



人工智能与自动化学院

模式识别



第三讲 概率密度函数的参数估计

贝叶斯决策：已知 $p(\omega_i)$ 和 $p(\mathbf{x}|\omega_i)$ ，对未知样本分类（设计分类器）。

实际问题：已知一定数目的样本，对未知样本分类（设计分类器）

怎么办？

① 根据样本估计 $p(\omega_i)$ 和 $p(\mathbf{x}|\omega_i)$ ，记为 $\hat{p}(\omega_i)$ 和 $\hat{p}(\mathbf{x}|\omega_i)$ 。

② 用估计的概率密度设计贝叶斯分类器。——（基于样本的）两步贝叶斯决策

希望：当样本数 $N \rightarrow \infty$ 时， $\hat{p}(\omega_i) \xrightarrow{N \rightarrow \infty} p(\omega_i)$ $\hat{p}(\mathbf{x}|\omega_i) \xrightarrow{N \rightarrow \infty} p(\mathbf{x}|\omega_i)$

此时，得到的分类器收敛于理论上的最优解。

重要前提：

- 训练样本的分布能代表样本的真实分布，所谓i.i.d条件
- 有充分的训练样本 **N要足够大**

独立同分布 (iid, independently identically distribution)
指随机过程中，任何时刻的取值都为随机变量，如果这些随机变量服从同一分布，并且互相独立，那么这些随机变量是独立同分布。

第三讲 概率密度函数的参数估计

如何利用样本集估计概率密度函数？

估计概率密度的两种基本方法：

- 参数方法 (Parametric Methods)

根据对问题的一般性认识，假设随机变量服从某种分布，其概率密度函数形式已知，只是表征函数的参数未知，通过训练数据来估计

训练样本：监督和非监督

估计方法：最大似然估计、Bayes估计

- 非参数方法 (Nonparametric Methods)

密度函数的形式未知，也不作假设，利用训练数据直接对概率密度进行估计

训练样本：监督

估计方法：Parzen窗法、 k_n -近邻法

第三讲 概率密度函数的参数估计

基本概念

参数估计(Parametric Estimation):

- 已知概率密度函数的形式，只是其中几个参数未知，目标是根据样本估计这些参数的值。

几个名词:

- 参数空间(Parametric Space): 待估计参数的取值空间 $\theta \in \Theta$
- 估计量(Estimation): $\hat{\theta}$

第三讲 概率密度函数的参数估计



- 3.1 最大似然估计基本原理
- 3.2 最大似然估计的求解
- 3.3 正态分布下的最大似然估计
- 3.4 贝叶斯估计
- 3.5 贝叶斯学习
- 3.6 正态分布下的贝叶斯估计
- 3.7 高斯混合模型与期望最大化算法
- 3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

3.1 最大似然估计基本原理

最大似然估计和Bayes估计的区别

两种方法估计的参数结果接近，但过程有区别：

- **最大似然估计**将未知参数看成是确定变量，在实际样本观察的概率最大的条件下，获得未知参数的最好估计。
- **Bayes估计**将未知参数看成是按某种分布的随机变量，样本的观察结果由先验分布转化为后验分布，再由后验分布修正参数的估计值。

3.1 最大似然估计基本原理

最大似然估计的思路

- 样本集可按类别分开，不同类别密度函数的参数分别用各类的样本集来估计。
- 概率密度函数的形式已知、参数未知，为了描述概率密度函数 $p(x/\omega_i)$ 与参数 θ 的依赖关系，我们用 $p(x/\omega_i, \theta)$ 表示概率密度函数。
- **最大似然估计**中待估计参数 θ 是确定而未知的“数”。
- 独立地按概率密度函数 $p(x/\theta)$ 抽取样本获得样本集 $\chi = \{x_1, x_2, \dots, x_N\}$ ，用 χ 估计未知参数 θ 。

3.1 最大似然估计基本原理

最大似然估计原理

似然函数：是关于统计模型参数 θ 的函数，写做 $L(\theta)$ 。观测结果 χ 在参数集合 θ 上的似然函数就是在给定参数值 θ 的基础上观察到结果 χ 的概率，故有 $L(\theta) = P(\chi|\theta)$ 。

似然函数：

$$l(\theta) = p(\chi | \theta) = p(x_1, x_2, \dots, x_N | \theta) = \prod_{k=1}^N p(x_k | \theta)$$

假设条件：

- ① 参数 θ 是确定的未知量（不是随机量）
- ② 样本集 $\chi_i (i = 1, \dots, c)$ 中的样本都是从密度为 $p(x | \omega_i)$ 的总体中独立抽取出来的(独立同分布, *i.i.d.*)
- ③ $p(x | \omega_i)$ 具有某种确定的函数形式，只其参数 θ 未知
- ④ 各类样本只包含本类分布的信息

注意：参数 θ 通常是向量，比如一维正态分布 $N(\mu_i, \sigma_i^2)$ ，未知参数可能是 $\theta_i = \begin{bmatrix} \mu_i \\ \sigma_i^2 \end{bmatrix}$ ，此时， $p(x | \omega_i)$ 可写成 $p(x | \theta_i)$ 或 $p(x | \omega_i, \theta_i)$ 。

3.1 最大似然估计基本原理

最大似然估计原理

鉴于上述假设，我们可以只考虑一类样本，记已知样本为 $\mathcal{X} = \{x_1, x_2, \dots, x_N\}$

似然函数
$$l(\theta) = p(\mathcal{X} | \theta) = p(x_1, x_2, \dots, x_N | \theta) = \prod_{i=1}^N p(x_i | \theta)$$

—— 在参数 θ 下观测到样本集 \mathcal{X} 的概率（联合分布）密度函数

基本思想：

如果在参数 $\theta = \hat{\theta}$ 下 $l(\theta)$ 最大，则 $\hat{\theta}$ 应是“最可能”的参数值，称作最大似然估计量。

* 为了便于分析，还可以定义对数似然函数 $H(\theta) = \ln l(\theta)$ 。

第三讲 概率密度函数的参数估计



- 3.1 最大似然估计基本原理
- 3.2 最大似然估计的求解
- 3.3 正态分布下的最大似然估计
- 3.4 贝叶斯估计
- 3.5 贝叶斯学习
- 3.6 正态分布下的贝叶斯估计
- 3.7 高斯混合模型与期望最大化算法
- 3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

3.2 最大似然估计的求解

似然函数

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

—— 在参数 θ 下观测到样本集 \mathcal{X} 的概率（联合分布）密度函数

基本思想：

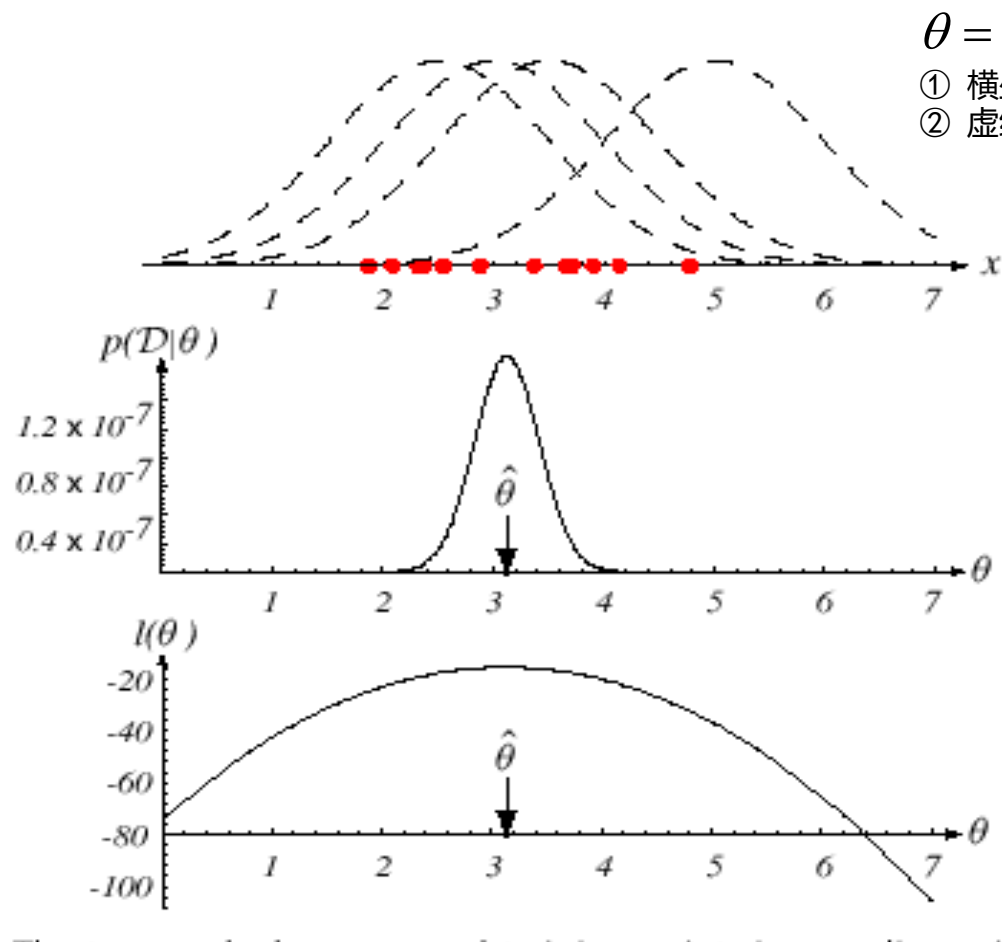
如果在参数 $\theta = \hat{\theta}$ 下 $l(\theta)$ 最大，则 $\hat{\theta}$ 应是“最可能”的参数值，称作最大似然估计量。

求解：

- 若似然函数满足连续、可微的条件，则最大似然估计量就是方程 $dl(\theta)/d\theta = 0$ 或 $dH(\theta)/d\theta = 0$ 的解（必要条件）。
- 若未知参数不止一个，即 $\theta = [\theta_1, \theta_2, \dots, \theta_s]^T$ ，记梯度算子 $\nabla_{\theta} = \left[\frac{\partial}{\partial \theta_1}, \frac{\partial}{\partial \theta_2}, \dots, \frac{\partial}{\partial \theta_s} \right]^T$ ，则最大似然估计量的必要条件由S个方程组成：

$$\nabla_{\theta} l(\theta) = 0 \quad \nabla_{\theta} H(\theta) = 0$$

3.2 最大似然估计的求解



最大似然估计示意图

$$\theta = \mu$$

- ① 横坐标为均值
- ② 虚线为每个可能的 $p(x_i | \theta)$

讨论：

- 如果 $l(\theta)$ 或 $H(\theta)$ 连续、可微，存在最大值，且上述必要条件方程组有唯一解，则其解就是最大似然估计量。（比如多元正态分布）。
- 如果必要条件有多解，则需从中求似然函数最大者。
- 若不满足条件，则无一般性方法，用其它方法求最大。

3.2 最大似然估计的求解

均匀分布的参数估计

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

例3.1 随机变量 x 服从均匀分布, 但参数 θ_1 和 θ_2 未知: $p(x|\theta) = \begin{cases} \frac{1}{\theta_2 - \theta_1} & \theta_1 < x < \theta_2 \\ 0 & \text{其它} \end{cases}$

计算: 均匀分布的对数似然函数为 $\ln(l(\theta_1, \theta_2)) = \ln\left(\frac{1}{(\theta_2 - \theta_1)^N}\right) = -N \ln(\theta_2 - \theta_1)$

$$\begin{cases} \frac{\partial \ln(l(\theta_1, \theta_2))}{\partial \theta_1} = \frac{N}{\theta_2 - \theta_1} \\ \frac{\partial \ln(l(\theta_1, \theta_2))}{\partial \theta_2} = -\frac{N}{\theta_2 - \theta_1} \end{cases}$$

观察, $\theta_2 - \theta_1$ 越小, 对数似然函数越大 $\Rightarrow \theta_2 = \max\{X_i\}, \theta_1 = \min\{X_i\}$

3.2 最大似然估计的求解

Bernouli分布的参数估计

例3.2 设 \mathbf{x} 为 d 维二值（0或1）向量，服从Bernouli分布 $p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^d \theta_i^{x_i} (1-\theta_i)^{1-x_i}$ ， $\boldsymbol{\theta} = [\theta_1, \theta_2, \dots, \theta_d]^T$ 为待估计参数向量， θ_i 为 $x_i = 1$ 的概率，证明其最大似然估计为 $\hat{\boldsymbol{\theta}} = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k$ 。

解：因 \mathbf{x} 服从Bernouli分布，故

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

$P(D|\boldsymbol{\theta})$ is called the likelihood of $\boldsymbol{\theta}$ w.r.t. the set of samples

$$p(\mathbf{x}^1, \dots, \mathbf{x}^n|\boldsymbol{\theta}) = \prod_{k=1}^n \prod_{i=1}^d \theta_i^{x_i^k} (1-\theta_i)^{1-x_i^k} \quad l(\boldsymbol{\theta}) = \ln P(D|\boldsymbol{\theta})$$

对数似然函数 $l(\boldsymbol{\theta}) = \sum_{k=1}^n \sum_{i=1}^d \left[x_i^k \ln \theta_i + (1-x_i^k) \ln (1-\theta_i) \right]$

$$\nabla_{\boldsymbol{\theta}} l = \mathbf{0} \quad \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_i} = \sum_{k=1}^n \left[\frac{x_i^k}{\theta_i} - \frac{1-x_i^k}{1-\theta_i} \right] = \sum_{k=1}^n \frac{x_i^k - \theta_i}{\theta_i(1-\theta_i)} = 0$$

$$\sum_{k=1}^n (x_i^k - \theta_i) = 0$$

$$\sum_{k=1}^n (x_i^k - \theta_i) = 0$$

$$\theta_i = \frac{1}{n} \sum_{k=1}^n x_i^k$$

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

第三讲 概率密度函数的参数估计



- 3.1 最大似然估计基本原理
- 3.2 最大似然估计的求解
- 3.3 正态分布下的最大似然估计
- 3.4 贝叶斯估计
- 3.5 贝叶斯学习
- 3.6 正态分布下的贝叶斯估计
- 3.7 高斯混合模型与期望最大化算法
- 3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

3.3 正态分布下的最大似然估计

单变量正态分布

$$p(x | \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right] \quad \theta = [\theta_1, \theta_2]^T \quad \begin{matrix} \theta_1 = \mu \\ \theta_2 = \sigma^2 \end{matrix} \quad \text{样本集 } \mathcal{X} = \{x_1, x_2, \dots, x_N\}$$

似然函数 $l(x) = p(\mathcal{X} | \theta) = \prod_{k=1}^N p(x_k | \theta)$

对数似然函数 $H(\theta) = \ln l(x) = \sum_{k=1}^N \ln p(x_k | \theta)$

最大似然估计量 $\hat{\theta}$ 满足方程 $\nabla_{\theta} H(\theta) = \sum_{k=1}^N \nabla_{\theta} \ln p(x_k | \theta) = 0$

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

3.3 正态分布下的最大似然估计

单变量正态分布

$$\nabla_{\theta} H(\theta) = \sum_{k=1}^N \nabla_{\theta} \ln p(x_k | \theta) = 0$$

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

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

$$\begin{cases} \sum_{k=1}^N \frac{1}{\hat{\theta}_2} (x_k - \hat{\theta}_1) = 0 \\ -\sum_{k=1}^N \frac{1}{\hat{\theta}_2} + \sum_{k=1}^N \frac{(x_k - \hat{\theta}_1)^2}{\theta_2^2} = 0 \end{cases}$$

$$\begin{cases} \hat{\mu} = \hat{\theta}_1 = \frac{1}{N} \sum_{k=1}^N x_k \\ \hat{\sigma}^2 = \hat{\theta}_2 = \frac{1}{N} \sum_{k=1}^N (x_k - \hat{\mu})^2 \end{cases}$$

3.3 正态分布下的最大似然估计

多变量正态分布

对于一般的多元正态分布，计算方法完全类似，且有

$$\begin{cases} \hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{k=1}^N \mathbf{x}_k \\ \hat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{k=1}^N (\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^T \end{cases}$$

- 均值估计是无偏的，协方差矩阵估计是有偏的。
- 协方差矩阵的无偏估计是：

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N-1} \sum_{k=1}^N (\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^T$$

单变量正态分布

$$\begin{cases} \hat{\mu} = \hat{\theta}_1 = \frac{1}{N} \sum_{k=1}^N x_k \\ \hat{\sigma}^2 = \hat{\theta}_2 = \frac{1}{N} \sum_{k=1}^N (x_k - \hat{\mu})^2 \end{cases}$$

第三讲 概率密度函数的参数估计

- 3.1 最大似然估计基本原理
- 3.2 最大似然估计的求解
- 3.3 正态分布下的最大似然估计
- 3.4 贝叶斯估计
- 3.5 贝叶斯学习
- 3.6 正态分布下的贝叶斯估计
- 3.7 高斯混合模型与期望最大化算法
- 3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

3.4 贝叶斯估计



最大似然估计和Bayes估计的区别

两种方法估计的参数结果接近，但过程有区别：

- 最大似然估计将未知参数看成是确定变量，在实际样本观察的概率最大的条件下，获得未知参数的最好的估计；
- Bayes估计将未知参数看成是按某种分布的随机变量，样本的观察结果由先验分布转化为后验分布，再由后验分布修正参数的估计值。

3.4 贝叶斯估计

贝叶斯估计思路与贝叶斯决策类似，只是离散的决策状态变成了连续的估计。

基本思想：

是变量就有分布

把待估计参数 θ 看作具有先验概率密度函数 $p(\theta)$ 的随机变量，其取值与样本集 \mathcal{X} 有关，根据样本集 $\mathcal{X} = \{x_1, x_2, \dots, x_N\}$ 估计 θ 。

损失函数：把 θ 估计为 $\hat{\theta}$ 所造成的损失，记为 $\lambda(\hat{\theta}, \theta)$

3.4 贝叶斯估计

$$\mathbf{x} \in E^d, \quad \theta \in \Theta$$

期望风险: $R = \int_{E^d} \int_{\Theta} \lambda(\hat{\theta}, \theta) p(\mathbf{x}, \theta) d\theta d\mathbf{x}$

$$= \int_{E^d} \int_{\Theta} \lambda(\hat{\theta}, \theta) p(\theta | \mathbf{x}) p(\mathbf{x}) d\theta d\mathbf{x}$$

$$= \int_{E^d} R(\hat{\theta} | \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

条件风险: $R(\hat{\theta} | \mathbf{x}) = \int_{\Theta} \lambda(\hat{\theta}, \theta) p(\theta | \mathbf{x}) d\theta$

是变量就有分布

$$P(AB) = P(A|B)P(B)$$

$$R(\alpha) = \int R(\alpha(x) | x) p(x) dx$$

$$R(\alpha_i | x) = E[\lambda(\alpha_i, \omega_j) | x] = \sum_{j=1}^c \lambda(\alpha_i, \omega_j) P(\omega_j | x), \quad i = 1, \dots, k$$

目标: 最小化期望风险

- 期望风险是在所有可能的 x 情况下条件风险的积分
- 条件风险都是非负

最小化期望风险 \Rightarrow 最小化条件风险 (对所有可能的 \mathbf{x})

3.4 贝叶斯估计

贝叶斯估计：（在样本集 \mathcal{X} 下）使条件风险 $R(\hat{\theta} | \mathcal{X}) = \int_{\Theta} \lambda(\hat{\theta}, \theta) p(\theta | \mathcal{X}) d\theta$ 最小的估计量 $\hat{\theta}$

损失：

- 离散情况：损失函数表（决策表）
- 连续情况：损失函数

常用的损失函数： $\lambda(\hat{\theta}, \theta) = (\theta - \hat{\theta})^2$ （平方误差损失函数）

定理3.1 如果采用平方误差损失函数，则 θ 的贝叶斯估计量 $\hat{\theta}$ 是在给定 x 时 θ 的条件期望，

$$\text{即 } \hat{\theta} = E[\theta | x] = \int_{\Theta} \theta p(\theta | x) d\theta。$$

⇒ 在给定样本集 \mathcal{X} 下， θ 的贝叶斯估计是 $\hat{\theta} = E[\theta | \mathcal{X}] = \int_{\Theta} \theta p(\theta | \mathcal{X}) d\theta$

最小风险贝叶斯估计

3.4 贝叶斯估计

定理3.1证明:

$$R = \int_{\Omega^N} \int_{\Theta} \lambda(\hat{\theta}, \theta) p(\theta|x) p(x) d\theta dx$$

$$\lambda(\hat{\theta}, \theta) = (\theta - \hat{\theta})^T (\theta - \hat{\theta})$$

$$R = \int_{\Omega^N} \int_{\Theta} \lambda(\hat{\theta}, \theta) p(\theta|\chi) p(\chi) d\theta d\chi$$

$$= \int_{\Omega^N} \int_{\Theta} (\theta - \hat{\theta})^T (\theta - \hat{\theta}) p(\theta|\chi) p(\chi) d\theta d\chi$$

$$= \int_{\Omega^N} \left[\int_{\Theta} (\theta - \hat{\theta})^T (\theta - \hat{\theta}) p(\theta|\chi) d\theta \right] p(\chi) d\chi$$

$$R(\hat{\theta}|\chi) = \int_{\Theta} (\theta - \hat{\theta})^T (\theta - \hat{\theta}) p(\theta|\chi) d\theta$$

$$\min R \Leftrightarrow \min R(\hat{\theta}|\chi)$$

$$\min R(\hat{\theta}|\chi) \Leftrightarrow \frac{\partial R(\hat{\theta}|\chi)}{\partial \hat{\theta}} = -2 \int_{\Theta} (\theta - \hat{\theta}) p(\theta|\chi) d\theta = 0$$

$$\int_{\Theta} (\theta - \hat{\theta}) p(\theta|\chi) d\theta = \int_{\Theta} \theta p(\theta|\chi) d\theta - \hat{\theta} \int_{\Theta} p(\theta|\chi) d\theta$$

$$= \int_{\Theta} \theta p(\theta|\chi) d\theta - \hat{\theta}$$

$$\int_{\Theta} p(\theta|\chi) d\theta = 1$$

最小方差Bayes估计是在观测 χ 条件下 θ 的条件期望。

$$-2 \int_{\Theta} (\theta - \hat{\theta}) p(\theta|\chi) d\theta = 0 \Leftrightarrow \hat{\theta} = \int_{\Theta} \theta p(\theta|\chi) d\theta = E[\theta|\chi]$$

3.4 贝叶斯估计

贝叶斯估计方法

(平方误差损失 $\lambda(\hat{\theta}, \theta) = (\theta - \hat{\theta})^2$ 条件下)

① 确定 θ 的先验分布概密 $p(\theta)$

θ 是变量

② 求样本集的联合分布概密 $p(\mathcal{X} | \theta) = \prod_{i=1}^N p(\mathbf{x}_i | \theta)$

③ 求 θ 的后验概率分布密度 $p(\theta | \mathcal{X}) = \frac{p(\mathcal{X} | \theta) p(\theta)}{\int_{\Theta} p(\mathcal{X} | \theta) p(\theta) d\theta}$

④ 求 θ 的贝叶斯估计量 $\hat{\theta} = \int_{\Theta} \theta p(\theta | \mathcal{X}) d\theta$

最小风险贝叶斯估计

3.4 贝叶斯估计

贝叶斯估计方法

(平方误差损失 $\lambda(\hat{\theta}, \theta) = (\theta - \hat{\theta})^2$ 条件下)

Q1: 如何求样本集的联合分布概密 $p(\mathcal{X} | \theta) = \prod_{i=1}^N p(\mathbf{x}_i | \theta)$

θ 是变量, $p(x_i | \theta)$ 是带变量 θ 的条件概率密度函数, 与概率密度函数 $p(x_i)$ 的形式相同, 所以样本集的联合分布概密求出来是带变量 θ 概率密度函数 $p(x_i)$ 的联乘形式。

3.4 贝叶斯估计

贝叶斯估计方法

(平方误差损失 $\lambda(\hat{\theta}, \theta) = (\theta - \hat{\theta})^2$ 条件下)

T2: 如何求后验概率密度函数 $p(\theta | \mathcal{X}) = \frac{p(\mathcal{X} | \theta) p(\theta)}{\int_{\Theta} p(\mathcal{X} | \theta) p(\theta) d\theta}$

- 需要先求出全概率密度部分 $p(\mathcal{X}) = \int_{\Theta} p(\mathcal{X} | \theta) p(\theta) d\theta$
- 后验概率通常是很难计算的，因为要对所有参数进行积分，不能找到一个典型的闭合解（解析解），可以采用一种近似的方法求后验概率，这就是**最大后验概率估计**。

$$\theta_{MAP} = \arg \max_{\theta} P(\mathcal{X} | \theta) P(\theta)$$

- 整个贝叶斯估计领域的核心技术就是要近似的计算 $p(\theta | \mathcal{X})$ 。

3.4 贝叶斯估计

贝叶斯估计的另一种方法

(平方误差损失 $\lambda(\hat{\theta}, \theta) = (\theta - \hat{\theta})^2$ 条件下)

对于一个训练集合 χ , 贝叶斯公式就变成了:

$$P(\omega_i | x, \chi) = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}} = \frac{p(x|\omega_i, \chi)P(\omega_i|\chi)}{\sum_{j=1}^c p(x|\omega_j, \chi)P(\omega_j|\chi)}$$

可以假定先验概率满足: $P(\omega_i|\chi) = P(\omega_i)$ 。同时, 假定函数独立:

$$P(\omega_i | x, \chi) = \frac{p(x|\omega_i, \chi_i)P(\omega_i)}{\sum_{j=1}^c p(x|\omega_j, \chi_j)P(\omega_j)}$$

先验概率事先知道
或简单计算可得到

χ_i 是属于第 i 类的样本

因此, 我们处理的核心问题, 实际上是根据一组训练样本 χ_i , 估计分布 $p(x|\chi_i)$, 简单记 χ_i 为 χ , $p(x|\chi_i)$ 为 $p(x|\chi)$ 。

3.4 贝叶斯估计

贝叶斯估计的另一种方法

(平方误差损失 $\lambda(\hat{\theta}, \theta) = (\theta - \hat{\theta})^2$ 条件下)

直接估计条件概率密度函数：

$$\begin{aligned} p(x|\chi) &= \int p(x, \theta|\chi) d\theta \\ &= \int p(x|\theta, \chi) p(\theta|\chi) d\theta \\ &= \int p(x|\theta) p(\theta|\chi) d\theta \end{aligned}$$

$$p(\theta|\chi) = \frac{p(\chi|\theta)p(\theta)}{\int_{\Theta} p(\chi|\theta)p(\theta)d\theta}$$

$$p(\chi|\theta) = \prod_{i=1}^N p(x_i|\theta)$$

θ 是有一定分布的随机变量

贝叶斯估计
的核心公式

将观测样本转化为后
验概率 $p(\theta|\chi)$ ，希望
其尖峰值逼近真值

思路：如果条件概率密度形式已知，则利用已有训练样本，可通过 $p(\theta|\chi)$ 对 $p(x|\chi)$ 进行估计。

3.4 贝叶斯估计

贝叶斯估计与贝叶斯决策

Bayes决策

确定 x 的真实状态 ω_i (模式类)

Bayes估计

根据一样本集 $\mathcal{X} = \{x_1, x_2, \dots, x_N\}$

找出估计量 $\hat{\theta}$, 估计 \mathcal{X} 所属总体分布的某个真实参数 θ , 使带来的 Bayes 风险最小。

<u>Bayes 决策问题</u>	<u>Bayes 估计问题</u>
样本 x	样本集 $\mathcal{X} = \{x_1, x_2, \dots, x_N\}$
决策 α_i	估计量 $\hat{\theta}$
真实状态 ω_j	真实参数 θ
状态空间 A 是离散空间	参数空间 Θ 是连续空间
先验概率 $P(\omega_j)$	参数的先验分布 $p(\theta)$

3.4 贝叶斯估计

贝叶斯估计与最大似然估计

$$\hat{\theta} = \int_{\Theta} \theta p(\theta | \mathcal{X}) d\theta$$

贝叶斯估计	最大似然估计
已知概率密度函数 $p(x)$ 的模型形式，都是为了估计模型的参数 θ 。 当训练样本趋于无穷多时，两种估计效果相同。	
小样本情况下的优良估计方法	大样本情况下才具有一定的优良性质
将未知参数看成是按某种分布的随机变量， 样本的观察结果由先验分布转化为后验分布， 再由后验分布修正参数的估计值	在实际样本观察的概率最大的条件下，获得 未知参数的最好的估计
能利用更多的信息，如果这些信息是可靠的， 则更准确。	计算复杂度小、更易理解和掌握

第三讲 概率密度函数的参数估计

- 3.1 最大似然估计基本原理
- 3.2 最大似然估计的求解
- 3.3 正态分布下的最大似然估计
- 3.4 贝叶斯估计
- 3.5 贝叶斯学习
- 3.6 正态分布下的贝叶斯估计
- 3.7 高斯混合模型与期望最大化算法
- 3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

3.5 贝叶斯学习

递推的贝叶斯估计

考虑贝叶斯估计的收敛性：记学习样本个数为 N ，样本集 $\mathcal{X}^N = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ ， $N > 1$ 时有：

$$p(\mathcal{X}^N | \theta) = p(\mathbf{x}_N | \theta) p(\mathcal{X}^{N-1} | \theta)$$

而贝叶斯估计关于待估计参数的后验概率公式为

$$\begin{aligned} p(\theta | \mathcal{X}) &= \frac{p(\mathcal{X} | \theta) p(\theta)}{\int_{\Theta} p(\mathcal{X} | \theta) p(\theta) d\theta} \quad \longrightarrow \quad p(\theta | \mathcal{X}^N) = \frac{p(\mathcal{X}^N | \theta) p(\theta)}{\int_{\Theta} p(\mathcal{X}^N | \theta) p(\theta) d\theta} = \frac{p(x_N | \theta) p(\mathcal{X}^{N-1} | \theta) p(\theta)}{\int_{\Theta} p(x_N | \theta) p(\mathcal{X}^{N-1} | \theta) p(\theta) d\theta} \\ p(\theta | \mathcal{X}^{N-1}) &= \frac{p(\mathcal{X}^{N-1} | \theta) p(\theta)}{\int_{\Theta} p(\mathcal{X}^{N-1} | \theta) p(\theta) d\theta} = \frac{p(x_N | \theta) p(\theta | \mathcal{X}^{N-1})}{\int_{\Theta} p(x_N | \theta) p(\theta | \mathcal{X}^{N-1}) d\theta} \end{aligned}$$

设 $p(\theta | \mathcal{X}^0) = p(\theta)$ ，则随着样本数增多，可得后验概率密度函数序列：

$$p(\theta), p(\theta | \mathbf{x}_1), p(\theta | \mathbf{x}_1, \mathbf{x}_2), \dots$$

—— 参数估计的递推贝叶斯方法

3.5 贝叶斯学习

贝叶斯学习

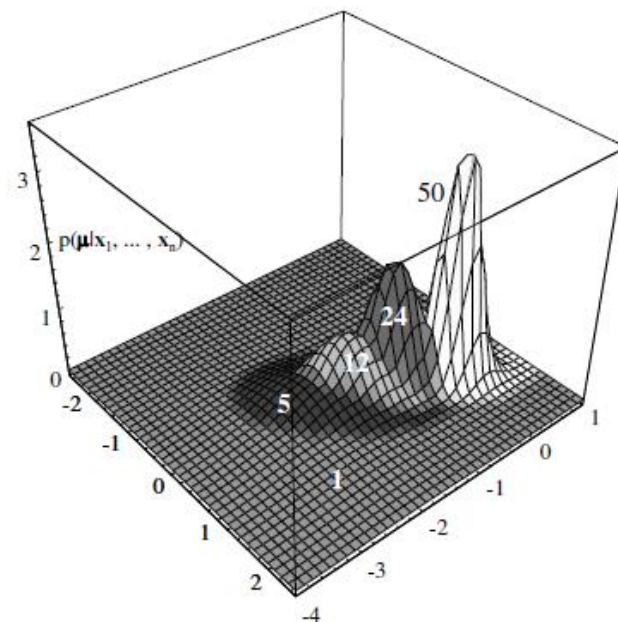
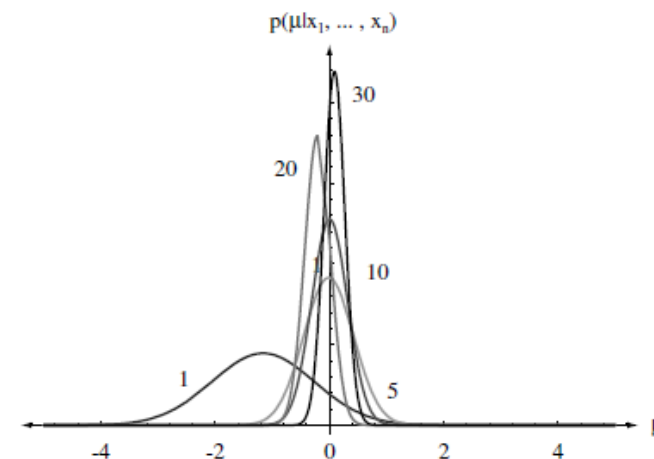
贝叶斯估计关于待估计参数的后验概率公式

$$p(\theta | \mathcal{X}^N) = \frac{p(x_N | \theta) p(\theta | \mathcal{X}^{N-1})}{\int p(x_N | \theta) p(\theta | \mathcal{X}^{N-1}) d\theta}$$

贝叶斯估计关于待估计参数的后验概率序列

$$p(\theta), p(\theta | x_1), p(\theta | x_1, x_2), \dots$$

如果随着样本数的增加，待估计参数后验概率序列逐渐尖锐，趋向于以参数真实值为中心的一个尖峰，当样本无穷多时收敛于在参数真实值上的脉冲函数，则称这一过程为**贝叶斯学习**。



第三讲 概率密度函数的参数估计

- 3.1 最大似然估计基本原理
- 3.2 最大似然估计的求解
- 3.3 正态分布下的最大似然估计
- 3.4 贝叶斯估计
- 3.5 贝叶斯学习
- 3.6 正态分布下的贝叶斯估计
- 3.7 高斯混合模型与期望最大化算法
- 3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

3.6 正态分布下的贝叶斯估计

单变量正态分布 $p(x|\mu) \sim N(\mu, \sigma^2)$ ，已知 σ^2 ，估计 μ 。假设先验分布 $p(\mu) \sim N(\mu_0, \sigma_0^2)$ 。

平方误差损失函数条件下求解 Bayes 估计步骤：

(1) 确定 θ 的先验分布 $p(\theta)$ ；

(2) 由样本集 $\chi = \{x_1, x_2, \dots, x_N\}$ 求样本联合分布 $p(\chi|\theta)$

(3) 求 θ 的后验分布
$$p(\theta|\chi) = \frac{p(\chi|\theta)p(\theta)}{\int_{\Theta} p(\chi|\theta)p(\theta)d\theta}$$

(4)
$$\hat{\theta} = \int_{\Theta} \theta p(\theta|\chi)d\theta$$

现 (1) 已知。 $\theta = \mu$

3.6 正态分布下的贝叶斯估计

单变量正态分布 $p(x | \mu) \sim N(\mu, \sigma^2)$ ，已知 σ^2 ，估计 μ 。假设先验分布 $p(\mu) \sim N(\mu_0, \sigma_0^2)$ 。

(2) 样本联合分布（样本独立抽取）

$$p(\chi | \theta = \mu) = \prod_{k=1}^n p(x_k | \mu) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2} \sum_{k=1}^n \left(\frac{x_k - \mu}{\sigma}\right)^2\right\}$$

(3) 计算后验分布 $p(\mathbf{x}_k | \mu) \sim N(\mu, \sigma^2)$ $p(\mu) \sim N(\mu_0, \sigma_0^2)$

$$\begin{aligned} p(\mu | \chi) &= \frac{p(\chi | \mu) p(\mu)}{\int_{\Theta} p(\chi | \mu) p(\mu) d\mu} = \alpha p(\chi | \mu) p(\mu) = \alpha \prod_{i=1}^n p(x_i | \mu) p(\mu) \\ &= \alpha \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2} \sum_{k=1}^n \left(\frac{x_k - \mu}{\sigma}\right)^2\right\} \cdot \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{1}{2} \left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right\} \end{aligned}$$

$$\text{here, } \alpha = 1 / \int_{\Theta} p(\chi | \mu) p(\mu) d\mu$$

3.6 正态分布下的贝叶斯估计

单变量正态分布 $p(x | \mu) \sim N(\mu, \sigma^2)$ ，已知 σ^2 ，估计 μ 。假设先验分布 $p(\mu) \sim N(\mu_0, \sigma_0^2)$ 。

(3) 计算后验分布

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

$$\text{其中: } \alpha' = \alpha \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \frac{1}{\sqrt{2\pi}\sigma_0}$$

3.6 正态分布下的贝叶斯估计

单变量正态分布 $p(x | \mu) \sim N(\mu, \sigma^2)$ ，已知 σ^2 ，估计 μ 。假设先验分布 $p(\mu) \sim N(\mu_0, \sigma_0^2)$ 。

(3) 计算后验分布

$$\begin{aligned}
 p(\mu | \chi) &= \alpha' \exp \left\{ -\frac{1}{2\sigma^2} \left(\sum_{k=1}^n x_k^2 + n\mu^2 - 2\mu \sum_{k=1}^n x_k \right) - \frac{1}{2\sigma_0^2} (\mu^2 + \mu_0^2 - 2\mu\mu_0) \right\} \\
 &= \alpha' \exp \left\{ -\frac{1}{2} \left[\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) \mu^2 - 2 \left(\frac{1}{\sigma^2} \sum_{k=1}^n x_k + \frac{\mu_0}{\sigma_0^2} \right) \mu \right] - \frac{1}{2} \left(\frac{1}{\sigma^2} \sum_{k=1}^n x_k^2 + \frac{1}{\sigma_0^2} \mu_0^2 \right) \right\} \\
 &= \alpha'' \exp \left\{ -\frac{1}{2} \left[\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) \mu^2 - 2 \left(\frac{1}{\sigma^2} \sum_{k=1}^n x_k + \frac{\mu_0}{\sigma_0^2} \right) \mu \right] \right\} \quad \text{无关项}
 \end{aligned}$$

由此可见 $p(\mu | \chi)$ 是 μ 的二次函数的指数函数，因此 $p(\mu | \chi)$ 满足正态分布

3.6 正态分布下的贝叶斯估计

单变量正态分布 $p(x | \mu) \sim N(\mu, \sigma^2)$ ，已知 σ^2 ，估计 μ 。假设先验分布 $p(\mu) \sim N(\mu_0, \sigma_0^2)$ 。

(3) 计算后验分布

$$p(\mu | \chi) = \alpha'' \exp \left\{ -\frac{1}{2} \left[\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \right) \mu^2 - 2 \left(\frac{1}{\sigma^2} \sum_{k=1}^N x_k + \frac{\mu_0}{\sigma_0^2} \right) \mu \right] \right\}$$

$$\because p(\mu | \chi) \sim N(\mu_N, \sigma_N^2)$$

$$\begin{aligned} p(\mu | \chi) &= \frac{1}{\sqrt{2\pi}\sigma_N} \exp \left(\left[-\frac{1}{2} \left(\frac{\mu - \mu_N}{\sigma_N} \right)^2 \right] \right) \\ &= \alpha'' \exp \left\{ -\frac{1}{2} \left[\frac{1}{\sigma_N^2} \mu^2 - 2 \frac{\mu_N}{\sigma_N^2} \mu \right] \right\} \end{aligned}$$

$$\begin{cases} \frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \\ \frac{\mu_N}{\sigma_N^2} = \frac{1}{\sigma^2} \sum_{k=1}^N x_k + \frac{\mu_0}{\sigma_0^2} = \frac{N}{\sigma^2} m_N + \frac{\mu_0}{\sigma_0^2} \end{cases}$$

$$\begin{cases} \mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} m_N + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 \\ \sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2} \end{cases}$$

$$m_N = \frac{1}{N} \sum_{k=1}^N x_k$$

3.6 正态分布下的贝叶斯估计

单变量正态分布 $p(x | \mu) \sim N(\mu, \sigma^2)$ ，已知 σ^2 ，估计 μ 。假设先验分布 $p(\mu) \sim N(\mu_0, \sigma_0^2)$ 。

$$(4) \quad \hat{\theta} = \int_{\Theta} \theta p(\theta | \chi) d\theta$$

$$\begin{cases} \mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} m_N + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 & \boxed{N \rightarrow \infty \quad \mu_N \rightarrow m_N} \\ \sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2} & \boxed{N \rightarrow \infty \quad \sigma_N^2 \rightarrow 0} \end{cases}$$

$$\hat{\mu} = \int_{\Theta} \mu p(\mu | \chi) d\mu = \int_{\Theta} \mu \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_N}{\sigma_N}\right)^2\right] d\mu$$

$$= \mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} m_N + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0$$

$$m_N = \frac{1}{N} \sum_{k=1}^N x_k$$

3.6 正态分布下的贝叶斯估计

单变量正态分布 $p(x | \mu) \sim N(\mu, \sigma^2)$ ，已知 σ^2 ，估计 μ 。假设先验分布 $p(\mu) \sim N(\mu_0, \sigma_0^2)$ 。

$$\hat{\theta} \longrightarrow \hat{\mu} = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} m_N + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 \quad \text{其中 } m_N = \frac{1}{N} \sum_{i=1}^N x_i$$

----- 样本信息与先验知识的线性组合

讨论：

- $N = 0$ 时， $\hat{\mu} = \mu_0$ ； $N \rightarrow \infty$ 时， $\hat{\mu} \rightarrow m_N$
- 若 $\sigma_0^2 = 0$ ，则 $\hat{\mu} \equiv \mu_0$ （先验知识可靠，样本不起作用）
- 若 $\sigma_0 \gg \sigma$ ，则 $\hat{\mu} = m_N$ （先验知识十分不确定，完全依靠样本信息）

3.6 正态分布下的贝叶斯估计

正态分布的贝叶斯学习

$$\begin{cases} \mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} m_N + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 \\ \sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2} \end{cases}$$

$N \rightarrow \infty \quad \mu_N \rightarrow m_N$

$N \rightarrow \infty \quad \sigma_N^2 \rightarrow 0$

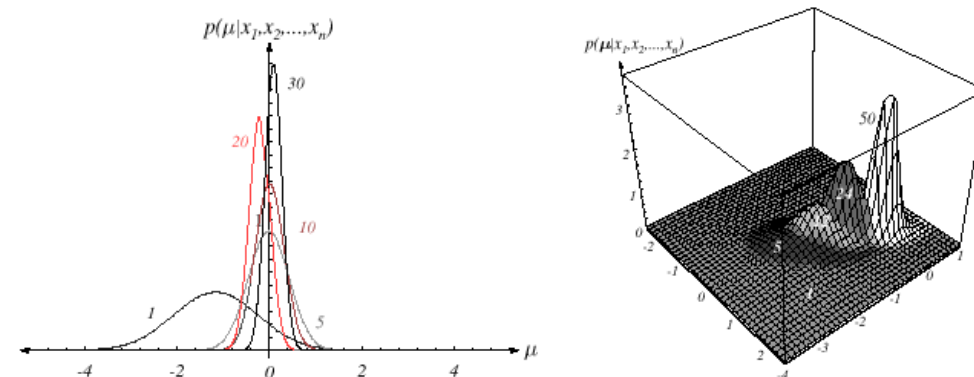


FIGURE 3.2. Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- 当观察一个样本时, $N=1$ 就会有一个 μ 的估计值的修正值;当观察 $N=4$ 时, 对 μ 进行修正, 向真正的 μ 靠近; 当观察 $N=9$ 时, 对 μ 进行修正, 向真正的 μ 靠的更近
- 当 $N \uparrow$, μ_N 就反映了观察到 N 个样本后对 μ 的最好推测, 而 σ_N^2 反映了这种推测的不确定性, $N \uparrow$, $\sigma_N^2 \downarrow$, σ_N^2 随观察样本增加而单调减小, 且当 $N \rightarrow \infty$, $\sigma_N^2 \rightarrow 0$
- 当 $N \uparrow$, $P(\mu|\chi)$ 越来越尖峰突起。 $N \rightarrow \infty$, $P(\mu|\chi) \rightarrow$ 冲击函数, 这个过程称为贝叶斯学习。

3.6 正态分布下的贝叶斯估计

单变量正态分布 $p(x|\mu) \sim N(\mu, \sigma^2)$ ，已知 σ^2 ，估计 μ 。假设先验分布 $p(\mu) \sim N(\mu_0, \sigma_0^2)$ 。

$$p(\mu|\chi) \sim N(\mu_N, \sigma_N^2)$$

Bayes 学习是利用 θ 的先验分布及样本提供的信息求出 θ 的后验分布 $p(\theta|\chi)$ ，然后直接求总体分布。

$$\begin{aligned} p(x|\chi) &= \int_{\theta} p(x|\theta) p(\theta|\chi) d\theta = \int_{\theta} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \times \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2\right] d\mu \\ &= \frac{1}{2\pi\sigma\sigma_N} f(\sigma, \sigma_N) \exp\left[-\frac{1}{2}\left(\frac{x-\mu_N}{\sqrt{\sigma^2+\sigma_N^2}}\right)^2\right] \Rightarrow N(\mu_N, \sigma^2+\sigma_N^2) \end{aligned}$$

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

第三讲 概率密度函数的参数估计



- 3.1 最大似然估计基本原理
- 3.2 最大似然估计的求解
- 3.3 正态分布下的最大似然估计
- 3.4 贝叶斯估计
- 3.5 贝叶斯学习
- 3.6 正态分布下的贝叶斯估计
- 3.7 高斯混合模型与期望最大化算法
- 3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

3.7 高斯混合模型与期望最大化算法

高斯混合模型

之前算法都是假定类条件概率密度可参数化，如满足正态分布。如果不满足这样的假定呢？

- 非参数估计法
- 混合模型

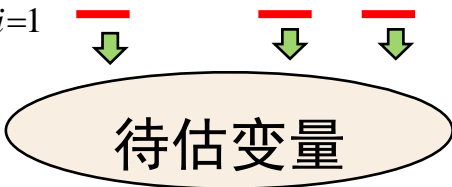
$$p(x|\theta) = \sum_{m=1}^M a_m p_m(x|\theta_m)$$

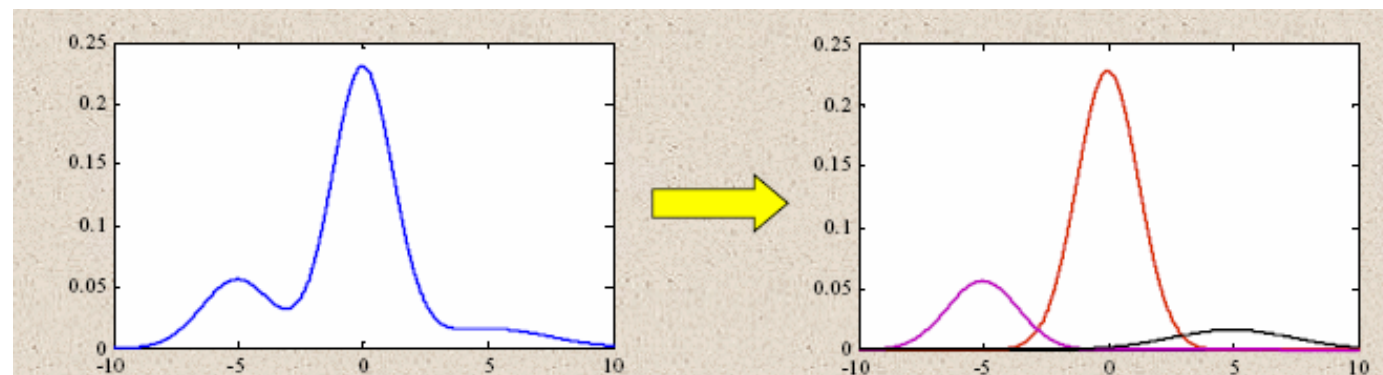
$$\sum_{m=1}^M a_m = 1$$

一个复杂概率密度分布函数可由多个简单的密度函数混合构成

高斯混合模型 (Gauss Mixture Model , GMM)

$$p(x) = \sum_{i=1}^M a_i N(x; \mu_i, \Sigma_i)$$





3.7 高斯混合模型与期望最大化算法

高斯混合模型

$$p(x) = \sum_{i=1}^M a_m N(x; \mu_m, \Sigma_m)$$

混合密度模型的参数可以表示为：

$$\theta = (a_1, a_2, \dots, a_M, \theta_1, \theta_2, \dots, \theta_M)$$

参数的估计方法：

- 利用最优化方法直接对似然函数进行优化，如梯度下降法；
- 引入未知隐变量Y对问题进行简化，将Y看作丢失的数据，使用**期望最大化算法**进行优化。

3.7 高斯混合模型与期望最大化算法

期望最大化算法 (Expectation-Maximization algorithm, EM)

EM算法是一类通过迭代进行最大似然估计的优化算法，通常作为牛顿迭代法的替代，用于对包含隐变量或缺失数据的情况进行参数估计。

EM核心思想：根据已有的数据来递归估计似然函数。

3.7 高斯混合模型与期望最大化算法

两枚硬币A和B，假定随机抛掷后正面朝上概率分别为 P_A ， P_B 。为了估计这两个硬币朝上的概率，咱们轮流抛硬币A和B，每一轮都连续抛5次，总共5轮：

硬币	结果	统计
A	正正反正反	3正-2反
B	反反正正反	2正-3反
A	正反反反反	1正-4反
B	正反反正正	3正-2反
A	反正正反反	2正-3反

$$P_A = (3+1+2) / 15 = 0.4$$

$$P_B = (2+3) / 10 = 0.5$$

3.7 高斯混合模型与期望最大化算法

如果我们不知道抛的硬币是A还是B呢（即硬币种类是隐变量），然后再轮流抛五轮，得到如下结果：

硬币	结果	统计
Unknown	正正反正反	3正-2反
Unknown	反反正正反	2正-3反
Unknown	正反反反反	1正-4反
Unknown	正反反正正	3正-2反
Unknown	反正正反反	2正-3反

$P_A = ?$

$P_B = ?$

- 多了一个硬币种类的隐变量，设为 z ，是一个5维的向量 $(z_1, z_2, z_3, z_4, z_5)$ ，代表每次投掷时所使用的硬币，比如 z_1 ，就代表第一轮投掷时使用的硬币是A还是B。
- 先随机初始化一个 P_A 和 P_B ，用其来估计 z ；然后基于 z ，还是按照最大似然概率法则去估计新的 P_A 和 P_B ；然后依次循环，直到 **P_A 和 P_B 收敛到真实值为止**。

3.7 高斯混合模型与期望最大化算法

给PA和PB赋一个值，比如：硬币A正面朝上的概率 $P_A = 0.2$ 硬币B正面朝上的概率 $P_B = 0.7$

轮数	若是硬币A	若是硬币B	最大似然法则
1	0.00512, 即0.2 0.2 0.2 0.8 0.8, 3正-2反	0.03087, 3正-2反	B
2	0.02048, 即0.2 0.2 0.8 0.8 0.8, 2正-3反	0.01323, 2正-3反	A
3	0.08192, 即0.2 0.8 0.8 0.8 0.8, 1正-4反	0.00567, 1正-4反	A
4	0.00512, 即0.2 0.2 0.2 0.8 0.8, 3正-2反	0.03087, 3正-2反	B
5	0.02048, 即0.2 0.2 0.8 0.8 0.8, 2正-3反	0.01323, 2正-3反	A

- $P_A = (2+1+2) / 15 = 0.33$ $P_B = (3+3) / 10 = 0.6$
- 用估计出的PA和PB再来估计z，再用z来估计新的PA和PB，反复迭代下去，就可以最终得到
 $P_A = 0.4$, $P_B = 0.5$

3.7 高斯混合模型与期望最大化算法

EM算法原理

设 X 是观察到的样本数据集合, Y 为丢失的数据集合或模型中任何无法直接观测的随机变量, 则完整的样本集合为 $D = X \cup Y$ 。

$$p(D|\theta) = p(X, Y|\theta)$$

由于 Y 未知, 在给定参数 θ 时, 似然函数可以看作 Y 的函数:

$$l(\theta) = l(\theta|D) = l(\theta|X, Y) = \ln p(X, Y|\theta) \quad \text{或} \log p(X, Y|\theta)$$

由于 Y 未知, 因此需要寻找在 Y 的所有可能情况下平均意义上的似然函数最大值, 即似然函数对 Y 的期望的最大值:

$$Q(\theta, \theta^{i-1}) = E_Y(l(\theta|X, Y)|X, \theta^{i-1}) = E_Y(\ln p(X, Y|\theta)|X, \theta^{i-1})$$

则

$$\theta^i = \arg \max_{\theta} Q(\theta, \theta^{i-1})$$

3.7 高斯混合模型与期望最大化算法

EM的基本算法

1. *begin initialize* $\theta^0, T, i \leftarrow 0$
2. *do* $i \leftarrow i+1$
3. *E step*: compute $Q(\theta, \theta^{i-1})$
4. *M step*: $\theta^i = \arg \max_{\theta} Q(\theta, \theta^{i-1})$
5. *until* $Q(\theta^{i+1}, \theta^i) - Q(\theta^i, \theta^{i-1}) \leq T$
6. *return*
7. *end* $\hat{\theta} = \theta^{i+1}$

$$\begin{aligned} Q(\theta, \theta^{i-1}) &= E_Y \left(l(\theta | X, Y) | X, \theta^{i-1} \right) \\ &= E_Y \left(\ln p(X, Y | \theta) | X, \theta^{i-1} \right) \\ \theta^i &= \arg \max_{\theta} Q(\theta, \theta^{i-1}) \end{aligned}$$

3.7 高斯混合模型与期望最大化算法

EM的基本算法

EM—Expectation

- 观测数据 X 已知，参数 θ 的当前值 θ^{i-1} 已知，在完整似然函数中，缺失数据(隐含变量) Y 未知，完整对数似然函数对 Y 求期望。
- 设 $Q(\theta, \theta^{i-1}) = E_Y \left(l(\theta | X, Y) | X, \theta^{i-1} \right) = E_Y [\ln p(X, Y | \theta) | X, \theta^{i-1}] = \int_{y \in \gamma} \ln p(X, Y | \theta) p(y | X, \theta^{i-1}) dy$
- 通过求期望，去掉了完整似然函数中的变量 Y 。

EM—Maximization

- 对 E 步计算得到的完整似然函数的期望求极大值，得到参数新的估计值，即

$$\theta^i = \arg \max_{\theta} Q(\theta, \theta^{i-1})$$

- 每次参数更新会增加非完整似然值
- 反复迭代后，会收敛到似然的局部最大值

3.7 高斯混合模型与期望最大化算法

EM算法的收敛性

$$Q(\theta, \theta^i) = E_Y[l(\theta | X, Y) | X, \theta^i]$$

$$= \int_{y \in \gamma} l(\theta | X, y) p(y | X, \theta^i) dy$$

$$= \int_{y \in \gamma} \ln[p(y | X, \theta) \cdot p(X | \theta)] \cdot p(y | X, \theta^i) dy$$

$$= \int_{y \in \gamma} \ln(p(y | X, \theta)) p(y | X, \theta^i) dy + l(\theta | X)$$

$$l(\theta | X) = \ln p(X | \theta)$$

$$\int_{y \in \gamma} p(y | X, \theta^i) dy = 1$$

$$l(\theta^{i+1} | X) - l(\theta^i | X) = Q(\theta^{i+1}, \theta^i) - Q(\theta^i, \theta^i) + D(\theta^i, \theta^{i+1})$$

$$\text{其中 } D(\theta^i, \theta^{i+1}) = \int_{y \in \gamma} \ln \frac{p(y | X, \theta^i)}{p(y | X, \theta^{i+1})} p(y | X, \theta^i) dy \geq 0, \quad \text{Jensen's inequality}$$

- 当 Q 取极大值时，观测数据的似然也在相同点取极大值
- EM 算法会收敛到似然函数的局部极大值，不能保证收敛于全局最优点。

3.7 高斯混合模型与期望最大化算法

高斯混合模型的EM估计

高斯混合模型为 $p(x) = \sum_{m=1}^M a_m N(x; \mu_m, \Sigma_m)$, 已知观测到随机变量 X 的IID样本 $X = \{x_1, x_2, \dots, x_n\}$,

隐含数据集合 $Y = \{y_1, y_2, \dots, y_n\}$, $y_s \in \{1, \dots, M\}$, 代表第 i 个训练样本是由第 y_i 个高斯函数产生的, 待估计参数 $\theta = (a_1, a_2, \dots, a_M, \theta_1, \theta_2, \dots, \theta_M)$, $\theta_m = (\mu_m, \Sigma_m)$ 。采用EM算法进行迭代估计。

- 设混合模型数为 M , 初始化模型参数 θ^0 , 阈值 Th , $i \leftarrow 0$ 。
- 迭代计算模型参数, 直到似然函数 l 变化小于阈值 Th (或待估参数不再发生变化) 为止

3.7 高斯混合模型与期望最大化算法

高斯混合模型的EM估计

E Step 求期望

$$\begin{aligned}
 Q(\theta, \theta^{i-1}) &= E_Y \left(l(\theta | X, Y) | X, \theta^{i-1} \right) \\
 &= E_Y [\ln p(X, Y | \theta) | X, \theta^{i-1}] \\
 &= \sum_{y \in \gamma} \ln p(X, Y | \theta) p(y | X, \theta^{i-1}) \\
 &= \sum_{s=1}^n \sum_{m=1}^M p(y_s = m | x_s, \theta^{i-1}) \ln p(x_s, y_s = m | \theta) \\
 &= \sum_{s=1}^n \sum_{m=1}^M r_{sm} (\ln p(y_s = m | \theta) + \ln p(x_s | y_s = m, \theta))
 \end{aligned}$$

其中，需要计算丢失数据的后验概率

$$\begin{aligned}
 r_{sm} &= p(y_s = m | x_s, \theta^i) \\
 &= p(m | x_s, \theta^i) \\
 &= \frac{a_m^i p_m(x_s | \theta_m^i)}{\sum_{j=1}^M a_j^i p_j(x_s | \theta_j^i)} \quad s = 1, 2, \dots, n
 \end{aligned}$$

样本 x_s 在第 m 个 Gauss 模型的类概率密度, θ_m^j 已知时该值为已知

- [1] J. A. Bilmes (1998). A General Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models.
- [2] <http://www.it.lut.fi/publications/files/publications/192/laivosrap95.pdf>

- 样本 x_s 属于第 m 个 Gauss 模型的后验概率

3.7 高斯混合模型与期望最大化算法

高斯混合模型的EM估计

M Step 求得最大化 $Q(\theta, \theta^{i-1})$ 下的 θ

令 $Q(\theta, \theta^{i-1})$ 梯度等于0，求极值条件下的参数，更新迭代待估计参数，可得：

$$\begin{aligned} a_m^{i+1} &= \frac{1}{n} \sum_{s=1}^n p(m|x_s, \theta^i) \\ \mu_m^{i+1} &= \frac{\sum_{s=1}^n x_s p(m|x_s, \theta^i)}{\sum_{s=1}^n p(m|x_s, \theta^i)} \\ \Sigma_m^{i+1} &= \frac{\sum_{s=1}^n p(m|x_s, \theta^i) (x_s - \mu_m^{i+1})(x_s - \mu_m^{i+1})^T}{\sum_{s=1}^n p(m|x_s, \theta^i)} \end{aligned}$$

3.7 高斯混合模型与期望最大化算法

例3.1 设样本服从二维高斯分布，样本集包括4个样本点 $D = \{x_1, x_2, x_3, x_4\} = \left\{ \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} * \\ 4 \end{pmatrix} \right\}$

其中，*表示丢失的特征值。设二维高斯分布协方差矩阵为对角阵，试估计均值和协方差。

解：坏数据集 D_b 由特征 x_{41} 组成，好的数据集由其余特征组成。

待估计参数向量为 $\theta^T = (\mu_1, \mu_2, \sigma_1, \sigma_2)$ ，设参数向量初始值为 $\theta^0 = (0, 0, 1, 1)$

3.7 高斯混合模型与期望最大化算法

计算给定初始值条件下的 $Q(\theta, \theta^0)$

$$Q(\theta, \theta^0) = E[\ln p(D_g, D_b; \theta | \theta^0; D_g)]$$

$$= \int p(D_b | \theta^0, D_g) \ln p(D_g, D_b; \theta | \theta^0; D_g) dD_b$$

$$= \int p(x_{41} | \theta^0; x_{42} = 4) \sum_{k=1}^4 \ln p(x_k | \theta) dx_{41}$$

$$= \sum_{k=1}^3 \ln p(x_k | \theta) \int p(x_{41} | \theta^0; x_{42} = 4) dx_{41} + \int \ln p(x_4 | \theta) p(x_{41} | \theta^0; x_{42} = 4) dx_{41}$$

$$= \sum_{k=1}^3 \ln p(x_k | \theta) + \int \ln p(x_4 | \theta) \frac{p(x_4 | \theta^0)}{\underbrace{\left(\int_{-\infty}^{+\infty} p\left(\begin{pmatrix} x'_{41} \\ 4 \end{pmatrix} | \theta^0 \right) dx'_{41} \right)}_K} dx_{41}$$

$$Q(\theta, \theta^{i-1}) = E_Y \left(l(\theta | X, Y) | X, \theta^{i-1} \right)$$

$$= E_Y [\ln p(X, Y | \theta) | X, \theta^{i-1}]$$

$$= \sum_{y \in \gamma} \ln p(X, Y | \theta) p(y | X, \theta^{i-1})$$

$$= \sum_{s=1}^n \sum_{m=1}^M p(y_s = m | x_s, \theta^{i-1}) \ln p(x_s, y_s = m | \theta)$$

$$= \sum_{s=1}^n \sum_{m=1}^M r_{sm} (\ln p(y_s = m | \theta) + \ln p(x_s | y_s = m, \theta))$$

$$P(A|B) = P(AB) / P(B)$$

3.7 高斯混合模型与期望最大化算法



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

$$Q(\theta; \theta^0) = \sum_{k=1}^3 [\ln p(x_k | \theta)] + \frac{1}{K} \int \ln p(x_4 | \theta) \frac{1}{2\pi \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix}} \exp\left(-\frac{1}{2}(x_{41}^2 + 4^2)\right) dx_{41}$$

$$= \sum_{k=1}^3 [\ln p(x_k | \theta)] - \frac{1 + \mu_1^2}{2\sigma_1^2} - \frac{(4 - \mu_2)^2}{2\sigma_2^2} - \ln(2\pi\sigma_1\sigma_2)$$

求最大化 Q 时的 θ 值即为 θ 的更新值：

$$\theta^1 = \begin{pmatrix} 0.75 \\ 2.0 \\ 0.938 \\ 2.0 \end{pmatrix}$$

进行迭代计算可得：

$$\mu = \begin{pmatrix} 1.0 \\ 2.0 \end{pmatrix} \text{ 和 } \Sigma = \begin{pmatrix} 0.667 & 0 \\ 0 & 2.0 \end{pmatrix}$$

第三讲 概率密度函数的参数估计



- 3.1 最大似然估计基本原理
- 3.2 最大似然估计的求解
- 3.3 正态分布下的最大似然估计
- 3.4 贝叶斯估计
- 3.5 贝叶斯学习
- 3.6 正态分布下的贝叶斯估计
- 3.7 高斯混合模型与期望最大化算法
- 3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

隐马尔可夫模型

问题 与时间相关的问题，即过程随着时间而进行， t 时刻发生的事件要受之前时刻发生事件的直接影响，如何识别？

例如：语音识别或手势识别等问题。

隐马尔可夫模型 (Hidden Markov Models, HMMs)

- 数学中具有马尔可夫性质的离散时间随机过程，是用于描述随机过程统计特征的概率模型。
- 具有一组已经设置好的参数，它们可以很好地解释特定类别中的样本。在使用时，一个测试样本被归类为能产生最大后验概率的模型对应的那个类别。

3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

隐马尔可夫模型

考虑: 对于连续时间内的一系列状态, 设 $\omega(t)$ 表示 t 时刻的状态, 那么一个长度为 T 的特定状态序列可设为

$$\omega^T = \{\omega(1), \omega(2), \dots, \omega(T)\}$$

- 系统可以在不同的步骤中重新访问一个状态。

例如: $\omega^6 = \{\omega_1, \omega_4, \omega_2, \omega_2, \omega_1, \omega_4\}$.

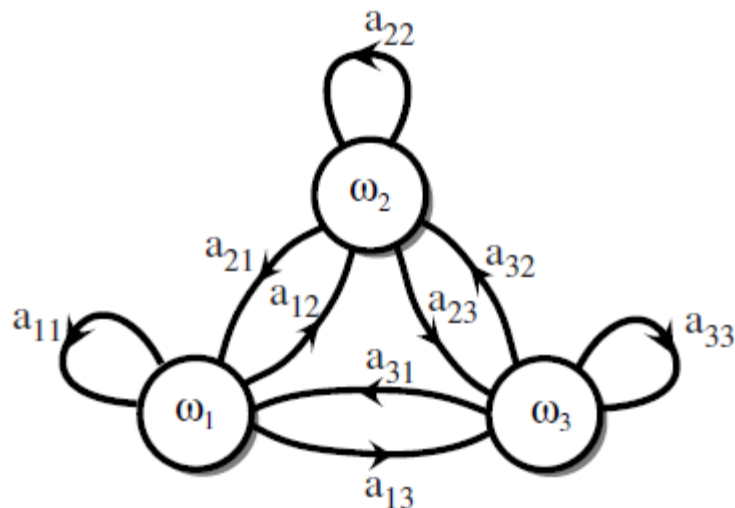
转移概率: 即系统在某一时刻 t 处于状态 ω_i 的条件下, 在时间 $t+1$ 时变为状态 ω_j 的概率, 该概率与具体的时刻无关, 记为

$$P(\omega_j(t+1)|\omega_i(t)) = a_{ij}$$

- 没有要求转移概率是对称的 (通常 $a_{ij} = a_{ji}$)
- 一个特定的状态可能会被连续访问

3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

隐马尔可夫模型



基本的马尔可夫模型

- 节点代表离散状态 ω_i
- 连线代表转移概率 a_{ij}

假设 已知一个特定的马尔可夫模型 θ (转移概率 a_{ij} 的完整集合) 和一个特定状态序列 ω^T 。

若要该马尔可夫模型生成特定序列的概率，只需将连续的状态转移概率相乘即可。

例如，求某个特定马尔可夫模型生成序列 $\omega^6 = \{\omega_1, \omega_4, \omega_2, \omega_2, \omega_1, \omega_4\}$ 的概率：

$$P(\omega_j(t+1) | \omega_i(t)) = a_{ij}$$

$$P(\omega^T | \theta) = a_{14} a_{42} a_{22} a_{21} a_{14}$$

- 如果初始状态 $P(\omega(1) = \omega_i)$ 有一个先验概率，我们也可以包括这个因子。

一阶离散时间的马尔可夫模型: $t + 1$ 时刻的概率只取决于 t 时刻的状态。

3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

隐马尔可夫模型

在生成语音的马尔可夫模型中，状态代表音素。然而，在语音识别中，接收器只能测量声音的特性，无法直接测量音素。考虑扩充马尔可夫模型：

可见状态：可直接进行外部测量的状态，记为 $v(t)$ 。

隐状态：不能被直接测量得到的内部状态，记为 $\omega(t)$ 。

假设在某一时刻 t ，系统处于隐状态 $\omega(t)$ ，同时，系统激发特定可见符号 $v(t)$ 。

- 马尔可夫模型允许激发的可见状态为连续函数（比如功率谱），本课程只考虑离散符号的情形。

考虑：与状态一样，我们定义了一个特定的可见状态序列，记为 $V^T = \{v(1), v(2), \dots, v(T)\}$ 。

例如： $V^6 = \{v_5, v_1, v_1, v_5, v_2, v_3\}$ 。

发射概率：在 某 t 时刻的状态 $\omega(t)$ 下，可见状态 $v_k(t)$ 激发的概率，该概率同样与具体时刻无关，记为：

$$p(v_k(t) | \omega_j(t)) = b_{jk}.$$

3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

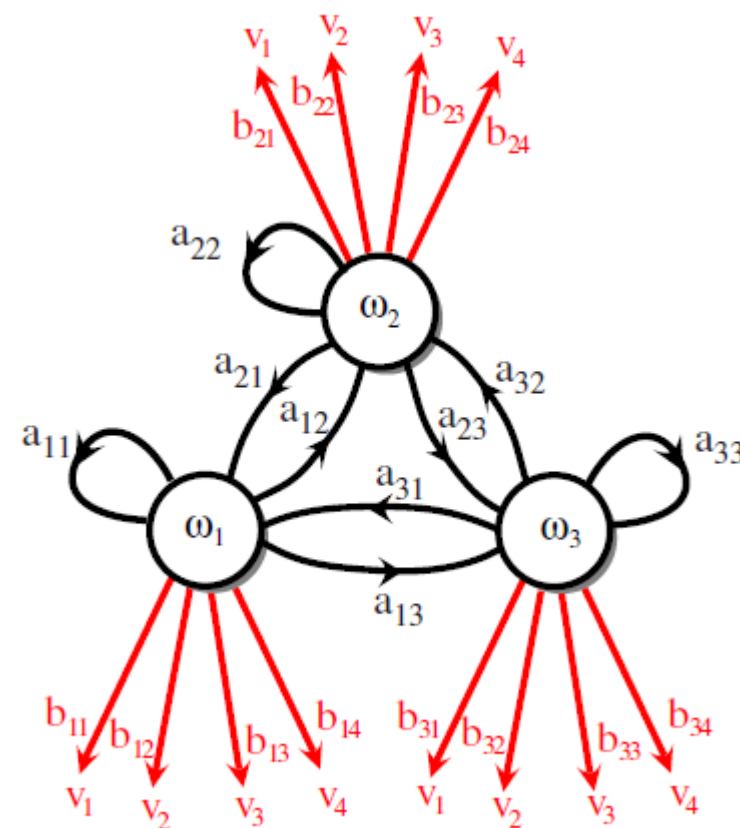
隐马尔可夫模型

因为我们只能观测到可见的状态，而不能直接知道 ω_j 处于什么内部状态，故模型就被称为“**隐马尔可夫模型**”。

转移概率: $P(\omega_j(t+1)|\omega_i(t)) = a_{ij}$ ($\sum_j a_{ij} = 1$ 对于所有的 i)

发射概率: $p(v_k(t)|\omega_j(t)) = b_{jk}$ ($\sum_k b_{jk} = 1$ 对于所有的 j)

- 每一时刻都必须转移到下一时刻，同时，发出一个可见符号，故有归一化条件。
- 马尔可夫网络是严格因果的，下一时刻状态的概率只取决于上一时刻的状态。
- 最终状态或吸收状态 ω_0 指系统一旦进入这个状态，就再也无法离开 (即: $a_{00} = 1$)。
- 该模型展示了任何方式的状态转移都是可能的，然而在一般的隐马尔可夫模型中，这种的任意的状态转移并不能得到保证。



隐马尔可夫模型

黑色连线表示隐状态之间的转移概率，红色的字符表示在每一隐状态产生的可见状态。

3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

隐马尔可夫模型的三个核心问题

估值问题： 假设有一个HMM，转移概率 a_{ij} 和 b_{jk} 已知，计算该模型生成的特定的可见状态序列 V^T 的概率。

解码问题： 假设有一个HMM及一组观察值 V^T ，决定最有可能产生这些观察结果的隐状态序列 ω^T 。

学习问题： 假设已知模型的大致结构（比如隐状态数和可见状态数），但没有给出转移概率 a_{ij} 和 b_{jk} ，如何从给定的一组训练样本中确定这些参数。

3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

估值问题

隐马尔可夫模型产生的可见状态序列 V^T 的概率为：

$$P(V^T) = \sum_{r=1}^{r_{max}} P(V^T | \omega_r^T) P(\omega_r^T)$$

其中， r 是每个特定长度为 T 的隐状态序列的下标：

$$\omega_r^T = \{\omega(1), \omega(2), \dots, \omega(T)\}$$

- 在有 c 个不同隐状态的情况下， $P(V^T)$ 共有 $r_{max} = c^T$ 项
- 为了计算模型产生特定的可见状态序列 V^T 的概率，应该考虑每一种可能的隐状态序列，计算它们生成 V^T 的概率，然后将这些概率相加。
- 特定可见序列的概率就是对应（隐）转移概率 a_{ij} 和（可见）发射概率 b_{jk} 的乘积结果。

3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

估值问题

$$P(V^T) = \sum_{r=1}^{r_{max}} P(V^T | \omega_r^T) P(\omega_r^T)$$

① 描述隐状态转移概率的第二项 $P(\omega_r^T)$ 可以改写为：

$$P(\omega_r^T) = \prod_{t=1}^T P(\omega(t) | \omega(t-1))$$

- $P(\omega_r^T)$ 实际上就是 a_{ij} 的乘积。
- $\omega(T) = \omega_0$ 表示最终的吸收态，它产生唯一独特的可见符号 v_0 。比如在语音识别应用中， ω_0 通常表示零状态或者无话语，而 v_0 则表示静音。

② 可设每个时刻发出可见符号的概率仅取决于这个时刻的隐状态，因此，可将第一项写为：

$$P(V^T | \omega_r^T) = \prod_{t=1}^T P(v(t) | \omega(t))$$

3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

估值问题

$$P(V^T) = \sum_{r=1}^{r_{max}} P(V^T | \omega_r^T) P(\omega_r^T) \quad \longrightarrow \quad P(\mathbf{v}^T) = \sum_{r=1}^{r_{max}} \prod_{t=1}^T P(v(t) | \omega(t)) P(\omega(t) | \omega(t-1))$$

- 我们观察到某一特定可见状态序列 V^T 的概率等于所有可能产生这个可见状态序列的隐状态序列的情况的相加，而每一种可能的隐状态序列的情况发生的概率都是隐状态之间转移概率和产生可见符号发射概率依次相乘得到。
- 估值问题的计算复杂度是 $O(c^T T)$ ，这么大的计算量在实际上是非常不现实的。

比如，当 $c=10$, $T=20$ 时，我们需要进行 10^{21} 次基本运算！

3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

估值问题的前向算法

$$P(\mathbf{V}^T) = \sum_{r=1}^{r_{\max}} \prod_{t=1}^T P(v(t) | \omega(t)) P(\omega(t) | \omega(t-1))$$

思路： 我们可以递归计算 $P(\mathbf{V}^T)$ ，因为每一项 $P(v(t) | \omega(t)) P(\omega(t) | \omega(t-1))$ 均只涉及 $v(t)$ ， $\omega(t)$ ， $\omega(t-1)$ 三项。

定义：

$$\alpha_j(t) = \begin{cases} 0 & t = 0 \text{ 且 } j \neq \text{初始状态} \\ 1 & t = 0 \text{ 且 } j = \text{初始状态} \\ \sum_i \alpha_i(t-1) a_{ij} b_{jk} v(t) & \text{其他} \end{cases}$$

- $b_{jk} v(t)$ 表示由 t 时刻的可见状态 $v(t)$ 确定的发射概率 b_{jk} 。
- $\alpha_j(t)$ 表示HMM在 t 时刻位于隐状态 ω_j 且已产生了可见状态序列 \mathbf{V}^T 前 t 个符号的概率。

3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

估值问题的前向算法

$$P(\mathbf{V}^T) = \sum_{r=1}^{r_{\max}} \prod_{t=1}^T P(v(t) | \omega(t)) P(\omega(t) | \omega(t-1))$$

HMM Forward Algorithm

```
1 initialize  $\omega(1), t = 0, a_{ij}, b_{jk}$ , visible sequence  $\mathbf{V}^T, \alpha(0) = 1$   
2 for  $t \leftarrow t + 1$   
3      $\alpha_j(t) \leftarrow \sum_{i=1}^c \alpha_i(t-1) a_{ij} b_{jk}$   
4 until  $t = T$   
5 return  $P(\mathbf{V}^T) \leftarrow \alpha_0(T)$   
6 end
```

$$\alpha_i(t) = \begin{cases} 0 & t = 0 \text{ 且 } j \neq \text{初始状态} \\ 1 & t = 0 \text{ 且 } j = \text{初始状态} \\ \sum_i \alpha_i(t-1) a_{ij} b_{jk} v(t) & \text{其他} \end{cases}$$

- 在第5行中， α_0 表示序列的结束。
- 前向算法的计算复杂度为 $O(c^2T)$ ，这比穷举法的效率要高得多。如果同样当 $c = 10, T = 20$ 的情况，前向算法只需要执行2000次操作，这几乎比穷举法快 10^{17} 倍！

3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

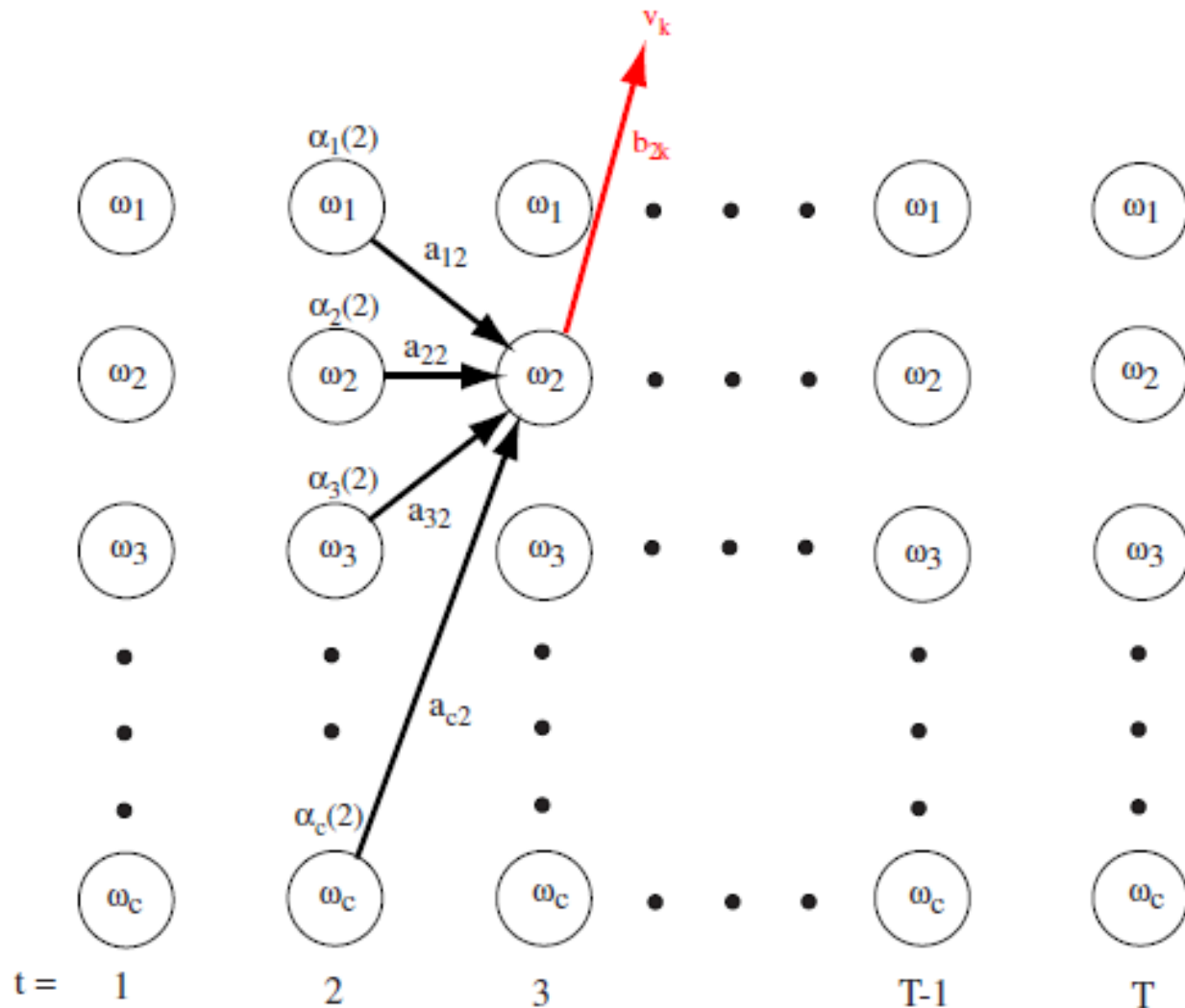
估值问题的前向算法

求 $\alpha_2(3)$ ($t = 3$ 时, 系统位于状态 ω_2 并生成规定可见状态序列的概率)

在 $t = 2$ 时且位于状态 ω_i 的概率为 $\alpha_i(2)$, 其中 $i = 1, 2, \dots, c$ 。为了求 $\alpha_2(3)$, 必须把这些项相加, 同时乘以发出字符 v_k 的概率,

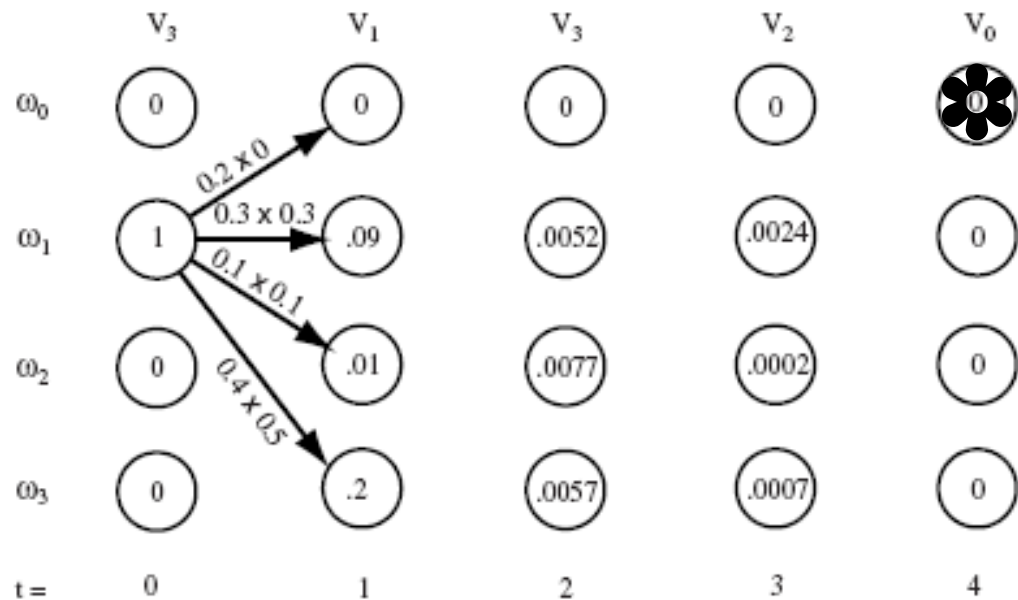
即:
$$\alpha_2(3) = b_{2k} \sum_i \alpha_i(2) a_{i2}$$

$$\alpha_i(t) = \begin{cases} 0 & t = 0 \text{ 且 } j \neq \text{初始状态} \\ 1 & t = 0 \text{ 且 } j = \text{初始状态} \\ \sum_i \alpha_i(t-1) a_{ij} b_{jk} v(t) & \text{其他} \end{cases}$$



前向算法网格说明: 每一项是按时间对HMM的“展开”。

3.8 隐马尔可夫模型与维特比 (Viterbi) 方法



$$a_{ij} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.2 & 0.3 & 0.1 & 0.4 \\ 0.2 & 0.5 & 0.2 & 0.1 \\ 0.8 & 0.1 & 0.0 & 0.1 \end{pmatrix} \text{ and } b_{jk} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0.3 & 0.4 & 0.1 & 0.2 \\ 0 & 0.1 & 0.1 & 0.7 & 0.1 \\ 0 & 0.5 & 0.2 & 0.1 & 0.2 \end{pmatrix}.$$

- The calculation requires a sum over all hidden states at the previous time.

$$\alpha_j(t) = \begin{cases} 0 & t = 0 \text{ and } j \neq \text{initial state} \\ 1 & t = 0 \text{ and } j = \text{initial state} \\ \sum_i \alpha_i(t-1) a_{ij} b_{jk} v(t) & \text{otherwise} \end{cases}$$

- The probability shown in the final (absorbing) state gives the probability of the full sequence observed, $P(V^T | \theta) = ?$.

3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

估值问题

如果把隐马尔可夫模型中的转移概率和发射概率 (a 和 b) 采用参数向量 θ 表示, 那么根据贝叶斯公式, 在已知观测序列的情况下, 模型的概率为:

$$P(\theta | V^T) = \frac{P(V^T | \theta)P(\theta)}{P(V^T)}$$

在隐马尔可夫模式识别中, 我们可能会有多个HMM, 每个模型代表一个类别。对测试样本进行分类, 就是计算哪一个模型产生这个测试样本的概率最大。

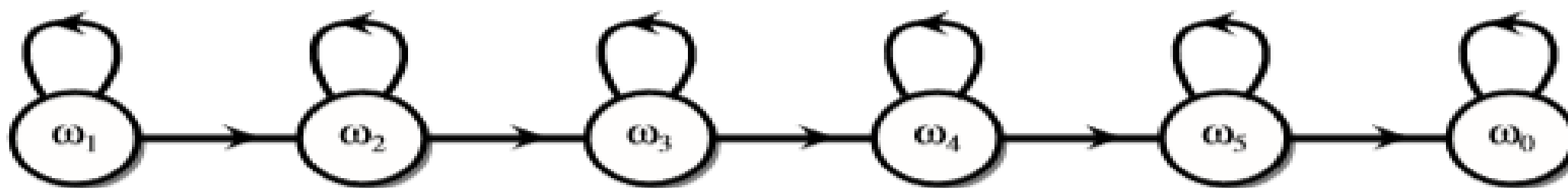
- 前向算法使我们能够计算 $P(V^T | \theta)$
- 模型的先验概率 $P(\theta)$ 由外部的知识确定 (比如在语音识别中, 可能是一个语言模型), 这个先验概率可能依赖于上下文语义, 或者是前面的单词等。

3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

估值问题

在语音识别领域，通常使用一个从左向右的隐马尔可夫模型。实际上，几乎所有的隐马尔可夫模型都是从左向右递推的模型。

例如，在隐马尔可夫语音识别中，我们有两个模型，其中一个用来产生发音“stand”，另一个用来产生发音“plate”。现在有一个用来测试的未知发音，需要确定哪个模型产生该发音的可能性更大。



这样一个模型可以描述发音“stand”，其中， ω_1 代表音素/s/， ω_2 代表音素/t/...，直到 ω_0 代表最终状态。

3.8 隐马尔可夫模型与维特比 (Viterbi) 方法

解码问题

假设有一个HMM及一组观察值 V^T ，决定最有可能产生这些观察结果的隐状态序列 ω^T 。

3.8 隐马尔可夫过程与维特比 (Viterbi) 方法

维特比 (Viterbi) 方法

“最有可能”（概率最大）的隐状态序列： $\omega^{T*} = \operatorname{argmax}_{\omega^T} P(\omega^T | V^T, \theta)$

设有Viterbi变量： $\delta_i(t) = \max_{\omega^t} P(\omega^t, V^t | \theta)$

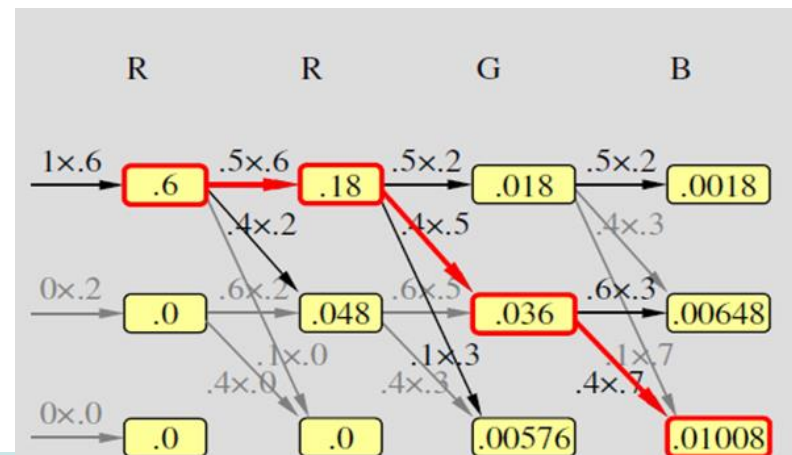
思想：利用动态规划求解，复杂性 $O(c^2T)$ 。

递归关系： $\delta_j(t+1) = \max_i \delta_i(t) a_{ij} b_{jk} \quad v(t+1) = \max_i [\delta_i(t) a_{ij}] b_{jk} \quad v(t+1)$

$$\varphi_j(t+1) = \operatorname{argmax}_i [\delta_i(t) a_{ij}]$$

记忆变量： $\varphi_i(t)$ 记录概率最大路径上当前状态的前一个状态。

目标：找到T时刻最大的 $\delta_i(T)$ 代表的那个隐状态序列。



3.8 隐马尔可夫过程与维特比 (Viterbi) 方法

维特比 (Viterbi) 方法

HMM Viterbi Algorithm

1. *begin initialize* $\delta_j(1) = \beta_j b_{jk}, \varphi_j(1)=0$
2. *do* $j \leftarrow j+1, t \leftarrow t+1$
3. *compute* $\delta_j(t), \varphi_j(t)$
4. *until* $j=c, t=T$
5. *Return* $\omega^{T*} \leftarrow \arg \max_j [\delta_j(T)]$
6. *end*

$$\begin{aligned}\delta_j(t+1) &= \max_i \delta_i(t) a_{ij} b_{jk} v(t+1) \\ &= \max_i [\delta_i(t) a_{ij}] b_{jk} v(t+1)\end{aligned}$$

$$\varphi_j(t+1) = \arg \max_i [\delta_i(t) a_{ij}]$$

3.8 隐马尔可夫过程与维特比 (Viterbi) 方法

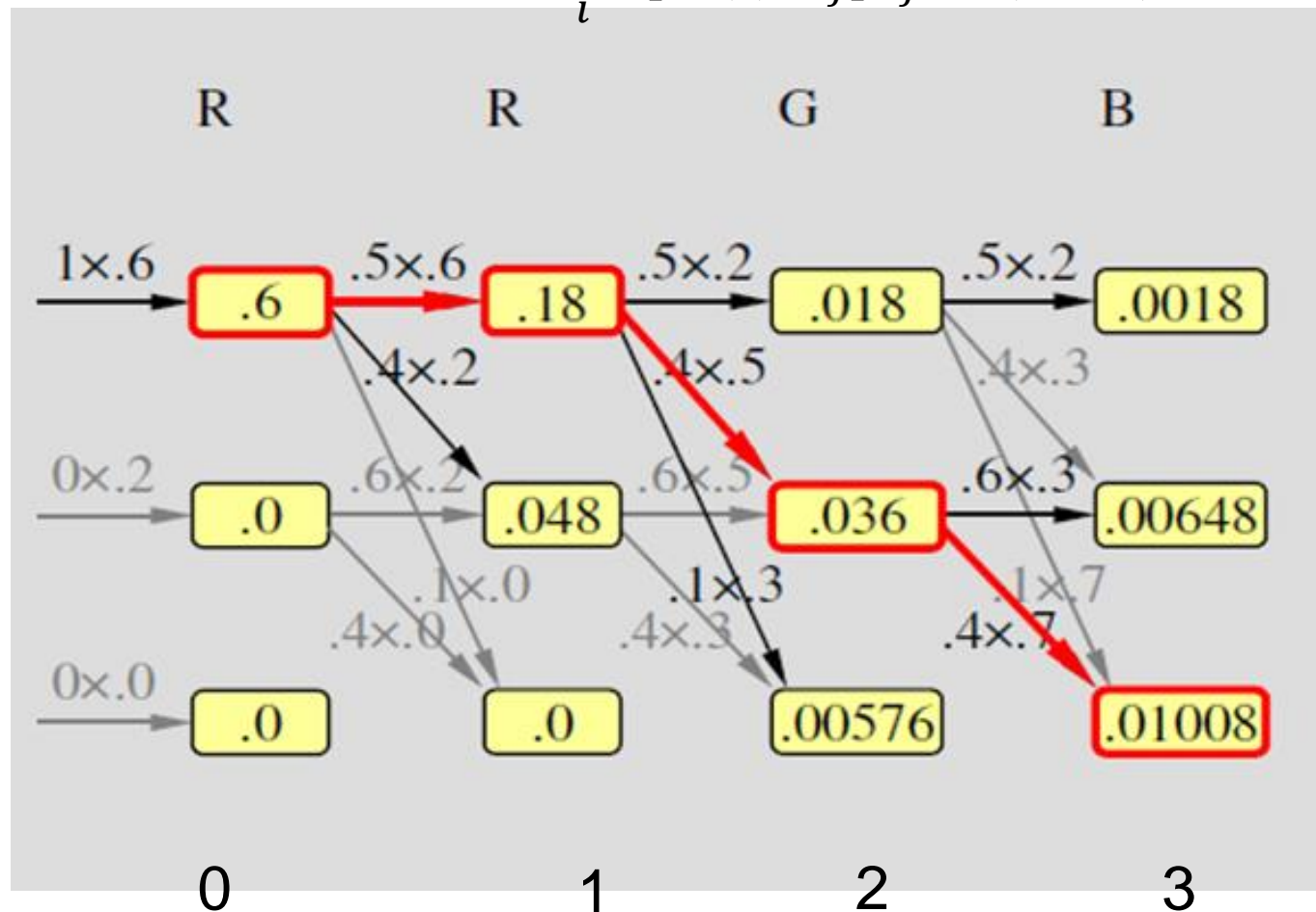
维特比 (Viterbi) 方法

已知 $t = 0$ 时刻, 系统的初始隐状态为 ω_1 , $\delta_j(t)$ 在每个单元内表示。

$$a = \begin{bmatrix} 0.5 & 0.4 & 0.1 \\ 0 & 0.6 & 0.4 \\ 0 & 0 & 0 \end{bmatrix}$$

$$b = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.2 & 0.5 & 0.3 \\ 0 & 0.3 & 0.7 \end{bmatrix}$$

$$\begin{aligned} \delta_j(t+1) &= \max_i \delta_i(t) a_{ij} b_{jk} v(t+1) \\ &= \max_i [\delta_i(t) a_{ij}] b_{jk} v(t+1) \end{aligned}$$





Ending

