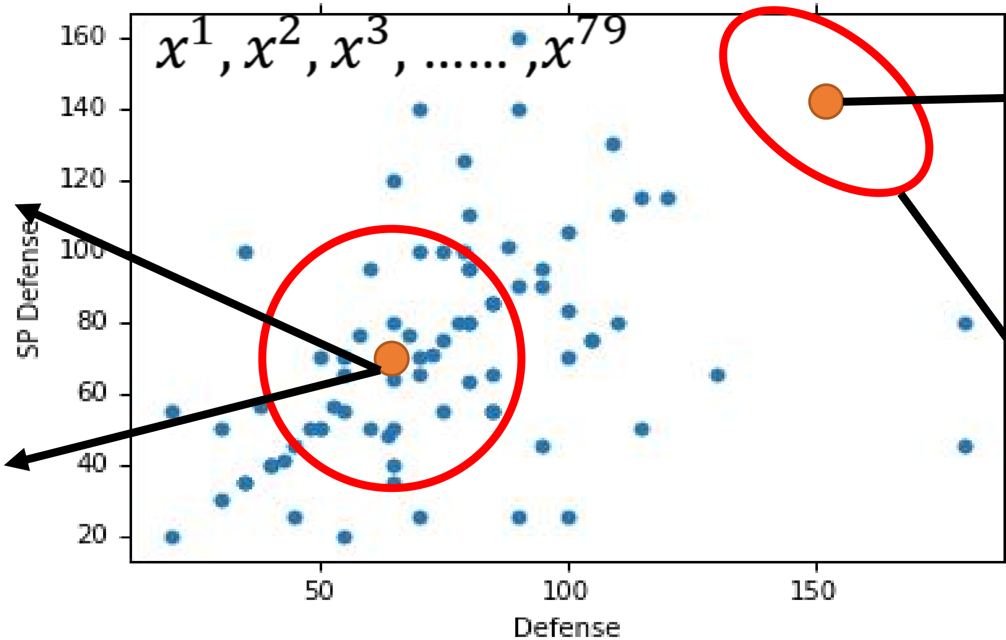
参考自李宏毅老师课件

概率分类

Maximum Likelihood

$$\Sigma = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 75 \\ 75 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 150 \\ 140 \end{bmatrix}$$

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

The Gaussian with any mean μ and covariance matrix Σ can generate these points. → Different Likelihood

Likelihood of a Gaussian with mean μ and covariance matrix Σ

= the probability of the Gaussian samples $x^1, x^2, x^3, \ldots\ldots, x^{79}$

$$L(\mu, \Sigma) = f_{\mu,\Sigma}(x^1) f_{\mu,\Sigma}(x^2) f_{\mu,\Sigma}(x^3) \ldots\ldots f_{\mu,\Sigma}(x^{79})$$

参考自李宏毅老师课件

概率分类

Maximum Likelihood

We have the “Water” type : $x^1, x^2, x^3, \ldots\ldots, x^{79}$

We assume $x^1, x^2, x^3, \ldots\ldots, x^{79}$ generate from the Gaussian (μ^*, Σ^*) with the ***maximum likelihood***

$$L(\mu, \Sigma) = f_{\mu,\Sigma}(x^1) f_{\mu,\Sigma}(x^2) f_{\mu,\Sigma}(x^3) \ldots\ldots f_{\mu,\Sigma}(x^{79})$$

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right\}$$

$$\mu^*, \Sigma^* = arg\max_{\mu,\Sigma} L(\mu, \Sigma)$$

$$\mu^* = \frac{1}{79} \sum_{n=1}^{79} x^n \qquad \Sigma^* = \frac{1}{79} \sum_{n=1}^{79} (x^n - \mu^*)(x^n - \mu^*)^T$$

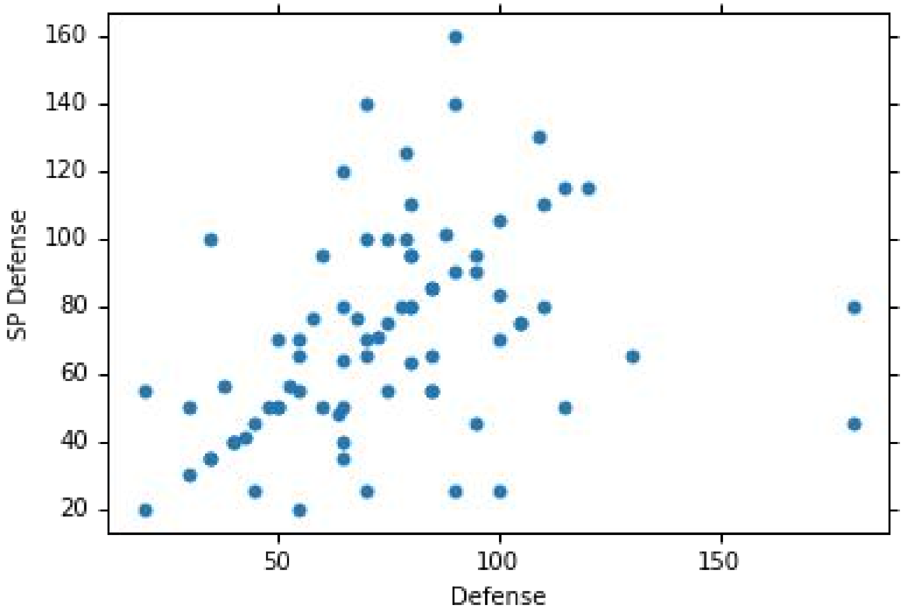
average

参考自李宏毅老师课件

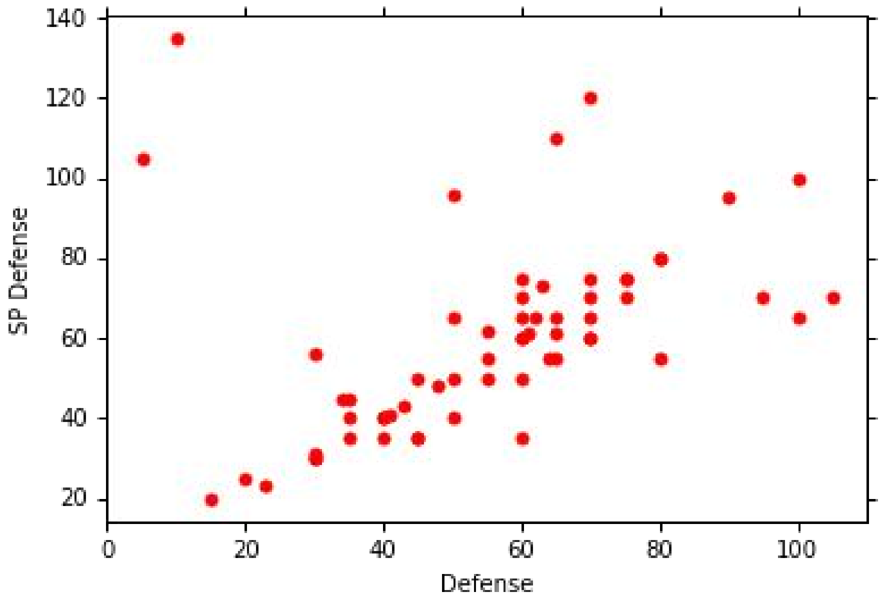
概率分类

Maximum Likelihood

Class 1: Water



Class 2: Normal



$$\mu^1 = \begin{bmatrix} 75.0 \\ 71.3 \end{bmatrix} \quad \Sigma^1 = \begin{bmatrix} 874 & 327 \\ 327 & 929 \end{bmatrix}$$

$$\mu^2 = \begin{bmatrix} 55.6 \\ 59.8 \end{bmatrix} \quad \Sigma^2 = \begin{bmatrix} 847 & 422 \\ 422 & 685 \end{bmatrix}$$

参考自李宏毅老师课件

概率分类

Now we can do classification

$$f_{\mu^1,\Sigma^1}(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma^1|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1)\right\}$$

$$\mu^1 = \begin{bmatrix} 75.0 \\ 71.3 \end{bmatrix} \quad \Sigma^1 = \begin{bmatrix} 874 & 327 \\ 327 & 929 \end{bmatrix}$$

P(C1)
= 79 / (79 + 61) =0.56

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

$$f_{\mu^2,\Sigma^2}(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma^2|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)\right\}$$

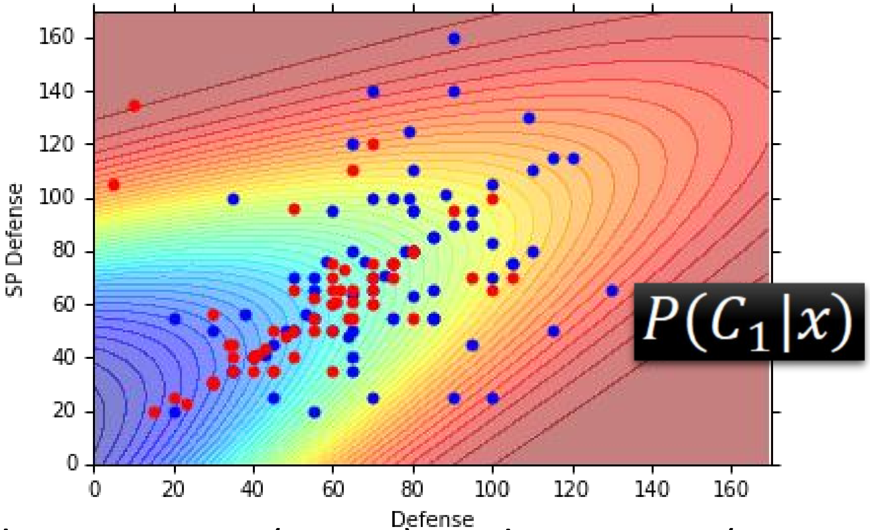
$$\mu^2 = \begin{bmatrix} 55.6 \\ 59.8 \end{bmatrix} \quad \Sigma^2 = \begin{bmatrix} 847 & 422 \\ 422 & 685 \end{bmatrix}$$

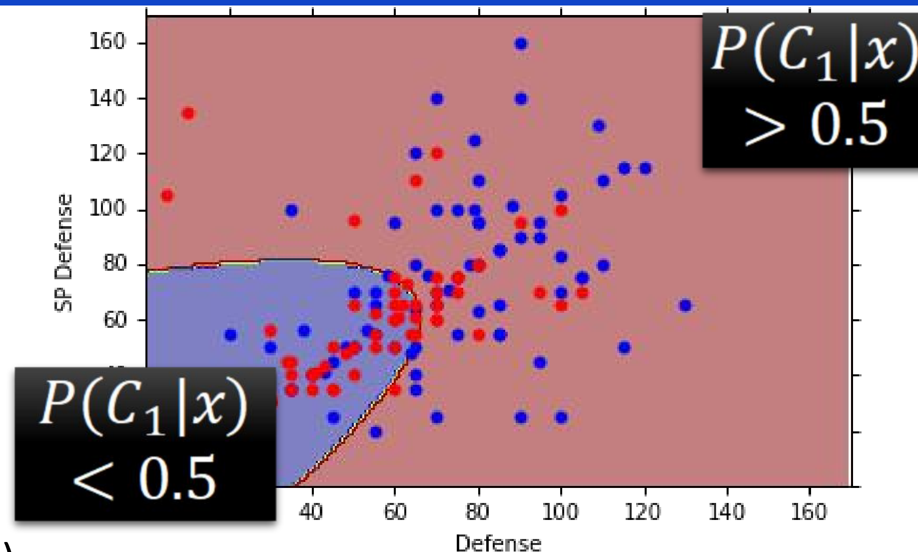
P(C2)
= 61 / (79 + 61)
=0.44

If $P(C_1|x) > 0.5$ ➡ x belongs to class 1 (Water)

参考自李宏毅老师课件

概率分类





Blue points: C_1 (Water), Red points: C_2 (Normal)

How's the results?

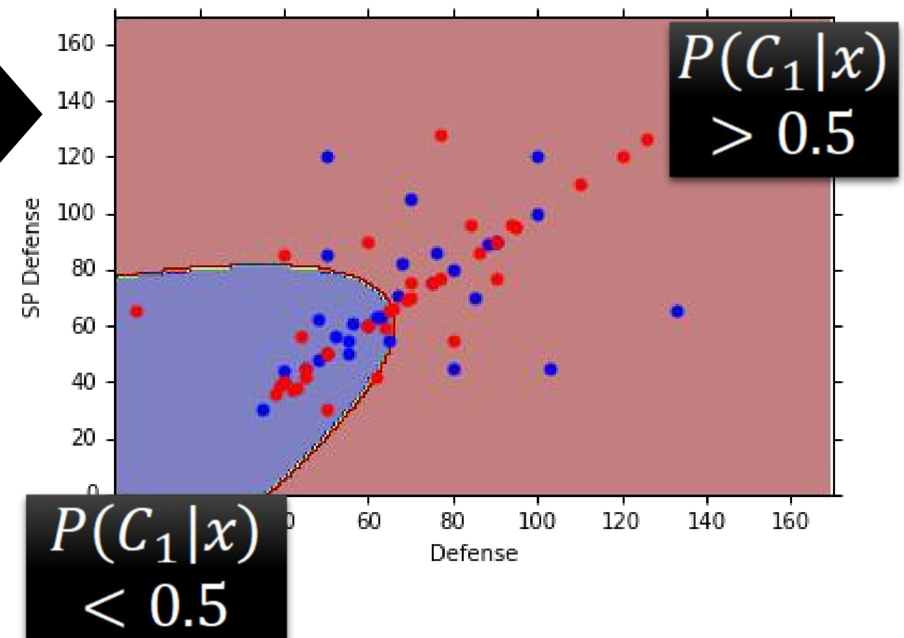
Testing data: 47% accuracy ☹

All: hp, att, sp att,

de, sp de, speed (6 features)

μ^1, μ^2: 6-dim vector

Σ^1, Σ^2: 6 x 6 matrices

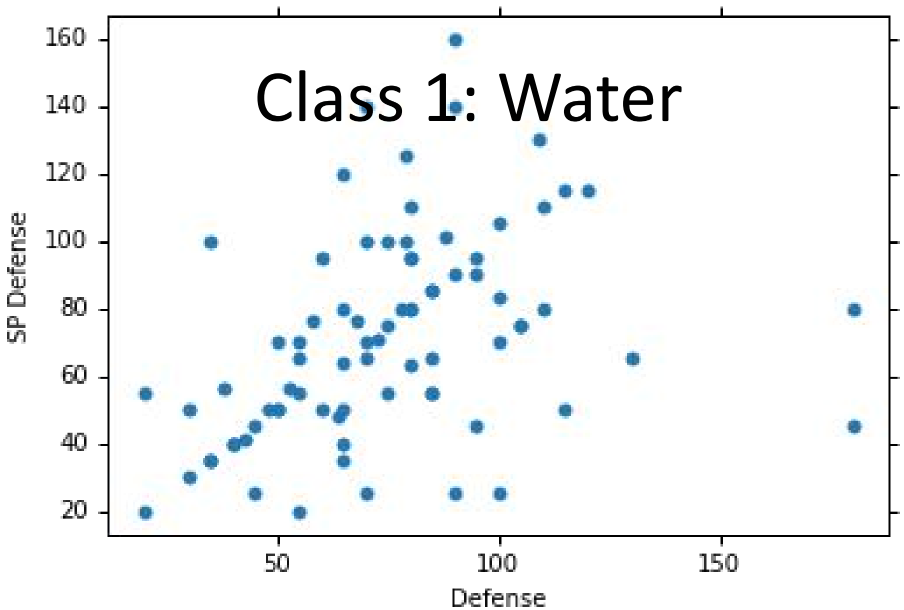


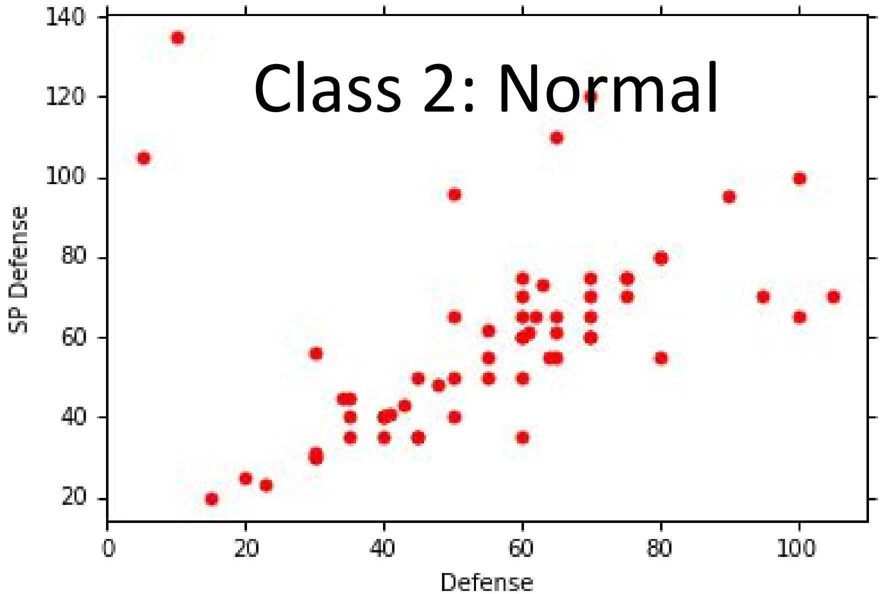
54% accuracy …

参考自李宏毅老师课件

概率分类

Modifying Model





$$\mu^1 = \begin{bmatrix} 75.0 \\ 71.3 \end{bmatrix} \quad \Sigma^1 = \begin{bmatrix} 874 & 327 \\ 327 & 929 \end{bmatrix} \qquad \mu^2 = \begin{bmatrix} 55.6 \\ 59.8 \end{bmatrix} \quad \Sigma^2 = \begin{bmatrix} 847 & 422 \\ 422 & 685 \end{bmatrix}$$

The same Σ

Less parameters

参考自李宏毅老师课件

概率分类

Modifying Model

- Maximum likelihood

"Water" type Pokémons: $x^1, x^2, x^3, \ldots\ldots, x^{79}$

"Normal" type Pokémons: $x^{80}, x^{81}, x^{82}, \ldots\ldots, x^{140}$

μ^1 $\quad\Sigma\quad$ μ^2

Find μ^1, μ^2, Σ maximizing the likelihood $L(\mu^1,\mu^2,\Sigma)$

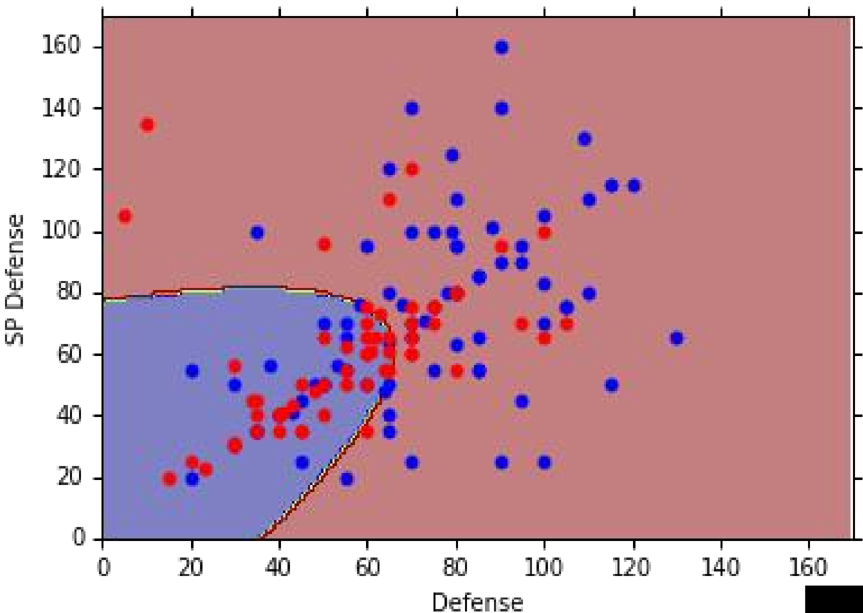
$$L(\mu^1,\mu^2,\Sigma) = f_{\mu^1,\Sigma}(x^1) f_{\mu^1,\Sigma}(x^2) \cdots f_{\mu^1,\Sigma}(x^{79}) \times f_{\mu^2,\Sigma}(x^{80}) f_{\mu^2,\Sigma}(x^{81}) \cdots f_{\mu^2,\Sigma}(x^{140})$$

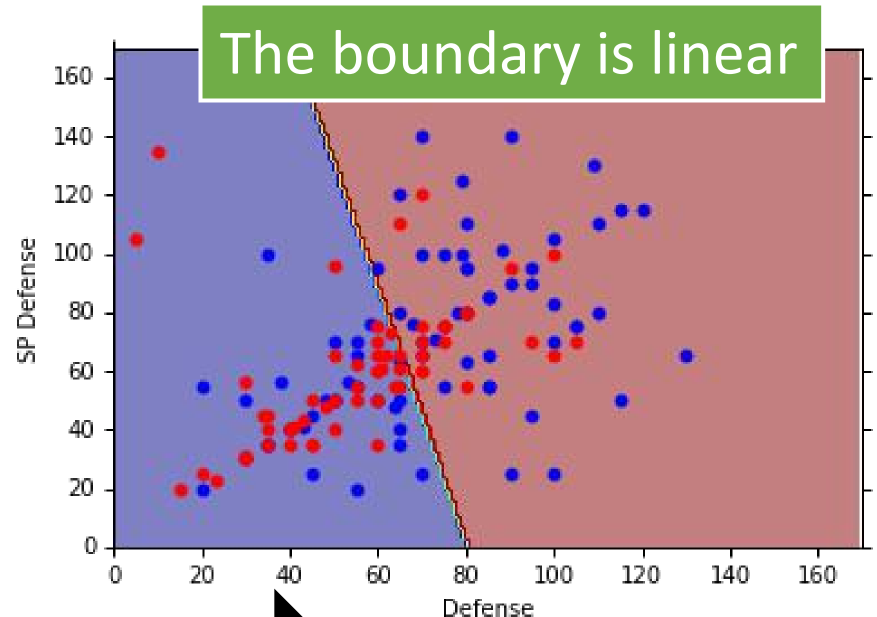
μ^1 and μ^2 is the same

$$\Sigma = \frac{79}{140}\Sigma^1 + \frac{61}{140}\Sigma^2$$

概率分类

Modifying Model





The same covariance matrix

All: hp, att, sp att, de, sp de, speed

54% accuracy → 73% accuracy

概率分类

Three Step

- Function Set (Model):

$x \rightarrow$

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

If $P(C_1|x) > 0.5$, output: class 1

Otherwise, output: class 2

- Goodness of a function:
  - The mean μ and covariance Σ that maximizing the likelihood (the probability of generating data)
- Find the best function: easy

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

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

等价于

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

最小错误率判决规则

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

贝叶斯统计决策

- **最小风险**
  - 目标：判定过程中，考虑不同判决时产生代价不同
  - 风险：对于某一样本$\mathbf{x} \in \omega_j$，若采取判决α_i，则招致损失 $L(\alpha_i \mid \omega_j)$，简记为$L_{ji}$，称 $R(\alpha_i \mid \mathbf{x})$ 为取行动α_i时的条件风险

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

  - 总风险 $R = \int R(\alpha(\mathbf{x}) \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$

  $\alpha(\mathbf{x})$是对每一个$\mathbf{x}$所可能取得行动 α_1 α_2 α_R中的一个

贝叶斯统计决策

■ 最小风险

➢两类问题

决策1： α_1 ----〉决策为 ω_1类

决策2： α_2 ----〉决策为ω_2类

L_{ij}表示真类别为 ω_i，判决为 ω_j所招致的损失

则

$$L_{ij} = L(\alpha_j \mid \omega_i)$$

$$R(\alpha_1 \mid \mathbf{x}) = L_{11} p(\omega_1 \mid \mathbf{x}) + L_{21} p(\omega_2 \mid \mathbf{x})$$

$$R(\alpha_2 \mid \mathbf{x}) = L_{12} p(\omega_1 \mid \mathbf{x}) + L_{22} p(\omega_2 \mid \mathbf{x})$$

贝叶斯统计决策

■ 最小风险

➢判决准则

若 $R(\alpha_1 \mid \mathbf{x}) < R(\alpha_2 \mid \mathbf{x})$

则采取行动 α_1 的损失较小

（1）在给定样本x条件下，计算各类后验概率 $p(w_j \mid \mathbf{x})$
$j = 1, 2, \cdots\cdots, c$ 。

（2）求各种判决的条件平均风险 $R(\alpha_i \mid \mathbf{x})$ ， $j = 1, 2, \cdots\cdots, a$
此时需要知道风险矩阵。

（3）比较各种判决的条件平均风险，把样本x归属于条件平均风险最小的那一种判决。

贝叶斯统计决策

■ 最小风险

- 在例1的基础上，考虑决策风险

损失 状态 / 决策	ω_1	ω_2
α_1	**0.5**	**6**
α_2	**2**	**0.5**

试对该病人x进行分类。

贝叶斯统计决策

- **最小风险**

解：已知条件

$P(\omega_1) = 0.995$ $\quad$ $P(\omega_2) = 0.005$

$\rho(\mathbf{x} \mid \omega_1) = 0.01$ $\quad$ $\rho(\mathbf{x} \mid \omega_2) = 0.95$

$L_{11} = 0.5$ $\quad$ $L_{21} = 6$

$L_{12} = 2$ $\quad$ $L_{22} = 0.5$

后验概率 $P\left(\omega_1 \mid \mathbf{x}\right) = 0.677$ $\quad$ $P\left(\omega_2 \mid \mathbf{x}\right) = 0.323$

条件风险

$$R(\alpha_1 \mid \mathbf{x}) = L_{11} P(\omega_1 \mid \mathbf{x}) + L_{21} P(\omega_2 \mid \mathbf{x}) = 2.2765$$

$$R(\alpha_2 \mid \mathbf{x}) = L_{12} P(\omega_1 \mid \mathbf{x}) + L_{22} P(\omega_2 \mid \mathbf{x}) = 1.5155$$

$$\boxed{R(\alpha_2 \mid \mathbf{x}) < R(\alpha_1 \mid \mathbf{x})}$$

结论：决策x为癌症病人，与例1 的结论相反

贝叶斯统计决策

■ 最小风险

➢结论：在0 - 1损失函数情况下，最小风险判决规则退化为最小错误率判决规则。也就是说，最小错误率判决规则是最小风险判决规则的一个特例。

➢推导:

现假设正确判决损失为0，错误判决损失为1，且判决数目与类型数目相等。即有0 - 1损失函数:

$$L(\alpha_i \mid w_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

贝叶斯统计决策

■ Neyman-Pearson判决规则

➢问题

- 通常情况下，无法确定损失 $L_{11}, L_{12}, L_{21}, L_{22}$。
- **先验概率未知**

➢基本思想

- 限定或者约束某一个错误概率，与此同时通过一些计算，使得另一个错误概率最小

➢实现

- 设计一个辅助函数

$$\gamma = \varepsilon_1 + \mu\varepsilon_2 \qquad \varepsilon_1 = \int_{R_2} p(\mathbf{x} \mid \omega_1) d\mathbf{x} \qquad \varepsilon_2 = \int_{R_1} p(\mathbf{x} \mid \omega_2) d\mathbf{x}$$

贝叶斯统计决策

■ Neyman-Pearson判决规则

➢由于

$$\int_{R_2} \rho(\mathbf{x} \mid \omega_1)d\mathbf{x} + \int_{R_1} \rho(\mathbf{x} \mid \omega_1)d\mathbf{x} = 1$$

所以

$$\begin{aligned}
\gamma &= \int_{R_2} \rho(\mathbf{x} \mid \omega_1)d\mathbf{x} + \mu\int_{R_1} \rho(\mathbf{x} \mid \omega_2)d\mathbf{x} \\
&= \int_{R_2} \rho(\mathbf{x} \mid \omega_1)d\mathbf{x} + \int_{R_1} \rho(\mathbf{x} \mid \omega_1)d\mathbf{x} + \mu\int_{R_1} \rho(\mathbf{x} \mid \omega_2)d\mathbf{x} - \int_{R_1} \rho(\mathbf{x} \mid \omega_1)d\mathbf{x} \\
&= 1 + \int_{R_1} \left[\mu\rho(\mathbf{x} \mid \omega_2) - \rho(\mathbf{x} \mid \omega_1)\right]d\mathbf{x}
\end{aligned}$$

$\min(\gamma) \Longrightarrow \mu\rho(\mathbf{x} \mid \omega_2) < \rho(\mathbf{x} \mid \omega_1)$ 在R_1中

结论 $$\frac{\rho(\mathbf{x} \mid \omega_1)}{\rho(\mathbf{x} \mid \omega_2)} > \mu \Rightarrow \mathbf{x} \in \omega_1$$

贝叶斯统计决策

■ Neyman-Pearson判决规则

- 同样地，在R_2中

$$\gamma = 1 + \int_{R_2} \left[\rho(\mathbf{x} \mid \omega_1) - \mu\rho(\mathbf{x} \mid \omega_2) \right] d\mathbf{x}$$

$$\min(\gamma) \Rightarrow \mu\rho(\mathbf{x} \mid \omega_2) > \rho(\mathbf{x} \mid \omega_1)$$

$$\frac{\rho(\mathbf{x} \mid \omega_1)}{\rho(\mathbf{x} \mid \omega_2)} < \mu \Rightarrow \mathbf{x} \in \omega_2$$

贝叶斯统计决策

■ Neyman-Pearson判决规则

- 结论

Neyman-Pearson决策的任务就是寻找阈值常数 μ，当

$$\frac{\rho(\mathbf{x} \mid \omega_1)}{\rho(\mathbf{x} \mid \omega_2)} = \mu$$ 时

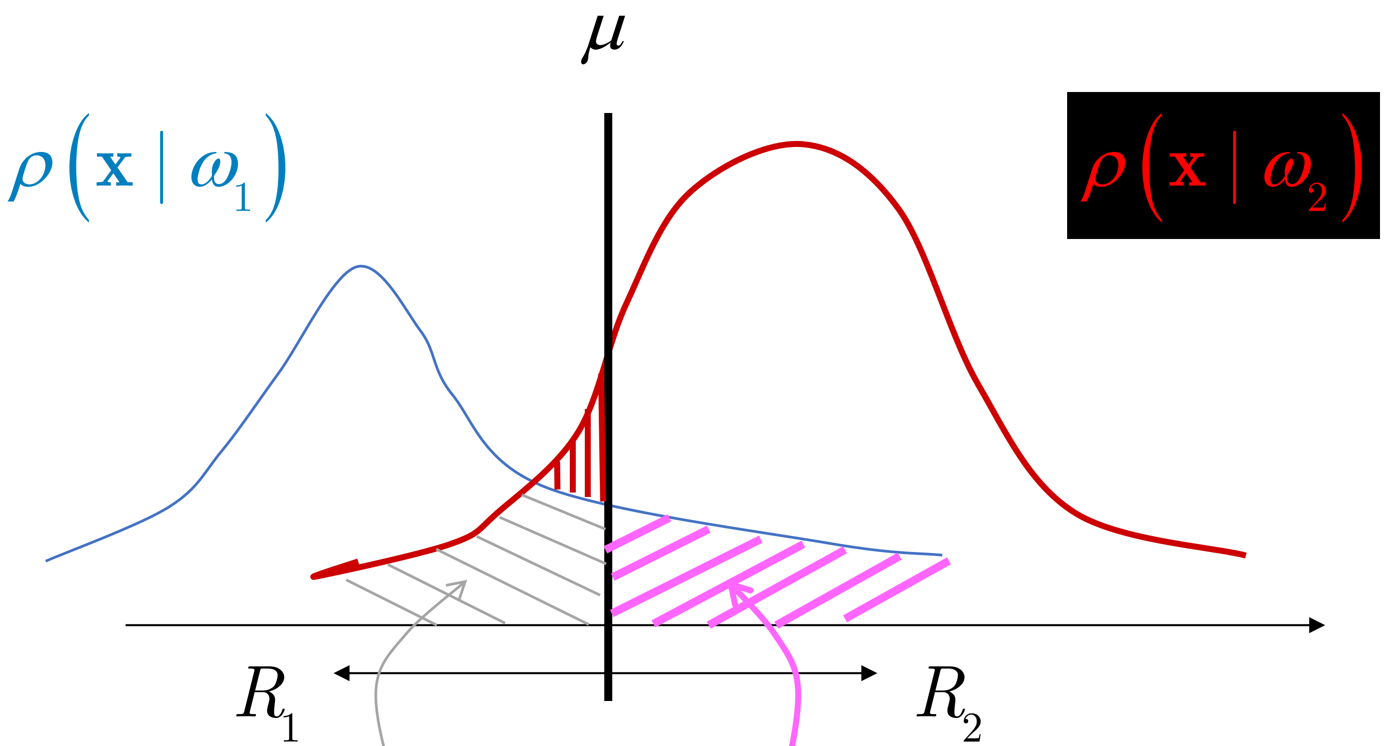
μ 就可以划分子空间R_1和R_2。而

$$\varepsilon_2 = \int_{-\infty}^{\mu} \rho(\mathbf{x} \mid \omega_2) d\mathbf{x}$$

ε_2 已知 ⇨ μ ⇨ 子空间R_1和R_2

贝叶斯统计决策

■ **Neyman-Pearson判决规则**



$$\varepsilon_2 = \int_{R_1} \rho\left(\mathbf{x} \mid \omega_2\right) dx \qquad \varepsilon_1 = \int_{R_2} \rho\left(\mathbf{x} \mid \omega_1\right) dx$$

贝叶斯统计决策

■ Neyman-Pearson判决规则

- 两类的概率密度函数是正态的，两类的均值向量分别为 $\mu_1=(-1,0)^T$ 和 $\mu_2=(0,1)^T$，协方差矩阵相等且为单位矩阵。给定 $\varepsilon_2=0.046$，试确定N-P判决门限t。

贝叶斯统计决策

Neyman-Pearson判决规则

- 根据给定的条件，写出两类的类概率密度函数，即：

和 $\rho(\mathbf{x} \mid w_1) \sim N(\mu_1, \Sigma)$ $\rho(\mathbf{x} \mid w_2) \sim N(\mu_2, \Sigma)$

$$\Sigma = E[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T]$$

而 $\mu_1 = (-1,0)^T$， $\mu_2 = (1,0)^T$， $\Sigma = I$， 得到

$$\rho(\mathbf{x} \mid w_1) = \frac{1}{(2\pi)^{d/2} \mid \Sigma \mid^{1/2}} \exp[-\frac{1}{2}(\mathbf{x} - \mu_1)^T \cdot \Sigma^{-1}(\mathbf{x} - \mu_1)]$$

$$= \frac{1}{2\pi} \exp(-\frac{1}{2}[(x_1 + 1)^2 + x_2^2]) = \frac{1}{2\pi} \exp(-\frac{1}{2}(x_1^2 + 2x_1 + 1 + x_2^2))$$

贝叶斯统计决策

■ Neyman-Pearson判决规则

$$\rho(\mathbf{x}\mid w2)=\frac{1}{2\pi}\exp(-\frac{1}{2}[(x_1-1)^2+x_2{}^2])=\frac{1}{2\pi}\exp(-\frac{1}{2}[(x_1{}^2-2x_1+x_2{}^2))$$

$$\frac{\rho(\mathbf{x}\mid w_1)}{\rho(\mathbf{x}\mid w_2)}=\exp(-2x_1)$$

- 故：$\mu(\mathbf{x})=\exp(-2x_1)$，$\mu$只是$x_1$的函数，与$x_2$无关。有 $x_1=-\frac{1}{2}\ln\mu$。又因为$\rho(\mathbf{x}\mid w_2)$ 的边缘密度为 $\rho(x_1\mid w_2)$

■ Neyman-Pearson判决规则

$$\rho(x_1 \mid w_2) = \int_{-\infty}^{\infty} \rho(\mathbf{x} \mid w_2) dx_2$$

$$= \int \frac{1}{2\pi} \exp[-\frac{1}{2}(x_1^2 - 2x_1 + 1 + x_2^2)] dx_2$$

$$= \frac{1}{2\pi} \exp[-\frac{1}{2}(x_1^2 - 2x_1 + 1)] \cdot \int_{-\infty}^{\infty} \exp(-\frac{1}{2} x_2^2) dx_2$$

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

对于给定的正数 ε_2，可由下式计算：

$$\varepsilon_2 = \int_{-\infty}^{-\frac{1}{2}\ln\mu} \frac{1}{\sqrt{2\pi}} \exp[-\frac{(x_1-1)^2}{2}] dx_1 = \int_{-\infty}^{-\frac{1}{2}\ln\mu - 1} \frac{1}{\sqrt{2\pi}} \exp(-\frac{y^2}{2}) dy$$

贝叶斯统计决策

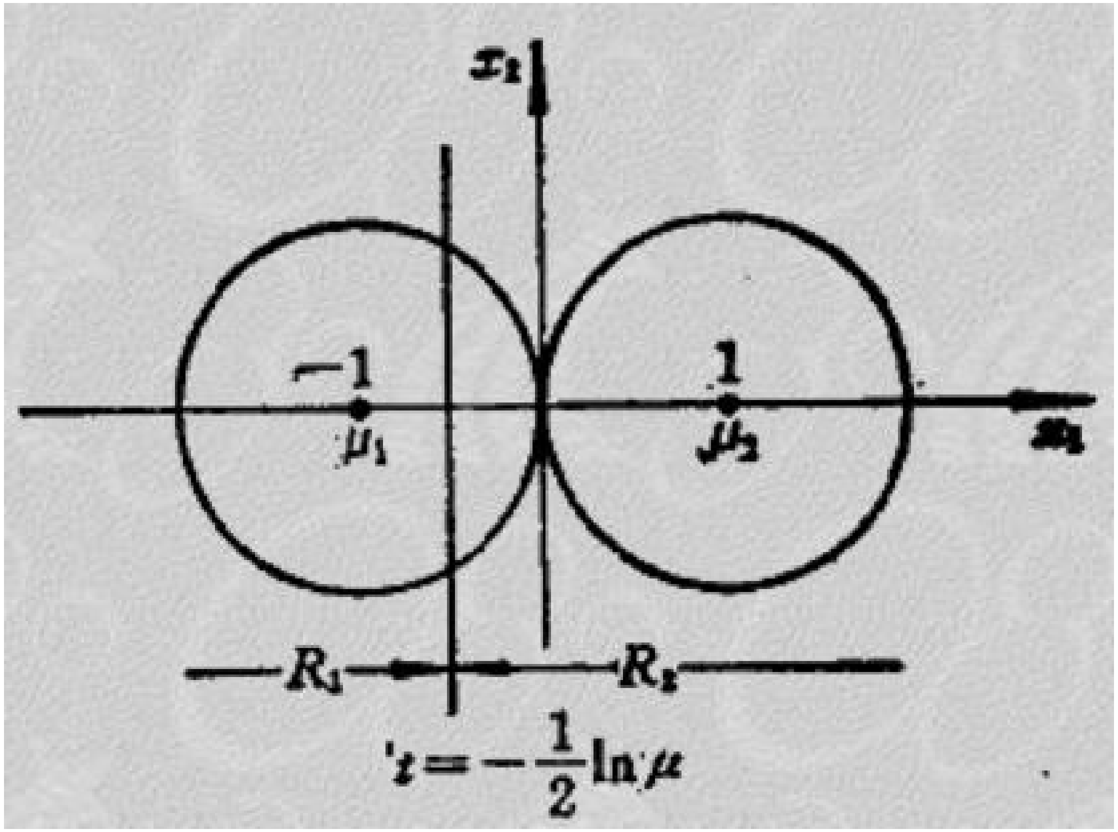
- **Neyman-Pearson判决规则**
  - y是服从标准正态分布的随机变量，令 $y_1 = -\frac{1}{2}\ln\mu - 1$，则 $\varepsilon_2 = \Phi(y_1)$ ：
  - ε_2 与y_1具有一一对应的关系，有表可查。当

$\varepsilon_2 = 0.046$ 时， $y_1 = -1.693$， $\mu = 4$ ，$x_1 = -0.693$， 因此判决门限 $t = x_1 = -0.693$ 。分区界线是 $x_1 = -0.693$ 的一条直线，对于样本 $\mathbf{x} = (x_1, x_2)^T$ 的分类判决，只需考察特征 x_1 ，

判决规则为，若：$x_1 < -0.693$，则 $\mathbf{x} \in w_1$ 。否则，x属于w_2。

贝叶斯统计决策

■ **Neyman-Pearson判决规则**



选择门限图

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
  - 似然函数： $p(\mathcal{X} \mid \theta)$

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

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

计算:

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

$$\nabla_\theta = \left\{ \begin{array}{c} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{array} \right\}$$

参数学习

- **贝叶斯估计**
  - Bayes估计与最大似然估计的区别

　　最大似然估计是把待估计的参数当作未知但固定的量；而贝叶斯估计则把待估计的参数本身看作是随机变量，要做的是根据观测数据对参数的分布进行估计，除了测量数据外，还可以考虑参数的先验分布

  – 目的：把待估参数 θ 看成具有先验分布密度$p(\theta)$的随机变量，其取值与样本集 $\mathcal{X} = \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$有关，我们要做的是根据 $\mathcal{X}$ 估计最优的 θ^* 。

参数学习

■ 贝叶斯估计

- 贝叶斯估计的**步骤**是：

1. 确定 θ 的先验分布 $p(\theta)$
2. 求出样本集的联合分布为 $p(\mathcal{X}\mid\theta)=\prod_{i=1}^{N}p(\mathbf{x}_i\mid\theta)$
3. 利用贝叶斯公式，求 θ 的后验概率分布：

$$p(\theta\mid\mathcal{X})=\frac{p(\mathcal{X}\mid\theta)p(\theta)}{\int_{\Theta}p(\mathcal{X}\mid\theta)p(\theta)d\theta}$$

4. θ 的贝叶斯估计量是：$\theta^{*}=\int_{\Theta}\theta p(\theta\mid\mathcal{X})d\theta$

参数学习

■ 贝叶斯估计

- 贝叶斯估计的**步骤**是：

1. 确定 θ 的先验分布 $p(\theta)$
2. 求出样本集的联合分布为 $p(\mathcal{X}\mid\theta)=\prod_{i=1}^{N}p(\mathbf{x}_i\mid\theta)$
3. 利用贝叶斯公式，求 θ 的后验概率分布：

$$p(\theta\mid\mathcal{X})=\frac{p(\mathcal{X}\mid\theta)p(\theta)}{\int_{\Theta}p(\mathcal{X}\mid\theta)p(\theta)d\theta}$$

4. θ 的贝叶斯估计量是： $\theta^{*}=\int_{\Theta}\theta p(\theta\mid\mathcal{X})d\theta$

参数学习

■ 正态分布下的贝叶斯估计

设 ω_j 类：$p(\mathbf{x} \mid \mu) \sim N(\mu, \sigma^2)$， μ 为未知随机参数

条件1： $\mathscr{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ 为已知类别为 ω_j的n个同类样本，并且是独立抽取的。

条件2：考虑 $\mathbf{x}$ 是一维的情况。

条件3：把 μ 看作是随机变量，遵循如下分布

$$p(\mathbf{x}_k \mid \mu) \sim N(\mu, \sigma^2)$$

$$p(\mu) \sim N(\mu_0, \sigma_0^2)$$

参数学习

- 正态分布下的贝叶斯估计

条件1

$$p(\mu \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \mu)p(\mu)}{\int p(\mathcal{X} \mid \mu)p(\mu)d\mu} = \alpha p(\mathcal{X} \mid \mu)p(\mu) = \alpha\left[\prod_{k=1}^{n} p(\mathbf{x}_k \mid \mu)\right]p(\mu)$$

条件3

$$p(\mu \mid \mathcal{X}) = \alpha\left\{\prod_{k=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(\mathbf{x}_k-\mu)^2}{2\sigma^2}\right]\right\}\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right]$$

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

$$= \alpha'\exp\left\{-\frac{1}{2}\left[\sum_{k=1}^{n}\left(\frac{\mathbf{x}_k^{\,2}}{\sigma^2}-\frac{2\mathbf{x}_k\mu}{\sigma^2}+\frac{\mu^2}{\sigma^2}\right)+\frac{\mu^2}{\sigma_0^{\,2}}-\frac{2\mu\mu_0}{\sigma_0^{\,2}}+\frac{\mu_0^{\,2}}{\sigma_0^{\,2}}\right]\right\}$$

$$= \alpha''\exp\left\{-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^{\,2}}\right)\mu^2-2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n}\mathbf{x}_k+\frac{\mu_0}{\sigma_0^{\,2}}\right)\mu\right]\right\}$$

参数学习

■ 正态分布下的贝叶斯估计

$p(\mu \mid \mathcal{X})$ 仍是一个正态函数，称为再生密度。

$$p(\mu \mid \mathcal{X}) = \alpha'' \exp\left\{-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^{\ 2}}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n}\mathbf{x}_k + \frac{\mu_0}{\sigma_0^{\ 2}}\right)\mu\right]\right\}$$

假设 $p(\mu \mid \mathcal{X}) \sim N(\mu_n, \sigma_n^{\ 2})$，即

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

$$= \frac{1}{\sqrt{2\pi}\sigma_n}\exp\left[-\frac{1}{2}\left(\frac{1}{\sigma_n^{\ 2}}\mu^2 - \frac{2\mu_n}{\sigma_n^{\ 2}}\mu + \frac{\mu_n^{\ 2}}{\sigma_n^{\ 2}}\right)\right]$$

比较

参数学习

■ 正态分布下的贝叶斯估计

- μ_n, σ_n 的求解

$$\begin{cases} \dfrac{1}{\sigma_n^2} = \dfrac{n}{\sigma^2} + \dfrac{1}{\sigma_0^2} \\ \dfrac{\mu_n}{\sigma_n^2} = \dfrac{1}{\sigma^2}\sum_{k=1}^{n}\mathbf{x}_k + \dfrac{\mu_0}{\sigma_0^2} = \dfrac{n}{\sigma^2} m_n + \dfrac{\mu_0}{\sigma_0^2} \end{cases}$$

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

$$\Downarrow$$

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} m_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

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

参数学习

■ 正态分布下的贝叶斯估计

- 根据 $\theta^* = \int_\Theta \theta p(\theta \mid \mathcal{X}) d\theta$

  计算 μ 的贝叶斯估计

$$\mu^* = \int \mu p(\mu \mid \mathcal{X}) d\mu = \int \mu \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\frac{(\mu-\mu_n)^2}{\sigma_n^2}\right] d\mu = \mu_n$$

$$= \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2} m_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0$$

参数学习

- 正态分布下的贝叶斯估计

• 分析： $\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2} m_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0 \quad \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}$

1. 再生密度的均值是样本均值和先验均值的线性组合。

2. 一般情况下 $\sigma_0 \neq 0$，则当 $n\to\infty, \mu_n \to m_n$。

极端情况1：$\sigma_0 = 0 \Rightarrow \mu_n = \mu_0, \forall n$，说明先验值 μ_0 十分可靠。

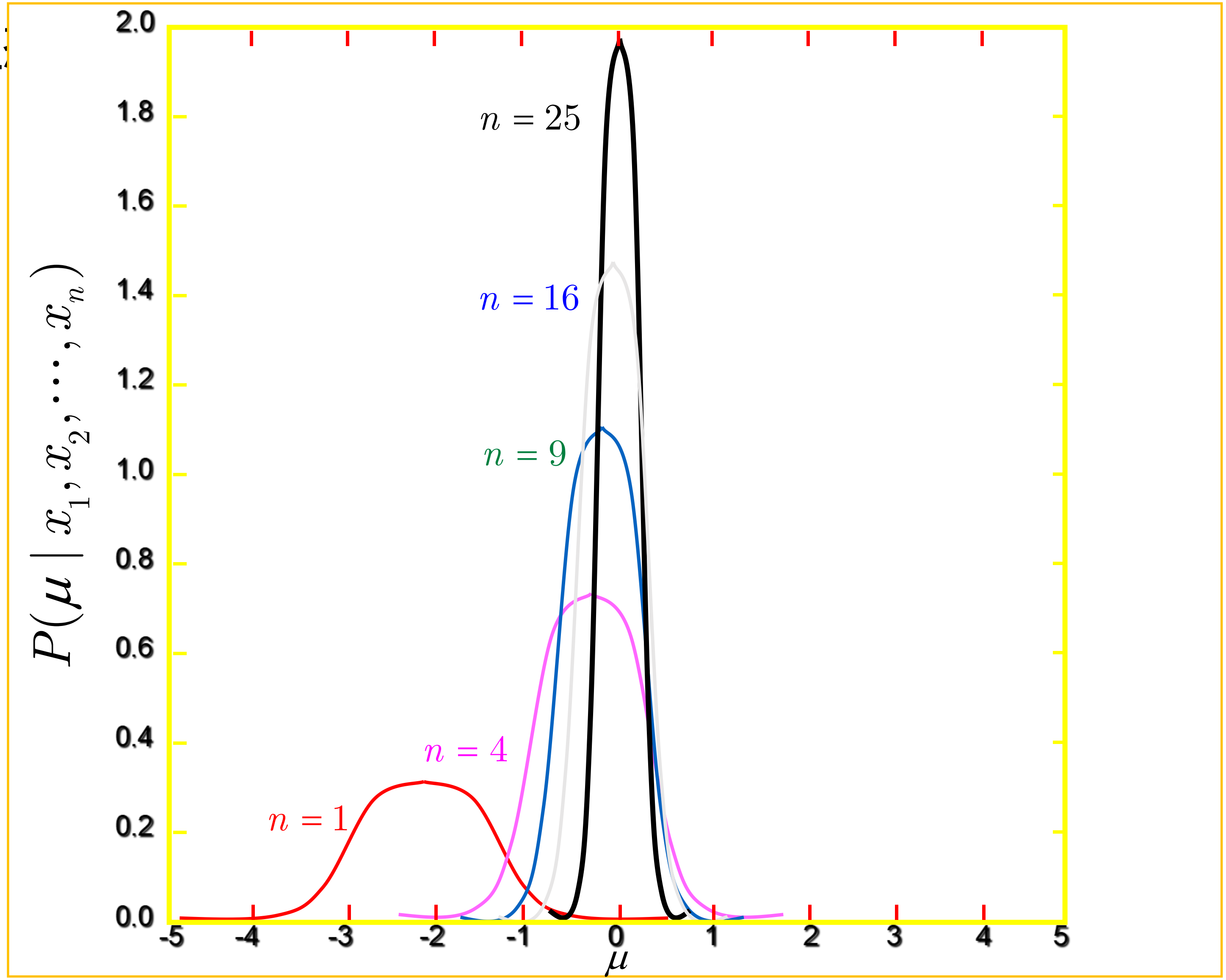
极端情况2：$\sigma_0 \gg \sigma \Rightarrow \mu_n = m_n$，说明先验值十分没有把握。

3. σ_n^2 随n的增加而减小，说明 σ_n^2 趋于 $\frac{\sigma^2}{n}$。

– 参见下页图示

参数学习

■ 正



$P(\mu \mid x_1, x_2, \cdots, x_n)$
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
$n = 25$
$n = 16$
$n = 9$
$n = 4$
$n = 1$
-5
-4
-3
-2
-1
0
1
2
3
4
5
μ

参数学习

- EM估计
  - EM算法和极大似然估计的前提是一样的，都要假设数据总体的分布
  - EM算法的前提会更复杂。男生和女生分别服从两种不同的正态分布。如何评估学生的身高分布呢?
    - 随便抽 100 个男生和 100 个女生，将男生和女生分开，对他们单独进行极大似然估计。分别求出男生和女生的分布。
    - 如果没有办法分开男生和女生?

参数学习

- EM估计
  - 对于每一个样本，有两个相互依赖的问题需要估计，一是这个人是男的还是女的？二是男生和女生对应的身高的正态分布的参数是多少？
    - 当已知每个人是男生还是女生，可以很容易的利用极大似然对男女各自的身高的分布进行估计。
    - 当已知男女身高的分布参数，可以知道每一个人更有可能是男生还是女生。例如我们已知男生的身高分布为 N(172, 5^2)， 女生的身高分布为 N[162, 5^2]， 一个学生的身高为180，我们可以推断出这个学生为男生的可能性更大。

参数学习

- EM估计
  - 先设定男生和女生的身高分布参数初始值（可能不准），如男生的身高分布为N(172, 5^2)， 女生的身高分布为 N[162, 5^2]，
  - Expectation步：计算出每个人更可能属于第一个还是第二个正态分布中的（如这个人的身高是180，他极大可能属于男生）；
  - Maximization步：按上面的方法将这 200 个人分为男生和女生两部分，根据极大似然估计分别对男生和女生的身高分布参数进行估计；

当更新这两个分布的时候，每一个学生属于女生还是男生的概率又变了，那么我们就再需要调整E步；如此往复，直到参数基本不再发生变化或满足结束条件为止。

参数学习

■ EM估计

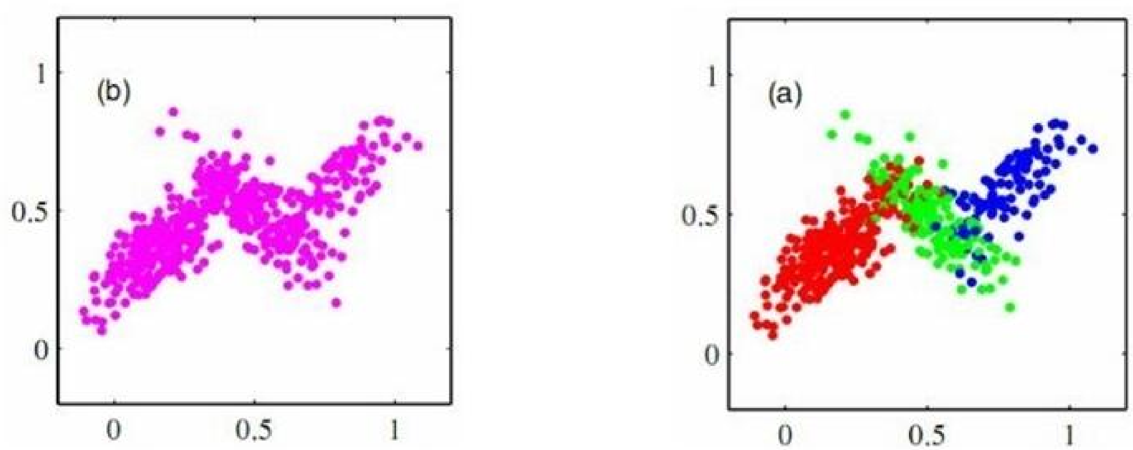
- 多维变量X服从高斯分布时，它的概率密度函数PDF为

$$N(\mathrm{x};\mu,\Sigma)=\frac{1}{\sqrt{2\pi|\Sigma|}}\exp[-\frac{1}{2}(\mathrm{x}-\mu)^T\Sigma^{-1}(\mathrm{x}-\mu)]$$

- 从几何上讲，单高斯分布模型在二维空间应该近似于椭圆，在三维空间上近似于椭球。遗憾的是在很多分类问题中，属于同一类别的样本点并不满足“椭圆”分布的特性。这就引入了高斯混合模型。

参数学习

- **EM估计**
  - 如下图所示，（a）中的数据显然不成“椭圆”形状，因此不能用单一的高斯模型去刻画，而（b）可以将数据分割为3个部分，每个部分都近似成“椭圆”形状，可以用高斯模型刻画，因此整个数据可以用高斯混合模型来刻画。



(a) (b)

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x} \mid \mu_k, \Sigma_k)$$

参数学习

■ 非参数估计

- 前几节的结论是基于概率密度的分布形式已知的假设。
- 实际问题并不一定满足这个假设。
  - 经典的参数密度是单峰的。
  - 实际的问题包含多峰的密度。
- 模式分类的非参数方法
  - 根据样品模式估计密度函数 $P(\mathbf{x} \mid \omega_j)$，然后利用Bayes公式；
  - 直接估计后验概率 $P(\omega_j \mid \mathbf{x})$。

参数学习

■ 非参数估计

➢一个样本X落在区域R里的概率P为

$$P = \int_R p(\mathbf{x})d\mathbf{x}$$

概率P是密度函数p($\mathbf{x}$)的一种经过平均后的形式，对P作估计就是估计出$p(\mathbf{x})$的这个平均值。

- 概率密度估计

➢设样本 $\mathbf{x}_1, \cdots, \mathbf{x}_n$ 是按照概率密度 $p(\mathbf{x})$ 独立抽取的，n个样本中有k个落在区域R里的概率符合二项定律。

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

其中，P是1个样本落在区域R里的概率。

参数学习

■ 非参数估计

k是一个随机变量，k的期望值是

$$E(k) = nP$$

由于k的二项分布在均值附近有一个峰值，所以k/n是P的一个很好的估计。

假设$p(\mathbf{x})$连续，且R小到$p(\mathbf{x})$在R上几乎没有什么变化，则，

$$P = \int_R p(\mathbf{x})d\mathbf{x} \approx p(\mathbf{x}) \cdot \int_R 1 d\mathbf{x} = p(\mathbf{x}) \cdot V$$

其中，$\mathbf{x}$是R中的一点，V是被R包围的体积。

$$p(\mathbf{x}) \approx \frac{k/n}{V}$$

参数学习

- 非参数估计

  - 理论上，假设可以利用的样本数无穷，可以利用极限的方法来研究密度函数的估计。即，
    构造一个包含$\mathbf{x}$在内的区域序列$R_1,\cdots,R_n$，设R_n的体积是V_n，其中的样本数为k_n，则

$$p_n(\mathbf{x})=\frac{k_n/n}{V_n}$$

什么条件？ $\xrightarrow{\text{make}}$ $p_n(\mathbf{x})\to p(\mathbf{x})$

参数学习

■ 非参数估计

- 三个条件：

$$
\left.\begin{aligned}
&1.\quad \lim_{n\to\infty} k_n = \infty \\
&2.\quad \lim_{n\to\infty} V_n = 0 \\
&3.\quad \lim_{n\to\infty} k_n / n = 0
\end{aligned}\right\} \Rightarrow \quad \lim_{n\to\infty} p_n(\mathbf{x}) = p(\mathbf{x})
$$

  - n增大时，落入V_n中样本数k_n也要增加；
  - 同时，V_n应不断减少，以使$p_n(\mathbf{x})$趋于$p(\mathbf{x})$；
  - 在小区域V_n中尽管落入了大量样本，但相对于样本总数，这个数量仍然很小；
  - 为了防止V_n下降太快，必须控制使之下降比V_n/n的下降慢一些，例如$V_n = 1/\sqrt{n}$。

参数学习

■ 非参数估计

- 概率密度估计的结论及方法的演变
  - Parzen窗：在具有一定数量的样本时，可以选定一个中心在$\mathbf{x}$处的体积V_n，然后计算落入其中的样本数k_n来估计局部密度$p_n(\mathbf{x})$的值。

  - k_n近邻估计：选定一个k_n值，以$\mathbf{x}$为中心建立一个体积V_n，让V_n不断增大，直到它能捕获k_n个样本，这是的体积V_n即用来计算$p_n(\mathbf{x})$的估值。
- 问题
  - 样本有限时，上述两种方法的性能难以估计。

参数学习

- 非参数估计
  - Parzen窗函数

$$\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)=\begin{cases}1 & |\mathbf{x}-\mathbf{x}_i|\le\frac{h_n}{2}, j=1,2,\cdots,d\\ 0 & \textbf{otherwise}\end{cases}$$

其中， $\mathbf{x}$是d维空间中要估计概率密度值$p_n(\mathbf{x})$的点，V_n是以$\mathbf{x}$为中心边长为h_n的超立方体。 $\mathbf{x}_i$是样本，落在V_n中的样本数k_n是

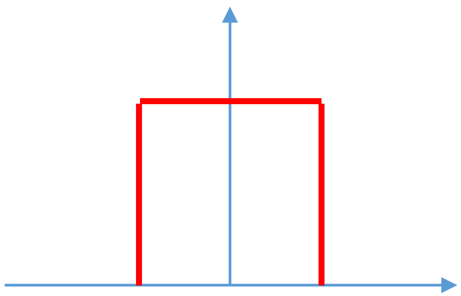
$$k_n=\sum_{i=1}^{n}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) \Rightarrow p_n(\mathbf{x})=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n^d}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)$$

参数学习

■ 非参数估计

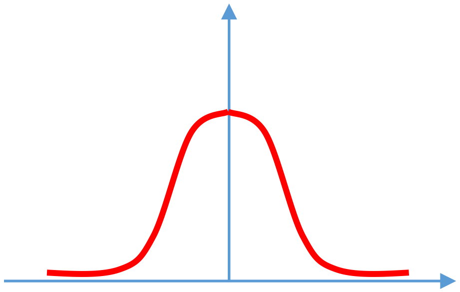
- 方窗函数

$$\varphi(u)=\begin{cases}1 & |u|\leq \frac{1}{2} \\ 0 & \textbf{otherwise}\end{cases}$$



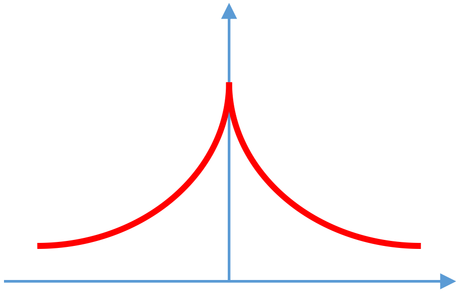
- 正态窗函数

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



- 指数窗函数

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



参数学习

Parzen窗估计法 h

- 例子：正态分布

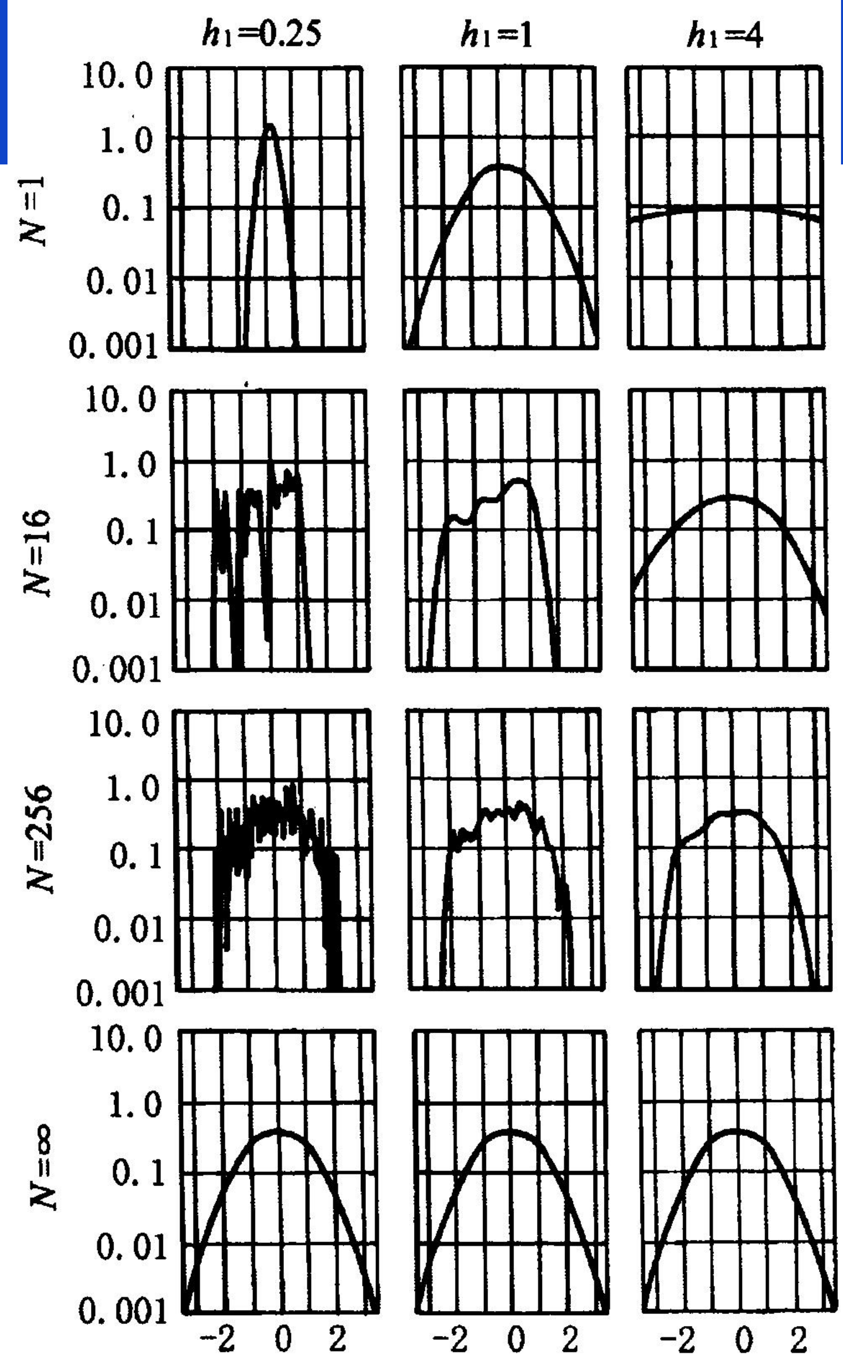
$$p(X) \sim N(0,1)$$

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

$$h_n = h_1 / \sqrt{n}$$

$$p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)$$

平滑的正态曲线



参数学习

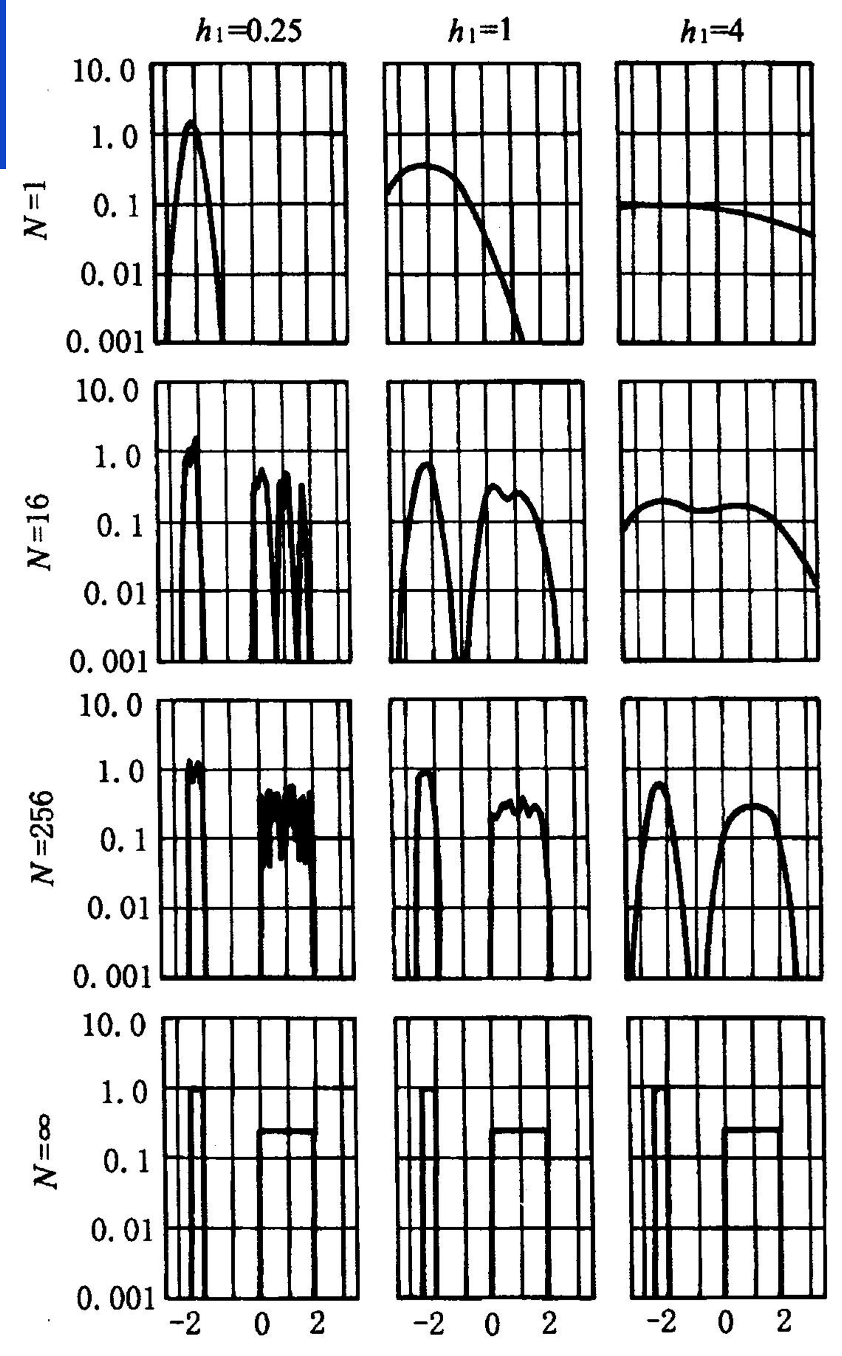
Parzen窗估计法 h

- 例子：二个均匀分布密度的混合。

$$p(\mathbf{x}) = \begin{cases} 1, & -2.5 < x < -2 \\ 0.25, & 0 < x < 2 \\ 0 & otherwise \end{cases}$$

$$p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)$$

混合的方波分布



参数学习

■ 非参数估计

Parzen窗估计法结论

- 优点
  - ➢一般情况下，无论对单峰分布或双峰混合情况，非参数方法都适用
- 缺点
  - ➢要求的样本数目很大，同时增加计算的成本
  - ➢如果特征维数增加，样本的数目按指数速度增长，导致“维数灾难”

参数学习

■ 非参数估计

近邻估计

- Parzen法的问题
  - 单元序列 $V_1, \cdots, V_n$ 的选择问题。对于某组数据适用的V不一定适用于其他数据。
- 策略
  - 建立单元序列和数据之间的函数关系，而不是简单地和样本数目相关。
  - 在数据X的周围建立一个单元并让它不断地增大直至捕获k_n个样品——$\mathbf{x}$的k_n个近邻。

参数学习

■ 非参数估计

近邻估计

- k_n近邻的选择方法

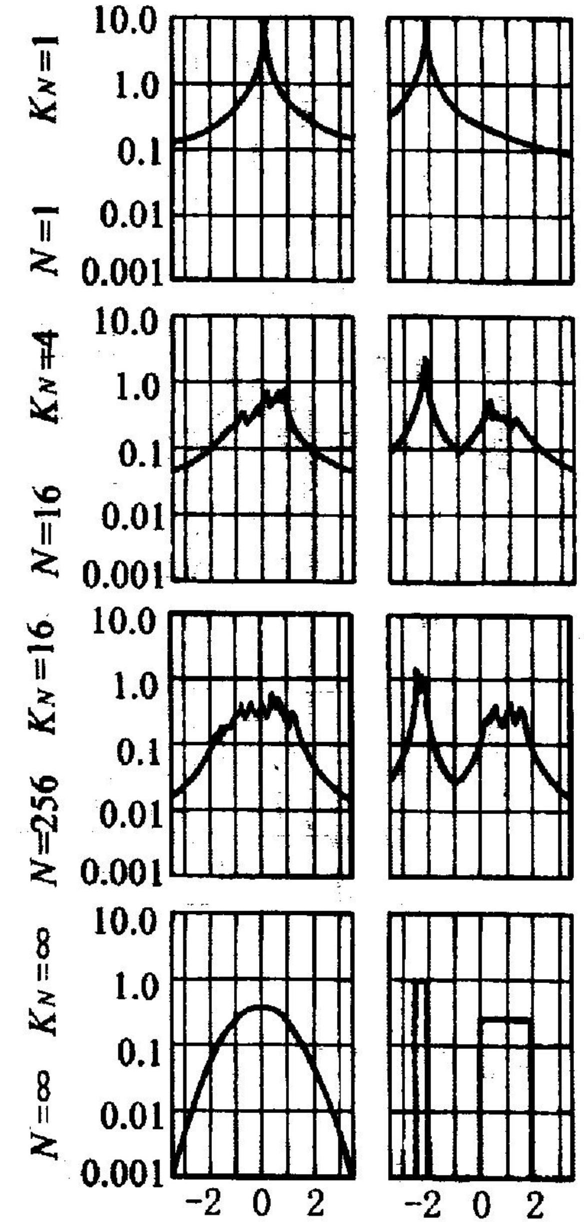
  利用欧氏距离作准则，根据$\mathbf{x}$到它的第k个近邻(k-NN)的欧氏距离$r_n(\mathbf{x})$来估计体积V_n

$$p_n(\mathbf{x}) = \frac{A}{\left[r_k(\mathbf{x})\right]^d}$$

d是特征空间的维数，A是常数，由k_n和n决定

参数学习

- 非参数估计
  - 例子
    - 单一正态分布
    - 两个均匀分布的估计



Thanks