

# 机器学习

## Machine learning

# 第二章 贝叶斯学习

## Bayesian Learning

授课人：周晓飞

zhouxiaofei@iie.ac.cn

2020-9-24

## 第二章 贝叶斯学习

2.1 概述

2.2 贝叶斯决策论

**2.3 贝叶斯分类器**

2.4 贝叶斯学习与参数估计问题

# 贝叶斯分类器

## 预备知识

**贝叶斯分类器**：基于 Bayesian 决策的分类器

**变量和参数**：

类别  $C$ :  $C=(c_1,c_2,\dots,c_M)$ ,

数据  $D$  和样本  $x$ :  $D=\{x_i\}$

**贝叶斯学习**

$$P(c_i|x) \propto P(x|c_i)P(c_i)$$

**核心是估计**

$$P(x|c_i)P(c_i)$$

# 贝叶斯分类器

## 预备知识

### 贝叶斯决策

类别相似性函数:

$$g_i(x) = p(c_i | x) = \frac{p(x | c_i) p(c_i)}{\sum_{j=1}^c p(x | c_j) p(c_j)}$$

$$g_i(x) = p(x | c_i) p(c_i)$$

$$g_i(x) = \ln p(x | c_i) + \ln p(c_i)$$

决策函数:

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

$$g(x) = p(c_1 | x) - p(c_2 | x)$$

$$g(x) = \ln \frac{p(x | c_1)}{p(x | c_2)} + \ln \frac{p(c_1)}{p(c_2)}$$

# 贝叶斯分类器

## 预备知识

### 贝叶斯分类器

- 朴素贝叶斯分类器：假设  $P(x|c)$  中  $x$  特征向量的各维属性独立；
- 半朴素贝叶斯分类器：假设  $P(x|c)$  中  $x$  的各维属性存在依赖；
- 正态分布的贝叶斯分类器：假设  $P(x|c(\theta))$  服从正态分布；

# 贝叶斯分类器

## 朴素贝叶斯分类器

采用了“属性条件独立性假设”

$$P(c|\mathbf{x}) = \frac{P(c)P(\mathbf{x}|c)}{P(\mathbf{x})} \propto P(c)P(\mathbf{x}|c) = P(c) \prod_{i=1}^d P(x_i|c)$$

关键问题：由训练样本学习类别条件概率和类别先验概率

$$P(x_i|c) \text{ 和 } P(c)$$

# 贝叶斯分类器

## 朴素贝叶斯分类器

采用了“属性条件独立性假设”

$$P(c|\mathbf{x}) = \frac{P(c)P(\mathbf{x}|c)}{P(\mathbf{x})} \propto P(c)P(\mathbf{x}|c) = P(c) \prod_{i=1}^d P(x_i|c)$$

关键问题：由训练样本学习类别条件概率和类别先验概率

$$P(x_i|c) \text{ 和 } P(c)$$

需要学习的概率分布？

$k$  个类别， $d$  个属性：  $p(c)$  和  $P(x_i|c_j)$ , ( $i=1, \dots, d$  个属性,  $j=1, \dots, k$ )

共  $1 + d*k$  个概率分布要统计.

# 贝叶斯分类器

## 朴素贝叶斯分类器

类别先验概率的估计  $P(c) = \frac{|D_c|}{|D|}$

### 类别概率密度估计

- $x_i$  离散情况:

$$P(x_i | c) = \frac{|D_{c,x_i}|}{|D_c|}$$

$D_{c,x_i}$  表示  $D_c$  中在第  $i$  个属性上取值为  $x_i$  的样本组成的集合;

- $x_i$  连续情况:

$$P(x_i | c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}^2}\right) \quad (\text{由某一概率分布估计类别概率})$$



# 贝叶斯分类器

## 朴素贝叶斯分类器

### 学习过程

(1) 类别先验估计  $P(c) = \frac{|D_c|}{|D|}$

(2) 类别条件概率估计  $P(x_i | c) = \frac{|D_{c,x_i}|}{|D_c|}$

# 贝叶斯分类器

## 朴素贝叶斯分类器

### 决策过程

(1) 类别先验估计  $P(c) = \frac{|D_c|}{|D|}$

(2) 类别条件概率估计  $P(\mathbf{x}|c) = \prod_{i=1}^d P(x_i|c)$

(3) 贝叶斯决策  $h(\mathbf{x}) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}} P(c) \prod_{i=1}^d P(x_i|c)$

# 贝叶斯分类器

## 朴素贝叶斯分类器

### 拉普拉斯平滑

避免因训练集样本不充分而导致概率估计值为零.

避免  $P(c|\mathbf{x}) \propto P(c) \prod_{i=1}^d P(x_i|c)$  中,  $P(c)$  或  $P(x_i|c)$  为 0 (即  $|D_c| = 0$  或  $|D_{c,x_i}| = 0$ )

进行拉普拉斯平滑

$$\hat{P}(c) = \frac{|D_c| + 1}{|D| + N},$$

$$\hat{P}(x_i | c) = \frac{|D_{c,x_i}| + 1}{|D_c| + N_i}$$

# 贝叶斯分类器

## 朴素贝叶斯分类器

**例子：《机器学习》教材 p.151**

- 数据： pp.84
- 离散属性： 色泽、根蒂、敲声、纹理、脐部、触感；  
连续属性： 密度、含糖绿  
类别属性： 好瓜？（是与否）

# 贝叶斯分类器

## 朴素贝叶斯分类器

- 训练数据

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

## 测试数据

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

# 贝叶斯分类器

## 朴素贝叶斯分类器

学习过程：

### (1) 类别先验估计

$$P(c) = \frac{|D_c|}{|D|} \Rightarrow \begin{cases} P(\text{好瓜} = \text{是}) = \frac{|D_{\text{好瓜}=\text{是}}|}{|D|} \\ P(\text{好瓜} = \text{否}) = \frac{|D_{\text{好瓜}=\text{否}}|}{|D|} \end{cases}$$

首先估计类别先验概率  $P(c)$ ，显然有

$$P(\text{好瓜} = \text{是}) = \frac{8}{17} \approx 0.471,$$

$$P(\text{好瓜} = \text{否}) = \frac{9}{17} \approx 0.529.$$

# 贝叶斯分类器

## 朴素贝叶斯分类器

### (2) 类别条件概率估计

对于离散值属性:  $P(x_i | c) = \frac{|D_{c,xi}|}{|D_c|} \Rightarrow \begin{cases} P(x_i | \text{好瓜} = \text{是}) = \frac{|D_{\text{好瓜}=\text{是},xi}|}{|D_c|} \\ P(x_i | \text{好瓜} = \text{否}) = \frac{|D_{\text{好瓜}=\text{否},xi}|}{|D_c|} \end{cases}$

对于连续值属性:

$$P(x_i | c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}^2}\right) \Rightarrow \begin{cases} P(x_i | \text{好瓜} = \text{是}) = \frac{1}{\sqrt{2\pi}\sigma_{\text{好瓜}=\text{是},i}} \exp\left(-\frac{(x_i - \mu_{\text{好瓜}=\text{是},i})^2}{2\sigma_{\text{好瓜}=\text{是},i}^2}\right) \\ P(x_i | \text{好瓜} = \text{否}) = \frac{1}{\sqrt{2\pi}\sigma_{\text{好瓜}=\text{否},i}} \exp\left(-\frac{(x_i - \mu_{\text{好瓜}=\text{否},i})^2}{2\sigma_{\text{好瓜}=\text{否},i}^2}\right) \end{cases}$$

均值、方差作为参数, 可用 ML 估计,

$$P_{\text{青绿|是}} = P(\text{色泽} = \text{青绿} | \text{好瓜} = \text{是}) = \frac{3}{8} = 0.375$$

$$P_{\text{青绿|否}} = P(\text{色泽} = \text{青绿} | \text{好瓜} = \text{否}) = \frac{3}{9} = 0.333$$

$$P_{\text{蜷缩|是}} = P(\text{根蒂} = \text{蜷缩} | \text{好瓜} = \text{是}) = \frac{5}{8} = 0.625$$

$$P_{\text{蜷缩|否}} = P(\text{根蒂} = \text{蜷缩} | \text{好瓜} = \text{否}) = \frac{3}{9} = 0.333$$

$$P_{\text{浊响|是}} = P(\text{敲声} = \text{浊响} | \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750$$

$$P_{\text{浊响|否}} = P(\text{敲声} = \text{浊响} | \text{好瓜} = \text{否}) = \frac{4}{9} = 0.444$$

$$P_{\text{清晰|是}} = P(\text{纹理} = \text{清晰} | \text{好瓜} = \text{是}) = \frac{7}{8} = 0.875$$

$$P_{\text{清晰|否}} = P(\text{纹理} = \text{清晰} | \text{好瓜} = \text{否}) = \frac{2}{9} = 0.222$$

$$P_{\text{凹陷|是}} = P(\text{脐部} = \text{凹陷} | \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750$$

$$P_{\text{凹陷|否}} = P(\text{脐部} = \text{凹陷} | \text{好瓜} = \text{否}) = \frac{2}{9} = 0.222$$

$$P_{\text{硬滑|是}} = P(\text{触感} = \text{硬滑} | \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750$$

$$P_{\text{硬滑|否}} = P(\text{触感} = \text{硬滑} | \text{好瓜} = \text{否}) = \frac{6}{9} = 0.667$$

为每个属性估计条件概率  $P(x_i | c)$

$$P_{\text{密度:0.697|是}} = p(\text{密度} = 0.697 | \text{好瓜} = \text{是})$$

$$= \frac{1}{\sqrt{2\pi} \cdot 0.129} \exp\left(-\frac{(0.697 - 0.574)^2}{2 \cdot 0.129^2}\right) \approx 1.959$$

$$P_{\text{密度:0.697|否}} = p(\text{密度} = 0.697 | \text{好瓜} = \text{否})$$

$$= \frac{1}{\sqrt{2\pi} \cdot 0.195} \exp\left(-\frac{(0.697 - 0.496)^2}{2 \cdot 0.195^2}\right) \approx 1.203$$

$$P_{\text{含糖:0.460|是}} = p(\text{含糖率} = 0.460 | \text{好瓜} = \text{是})$$

$$= \frac{1}{\sqrt{2\pi} \cdot 0.101} \exp\left(-\frac{(0.460 - 0.279)^2}{2 \cdot 0.101^2}\right) \approx 0.788$$

$$P_{\text{含糖:0.460|否}} = p(\text{含糖率} = 0.460 | \text{好瓜} = \text{否})$$

$$= \frac{1}{\sqrt{2\pi} \cdot 0.108} \exp\left(-\frac{(0.460 - 0.154)^2}{2 \cdot 0.108^2}\right) \approx 0.066$$



# 贝叶斯分类器

## 朴素贝叶斯分类器

### (3) 贝叶斯决策

$$\begin{aligned} &P(\text{好瓜} = \text{是}) \times P_{\text{青绿}|\text{是}} \times P_{\text{蜷缩}|\text{是}} \times P_{\text{浊响}|\text{是}} \times P_{\text{清晰}|\text{是}} \times P_{\text{凹陷}|\text{是}} \\ &\quad \times P_{\text{硬滑}|\text{是}} \times p_{\text{密度}:0.697|\text{是}} \times p_{\text{含糖}:0.460|\text{是}} \approx 0.038, \\ &P(\text{好瓜} = \text{否}) \times P_{\text{青绿}|\text{否}} \times P_{\text{蜷缩}|\text{否}} \times P_{\text{浊响}|\text{否}} \times P_{\text{清晰}|\text{否}} \times P_{\text{凹陷}|\text{否}} \\ &\quad \times P_{\text{硬滑}|\text{否}} \times p_{\text{密度}:0.697|\text{否}} \times p_{\text{含糖}:0.460|\text{否}} \approx 6.80 \times 10^{-5}. \end{aligned}$$

由于  $0.038 > 6.80 \times 10^{-5}$ ，因此，朴素贝叶斯分类器将测试样本“测 1”判别为“好瓜”。

*Have a break !*

# 贝叶斯分类器

## 正态密度的贝叶斯分类器

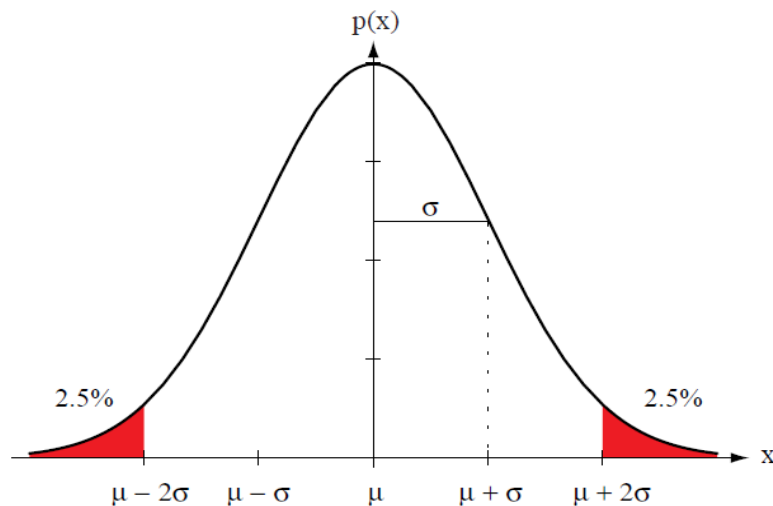
类别条件概率为正态分布

$$h(\mathbf{x}) = \operatorname{argmax}_{c \in \mathcal{Y}} P(c) \underbrace{P(\mathbf{x}|c)}_{\text{正态分布}}$$

正态分布

- 正态分布的概率密度  $N(\mu, \sigma^2)$  :

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



# 贝叶斯分类器

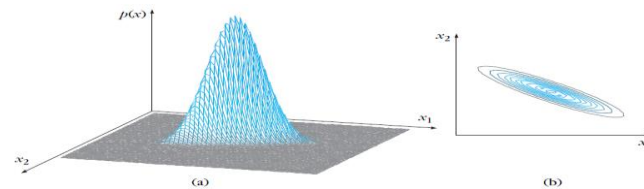
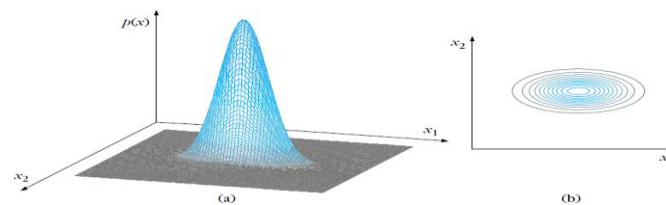
## 正态密度的贝叶斯分类器

- 多维正态分布的概率密度  $N(\mu, \Sigma)$  :

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

$$\mu \equiv \mathcal{E}[x] = \int x p(x) dx$$

$$\Sigma \equiv \mathcal{E}[(x - \mu)(x - \mu)^T] = \int (x - \mu)(x - \mu)^T p(x) dx$$



每个维度上都是正态分布  $\mu_i = \mathcal{E}[x_i]$ ;  $\sigma_{ij} = \mathcal{E}[(x_i - \mu_i)(x_j - \mu_j)]$

# 贝叶斯分类器

## 正态密度的贝叶斯分类器

贝叶斯分类：

- 贝叶斯学习（结果取对数）：

$$g_i(x) = \ln(p(x | \omega_i) p(\omega_i)) = \ln p(x | \omega_i) + \ln p(\omega_i)$$

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

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

# 贝叶斯分类器

## 正态密度的贝叶斯分类器

- 决策函数:

$$g_{ij}(x) \equiv g_i(x) - g_j(x)$$

$g_{ij}(x)=0$  为决策界

如果  $g_{ij}(x) \geq 0$  , 则归为  $i$  类

如果  $g_{ij}(x) < 0$  , 则归为  $j$  类

# 贝叶斯分类器

## 正态密度的贝叶斯分类器

### 不同高斯参数情况讨论

Case 1:  $\Sigma_i = \sigma^2 I$

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) + \ln p(\omega_i) + c_i$$

$$g_i(x) = -\frac{\|x - \mu_i\|^2}{2\sigma^2} + \ln p(\omega_i)$$

$$g_i(x) = -\frac{1}{2\sigma^2} [x^T x - 2\mu_i^T x + \mu_i^T \mu_i] + \ln p(\omega_i) \rightarrow \text{与类别无关, 可忽略}$$

$$\Rightarrow g_i(x) = w_i^T x + w_{i0}, \quad w_i = \frac{1}{\sigma^2} \mu_i, \quad w_{i0} = \frac{-1}{2\sigma^2} \mu_i^T \mu_i + \ln p(\omega_i)$$

# 贝叶斯分类器

## 正态密度的贝叶斯分类器

### 不同高斯参数情况讨论

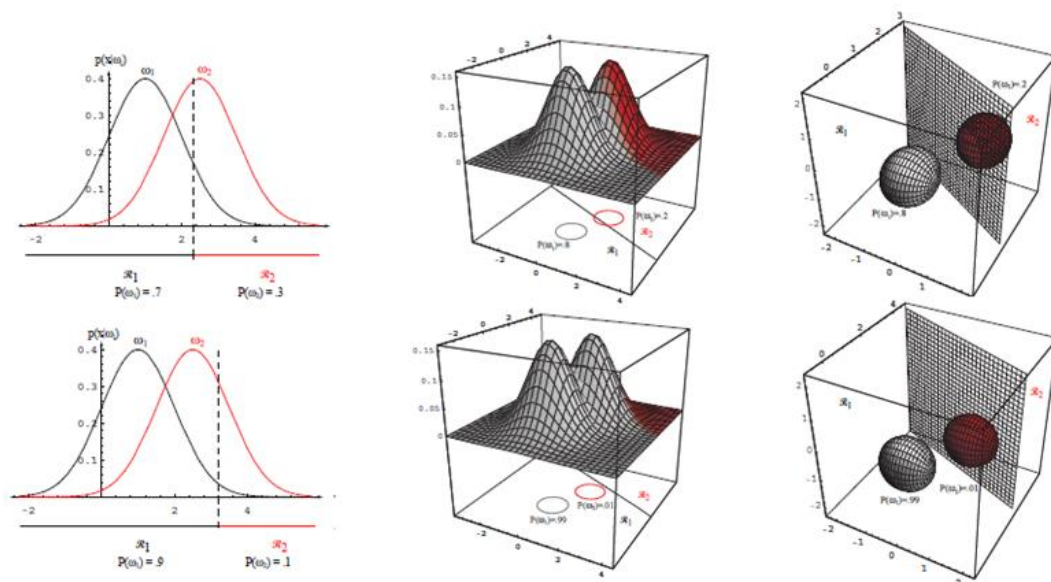
Case 1:  $\Sigma_i = \sigma^2 I$

决策界:  $g_i(x) - g_j(x) = 0$

$$w^T (x - x_0) = 0$$

$$w = \mu_i - \mu_j$$

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





# 贝叶斯分类器

## 正态密度的贝叶斯分类器

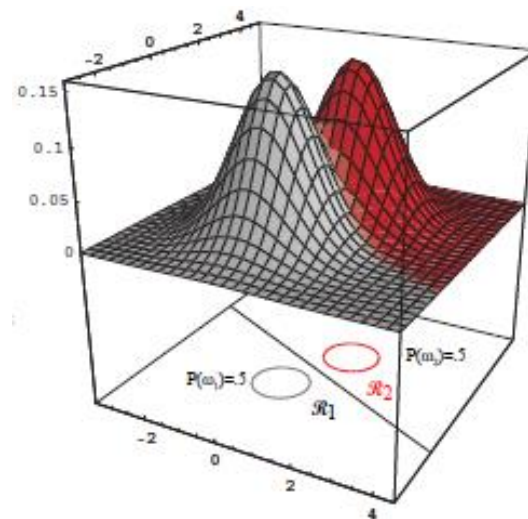
### 不同高斯参数情况讨论

*Case 1:  $\Sigma_i = \sigma^2 I$*

特殊情况，当各个类别先验相等时，退化为最小距离分类器。

$$W = \mu_i - \mu_j$$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j)$$



# 贝叶斯分类器

## 正态密度的贝叶斯分类器

### 不同高斯参数情况讨论

*Case 2:  $\Sigma_i = \Sigma$*

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

该式分解后:  $x^T \Sigma_i^{-1} x$  各类都相等, 可以忽略

$$g_i(x) = w_i^T x + w_{i0}, \quad w_i = \Sigma^{-1} \mu_i, \quad w_{i0} = -\frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \ln p(\omega_i)$$

# 贝叶斯分类器

## 正态密度的贝叶斯分类器

### 不同高斯参数情况讨论

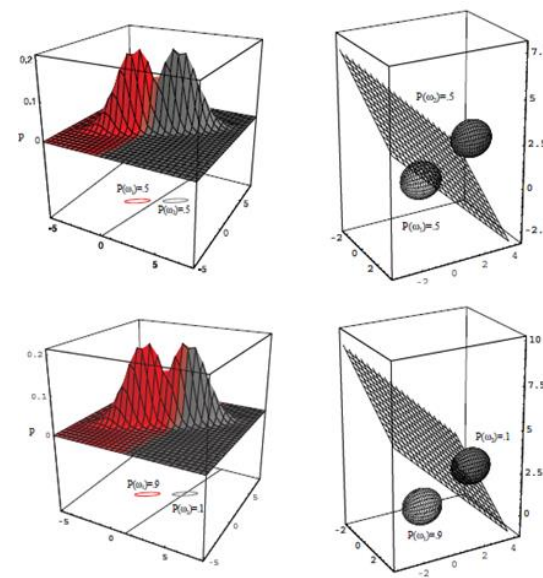
Case 2:  $\Sigma_i = \Sigma$

决策界:  $g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0$

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

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

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



# 贝叶斯分类器

## 正态密度的贝叶斯分类器

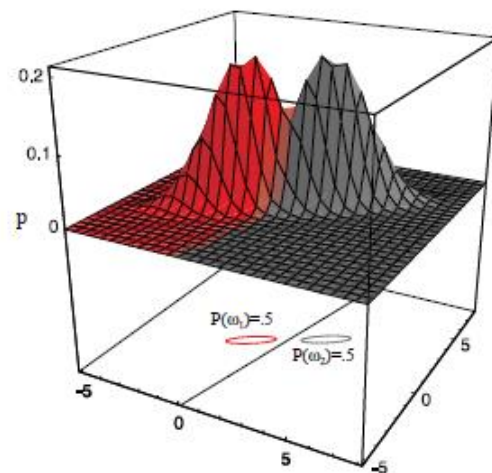
### 不同高斯参数情况讨论

Case 2:  $\Sigma_i = \Sigma$

当各个类别先验相等时,

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

$$x_0 = \frac{1}{2}(\mu_i + \mu_j)$$



# 贝叶斯分类器

## 正态密度的贝叶斯分类器

### 不同高斯参数情况讨论

*Case 3:  $\Sigma_i = \text{arbitrary}$*

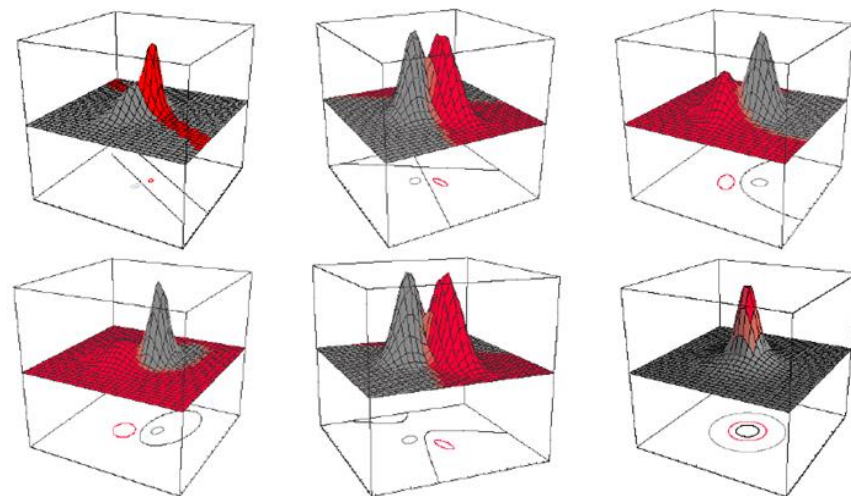
$$g_i(x) = x^T W_i x + w_i^T x + w_{i0}$$

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

$$w_i = \Sigma_i^{-1} \mu_i$$

$$w_{i0} = -\frac{1}{2} \mu_i^T \Sigma_i^{-1} \mu_i - \frac{1}{2} \ln |\Sigma_i| + \ln p(\omega_i)$$

决策界:  $g_i(x) - g_j(x) = 0$ , 情况比较复杂, 可能非线性。



# 贝叶斯分类器

## 正态密度的贝叶斯分类器

例子：

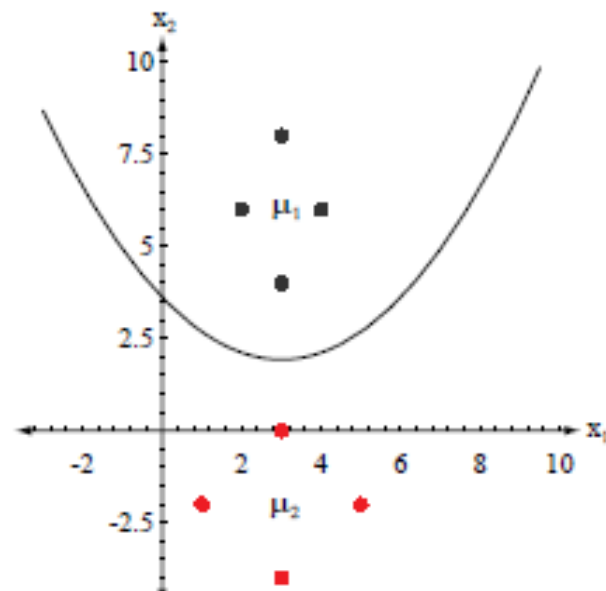
$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

$$\Sigma_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \text{ and } \Sigma_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

$$p(\omega_1) = p(\omega_2) = 0.5$$

决策界：  $g_1(x) \equiv g_2(x)$

$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2$$



*Have a break !*

## 第二章 贝叶斯学习

2.1 概述

2.2 贝叶斯决策论

2.3 贝叶斯分类器

2.4 贝叶斯学习与参数估计问题



# 贝叶斯学习与参数估计问题

## 问题描述

$\mathcal{D}$  — data set

$\mathcal{M}$  — models (or parameters)

The probability of a model  $\mathcal{M}$  given data set  $\mathcal{D}$  is:

$$P(\mathcal{M}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{M})P(\mathcal{M})}{P(\mathcal{D})}$$

$P(\mathcal{D}|\mathcal{M})$  is the *evidence* (or *likelihood*)

$P(\mathcal{M})$  is the *prior* probability of  $\mathcal{M}$

$P(\mathcal{M}|\mathcal{D})$  is the *posterior probability* of  $\mathcal{M}$

$$P(\mathcal{D}) = \int P(\mathcal{D}|\mathcal{M})P(\mathcal{M}) d\mathcal{M}$$

三个基本问题: **Bayes, MAP and ML**

### Bayesian Learning:



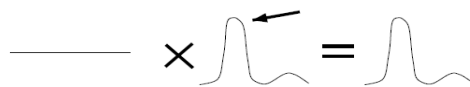
Assumes a prior over the model parameters. Computes the posterior distribution of the parameters:  $P(\theta|\mathcal{D})$ .

### Maximum a Posteriori (MAP) Learning:



Assumes a prior over the model parameters  $P(\theta)$ . Finds a parameter setting that maximises the posterior:  $P(\theta|\mathcal{D}) \propto P(\theta) P(\mathcal{D}|\theta)$ .

### Maximum Likelihood (ML) Learning:



Does not assume a prior over the model parameters. Finds a parameter setting that maximises the likelihood of the data:  $P(\mathcal{D}|\theta)$ .

# 贝叶斯学习与参数估计问题

## 贝叶斯学习

通过观测数据 likelihood 修正模型的先验，得到后验概率分布：

$$p(\theta|\mathcal{D}, \alpha) \propto p(\mathcal{D}|\theta)p(\theta|\alpha)$$

其中， $\alpha$  是超参数，不是估计的参数。

例子 1: Beta 先验分布

# 贝叶斯学习与参数估计问题

## 贝叶斯学习

### 例子 1: Beta 先验分布

- 观察数据

**Coin example:** we have a coin that can be biased

H H T T H H T H T H T T T H T H H H H T H H H H T  
1 1 0 0 1 1 0 1 0 1 0 0 0 1 0 1 1 1 1 0 1 1 1 0

**Data:**  $D$  a sequence of outcomes  $x_i$  such that

- **head**  $x_i = 1$
- **tail**  $x_i = 0$

**Model:** probability of a head  $\theta$   
probability of a tail  $(1 - \theta)$

# 贝叶斯学习与参数估计问题

## 贝叶斯学习

### 例子 1: Beta 先验分布

#### Binary Variables

$$x \in \{0, 1\}$$

$$p(x = 1 | \theta) = \theta$$

$$p(x = 0 | \theta) = 1 - \theta$$

- 贝努力分布(**Bernoulli**):

$$\text{Bern}(x | \theta) = \theta^x (1 - \theta)^{1-x}$$

# 贝叶斯学习与参数估计问题

## 贝叶斯学习

### 例子 1: Beta 先验分布

#### Binary Variables

$$x \in \{0, 1\}$$

$$p(x = 1 | \theta) = \theta$$

$$p(x = 0 | \theta) = 1 - \theta$$

- 贝努力分布(**Bernoulli**):

$$\text{Bern}(x | \theta) = \theta^x (1 - \theta)^{1-x}$$

- **Likelihood** (观察似然) :

H H T T H H T H T H T T T H T H H H H T H H H H T  
1 1 0 0 1 1 0 1 0 1 0 0 0 1 0 1 1 1 1 0 1 1 1 0

$$P(D | \theta) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{(1-x_i)}$$

$$P(D | \theta) = \theta^{N_1} (1 - \theta)^{N_2} \quad (N_1 + N_2 = N)$$

# 贝叶斯学习与参数估计问题

## 贝叶斯学习

### 例子 1: Beta 先验分布

- Binary Variables

$$x \in \{0, 1\}$$

$$p(x = 1 | \theta) = \theta$$

$$p(x = 0 | \theta) = 1 - \theta$$

- 贝努力分布(**Bernoulli**):

$$\text{Bern}(x | \theta) = \theta^x (1 - \theta)^{1-x}$$

- **Likelihood** (观察似然) :

H H T T H H T H T H T T T H T H H H H T H H H H T  
1 1 0 0 1 1 0 1 0 1 0 0 0 1 0 1 1 1 1 0 1 1 1 0

$$P(D | \theta) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{(1-x_i)}$$

$$P(D | \theta) = \theta^{N_1} (1 - \theta)^{N_2} \quad (N_1 + N_2 = N)$$

(对比)二项式分布:

$$\text{Bin}(m | N, \theta) = \binom{N}{m} \theta^m (1 - \theta)^{N-m}$$

# 贝叶斯学习与参数估计问题

## 贝叶斯学习

### 例子 1: Beta 先验分布

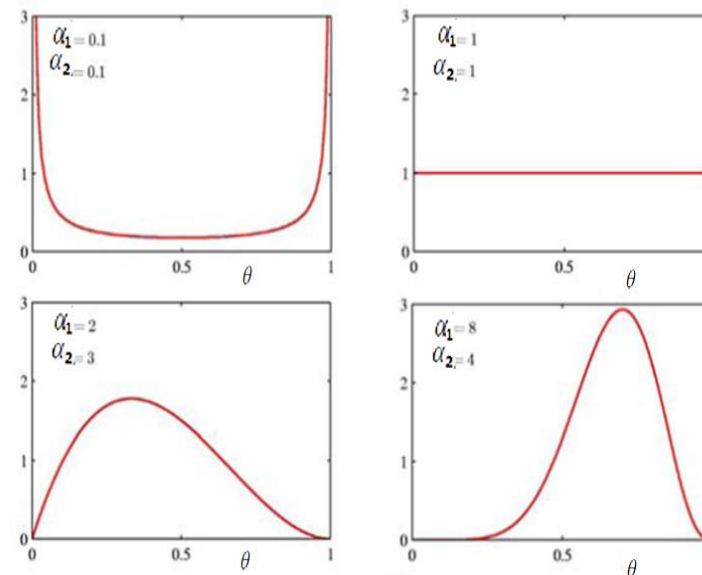
- **Prior**

Choice of prior: **Beta distribution**

$$p(\theta | \alpha) = \text{Beta}(\theta | \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1-1} (1-\theta)^{\alpha_2-1}$$

$\Gamma(x)$  - a Gamma function  $\Gamma(x) = (x-1)\Gamma(x-1)$   
For integer values of  $x$   $\Gamma(n) = (n-1)!$

**Beta distribution**



$$p(\theta | \alpha) = \text{Beta}(\theta | \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1-1} (1-\theta)^{\alpha_2-1}$$

# 贝叶斯学习与参数估计问题

## 贝叶斯学习

### 例子 1: Beta 先验分布

- **Prior**

Choice of prior: **Beta distribution**

$$p(\theta | \alpha) = \text{Beta}(\theta | \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1-1} (1-\theta)^{\alpha_2-1}$$

$\Gamma(x)$  - a Gamma function  $\Gamma(x) = (x-1)\Gamma(x-1)$   
For integer values of  $x$   $\Gamma(n) = (n-1)!$

#### Why to use Beta distribution?

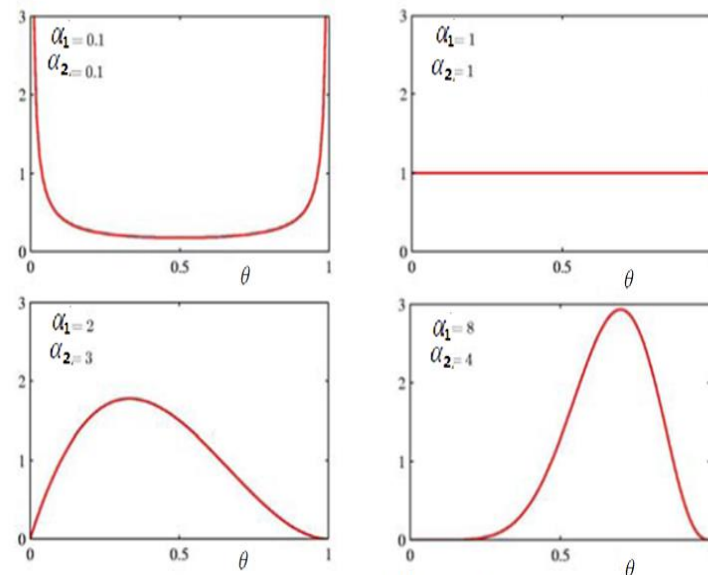
Beta distribution “fits” Bernoulli trials - **conjugate choices**

$$P(D | \theta) = \theta^{N_1} (1-\theta)^{N_2}$$

**Posterior distribution is again a Beta distribution**

$$p(\theta | D, \alpha) = \frac{P(D | \theta) \text{Beta}(\theta | \alpha_1, \alpha_2)}{P(D | \alpha)} = \text{Beta}(\theta | \alpha_1 + N_1, \alpha_2 + N_2)$$

#### Beta distribution



$$p(\theta | \alpha) = \text{Beta}(\theta | \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1-1} (1-\theta)^{\alpha_2-1}$$



# 贝叶斯学习与参数估计问题

## 贝叶斯学习

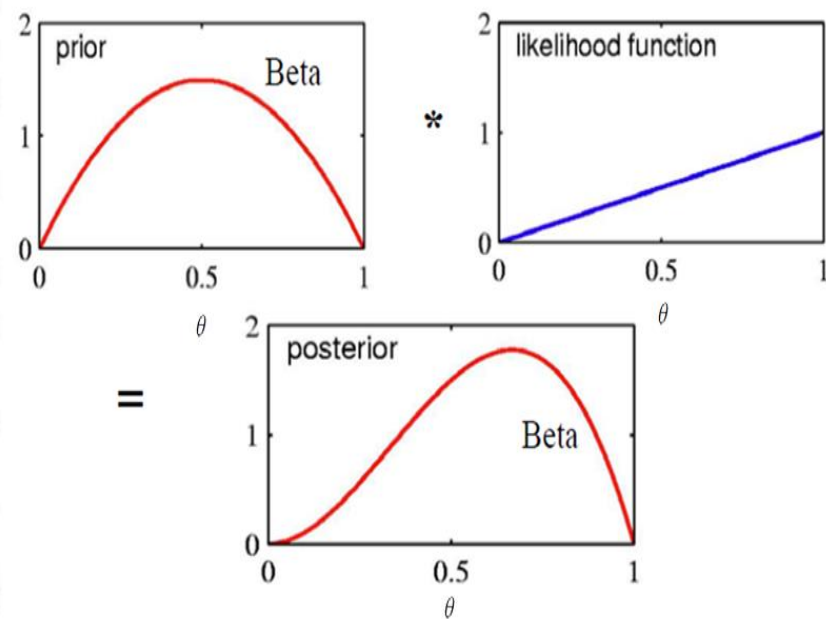
### 例子 1: Beta 先验分布

- **Posterior:**

$$p(\theta | D, \alpha) = \frac{P(D | \theta) \text{Beta}(\theta | \alpha_1, \alpha_2)}{P(D | \alpha)} = \text{Beta}(\theta | \alpha_1 + N_1, \alpha_2 + N_2)$$
$$= \frac{\Gamma(\alpha_1 + \alpha_2 + N_1 + N_2)}{\Gamma(\alpha_1 + N_1) \Gamma(\alpha_2 + N_2)} \theta^{N_1 + \alpha_1 - 1} (1 - \theta)^{N_2 + \alpha_2 - 1}$$

**Notice** that parameters of the prior  
act like counts of heads and tails  
(sometimes they are also referred to as **prior counts**)

### Posterior distribution



$$p(\theta | D, \alpha) = \frac{P(D | \theta) \text{Beta}(\theta | \alpha_1, \alpha_2)}{P(D | \alpha)} = \text{Beta}(\theta | \alpha_1 + N_1, \alpha_2 + N_2)$$

# 贝叶斯学习与参数估计问题

## 极大似然估计

### 问题描述

- 最大化观察数据的概率

$$p(\theta|\mathcal{D}, \alpha) \propto p(\mathcal{D}|\theta)p(\theta|\alpha)$$

← 最大化

似然函数 likelihood:

$$p(\mathcal{D}|\theta) = p(\mathbf{x}_1, \dots, \mathbf{x}_n|\theta) = \prod_{i=1}^n p(\mathbf{x}_i|\theta)$$

Maximum Likelihood

$$\hat{\theta} = \arg \max_{\theta} p(\mathcal{D}|\theta)$$

转化为求 log-likelihood 极大的问题

$$\hat{\theta} = \arg \max_{\theta} \sum_{i=1}^n \log p(\mathbf{x}_i|\theta)$$

求解过程

$$\sum_{i=1}^n \nabla_{\theta} \log p(\mathbf{x}_i|\theta) = 0$$

# 贝叶斯学习与参数估计问题

## 极大似然估计

### 例子 1：二项式分布的 ML

- likelihood

$$p(\mathcal{D}|\theta) = \prod_{n=1}^N p(x_n|\theta) = \prod_{n=1}^N \theta^{x_n} (1 - \theta)^{1-x_n}.$$

- Log-likelihood

$$\ln p(\mathcal{D}|\theta) = \sum_{n=1}^N \ln p(x_n|\theta) = \sum_{n=1}^N \{x_n \ln \theta + (1 - x_n) \ln(1 - \theta)\}$$

- 最优的参数

$$\theta_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n. \quad \theta_{\text{ML}} = \frac{m}{N}$$

# 贝叶斯学习与参数估计问题

## 极大似然估计

### 例子 1 : 二项式分布的 ML

- 实例:
  - Assume the unknown and possibly biased coin
  - Probability of the head is  $\theta$
  - **Data:**  
H H T T H H T H T T T H T H H H H T H H H H T
    - **Heads:** 15
    - **Tails:** 10

What is the ML estimate of the probability of head and tail ?

Likelihood:  $P(D | \theta) = \theta^{N_1} (1 - \theta)^{N_2}$

最优的参数:

$$\text{Head: } \theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2} = \frac{15}{25} = 0.6$$

$$\text{Tail: } (1 - \theta_{ML}) = \frac{N_2}{N} = \frac{N_2}{N_1 + N_2} = \frac{10}{25} = 0.4$$

# 贝叶斯学习与参数估计问题

## 极大似然估计

### 例子 2 : 高斯分布的 ML-估计 $\mu$

Let  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$  be vectors stemmed from a normal distribution with known covariance matrix and unknown mean, that is,

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

- Log-likelihood:

$$L(\boldsymbol{\mu}) \equiv \ln \prod_{k=1}^N p(\mathbf{x}_k; \boldsymbol{\mu}) = -\frac{N}{2} \ln((2\pi)^l |\Sigma|) - \frac{1}{2} \sum_{k=1}^N (\mathbf{x}_k - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

# 贝叶斯学习与参数估计问题

## 极大似然估计

### 例子 2 : 高斯分布的 ML-估计 $\mu$

Taking the gradient with respect to  $\mu$ , we obtain

$$\frac{\partial L(\mu)}{\partial \mu} \equiv \begin{bmatrix} \frac{\partial L}{\partial \mu_1} \\ \frac{\partial L}{\partial \mu_2} \\ \vdots \\ \frac{\partial L}{\partial \mu_d} \end{bmatrix} = \sum_{k=1}^N \Sigma^{-1}(\mathbf{x}_k - \mu) = 0$$

or

$$\hat{\mu}_{ML} = \frac{1}{N} \sum_{k=1}^N \mathbf{x}_k$$

# 贝叶斯学习与参数估计问题

## 极大似然估计

### 例子 3 : 高斯分布的 ML-估计方差

Assume that  $N$  data points,  $x_1, x_2, \dots, x_N$ , have been generated by a one-dimensional Gaussian pdf of known mean,  $\mu$ , but of unknown variance. Derive the ML estimate of the variance.

The log-likelihood function for this case is given by

$$\begin{aligned} L(\sigma^2) &= \ln \prod_{k=1}^N p(x_k; \sigma^2) = \ln \prod_{k=1}^N \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2}} \exp\left(-\frac{(x_k - \mu)^2}{2\sigma^2}\right) \\ &= -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{k=1}^N (x_k - \mu)^2 \end{aligned}$$

Taking the derivative of the above with respect to  $\sigma^2$  and equating to zero, we obtain

$$-\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{k=1}^N (x_k - \mu)^2 = 0 \qquad \hat{\sigma}_{ML}^2 = \frac{1}{N} \sum_{k=1}^N (x_k - \mu)^2$$

# 贝叶斯学习与参数估计问题

## 最大后验估计

### 问题描述

求使后验概率最大的模型或参数( $\theta$ )。

$$p(\theta | \mathcal{D}, \alpha) \propto p(\mathcal{D} | \theta) p(\theta | \alpha)$$

贝叶斯公式中

最大化

$$p(\theta | D, \alpha) = \frac{P(D | \theta) p(\theta | \alpha)}{P(D | \alpha)}$$

$$\hat{\theta}_{MAP} : \frac{\partial}{\partial \theta} p(\theta | D, \alpha) = 0 \quad \text{or} \quad \frac{\partial}{\partial \theta} P(D | \theta) p(\theta | \alpha) = 0$$



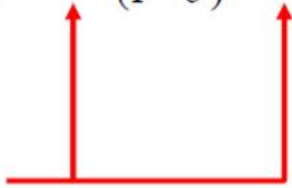
# 贝叶斯学习与参数估计问题

## 最大后验估计

### 例子 1 : Beta 先验分布的 MAP

#### Maximum a posteriori estimate

- Selects the mode of the **posterior distribution**

$$p(\theta | D, \alpha) = \frac{P(D | \theta) \text{Beta}(\theta | \alpha_1, \alpha_2)}{P(D | \alpha)} = \text{Beta}(\theta | \alpha_1 + N_1, \alpha_2 + N_2)$$
$$= \frac{\Gamma(\alpha_1 + \alpha_2 + N_1 + N_2)}{\Gamma(\alpha_1 + N_1) \Gamma(\alpha_2 + N_2)} \theta^{N_1 + \alpha_1 - 1} (1 - \theta)^{N_2 + \alpha_2 - 1}$$


**Notice** that parameters of the prior  
act like counts of heads and tails  
(sometimes they are also referred to as **prior counts**)

**MAP Solution:**

$$\theta_{MAP} = \frac{\alpha_1 + N_1 - 1}{\alpha_1 + \alpha_2 + N_1 + N_2 - 2}$$

# 贝叶斯学习与参数估计问题

## 最大后验估计

### 例子 1 : Beta 先验分布的 MAP

- 实例：
    - Assume the unknown and possibly biased coin
    - Probability of the head is  $\theta$
    - Data:**  
H H T T H H T H T H T T T H T H H H H T H H H H T
      - Heads:** 15
      - Tails:** 10
    - Assume  $p(\theta | \alpha) = \text{Beta}(\theta | 5, 5)$
- What is the MAP estimate ?

$$\theta_{MAP} = \frac{N_1 + \alpha_1 - 1}{N - 2} = \frac{N_1 + \alpha_1 - 1}{N_1 + N_2 + \alpha_1 + \alpha_2 - 2} = \frac{19}{33}$$

与 ML 比较:  $\theta_{ML} = 15/25 = 0.6$ ,  $\theta_{MAP} = 19/33 = 0.5758$

# 贝叶斯学习与参数估计问题

## 最大后验估计

### 例子 2 : 高斯分布的 MAP-估计 u

Let  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$  be vectors stemmed from a normal distribution with known covariance matrix and unknown mean, that is,

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

$$p(\boldsymbol{\mu}) = \frac{1}{(2\pi)^{J/2} \sigma_\mu^J} \exp\left(-\frac{1}{2} \frac{\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|^2}{\sigma_\mu^2}\right)$$

The MAP estimate is given by the solution of

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln\left(\prod_{k=1}^N p(\mathbf{x}_k | \boldsymbol{\mu}) p(\boldsymbol{\mu})\right) = 0$$

or, for  $\Sigma = \sigma^2 I$ ,

$$\sum_{k=1}^N \frac{1}{\sigma^2} (\mathbf{x}_k - \hat{\boldsymbol{\mu}}) - \frac{1}{\sigma_\mu^2} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)$$

$$\hat{\boldsymbol{\mu}}_{MAP} = \frac{\boldsymbol{\mu}_0 + \frac{\sigma_\mu^2}{\sigma^2} \sum_{k=1}^N \mathbf{x}_k}{1 + \frac{\sigma_\mu^2}{\sigma^2} N}$$

## 小 结

1. beyas 决策准则
2. 几种贝叶斯分类器
3. 贝叶斯学习与参数估计问题
  - Beyas Learning
  - M L 参数估计
  - M A P 参数估计

# 本讲参考文献

1. 周志华，机器学习，清华大学出版社，2016.
2. Duda, R.O. et al. Pattern classification. 2nd, 2003.
2. 边肇祺，张学工等编著，模式识别(第二版)，清华大学，1999。