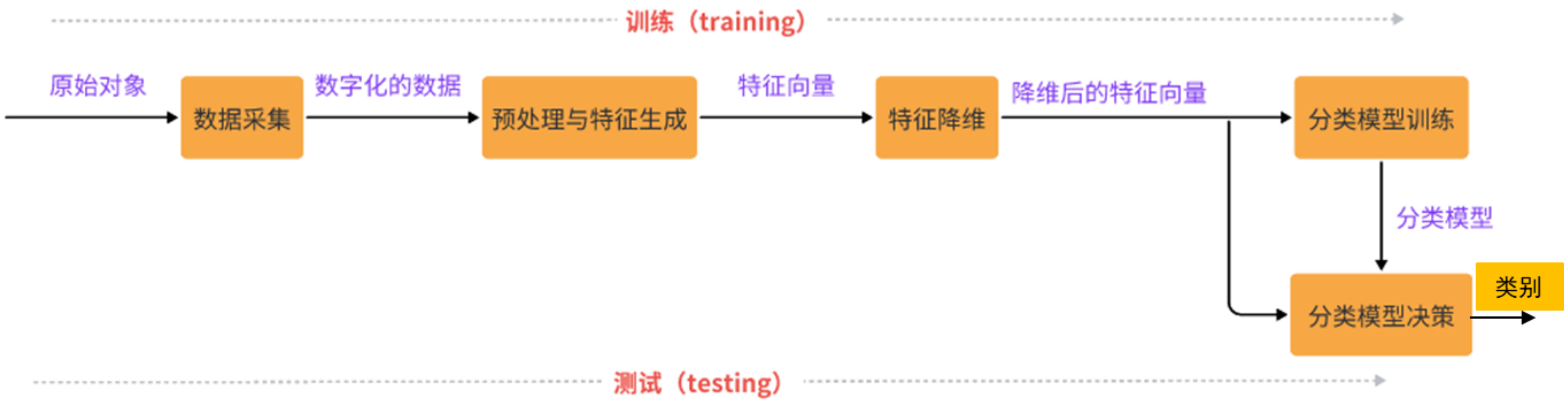
模式识别

（Pattern Recognition）

广东工业大学集成电路学院　邢延

回顾

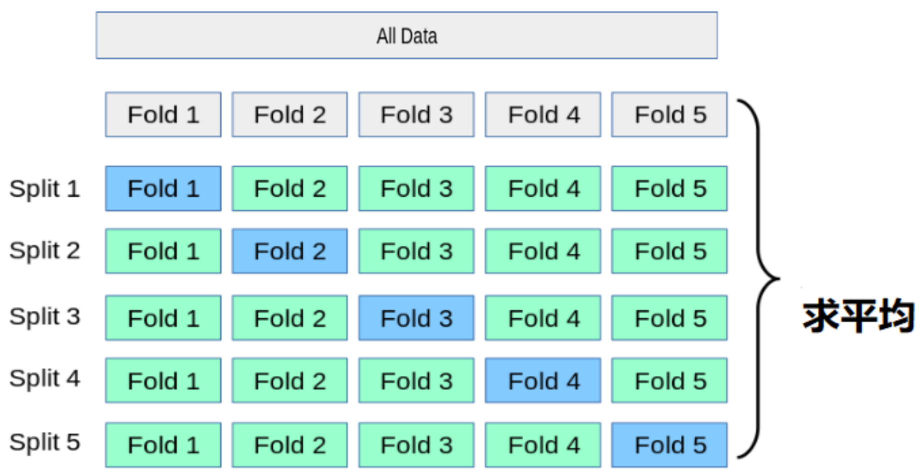
- **上次课内容**
  - **模式识别系统组成**



  - **基于距离的分类器**
    - 最小距离分类器
    - 最近邻分类器
    - K近邻分类器
  - **分类器性能评估**
    - 混淆矩阵、分类准确率、F1、AUC
    - 数据的划分：训练集和测试集

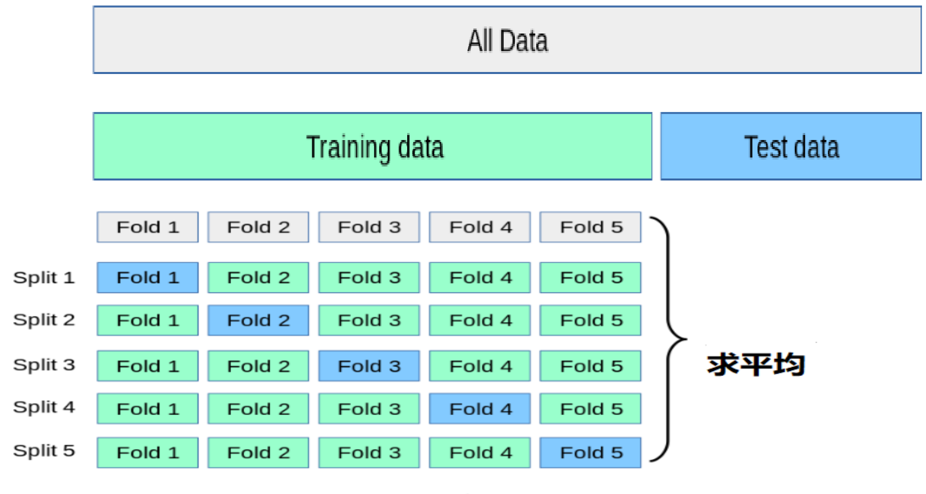
划分训练集和测试集的方法

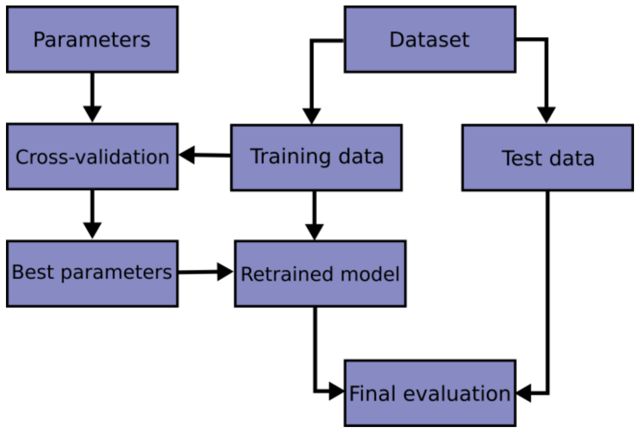
- **上一次课思考题**
  - ➢ **K折交叉验证，每一次训练的模型在参数上是不同的。如果有新数据需要测试，应该用哪一个模型呢?**



划分训练集和测试集的方法

- **思考题**
  - ➢ **答案：用所有数据重新训练一个模型，再用此模型去测试新数据**





模式识别算法编程实例演示

- **实例2：分类器的性能评估实例**
  - ➢ **算法编程实例/lesson02**

第四讲 朴素Bayes分类器
（04 Naïve Bayes Classifiers）

内容提要

- 朴素Bayes分类器
  - Bayes公式
  - 朴素Bayes分类器的基本原理
  - 朴素Bayes分类器的工作步骤
    - 高斯朴素Bayes分类器
  - 朴素Bayes分类器的特点
- 高斯过程分类器

Bayes公式

- **数学背景知识**
  - **频率**
    - ◆ **试验在相同的条件下重复N次，其中M次事件A 发生，则A 发生的频率为：**

$$f_N(A)=\frac{M}{N}$$

  - **概率**
    - ◆ **当N很大时，频率会趋向一个稳定值，称为A的概率：**

$$P(A)=\lim_{N\to\infty} f_N(A)$$

Bayes公式

- **数学背景知识**
  - **联合概率**
    - 设A,B是两个随机事件， A和B同时发生的概率称为联合概率，即$P(A,B)$
  - **条件概率**
    - 在B事件发生的条件下， A事件发生的概率称为条件概率，即$P(A|B)$
  - **乘法定理**

$$P(A|B)=\frac{P(A,B)}{P(B)}$$

Bayes公式

- **数学背景知识**
  - **概率分布函数**
    - ◆ **设X 为连续型随机变量，定义分布函数为：**

$$F(x) = P(X \le x)$$

  - **概率密度函数**
    - ◆ **给定X是随机变量，如果存在一个非负函数$f(x)$，使得对任意实数$a, b(a < b)$,有$P(a < X \le b) = \int_a^b f(x)dx$，则称$f(x)$ 为X 的概率密度函数。**

Bayes公式

- 数学背景知识
  - Bayes公式
    - 统计学中经典的概率公式
    - 事件A与事件B发生的概率用$P(A)$与$P(B)$来表示，先验概率转成后验概率的Bayes公式为：

$$P(A|B)=\frac{P(A,B)}{P(B)}=\frac{P(B,A)}{P(A)\times P(B)}\times P(A)=\frac{P(B|A)\times P(A)}{P(B)}$$

广东工业大学
GUANGDONG UNIVERSITY OF TECHNOLOGY

朴素Bayes分类器的基本原理

- "朴素" 的含义
  - 各观测量必须是相互独立的
    - 即：特征相互独立

朴素Bayes分类器的基本原理

- **基于Bayes公式**

假设：有$C_1, C_2, \cdots, C_m$类数据，和n维特征空间的随机向量$\mathbf{x}$。

对于给定类别$C_i, 1 \le i \le m$,存在类条件概率密度$p(\mathbf{x}|C_i)$,

$p(C_i)$代表$\mathbf{x}$属于类别C_i的先验概率，

则根据Bayes公式，$\mathbf{x}$属于类别C_i的后验概率为：

$$p(C_i|\mathbf{x}) = \frac{p(\mathbf{x}|C_i)p(C_i)}{p(\mathbf{x})}, 1 \le i \le m$$

其中，$p(x) = \sum_{i=1}^{m} p(\mathbf{x}|C_i)p(C_i)$，即$\mathbf{x}$的概率分布

朴素Bayes分类器的基本原理

对于随机向量$\mathbf{x}$, Bayes分类器的判别函数为:

$$\text{If } p(C_j \mid \mathbf{x}) = \max_{1\le i\le m}[p(C_i \mid \mathbf{x})]$$

$$\text{then } \mathbf{x} \in C_j$$

or

简化计算:

$$\text{If } p(C_j)p(\mathbf{x} \mid C_j) = \max_{1\le i\le m}[p(C_i)p(\mathbf{x} \mid C_i)]$$

$$\text{then } \mathbf{x} \in C_j$$

朴素Bayes分类器的基本原理

例子：医生要根据病人血液中白细胞的浓度来判断病人是否患血液病。

根据医学知识和以往的经验医生知道：

一般人群中，患血液病的人数比例为0.5%。

先验概率分布（prior）：没有获得观测数据（病人白细胞浓度）之前类别的分布：

$$p(C_1) = 0.005$$

$$p(C_2) = 0.995$$

C_1: 患病， C_2: 没患病

朴素Bayes分类器的基本原理

根据医学知识和以往的经验医生知道：

- 患血液病的人，白细胞的浓度服从均值=2000，方差=1000的正态分布；
- 未患血液病的人，白细胞的浓度服从均值=7000，方差=3000的正态分布。

类条件概率密度：

$$p(x|C_1) \sim N(2000,1000)$$
$$p(x|C_2) \sim N(7000,3000)$$

x: 白细胞的浓度，C_1: 患病， C_2: 没患病

朴素Bayes分类器的基本原理

一个人的白细胞浓度是3100，医生应该做出怎样的判断？

正态分布的概率密度函数：$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$

当$x=3100$时，类条件概率密度为：

$$p(x|C_1)=\frac{1}{1000\times\sqrt{2\pi}}\exp(-\frac{(3100-2000)^2}{2(1000)^2})=0.000218$$

$$p(x|C_2)=\frac{1}{3000\times\sqrt{2\pi}}\exp(-\frac{(3100-7000)^2}{2(3000)^2})=0.000057$$

则：

$$p(C_1)p(x|C_1)=0.005\times0.000218=0.0000011$$

$$p(C_2)p(x|C_2)=0.995\times0.000057=0.0000568$$

$\therefore$白细胞浓度为3100时被判为没有血液病

朴素Bayes分类器的基本原理

- **Bayes公式的物理含义**

$$p(C_i|\mathbf{x}) = \frac{p(\mathbf{x}|C_i)\,p(C_i)}{p(\mathbf{x})}, 1 \le i \le m$$

其中，$p(\mathbf{x}) = \sum_{i=1}^{m} p(\mathbf{x}|C_i)\,p(C_i)$，即$\mathbf{x}$的概率分布

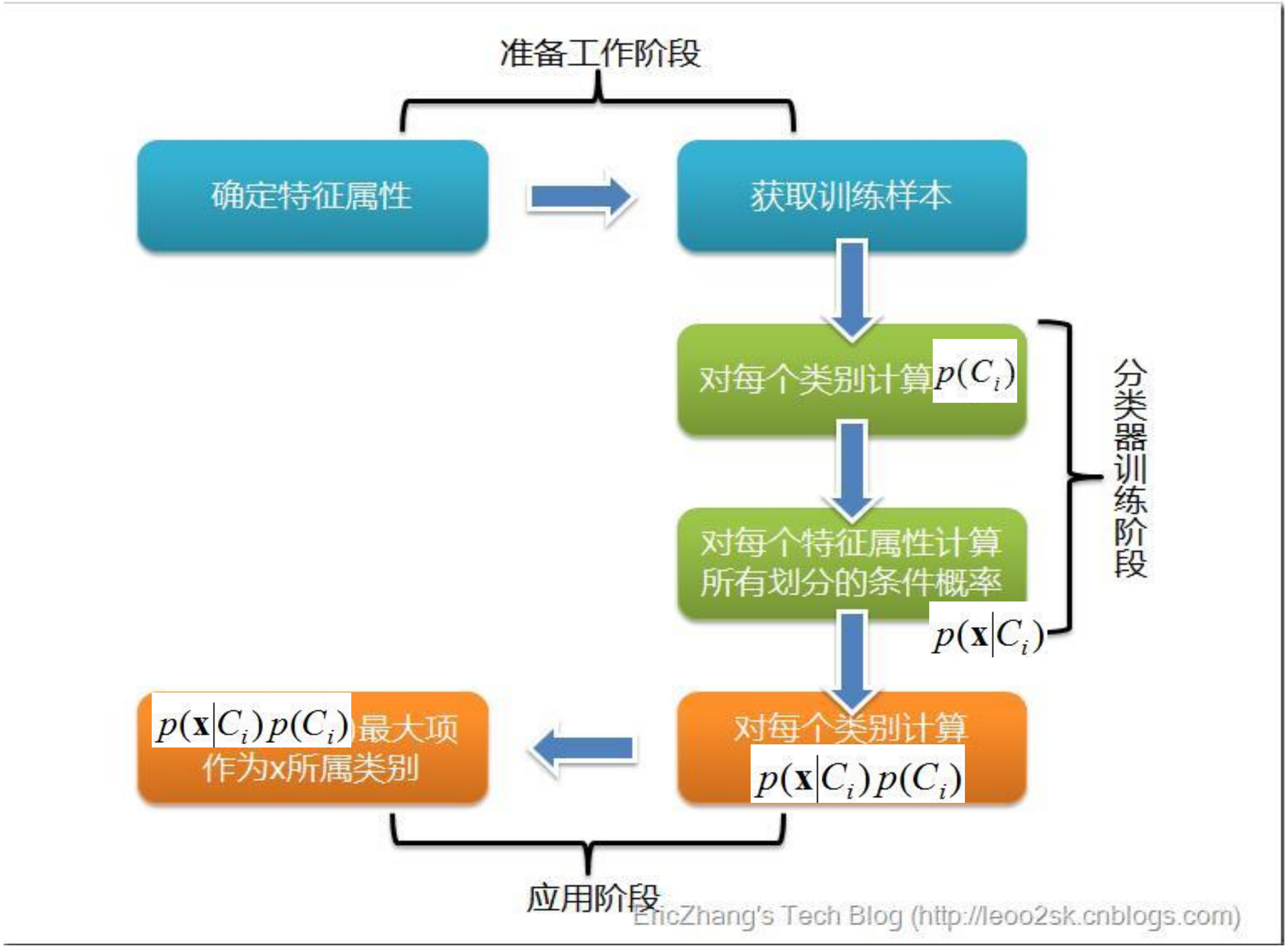
$$posterio = \frac{likelihood \times prior}{evidence}$$

后验概率主要由先验概率和似然函数的乘积所决定

给定训练数据集，evidence是个常数

朴素Bayes分类器的工作步骤

- **工作步骤**



朴素Bayes分类器的工作步骤

- 用Bayes分类器需要事先知道两种知识
  - 各类的先验概率；
  - 待分类向量的类条件概率密度。
- 知识的获取（估计—概率统计角度，训练—模式识别角度）
  - 对问题的一般性的认识；
  - 数据

朴素Bayes分类器的工作步骤

- **类的先验概率的估计**
  - 依靠经验；
  - 用训练数据中各类出现的频率估计
    - 用频率估计概率的优点：
      - 无偏性(No bias)；
      - 相合性(consistency)；
      - 收敛速度快。

朴素Bayes分类器的工作步骤

例子：男女19人进行体检，测量身高和体重，如下表。但事后发现4人忘了写性别，试问，这4人是男是女?

序号	身高（cm）	体重（kg）	性别	序号	身高（cm）	体重（kg）	性别
1	170	68	男	11	140	62	男
2	130	66	女	12	150	64	女
3	180	71	男	13	120	66	女
4	190	73	男	14	150	66	男
5	160	70	女	15	130	65	男
6	150	66	男	16	140	70	α ?
7	190	68	男	17	150	60	β ?
8	210	76	男	18	145	65	γ ?
9	100	58	女	19	160	75	δ ?
10	170	75	男				

训练（数据）集　　待分类（数据）集

朴素Bayes分类器的工作步骤

- **先验概率的估计 －－训练集中的频率**

$$C_1 \quad ---- \quad 男$$

$$C_2 \quad ---- \quad 女$$

$$P(C_1) = 10/15 = 2/3$$

$$P(C_2) = 5/15 = 1/3$$

朴素Bayes分类器的工作步骤

- **类条件概率密度的估计**
  - **非常难**
  - **概率密度函数包含了一个随机变量的全部信息**
  - **概率密度函数可以是满足下面条件的任何函数：**

$$p(x) \geq 0, (-\infty < x < \infty)$$

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

朴素Bayes分类器的工作步骤

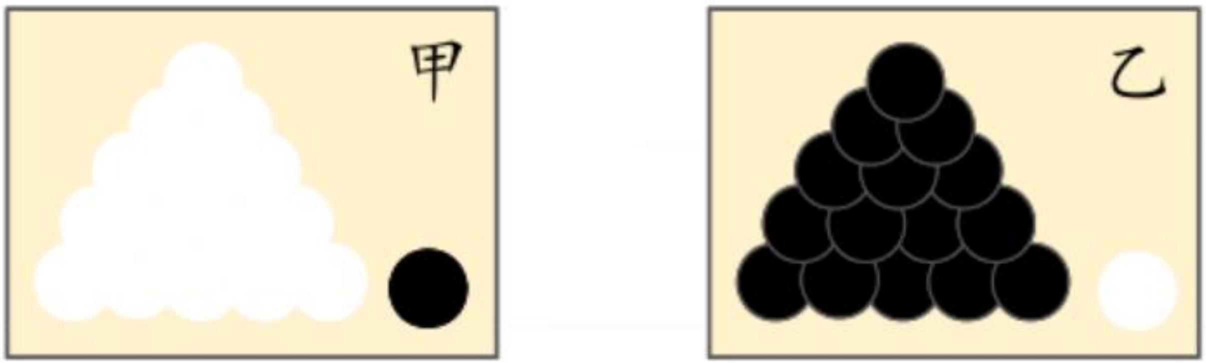
- **概率密度估计的两种主要思路**
  - **参数估计：**
    - 根据对问题的一般性的认识，假设随机变量服从某种概率分布，分布函数的参数通过训练数据来估计。
  - **非参数估计：**
    - 不用模型，而只利用训练数据本身对概率密度做估计。

朴素Bayes分类器的工作步骤

- **概率密度函数的参数估计(parametric methods)**
  - **前提**
    - **假设已知概率密度函数的形式（如正态分布、二项分布等），只需对函数中的参数（如均值、方差等）进行估计**
    - **方法**
      - **极大似然估计**

朴素Bayes分类器的工作步骤

- **极大似然估计(maximum likelihood Estimation)**
  - ➢ **最大似然原理**



- 例：有两个外形完全相同的箱子，甲箱中有99只白球，1只黑球；乙箱中有99只黑球，1只白球。一次试验取出一球，结果取出的是黑球。
- 问：黑球从哪个箱子中取出？

最大似然原理：“最像”就是“最大似然”

朴素Bayes分类器的工作步骤

- **极大似然估计(maximum likelihood Estimation)**
  - **极大似然估计的目的**
    - **利用已知的样本，反推最有可能（最大概率）导致这样结果的参数值。**

朴素Bayes分类器的工作步骤

- **极大似然估计(maximum likelihood Estimation)**
  - **原理**

考虑训练样本集D，各样本都是独立同分布：

$$D=\{x_1,x_2,\cdots,x_N\}$$

用D来估计参数向量θ。

参数向量θ的似然函数（linkehood function）：

$$l(\theta)=p(D|\theta)=p(x_1,x_2,\cdots,x_N|\theta)=\prod_{i=1}^{N}p(x_i|\theta)$$

求 $\hat{\theta}$ **使得** $l(\theta)$,

朴素Bayes分类器的工作步骤

- **原理**

$$\hat{\theta} = \arg\max_{\theta} l(\theta) = \arg\max_{\theta} \prod_{i=1}^{N} p(x_i \mid \theta)$$

为了简化计算，定义对数似然函数：

$$H(\theta) = \ln l(\theta)$$

则有：

$$\hat{\theta} = \arg\max_{\theta} H(\theta) = \arg\max_{\theta} \ln l(\theta) = \arg\max_{\theta} \sum_{i=1}^{N} \ln p(x_i \mid \theta)$$

存在以下两种情况：

朴素Bayes分类器的工作步骤

- **原理**

$$\hat{\theta} = \arg\max_{\theta} H(\theta) = \arg\max_{\theta} \ln l(\theta) = \arg\max_{\theta} \sum_{i=1}^{N} \ln p(x_i|\theta)$$

1. 未知参数只有一个（θ为标量）

在似然函数满足连续、可微的正则条件下，极大似然估计量是下面微分方程的解：

$$\frac{dl(\theta)}{d\theta} = 0 \quad \text{或者等价于} \quad \frac{dH(\theta)}{d\theta} = \frac{d\ln l(\theta)}{d\theta} = 0$$

朴素Bayes分类器的工作步骤

- **原理** $\hat{\theta} = \arg\max_{\theta} H(\theta) = \arg\max_{\theta} \ln l(\theta) = \arg\max_{\theta} \sum_{i=1}^{N} \ln p(x_i|\theta)$

2.未知参数有多个（θ为向量）

则θ可表示为具有S个分量的未知向量：

$$\theta = [\theta_1, \theta_2, \cdots, \theta_S]^T$$

记梯度算子：

$$\nabla_\theta = \left[\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}, \cdots, \frac{\partial}{\partial\theta_S}\right]^T$$

若似然函数满足连续可导的条件，则最大似然估计量就是如下方程的解。

$$\nabla_\theta H(\theta) = \nabla_\theta \ln l(\theta) = \sum_{i=1}^{N} \nabla_\theta \ln P(x_i \mid \theta) = 0$$

朴素Bayes分类器的工作步骤

- **例子**

例1：设样本服从正态分布 $N(\mu,\sigma^2)$ ，则似然函数为：

$$L(\mu,\sigma^2)=\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}=(2\pi\sigma^2)^{-\frac{n}{2}}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

它的对数：

$$\ln L(\mu,\sigma^2)=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

求导，得方程组：

$$\begin{cases}\dfrac{\partial\ln L(\mu,\sigma^2)}{\partial\mu}=\dfrac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0\\[2ex]\dfrac{\partial\ln L(\mu,\sigma^2)}{\partial\sigma^2}=-\dfrac{n}{2\sigma^2}+\dfrac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2=0\end{cases}$$

联合解得：

$$\begin{cases}\mu^*=\overline{x}=\dfrac{1}{n}\sum_{i=1}^{n}x_i\\[2ex]\sigma^{*2}=\dfrac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\end{cases}$$

朴素Bayes分类器的工作步骤

- **正态分布下的最大似然参数估计**
  - **单变量**

$$\mu^* = \bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

$$\sigma^{*2} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})^2$$

  - **多变量**

$$\boldsymbol{\mu}^* = \bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

$$\mathbf{C}^* = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

朴素Bayes分类器的工作步骤

例子：男女19人进行体检，测量身高和体重，如下表。但事后发现4人忘了写性别，试问，这4人是男是女?

序号	身高（cm）	体重（kg）	性别	序号	身高（cm）	体重（kg）	性别
1	170	68	男	11	140	62	男
2	130	66	女	12	150	64	女
3	180	71	男	13	120	66	女
4	190	73	男	14	150	66	男
5	160	70	女	15	130	65	男
6	150	66	男	16	140	70	α ?
7	190	68	男	17	150	60	β ?
8	210	76	男	18	145	65	γ ?
9	100	58	女	19	160	75	δ ?
10	170	75	男				

训练（数据）集　　待分类（数据）集

朴素Bayes分类器的工作步骤

- **类条件概率密度的最大似然估计**

设类条件概率密度函数为正态分布函数

C_1 ---- 男

C_2 ---- 女

$$\mathbf{x}=\{x_1,x_2\}=\{\text{身高，体重}\}$$

对于男性：

$$\boldsymbol{\mu}=\{168,69\}$$

$$\boldsymbol{\Sigma}=\begin{bmatrix}576 & 85\\ 85 & 19\end{bmatrix}$$

对于女性：

$$\boldsymbol{\mu}=\{132,64.8\}$$

$$\boldsymbol{\Sigma}=\begin{bmatrix}570 & 66.4\\ 66.4 & 19.2\end{bmatrix}$$

朴素Bayes分类器的工作步骤

- **高斯朴素Bayes分类器（Gaussian NB）**
  - **设特征为正态分布的朴素bayes分类器就是高斯NB分类器**
  - **此外，根据不同的概率分布假设**
    - **伯努利朴素贝叶斯分类器**
    - **多项式朴素贝叶斯分类器**
    - **可能会是翻转课堂题目**

朴素Bayes分类器的工作步骤

- **极大似然估计的特点**
  - ➢ 比其他估计方法更加简单；
  - ➢ 收敛性：无偏或者渐近无偏，当样本数目增加时，收敛性质会更好；
  - ➢ 如果假设的类条件概率模型正确，则通常能获得较好的结果。但如果假设模型出现偏差，将导致非常差的估计结果。

朴素Bayes分类器的工作步骤

- **概率密度估计的两种主要思路：**
  - **参数估计：**
    - 根据对问题的一般性的认识，假设随机变量服从某种分布，分布函数的参数通过训练数据来估计。
  - **非参数估计：**
    - 不用模型，而只利用训练数据本身对概率密度做估计。

朴素Bayes分类器的工作步骤

- **概率密度的非参数估计(non-parametric methods)**

  - **特点：**
    - 不假设概率密度函数的形式
    - 不假设参数的形式
    - 能处理任意的概率分布
    - 需要样本的数量足够大

  - **典型的非参数估计算法**
    - K近邻分类器

朴素Bayes分类器的特点

- 朴素Bayes分类器所需的条件是
  - 必须知道各类先验概率；
  - 必须知道观测量的类条件概率密度，否则要估计；
  - 各观测量必须是相互独立的；
  - 各样本必须是相互独立的。

模式识别算法编程实例演示

- **实例3：朴素Bayes分类器**
  - ➢ **参见Jupyter Notebook文档目录：lesson03**