贝叶斯决策、参数估计

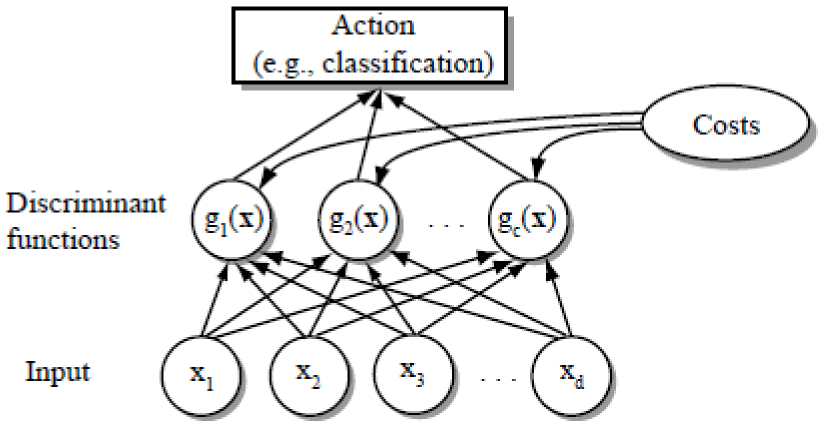
刘成林(liucl@nlpr.ia.ac.cn)

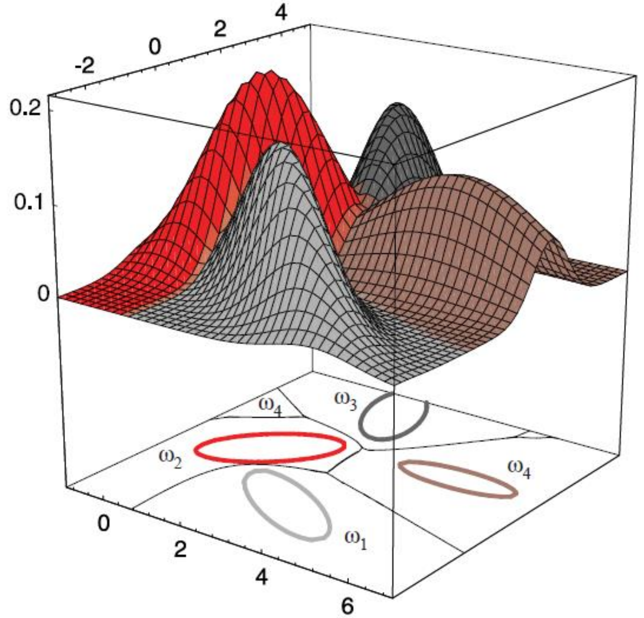
2017年9月27日

助教：何文浩(wenhao.he@nlpr.ia.ac.cn)
杨红明(hongming.yang@nlpr.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): $\mathbf{x} \rightarrow g_i(\mathbf{x})$





上次课主要内容回顾

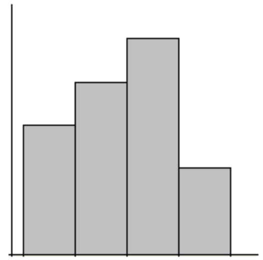
- 贝叶斯决策
  - 最小风险决策
  - (0-1 loss)最小错误率决策（最大后验概率决策）
- 高斯概率密度（正态分布）
  - 1D，多维（记住了？）
  - 协方差矩阵特性
    - 等密度点轨迹、马氏距离、特征值分解、正交化
  - 线性变换的高斯密度？
- 高斯密度下的判别函数
  - Quadratic discriminant function (QDF)
  - Three cases, linear discriminant function (LDF)
- 贝叶斯决策的错误率

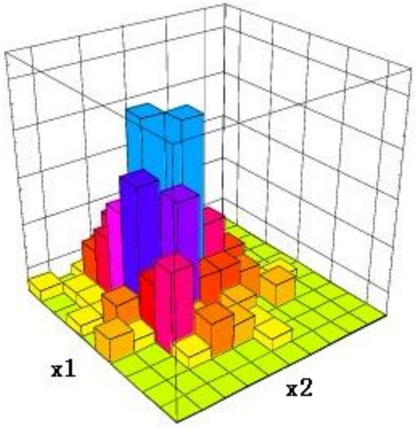
提 纲

- 第2章
  - 离散变量的贝叶斯决策
  - 复合模式分类
- 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 $R(\alpha_i|\mathrm{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathrm{x})$
  - 最小错误率(MAP): max $P(\omega_i|\mathbf{x})$
- 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 $p(\mathbf{x}|\omega_i) = p(x_1 x_2 \cdots x_d | \omega_i)$（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

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

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

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

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

  - Likelihood ratio

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

- 独立二值特征(Binary features)
  - Discriminant 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

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

w_i表征每个特征的判别性

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

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

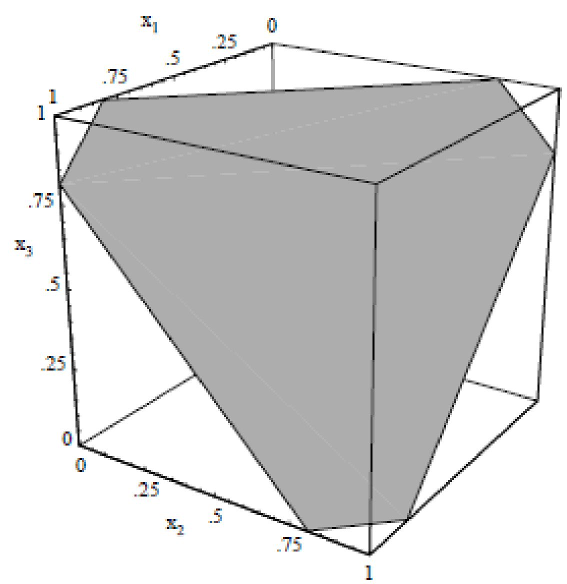
- 一个例子: 3D binary data
  - $P(\omega_1)$=0.5, $P(\omega_2)$=0.5
  - p_i=0.8, q_i=0.5, i=1,2,3

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

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

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

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

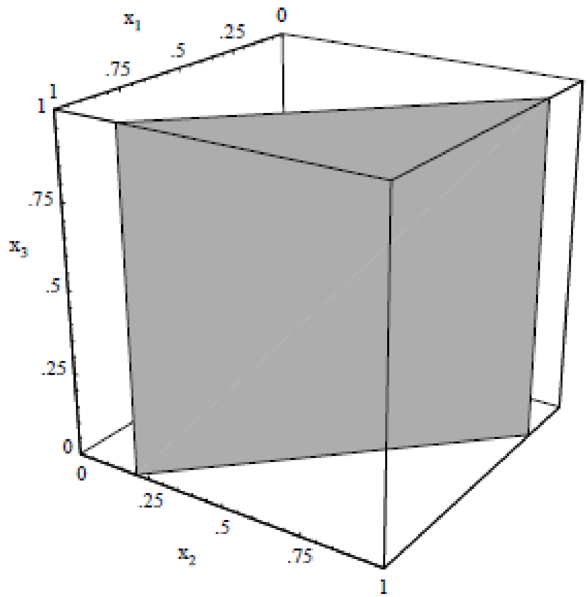


- 另一个例子: 3D binary data
  - $P(\omega_1)$=0.5, $P(\omega_2)$=0.5
  - p_1=p_2=0.8, p_3=0.5; q_i=0.5, i=1,2,3

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

$$w_3 = 0$$

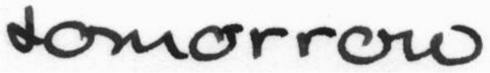
$$w_0 = 2\ln\frac{1-0.8}{1-0.5} = -1.8326$$



复合模式分类

(*2.12 Compound Bayesian Decision Theory and Context)

- 多个模式同时分类 $\mathbf{X} = \mathbf{x}_1\mathbf{x}_2\cdots\mathbf{x}_n \quad \boldsymbol{\omega} = \omega(1)\omega(2)\cdots\omega(n)$
  - 比如：字符串识别



  - 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：**ω**类别数巨大，p(X| **ω**)存储和估计困难
  - Conditionally independent

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

  - 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)

复合模式识别的另一个用途：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 $e_k \in \{\omega_1, \ldots, \omega_M\}$
    联合输出空间(又称为Behavior knowledge space)的后验概率

$$P(\omega_i|e_1,\ldots,e_K) = \frac{\boxed{P(e_1,\ldots,e_K|\omega_i)}P(\omega_i)}{P(e_1,\ldots,e_K)}, \quad i = 1,\ldots,M$$

  - 在验证(validation)数据集上估计离散空间的条件概率密度
    $P(e_1,\ldots,e_K|\omega_i)$ 指数级复杂度，需要大量样本
  - Naïve Bayes

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

  - Dependency tree approximation

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

多分类器融合：也可以分类器输出连续值
或排序作为重新分类的特征

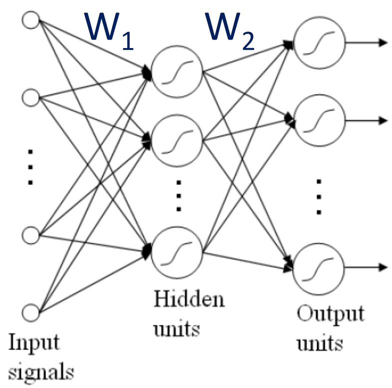
中国科学院大学
University of Chinese Academy of Sciences

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

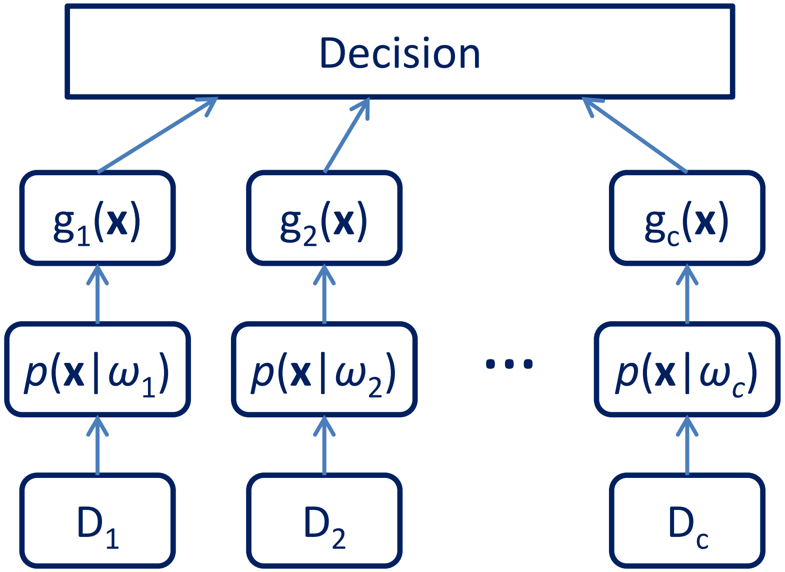
$$g_i(\mathbf{x}) = f(\mathbf{x}, W_1, W_{2,i})$$



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



University of Chinese Academy of Sciences

- ML估计一类模型的参数
  - Likelihood

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

  - Maximization

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

  - Gradient: vector in parameter space

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

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

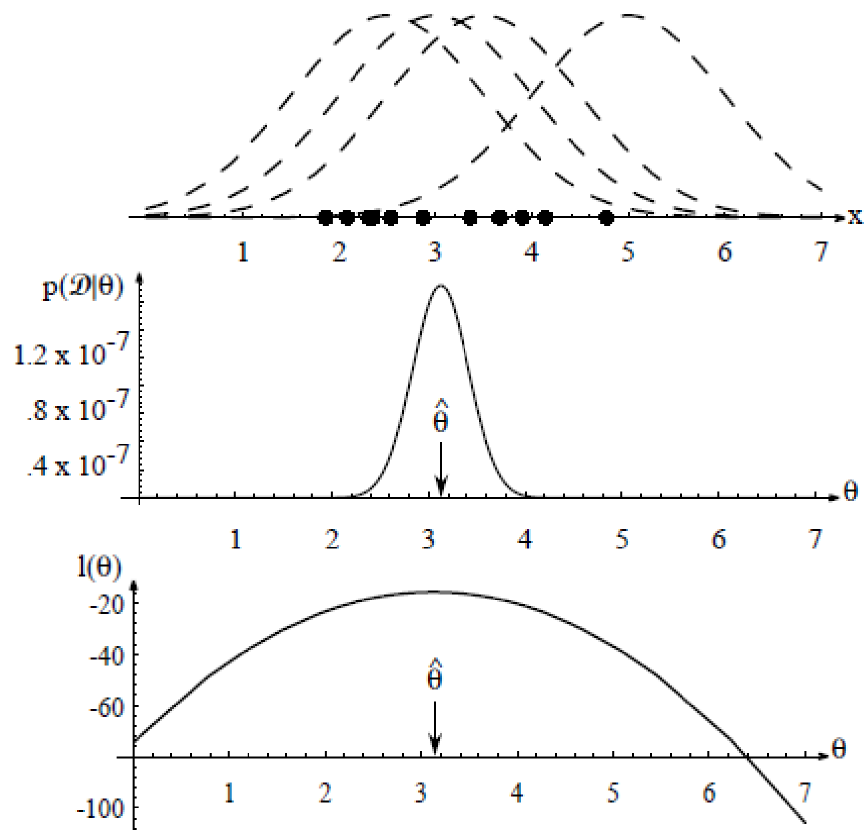
  - $\nabla_{\boldsymbol{\theta}} p(D\,|\,\boldsymbol{\theta}) = 0$ 的解：可能有解析解，也可能需要迭代求解（如梯度下降）

- ML参数估计：一个例子
  - 1D高斯密度，假设σ^2已知，μ未知

10个样本点，
4个假设的高斯密度函数
(Likelihood不同)

Likelihood: μ的函数

Log-likelihood



- Log-likelihood比较容易计算

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

  - ML estimate

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

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

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

- Maximum a posteriori (MAP) estimator

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

  - Equivalent to ML when $p(\boldsymbol{\theta})$ is uniform
  - A case of Bayesian estimation

- Gaussian case: unknown μ
  - Log-likelihood of a single point

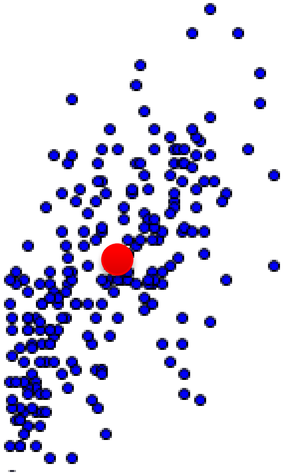
$$\ln p(\mathbf{x}_k|\boldsymbol{\mu}) = -\frac{1}{2}\ln\left[(2\pi)^d|\boldsymbol{\Sigma}|\right] - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

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

  - ML solution: sample mean

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

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



- Gaussian case: unknown μ and Σ
  - 1D case, $\theta_1 = \mu$ and $\theta_2 = \sigma^2$
    - Log-likelihood

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

    - ML solution

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

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

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

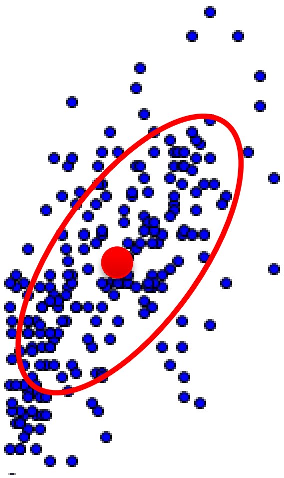
- Gaussian case: unknown μ and Σ
  - Multivariate case (Problem 6, Chapter 3)

    记住结论即可

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

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

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



- 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

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

  - 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D_i估计参数θ_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 $p(\boldsymbol{\theta}|\mathrm{D})$
  - Estimated density

$$p(\mathrm{x}|\mathcal{D}) = \int p(\mathrm{x}, \theta|\mathcal{D})\, d\theta$$

$$= \int p(\mathrm{x}|\theta)\underline{p(\theta|\mathcal{D})}\, d\theta$$

  - Model usage
    - Model average (weighting density functions)

$$p(\mathbf{x}\,|\,D) \propto \frac{1}{M}\sum_{i=1}^{M} p(\mathbf{x}\,|\,\theta_i)$$

Sampling parameter $\theta_i \sim p(\theta\,|\,D)$

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

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 σ^2
  - Assume prior density $p(\mu) \sim N(\mu_0, \sigma_0^2)$
  - Posterior density

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

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

α: normalization factor

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

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

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

$$
\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \qquad \frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2} \longleftarrow \hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k
$$

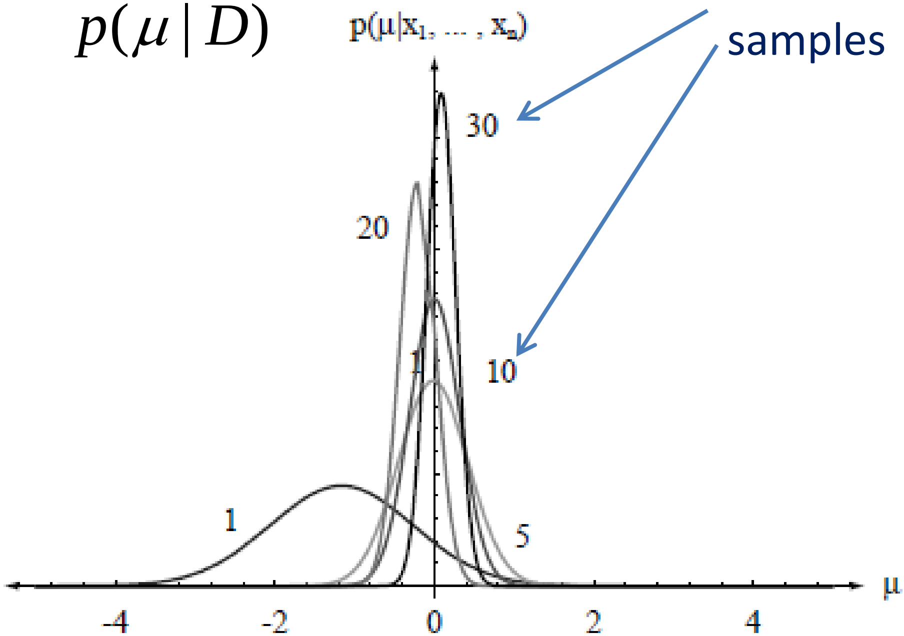
$$
\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2} \longrightarrow \mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0
$$

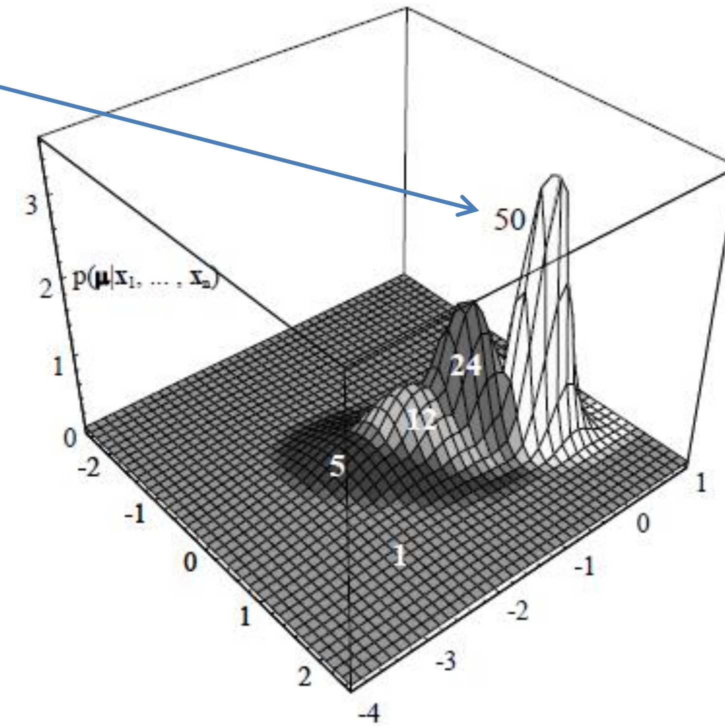
当n增大，μ_n趋近 $\hat{\mu}_n$， σ_n^2趋近σ^2/n

例子：Bayesian learning in 1D/2D space

$p(\mu \mid D)$

number of samples





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

University of Chinese Academy of Sciences

- 1D case: class-conditional density

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

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

(与x无关，可看作是常数)

  - Bayesian estimation

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

  - C.f. ML estimation

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

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

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

- Multivariate case, with Σ known

$$p(\mathrm{x}|\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \text{and} \quad p(\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$ 注意：不同空间！

  - Parameter posterior distribution

$$\begin{aligned} p(\boldsymbol{\mu}|\mathcal{D}) &= \alpha \prod_{k=1}^{n} p(\mathrm{x}_k|\boldsymbol{\mu}) p(\boldsymbol{\mu}) \\ &= \alpha' \exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}^t(n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1})\boldsymbol{\mu} - 2\boldsymbol{\mu}^t\left(\boldsymbol{\Sigma}^{-1}\sum_{k=1}^{n}\mathrm{x}_k + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)\right)\right] \\ &= \alpha'' \exp\left[-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_n)^t \boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_n)\right] \sim N(\mu_n, \Sigma_n) \end{aligned}$$

$$\boldsymbol{\Sigma}_n^{-1} = n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1} \qquad \boldsymbol{\Sigma}_n^{-1}\boldsymbol{\mu}_n = n\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_n + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 \qquad \hat{\boldsymbol{\mu}}_n = \frac{1}{n}\sum_{k=1}^{n}\mathrm{x}_k$$

$$\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_0\left(\boldsymbol{\Sigma}_0 + \frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\frac{1}{n}\boldsymbol{\Sigma} \longrightarrow \boldsymbol{\mu}_n = \boldsymbol{\Sigma}_0\left(\boldsymbol{\Sigma}_0 + \frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\hat{\boldsymbol{\mu}}_n + \frac{1}{n}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}_0 + \frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu}_0$$

  - Data (feature) posterior distribution

$$p(\mathrm{x}|\mathcal{D}) = \int p(\mathrm{x}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\mathcal{D})\, d\boldsymbol{\mu} \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_n)$$

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

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)\ d\theta} \qquad p(\mathcal{D}|\theta) = \prod_{k=1}^{n} p(\mathrm{x}_k|\theta)$$

  - Posterior data distribution

$$p(\mathrm{x}|\mathcal{D}) = \int p(\mathrm{x}|\theta)p(\theta|\mathcal{D})\ d\theta$$

  - Model usage: parameter sampling or MAP

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

University of Chinese Academy of Sciences

- Recursive Bayes Learning
  - Incremental data $\mathcal{D}^n = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$

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

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

  - Recursive update of posterior parameter density

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

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

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

  - Need to retain all samples 1…*n*-1?
    - **Sufficient statistics:** contain all needed information for parameter.
      e.g., in Gaussian case

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

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

University of Chinese Academy of Sciences

- 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

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

θ>=x!

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

n=3?

n=4?

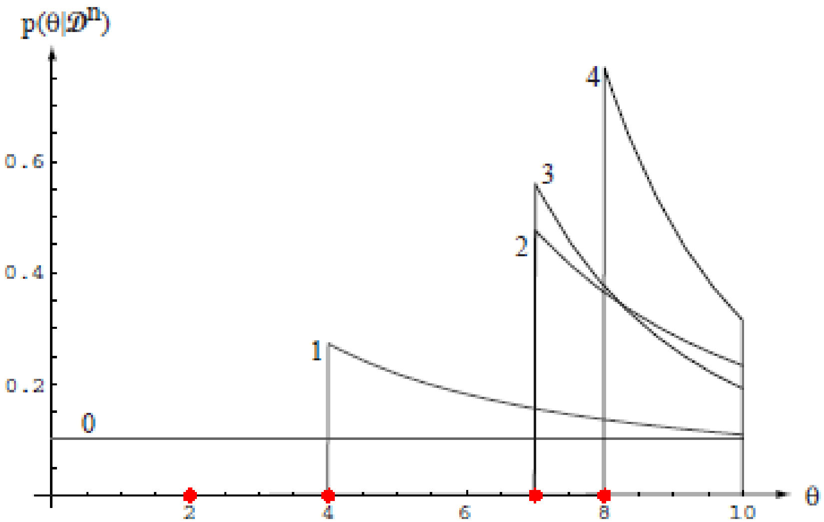
$$p(\theta|D^3) \propto p(x|\theta)p(\theta|D^2) = \begin{cases} 1/\theta^3 & \text{for } 7 \leq \theta \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

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

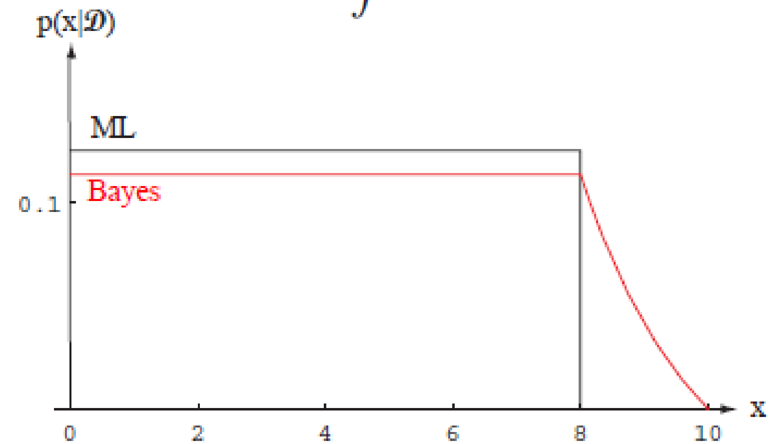
中国科学院大学
University of Chinese Academy of Sciences

– 分布函数图

- Parameter distribution vs feature distribution



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



$$p(\theta \mid D^4) \propto \begin{cases} 1/\theta^4 \text{ for } 8 \le \theta \le 10 \\ 0 \quad \text{otherwise} \end{cases}$$

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

$$p(x \mid D^4) \propto \int p(x \mid \theta) p(\theta \mid D^4) d\theta = \begin{cases} 8^{-4} - 10^{-4}, \; x \le 8 \\ f(x), \quad 8 < x \le 10 \\ 0, \qquad \text{otherwise} \end{cases}$$

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

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

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