

# 贝叶斯决策、参数估计

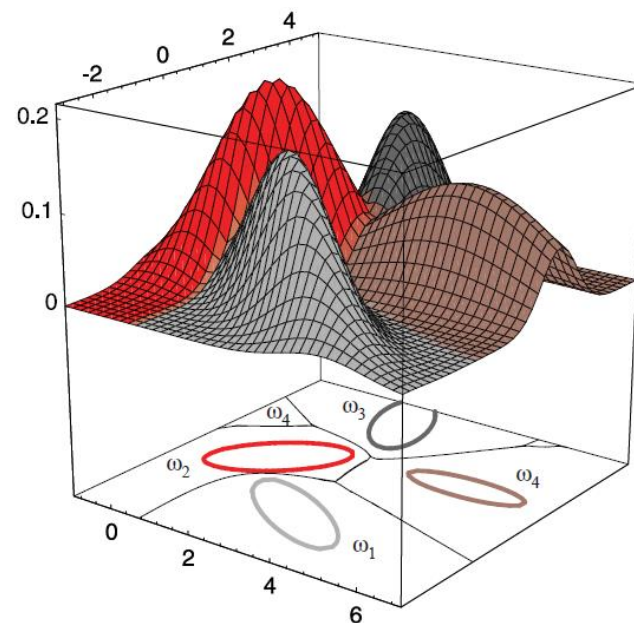
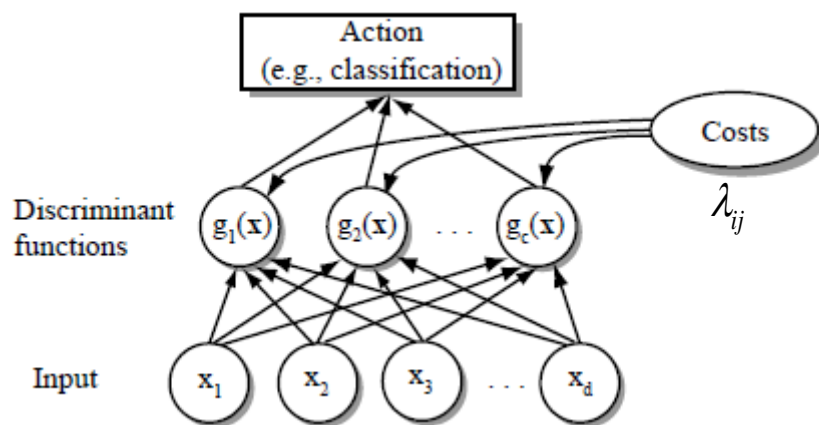
刘成林(liucl@nlpr.ia.ac.cn)

2022年9月14日

助教：张明亮(zhangmingliang2018@ia.ac.cn)  
程 真(chengzhen2019@nlpr.ia.ac.cn)  
张 姣(zhangjiao2019@ia.ac.cn)

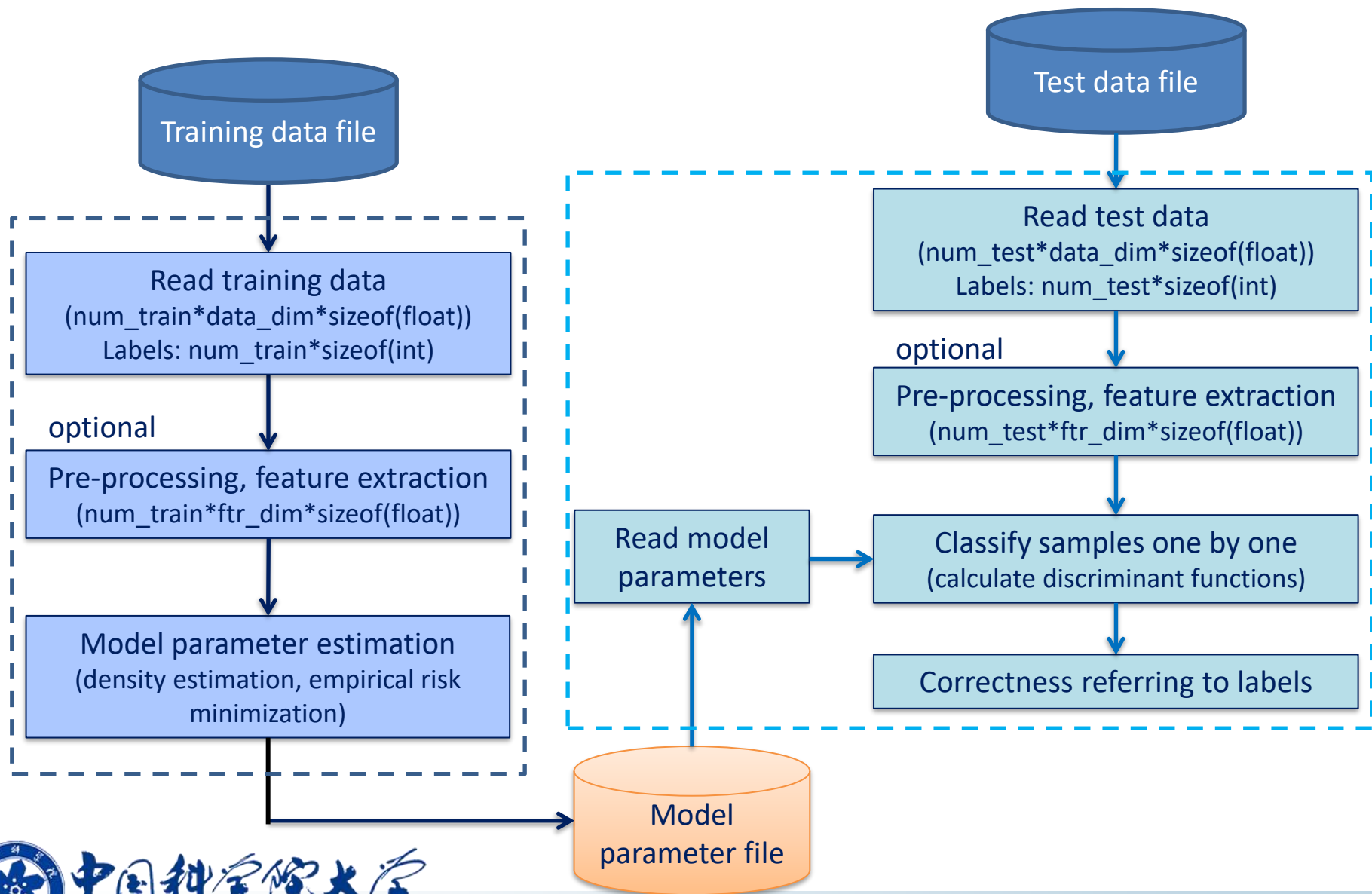
# 统计模式分类的基本框架

- 特征空间划分
  - 判别函数(Discriminant function)、决策面(Decision surface)
  - 生成模型(Generative model):  $\mathbf{x} \rightarrow p(\mathbf{x} | \omega_i) \rightarrow g_i(\mathbf{x})$
  - 判别模型(Discriminative model):  $\mathbf{x} \rightarrow g_i(\mathbf{x})$



# 如何编程实现？

两个程序，一个训练，一个测试。  
三个文件，存在硬盘里。模型文件是训练后生成的。



# 上次课主要内容回顾

- 贝叶斯决策
  - 最小风险决策
  - (0-1 loss)最小错误率决策（最大后验概率决策）
- 高斯概率密度（正态分布）
  - 1D, 多维（记住了？）
  - 协方差矩阵特性
    - 等密度点轨迹、马氏距离、特征值分解、正交化
  - 线性变换的高斯密度？
- 高斯密度下的判别函数
  - Quadratic discriminant function (QDF)
  - Three cases, linear discriminant function (LDF)
- 贝叶斯决策的错误率

# 提 纲

- 第2章
  - 离散变量的贝叶斯决策
  - 复合模式分类
    - 类似问题：多分类器融合
- 第3章
  - 导论：关于参数估计
  - 最大似然参数估计
  - 贝叶斯估计
  - 贝叶斯估计：高斯密度的情况
  - 贝叶斯估计：一般情况

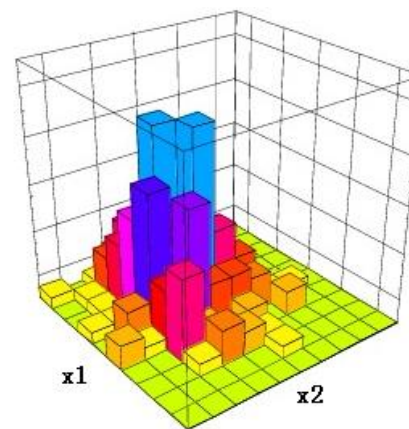
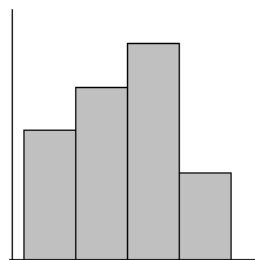
# 离散变量贝叶斯决策

- 贝叶斯决策

- 最小风险:  $\min R(\alpha_i|\mathbf{x}) = \sum_{j=1}^c \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x})$
- 最小错误率(MAP):  $\max P(\omega_j|\mathbf{x})$

- 离散特征变量

- 例如：问卷调查，每个问题2个或多个选项；  
医疗诊断：是否有某个症状
- 概率密度函数  $p(\mathbf{x}|\omega_i)=p(x_1x_2\cdots x_d|\omega_i)$   
(非参数、直方图表示)



- 独立二值特征(Binary features)

- 独立

$$p(\mathbf{x}) = p(x_1 x_2 \cdots x_d) = \prod_{i=1}^d p(x_i)$$

- Binary, 概率密度: d个参数

$$p_i = \text{Prob}(x_i=1|\omega_1) \quad i = 1, \dots, d$$

- 2-class  $q_i = \text{Prob}(x_i=1|\omega_2) \quad i = 1, \dots, d$

$$P(\mathbf{x}|\omega_1) = \prod_{i=1}^d p_i^{x_i} (1 - p_i)^{1-x_i}$$

$$P(\mathbf{x}|\omega_2) = \prod_{i=1}^d q_i^{x_i} (1 - q_i)^{1-x_i}$$

- Likelihood ratio

$$\frac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)} = \prod_{i=1}^d \left( \frac{p_i}{q_i} \right)^{x_i} \left( \frac{1 - p_i}{1 - q_i} \right)^{1-x_i}$$

- 独立二值特征(Binary features)

- Discriminant/decision function

$$g(\mathbf{x}) = \log \frac{p(\mathbf{x} | \omega_1) P(\omega_1)}{p(\mathbf{x} | \omega_2) P(\omega_2)} = \sum_{i=1}^d \left[ x_i \ln \frac{p_i}{q_i} + (1 - x_i) \ln \frac{1 - p_i}{1 - q_i} \right] + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

- 为线性判别函数

$$g(\mathbf{x}) = \sum_{i=1}^d w_i x_i + w_0$$

$w_i$  表征每个特征的判别性

$$w_i = \ln \frac{p_i(1 - q_i)}{q_i(1 - p_i)} \quad i = 1, \dots, d$$

跟  $p_i, q_i$  什么关系?

$$w_0 = \sum_{i=1}^d \ln \frac{1 - p_i}{1 - q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$



- 一个例子: 3D binary data

- $P(\omega_1)=0.5, P(\omega_2)=0.5$

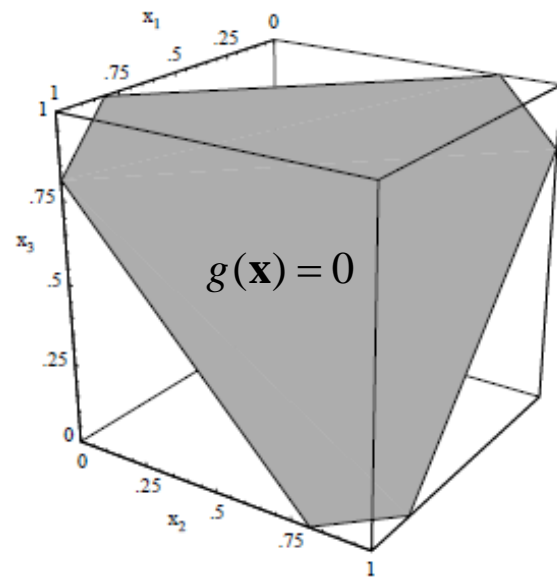
- $p_i=0.8, q_i=0.5, i=1,2,3$

$$P(\mathbf{x}|\omega_1) = \prod_{i=1}^d p_i^{x_i} (1 - p_i)^{1-x_i} \quad P(\mathbf{x}|\omega_2) = \prod_{i=1}^d q_i^{x_i} (1 - q_i)^{1-x_i}$$

$$g(\mathbf{x}) = \sum_{i=1}^d w_i x_i + w_0$$

$$w_i = \ln \frac{.8(1 - .5)}{.5(1 - .8)} = 1.3863$$

$$w_0 = \sum_{i=1}^3 \ln \frac{1 - .8}{1 - .5} + \ln \frac{.5}{.5} = -2.7489$$

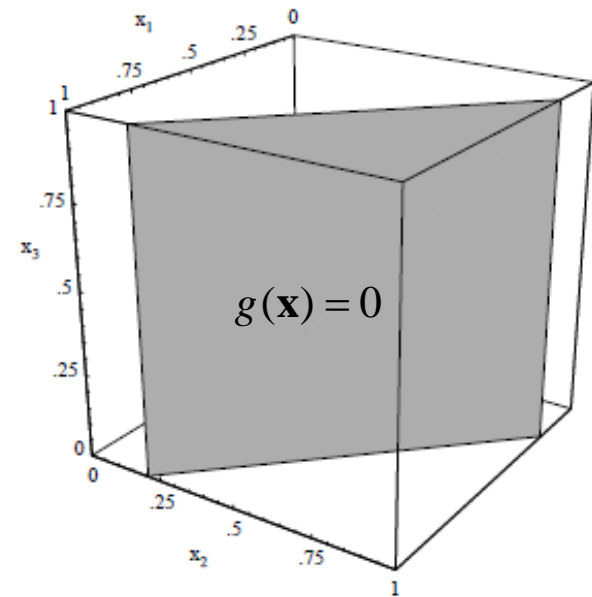


- 另一个例子: 3D binary data
  - $P(\omega_1)=0.5, P(\omega_2)=0.5$
  - $p_1=p_2=0.8, p_3=0.5; q_i=0.5, i=1,2,3$

$$w_1 = w_2 = \ln \frac{.8(1 - .5)}{.5(1 - .8)} = 1.3863$$

$$w_3 = 0$$

$$w_0 = 2 \ln \frac{1-0.8}{1-0.5} = -1.8326$$



# 复合模式分类

(\*2.12 Compound Bayesian Decision Theory and Context)

- 多个模式同时分类  $\mathbf{X} = \mathbf{x}_1 \mathbf{x}_2 \cdots \mathbf{x}_n$   $\omega = \omega(1)\omega(2)\cdots\omega(n)$

- 比如：字符串识别

tomorrow

- Bayesian decision

$$P(\omega|X) = \frac{p(X|\omega)P(\omega)}{p(X)} = \frac{p(X|\omega)P(\omega)}{\sum_{\omega} p(X|\omega)P(\omega)}$$

- 注意： $\omega$ 类别数巨大， $p(X|\omega)$ 存储和估计困难

- Conditionally independent

$$p(X|\omega) = \prod_{i=1}^n p(\mathbf{x}_i|\omega(i))$$

- Prior assumption

- Markov chain

$$P[\omega(1)\omega(2)\cdots\omega(n)] = P[\omega(1)] \prod_{j=2}^n P[\omega(j) | \omega(j-1)]$$

- Hidden Markov model (Chapter 3)

# 与复合模式识别类似的问题：多分类器融合

- 多个分类器的决策当作多维特征，Bayes方法重新分类

- 一个分类器的输出：离散变量  $e_k \in \{\omega_1, \dots, \omega_c\}$

联合输出空间(又称为Behavior knowledge space)的后验概率

$$P(\omega_i | e_1, \dots, e_K) = \frac{P(e_1, \dots, e_K | \omega_i) P(\omega_i)}{P(e_1, \dots, e_K)}, \quad i = 1, \dots, c$$

- 在验证(validation)数据集上估计离散空间的条件概率密度

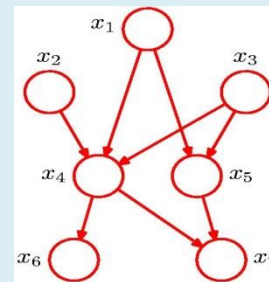
$P(e_1, \dots, e_K | \omega_i)$  指数级复杂度，需要大量样本

- Naïve Bayes

$$P(e_1 = \omega_{j_1}, \dots, e_K = \omega_{j_K} | \omega_i) = \prod_{k=1}^K P(e_k = \omega_{j_k} | \omega_i)$$

- Dependency tree approximation

$$P(e_1, \dots, e_K | \omega_i) = \prod_{k=1}^K P\{e_k | \omega_i, Pa(e_k)\}$$



多分类器融合：也可以分类器输出的判别函数值或排序作为重新分类的特征，设计新分类器(Meta-classifier)

# 第3章

## 最大似然和贝叶斯参数估计

# 关于参数估计

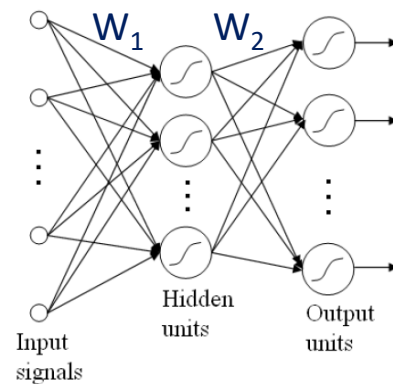
- 分类器设计

- 给定分类器结构/函数形式，从训练样本估计参数
- Statistical generative: density estimation
  - 参数法  $p(\mathbf{x}|\omega_i, \theta_i)$ , e.g.,  $N(\mu_i, \Sigma_i)$
- Statistical discriminative: discriminant function, e.g., neural network

$$g_i(\mathbf{x}) = f(\mathbf{x}, W_1, W_{2,i})$$

- 统计生成模型的参数估计

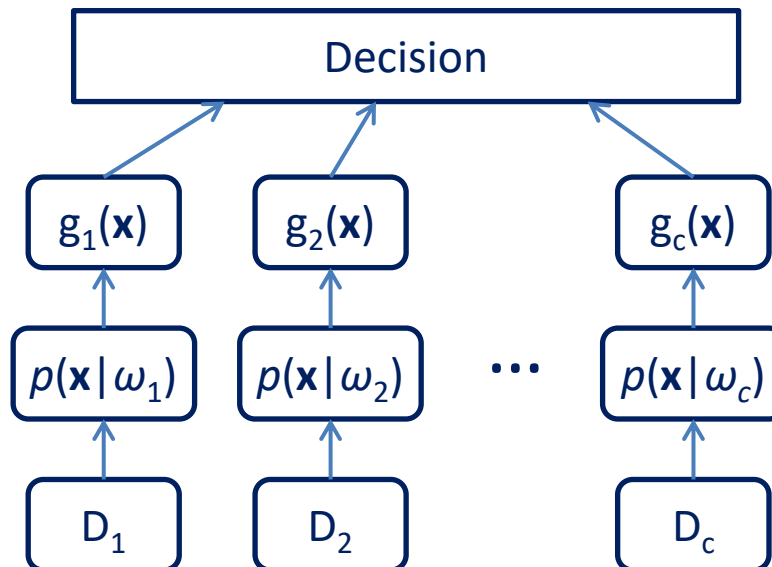
- Maximum likelihood (ML)
  - 假设参数为确定值，最优估计：似然度最大
- Bayesian estimation (Bayesian learning)
  - 假设参数为随机变量，估计其分布



# 最大似然估计

- 基本原理

- 假设概率密度函数  $p(\mathbf{x}|\omega_i, \theta_i)$ ,  $\theta_i$  to be estimated
- 样本数据  $D_1, \dots, D_c$ 
  - Samples in  $D_i$  assumed to be independent and identically distributed (*i.i.d.*)
  - $D_i$  used to estimate  $\theta_i$  disregarding the parameters of other classes



- ML估计一类模型的参数

- Likelihood

$$p(\mathcal{D}|\theta) = \prod_{k=1}^n p(\mathbf{x}_k|\theta)$$

- Maximization

$$\max_{\theta} p(D|\theta) \leftrightarrow \nabla_{\theta} p(D|\theta) = 0$$

- Gradient: vector in **parameter space**

Parameter space (p-D) versus feature space (d-D)

$$\nabla_{\theta} \equiv \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{bmatrix}$$

- $\nabla_{\theta} p(D|\theta) = 0$  的解：可能有解析解，也可能需要迭代求解（如梯度下降）

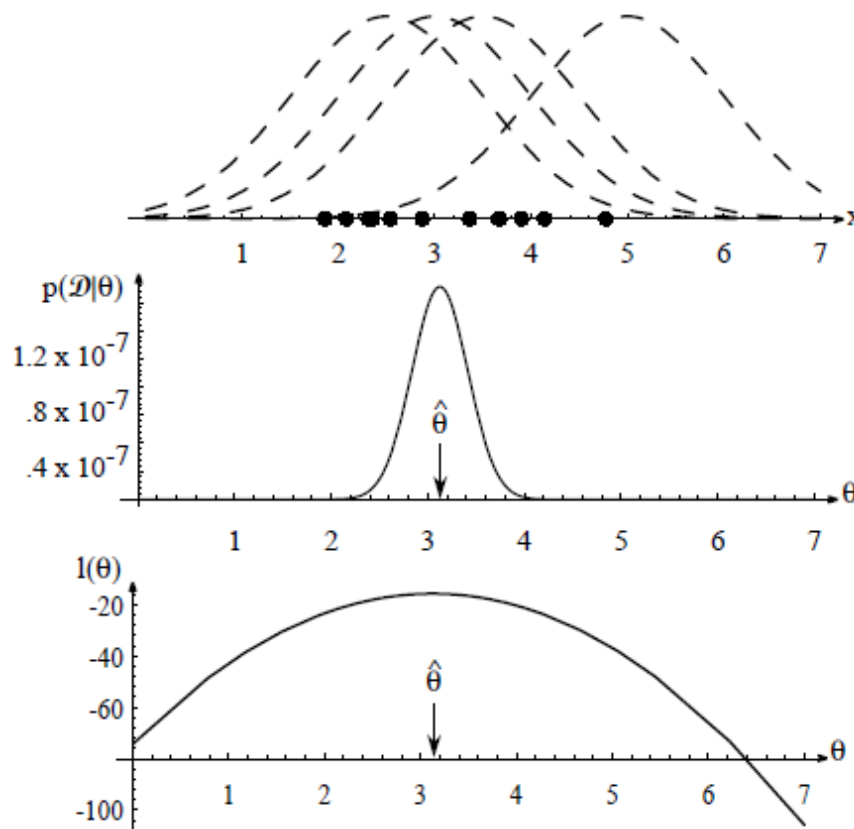


- ML参数估计：一个例子
  - 1D高斯密度，假设 $\sigma^2$ 已知， $\mu$ 未知

10个样本点，  
4个假设的高斯密度函数  
(Likelihood不同)

Likelihood:  $\mu$ 的函数

Log-likelihood



- Log-likelihood比较容易计算

$$l(\theta) \equiv \ln p(\mathcal{D}|\theta) \quad l(\theta) = \sum_{k=1}^n \ln p(\mathbf{x}_k|\theta)$$

- ML estimate

$$\hat{\theta} = \arg \max_{\theta} l(\theta)$$

$$\nabla_{\theta} l = \sum_{k=1}^n \nabla_{\theta} \ln p(\mathbf{x}_k|\theta) = 0$$

$$\frac{\partial l}{\partial \theta_j} = 0, \quad j = 1, \dots, p$$

- Maximum a posteriori (MAP) estimator

$$\max_{\theta} l(\theta) p(\theta)$$

- Equivalent to ML when  $p(\theta)$  is uniform
- Relation to Bayesian estimation?

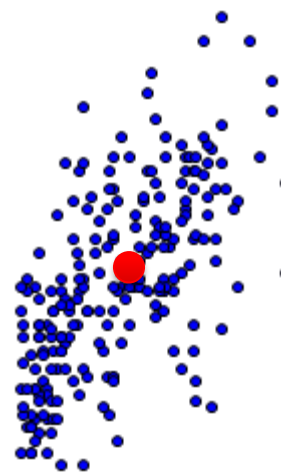
- Gaussian case: unknown  $\mu$ 
  - Log-likelihood of a single point

$$\ln p(\mathbf{x}_k | \mu) = -\frac{1}{2} \ln [(2\pi)^d |\Sigma|] - \frac{1}{2} (\mathbf{x}_k - \mu)^t \Sigma^{-1} (\mathbf{x}_k - \mu)$$

$$\nabla_{\theta} \ln p(\mathbf{x}_k | \mu) = \Sigma^{-1} (\mathbf{x}_k - \mu)$$

- ML solution: sample mean

$$\begin{aligned} \nabla_{\theta} l(\theta) = 0 &\Rightarrow \sum_{k=1}^n \Sigma^{-1} (\mathbf{x}_k - \hat{\mu}) = 0 \\ &\Rightarrow \hat{\mu} = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k \end{aligned}$$



- Gaussian case: unknown  $\mu$  and  $\Sigma$

- 1D case,  $\theta_1 = \mu$  and  $\theta_2 = \sigma^2$

- Log-likelihood

$$\ln p(x_k|\boldsymbol{\theta}) = -\frac{1}{2} \ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

- ML solution

$$\nabla_{\theta} l = \nabla_{\theta} \ln p(x_k | \theta) = \begin{bmatrix} -\frac{1}{\theta_2} (x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

$$\nabla_{\theta} l(\theta) = 0 \Rightarrow \sum_{k=1}^n \frac{1}{\hat{\theta}_2} (x_k - \hat{\theta}_1) = 0 \Rightarrow \hat{\mu} = \frac{1}{n} \sum_{k=1}^n x_k$$

$$-\sum_{k=1}^n \frac{1}{\hat{\theta}_2} + \sum_{k=1}^n \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \hat{\mu})^2$$

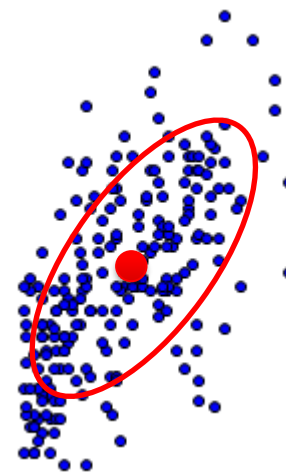
- Gaussian case: unknown  $\mu$  and  $\Sigma$ 
    - Multivariate case (Problem 6, Chapter 3)
- 记住结论即可

$$\nabla_{\theta} l = \sum_{k=1}^n \nabla_{\theta} \ln p(\mathbf{x}_k | \theta) = 0$$



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

$$\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^n (\mathbf{x}_k - \hat{\mu})(\mathbf{x}_k - \hat{\mu})^t$$



- ML estimate of variance/covariance is biased

$$\mathcal{E} \left[ \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right] = \frac{n-1}{n} \sigma^2 \neq \sigma^2$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- Unbiased estimate (sample covariance matrix)

$$\mathcal{E} \left[ \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right] = \sigma^2$$

$$\mathbf{C} = \frac{1}{n-1} \sum_{k=1}^n (\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t$$

- 不能说哪个对或错，实际使用中几乎没有区别

# Break

# 贝叶斯参数估计

- 贝叶斯估计
  - 参数被视为随机变量，估计其后验分布
  - 模型使用：MAP, sampled models combination
- Posterior probability from class-conditional densities

$$P(\omega_i|\mathbf{x}, \mathcal{D}) = \frac{p(\mathbf{x}|\omega_i, \mathcal{D})P(\omega_i|\mathcal{D})}{\sum_{j=1}^c p(\mathbf{x}|\omega_j, \mathcal{D})P(\omega_j|\mathcal{D})}$$

- Prior probabilities assumed known

$$P(\omega_i|\mathbf{x}, \mathcal{D}) = \frac{p(\mathbf{x}|\omega_i, \mathcal{D}_i)P(\omega_i)}{\sum_{j=1}^c p(\mathbf{x}|\omega_j, \mathcal{D}_j)P(\omega_j)}$$

- 用一类的数据 $\mathcal{D}_i$  估计参数 $\theta_i$  的分布



- Parameter distribution

- Assume known density function  $p(\mathbf{x}|\boldsymbol{\theta})$ , known prior density  $p(\boldsymbol{\theta})$
- To estimate posterior density  $p(\boldsymbol{\theta}|\mathcal{D})$  干什么用?
- Estimated density

$$\begin{aligned} p(\mathbf{x}|\mathcal{D}) &= \int p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \\ &= \int p(\mathbf{x}|\boldsymbol{\theta}) \underline{p(\boldsymbol{\theta}|\mathcal{D})} d\boldsymbol{\theta} \end{aligned}$$

- Model usage

- Model average (weighting density functions)

$$p(\mathbf{x}|\mathcal{D}) \propto \frac{1}{M} \sum_{i=1}^M p(\mathbf{x}|\theta_i) \quad \begin{array}{l} \text{Sampling} \\ \text{parameter} \end{array} \quad \theta_i \sim p(\boldsymbol{\theta}|\mathcal{D})$$

- If  $p(\boldsymbol{\theta}|\mathcal{D})$  peaks sharply, MAP  $p(\mathbf{x}|\mathcal{D}) \simeq p(\mathbf{x}|\hat{\boldsymbol{\theta}})$

# 高斯密度贝叶斯估计

- 1D case: to estimate  $p(\mu | D)$ 
  - Density function  $p(x|\mu) \sim N(\mu, \sigma^2)$  Assume known  $\sigma^2$
  - Assume prior density  $p(\mu) \sim N(\mu_0, \sigma_0^2)$
  - Posterior density

$$\begin{aligned} p(\mu | D) &= \frac{p(D|\mu)p(\mu)}{\int p(D|\mu)p(\mu) d\mu} \\ &= \alpha \prod_{k=1}^n p(x_k|\mu)p(\mu) \end{aligned}$$

$p(D | \mu) = \prod_{k=1}^n p(x_k | \mu)$

$\alpha$ : normalization factor

$$\begin{aligned}
p(\mu|\mathcal{D}) &= \alpha \prod_{k=1}^n \overbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp \left[ -\frac{1}{2} \left( \frac{x_k - \mu}{\sigma} \right)^2 \right]}^{p(x_k|\mu)} \overbrace{\frac{1}{\sqrt{2\pi}\sigma_0} \exp \left[ -\frac{1}{2} \left( \frac{\mu - \mu_0}{\sigma_0} \right)^2 \right]}^{p(\mu)} \\
&= \alpha' \exp \left[ -\frac{1}{2} \left( \sum_{k=1}^n \left( \frac{\mu - x_k}{\sigma} \right)^2 + \left( \frac{\mu - \mu_0}{\sigma_0} \right)^2 \right) \right] \\
&= \alpha'' \exp \left[ -\frac{1}{2} \left[ \left( \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) \mu^2 - 2 \left( \frac{1}{\sigma^2} \sum_{k=1}^n x_k + \frac{\mu_0}{\sigma_0^2} \right) \mu \right] \right]
\end{aligned}$$

$p(\mu|\mathcal{D})$ 仍为正态分布！  $p(\mu)$ : conjugate prior

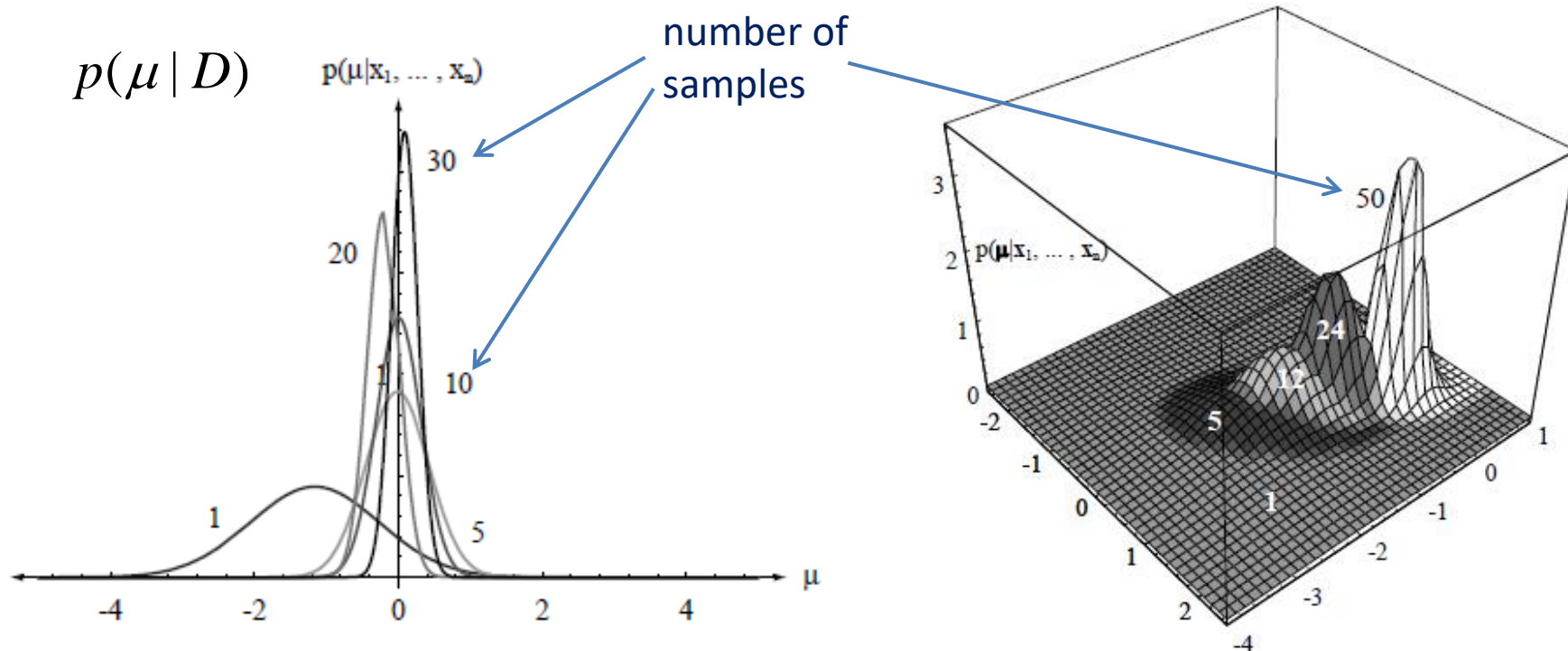
对照  $p(\mu|\mathcal{D}) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp \left[ -\frac{1}{2} \left( \frac{\mu - \mu_n}{\sigma_n} \right)^2 \right]$

$$\begin{aligned}
\frac{1}{\sigma_n^2} &= \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} & \frac{\mu_n}{\sigma_n^2} &= \frac{n}{\sigma^2} \hat{\mu}_n + \frac{\mu_0}{\sigma_0^2} \leftarrow \hat{\mu}_n = \frac{1}{n} \sum_{k=1}^n x_k \\
\downarrow & & \downarrow & \\
\sigma_n^2 &= \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2} \longrightarrow \mu_n = \left( \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \right) \hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0
\end{aligned}$$

当 $n$ 增大， $\mu_n$ 趋近 $\hat{\mu}_n$ ， $\sigma_n^2$ 趋近 $\sigma^2/n$



## 例子：Bayesian learning in 1D/2D space



$$\mu_n = \left( \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \right) \hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0$$

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

- 1D case: class-conditional density

$$\begin{aligned}
 p(x|\mathcal{D}) &= \int p(x|\mu)p(\mu|\mathcal{D}) d\mu \\
 &= \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] d\mu \\
 &= \frac{1}{2\pi\sigma\sigma_n} \exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2 + \sigma_n^2}\right] f(\sigma, \sigma_n),
 \end{aligned}$$

where  $f(\sigma, \sigma_n) = \int \exp\left[-\frac{1}{2}\frac{\sigma^2 + \sigma_n^2}{\sigma^2\sigma_n^2}\left(\mu - \frac{\sigma_n^2 x + \sigma^2 \mu_n}{\sigma^2 + \sigma_n^2}\right)^2\right] d\mu$

(概率密度函数的积分，结果为常数)

- Bayesian estimation

$$p(x|\mathcal{D}) \sim N(\mu_n, \sigma^2 + \sigma_n^2)$$

- C.f. ML estimation

$$p(x|\mathcal{D}) = N(\hat{\mu}_n, \sigma^2)$$

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right) \hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0$$

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

- Multivariate case, with  $\Sigma$  known

$$p(\mathbf{x}|\mu) \sim N(\mu, \Sigma) \quad \text{and} \quad p(\mu) \sim N(\mu_0, \Sigma_0) \quad \text{注意：不同空间！}$$

- Parameter posterior distribution

$$\begin{aligned} p(\mu|\mathcal{D}) &= \alpha \prod_{k=1}^n p(\mathbf{x}_k|\mu)p(\mu) \\ &= \alpha' \exp \left[ -\frac{1}{2} \left( \mu^t (n\Sigma^{-1} + \Sigma_0^{-1}) \mu - 2\mu^t \left( \Sigma^{-1} \sum_{k=1}^n \mathbf{x}_k + \Sigma_0^{-1} \mu_0 \right) \right) \right] \\ &= \alpha'' \exp \left[ -\frac{1}{2} (\mu - \mu_n)^t \Sigma_n^{-1} (\mu - \mu_n) \right] \sim N(\mu_n, \Sigma_n) \end{aligned}$$

$$\begin{aligned} \Sigma_n^{-1} &= n\Sigma^{-1} + \Sigma_0^{-1} & \Sigma_n^{-1} \mu_n &= n\Sigma^{-1} \hat{\mu}_n + \Sigma_0^{-1} \mu_0 & \hat{\mu}_n &= \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k \\ \Sigma_n &= \Sigma_0 \left( \Sigma_0 + \frac{1}{n} \Sigma \right)^{-1} \frac{1}{n} \Sigma & \mu_n &= \Sigma_0 \left( \Sigma_0 + \frac{1}{n} \Sigma \right)^{-1} \hat{\mu}_n + \frac{1}{n} \Sigma \left( \Sigma_0 + \frac{1}{n} \Sigma \right)^{-1} \mu_0 \end{aligned}$$

- Data (feature) posterior distribution

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\mu)p(\mu|\mathcal{D}) d\mu \sim N(\mu_n, \Sigma + \Sigma_n)$$

# 贝叶斯估计：一般情况

- 基本条件

- Known density function  $p(\mathbf{x}|\boldsymbol{\theta})$  with unknown parameters
- Prior parameter distribution  $p(\boldsymbol{\theta})$
- Dataset  $D$  of  $n$  samples independently drawn according to  $p(\mathbf{x})$

- Steps

- Posterior parameter distribution

$$p(\boldsymbol{\theta}|D) = \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \quad p(D|\boldsymbol{\theta}) = \prod_{k=1}^n p(\mathbf{x}_k|\boldsymbol{\theta})$$

- Posterior data distribution

$$p(\mathbf{x}|D) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|D) d\boldsymbol{\theta}$$

- Model usage: parameter sampling or MAP

If  $p(\boldsymbol{\theta}|D)$  peaks at  $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ ,  $p(\mathbf{x}|D)$  will be approximately  $p(\mathbf{x}|\hat{\boldsymbol{\theta}})$



- Recursive Bayes Learning

- Incremental data  $\mathcal{D}^n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$

$$p(\mathcal{D}^n | \theta) = p(\mathbf{x}_n | \theta) p(\mathcal{D}^{n-1} | \theta)$$

$$p(\mathcal{D}^n, \theta) = p(\mathbf{x}_n | \theta) p(\mathcal{D}^{n-1}, \theta) = p(\mathbf{x}_n | \theta) \underline{p(\theta | \mathcal{D}^{n-1})} p(\mathcal{D}^{n-1})$$

- Recursive update of posterior parameter density

$$p(\theta | \mathcal{D}) = \frac{p(\mathcal{D} | \theta) p(\theta)}{\int p(\mathcal{D} | \theta) p(\theta) d\theta} = \frac{p(\mathcal{D}, \theta)}{\int p(\mathcal{D}, \theta) d\theta} \longrightarrow p(\theta | \mathcal{D}^n) = \frac{p(\mathbf{x}_n | \theta) p(\theta | \mathcal{D}^{n-1})}{\int p(\mathbf{x}_n | \theta) p(\theta | \mathcal{D}^{n-1}) d\theta}$$

迭代更新

$p(\theta | \mathcal{D}^0) = p(\theta)$

$p(\mathcal{D}^{n-1})$   
cancelled out

- Need to retain all samples  $1 \dots n-1$ ?

- **Sufficient statistics:** contain all needed information for parameter.

e.g., in Gaussian case

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

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



- Recursive Bayes: An example

- Parametric density: uniform distribution

$$p(x|\theta) \sim U(0, \theta) = \begin{cases} 1/\theta & 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

- Parameter prior  $p(\theta|\mathcal{D}^0) = p(\theta) = U(0, 10)$

- Data samples  $\mathcal{D} = \{4, 7, 2, 8\}$

- Recursive

$$p(\theta|\mathcal{D}^1) \propto p(x|\theta)p(\theta|\mathcal{D}^0) = \begin{cases} 1/\theta & \text{for } 4 \leq \theta \leq 10 \\ 0 & \text{otherwise,} \end{cases} \quad \theta \geq x!$$

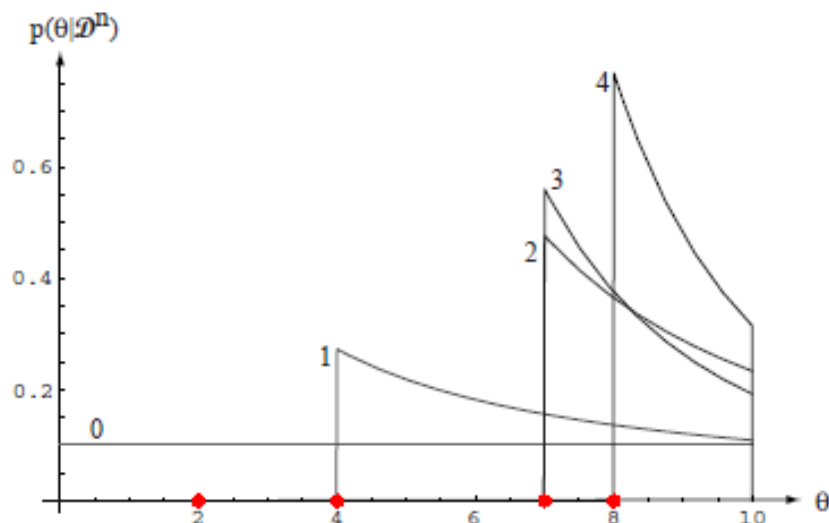
$$p(\theta|\mathcal{D}^2) \propto p(x|\theta)p(\theta|\mathcal{D}^1) = \begin{cases} 1/\theta^2 & \text{for } 7 \leq \theta \leq 10 \\ 0 & \text{otherwise,} \end{cases} \quad \begin{matrix} n=3? \\ n=4? \end{matrix}$$

$$p(\theta|\mathcal{D}^3) \propto p(x|\theta)p(\theta|\mathcal{D}^2) = \begin{cases} 1/\theta^3 & \text{for } 7 \leq \theta \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

$$p(\theta|\mathcal{D}^4) \propto p(x|\theta)p(\theta|\mathcal{D}^3) = \begin{cases} 1/\theta^4 & \text{for } 8 \leq \theta \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

## – 分布函数图

- Parameter distribution vs feature distribution

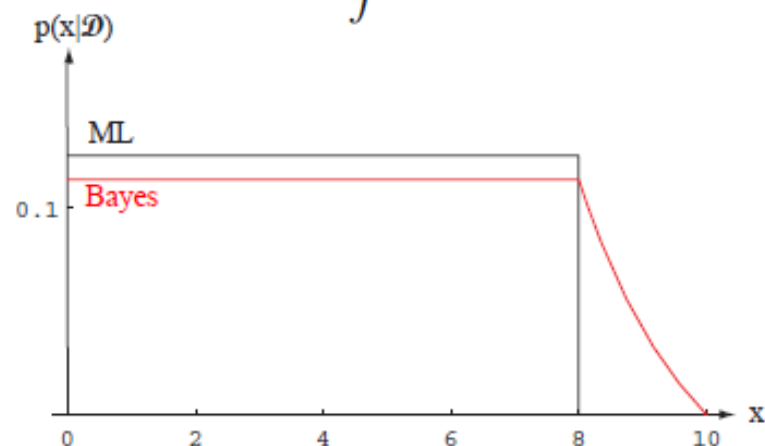


$$p(\theta | D^4) \propto \begin{cases} 1/\theta^4 & \text{for } 8 \leq \theta \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

$$p(x | D^4) \propto \int p(x | \theta) p(\theta | D^4) d\theta = \begin{cases} 8^{-4} - 10^{-4}, & x \leq 8 \\ f(x), & 8 < x \leq 10 \\ 0, & \text{otherwise} \end{cases}$$

$$f(x) \propto \int_x^{10} \frac{1}{\theta} \cdot \frac{1}{\theta^4} d\theta \propto x^{-4} - 10^{-4}$$

$$p(x | D) = \int p(x | \theta) p(\theta | D) d\theta$$



ML estimation:  $p(x | D) \sim U(0, 8)$  Why?

# 贝叶斯学习近似方法

- 复杂模型(高维参数空间)的参数后验分布和数据后验分布很难解析计算
- 采样: Markov Chain Monte Carlo (MCMC)
- 近似方法
  - Laplace method (Laplacian approximation): 用高斯分布近似后验分布
  - Variational Bayes: 引入潜变量(latent variable)  $Z$ , 最大化数据后验似然度的一个下界(lower bound)
- (参考: C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.)

# 讨论

- Maximum-likelihood (ML) vs Bayesian estimation (BL)
  - When  $n$  approaches infinite, ML and BL are equivalent
  - ML: computationally simple
  - BL: incorporating prior (sometime very informative), theoretically incremental, gives uncertainty of parameters
- BL for multi-variate parameter estimation
  - Usually assume Gaussian prior and posterior for parameters
  - Non-parametric Bayesian learning
  - Many issues in computation, approximation methods

# 下次课内容

- 第3章
  - 特征维数问题
  - 期望最大法
  - 隐马尔可夫模型