模式识别

Pattern Recognition

第6讲 统计分类器及其学习II

哈尔滨工业大学 计算机学院
模式识别与智能系统研究中心
ISBN 978-7-5603-4763-9

引言

进行Bayes决策需要事先知道两种知识：

- 各类的先验概率；
- 观测向量的类条件概率密度。

知识的获取（估计）：

- 一些训练数据；
- 对问题的一般性的认识

ISBN 978-7-5603-4763-9

引言

类的先验概率的估计（较容易）：

- 依靠经验；
- 用训练数据中各类出现的频率估计。
- 用频率估计概率的优点：
  - 无偏性；
  - 相合性；
  - 收敛速度快。

ISBN 978-7-5603-4763-9

引言

类条件概率密度的估计（非常难）：

- 概率密度函数包含了一个随机变量的全部信息；
- 概率密度函数可以是满足下面条件的任何函数：

$$p(\mathbf{x}) \geq 0 \qquad \int p(\mathbf{x})d\mathbf{x} = 1$$

- 问题可以表示为：已有c个类别的训练样本集合D1，D2，…，Dc，求取每个类别的类条件概率密度 $p(\mathbf{x}|\omega_i)$。

ISBN 978-7-5603-4763-9

引言

概率密度估计的两种主要思路：

参数估计：

根据对问题的一般性的认识，假设随机变量服从某种分布，分布函数的参数通过训练数据来估计。

例如，假定x服从正态分布$N(\mu,\Sigma)$，要估计的参数就是$\theta=(\mu,\Sigma)$

非参数估计：

不用模型，而只利用训练数据本身对概率密度做估计。

-----K近邻分类器

ISBN 978-7-5603-4763-9

本节主要内容

- 非参数估计
- 参数估计
- 混合高斯模型
- 隐马尔科夫模型

非参数估计的基本思想

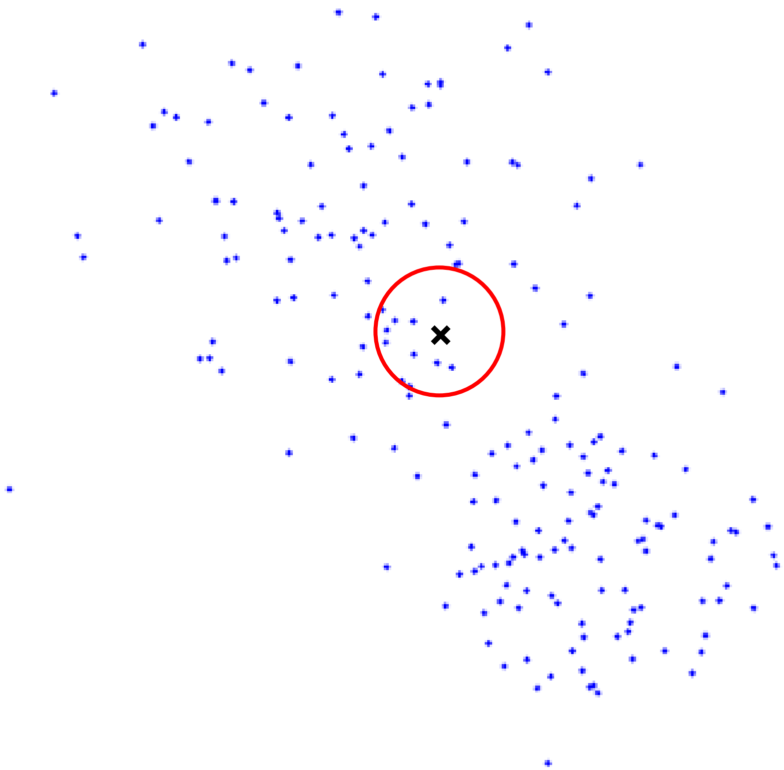
- 令R是包含样本点x的一个区域，其体积为V，设有n个训练样本，其中有k个落在区域R中，则可对概率密度作出一个估计：

$$p(\mathbf{x}) \approx \frac{k/n}{V}$$

相当于用**R**区域内的平均性质来作为一点x的估计，是一种数据的平滑。

非参数估计基本思想

根据训练样本，直接估计概率密度



当n固定时，V的大小对估计的效果影响很大：
过大则平滑过多，不够精确；
过小则可能导致在此区域内无样本点，k=0。

此方法的有效性取决于样本数量的多少，以及区域体积选择的合适。

哈尔滨工业大学 计算机学院
模式识别与智能系统研究中心
ISBN 978-7-5603-4763-9

区域选定的两个途径

- Parzen窗法：区域体积V是样本数n的函数，如：

$$V_n = \frac{1}{\sqrt{n}}$$

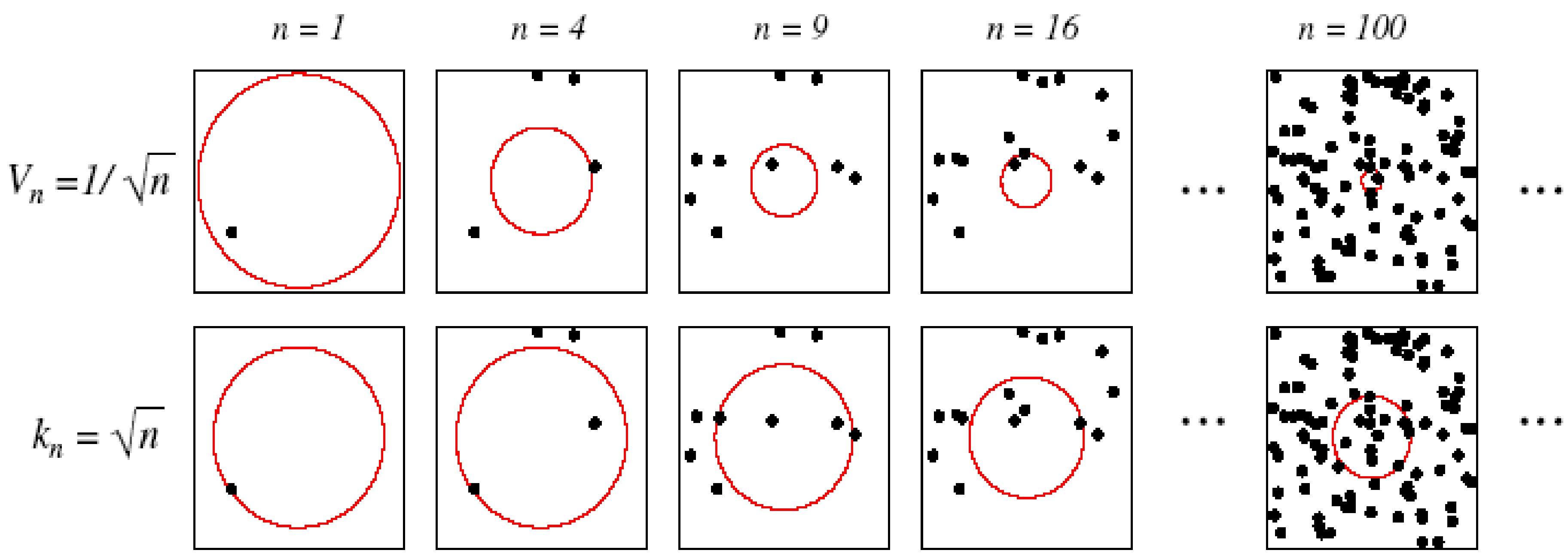
估计类条件概率密度

K-近邻法：落在区域内的样本数k是总样本数n的函数，如：

$$k_n = \sqrt{n}$$

直接估计后验概率

Parzen窗法和K-近邻法



ISBN 978-7-5603-4763-9

K近邻的估计原理

将一个体积为V的区域放到待识样本点x周围，包含k个训练样本点，其中k_i个属于ω_i类，总的训练样本数为n，则有：

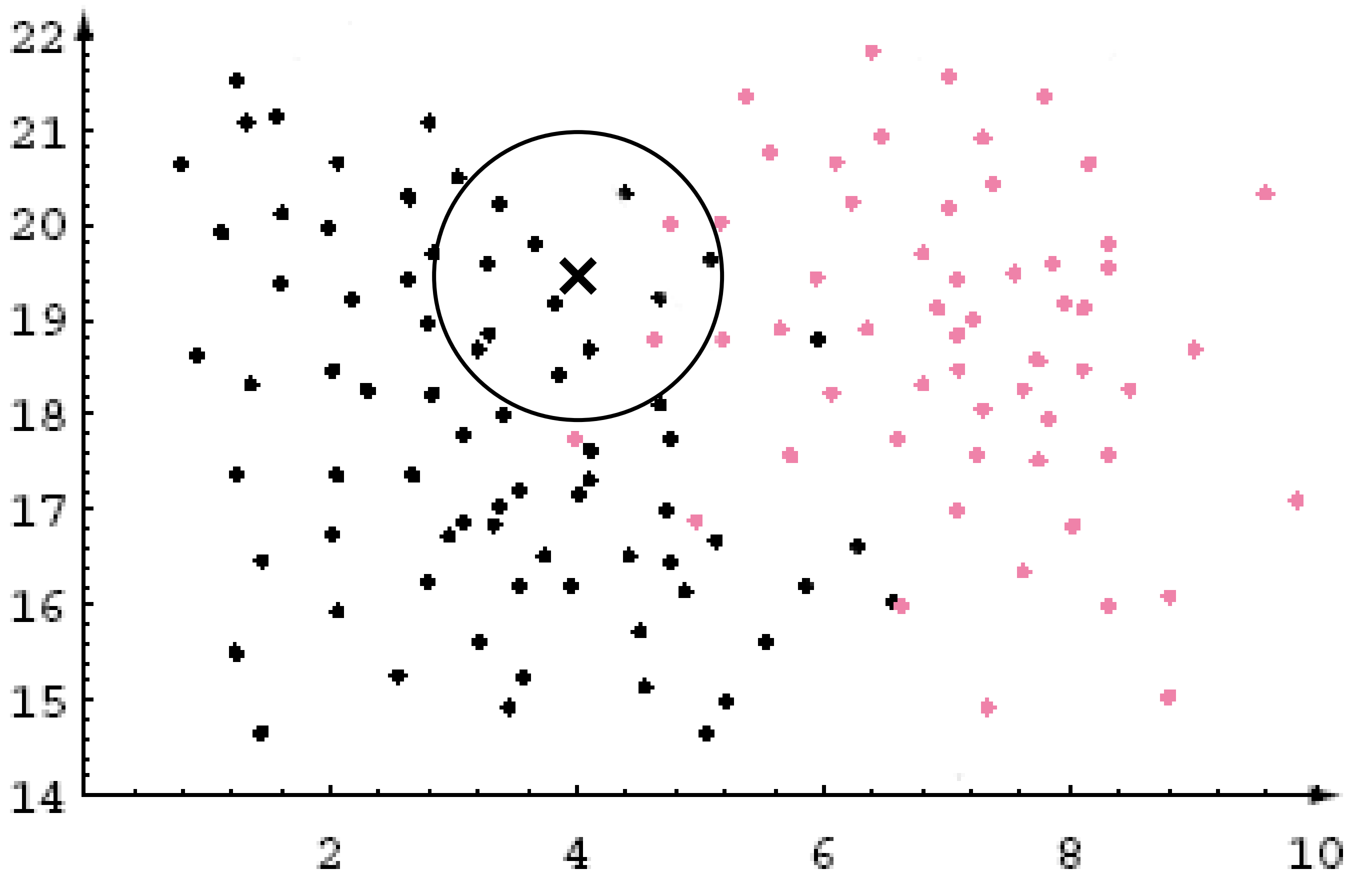
$$p_n(\mathbf{x},\omega_i)=\frac{k_i/n}{V}$$

$$p(\omega_i|\mathbf{x})=\frac{p_n(\mathbf{x},\omega_i)}{p_n(\mathbf{x})}=\frac{p_n(\mathbf{x},\omega_i)}{\sum_{j=1}^{c}p_n(\mathbf{x},\omega_j)}=\frac{k_i}{k}$$

k-近邻分类算法

1. 设置参数k，输入待识别样本$\mathbf{x}$；
2. 计算x与每个训练样本的距离；
3. 选取距离最小的前k个样本，统计其中包含各个类别的样本数k_i；
4. $class \leftarrow \arg\max_{1 \le i \le c} k_i$

k-近邻分类，k=13



ISBN 978-7-5603-4763-9

本节主要内容

非参数估计

参数估计

高斯混合模型

隐马尔科夫模型

哈尔滨工业大学 计算机学院
模式识别与智能系统研究中心
ISBN 978-7-5603-4763-9

概率密度函数的参数估计方法

预先假设每一个类别的概率密度函数的形式已知，而具体的参数未知；

- 最大似然估计(MLE, Maximum Likelihood Estimation)；
- 贝叶斯估计(Bayesian Estimation)。

ISBN 978-7-5603-4763-9

似然函数

样本集D中包含n个样本：$\mathbf{x}_1$，$\mathbf{x}_2$，…, $\mathbf{x}_n$，样本都是独立同分布的随机变量，样本集D出现的概率为：

$$p\left(D\middle|\boldsymbol{\theta}\right)=p\left(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n\middle|\boldsymbol{\theta}\right)=\prod_{i=1}^{n}p\left(\mathbf{x}_i\middle|\boldsymbol{\theta}\right)$$

定义对数似然函数：

$$l\left(\boldsymbol{\theta}\right)\equiv\ln p\left(D\middle|\boldsymbol{\theta}\right)=\sum_{i=1}^{n}\ln p\left(\mathbf{x}_i\middle|\boldsymbol{\theta}\right)$$

最大似然估计

- 寻找到一个最优矢量 $\hat{\boldsymbol{\theta}}$，使得似然函数 $l(\boldsymbol{\theta})$最大。

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

$$l(\boldsymbol{\theta}) \equiv \ln p(D|\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln p(\mathbf{x}_i|\boldsymbol{\theta})$$

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

$$\nabla_{\boldsymbol{\theta}} l = 0$$

ISBN 978-7-5603-4763-9

正态分布的似然估计

- Gauss分布的参数由均值矢量 $\boldsymbol{\mu}$ 和协方差矩阵 $\boldsymbol{\Sigma}$ 构成，最大似然估计结果为：

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

$$\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\underline{\left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}\right)\left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}\right)^t}$$

ISBN 978-7-5603-4763-9

设样本集$X=\{x_1,\cdots,x_n\}$服从 **Rayleigh** 分布：

$$p(x|\theta)=\begin{cases}2\theta x e^{-\theta x^2}, & x\ge 0\\ 0, & x<0\end{cases}$$

其中θ是一个未知参数，试推导参数θ的最大似然估计。

构造对数似然函数：

$$l(\theta)=\sum_{i=1}^{n}\ln p(x_i|\theta)=\sum_{i=1}^{n}\left(\ln 2+\ln\theta+\ln x_i-\theta x_i^2\right)$$

求对数似然函数关于θ的极值：

$$\frac{dl(\theta)}{d\theta}=\sum_{i=1}^{n}\left(\frac{1}{\theta}-x_i^2\right)=0$$，解得：$$\theta=\frac{1}{\frac{1}{n}\sum_{i=1}^{n}x_i^2}$$

贝叶斯估计

已有独立同分布训练样本集D；

已知类条件概率密度函数p(x|θ)的形式，但参数θ未知；

已知参数θ的先验概率密度函数p(θ)；

求在已有训练样本集D的条件下，类条件概率密度函数p(x|D)。

ISBN 978-7-5603-4763-9

贝叶斯估计与最大似然估计的差别

最大似然估计： θ是一个确定的未知矢量；

贝叶斯估计： θ是一个随机变量，θ以一定的概率分布 $p(\theta)$ 所有可能的值

ISBN 978-7-5603-4763-9

贝叶斯估计的一般理论

类条件概率：

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

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

X与D独立

参数θ的后验分布：

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

归一化系数

贝叶斯估计的一般理论

学习过程：计算参数的后验分布：

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

分类过程：将待识模式x和后验概率，计算x发生的概率

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

举例：单变量正态分布的贝叶斯估计

- 已知概率密度函数满足正态分布，其中方差 σ^2已知，均值 μ 未知，假设 μ 的先验概率满足正态分布，即：

$$p(x|\mu) \sim N(\mu, \sigma^2) \qquad p(\mu) \sim N(\mu_0, \sigma_0^2)$$

在已知训练样本集合$D=\{x_1,x_2,\ldots,x_n\}$的条件下，
估计x的概率密度函数：

$$p(x|D) = \int p(x|\mu)\, p(\mu|D)\, d\mu$$

独立同分布

其中：$p(\mu|D) = \dfrac{p(D|\mu)\,p(\mu)}{\int p(D|\mu)\,p(\mu)\,d\mu} = \alpha \prod_{i=1}^{n} p(x_i|\mu)\,p(\mu)$

归一化系数

计算在训练样本集D的条件下，参数μ的分布：

$$p(\mu|D)=\frac{p(D|\mu)p(\mu)}{\int p(D|\mu)p(\mu)d\mu}=\alpha\underbrace{\prod_{i=1}^{n}p(x_i|\mu)}\underbrace{p(\mu)}$$

$$=\underbrace{\alpha\prod_{k=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right]}\times\underbrace{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]}$$

$$=\alpha''\exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2-2\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i+\frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right]$$

$p(\mu|D)$是指数函数，且指数部分是μ的二次型，因此是正态分布

$$写作：p(\mu|D)\sim N(\mu_n,\sigma_n)=\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]$$

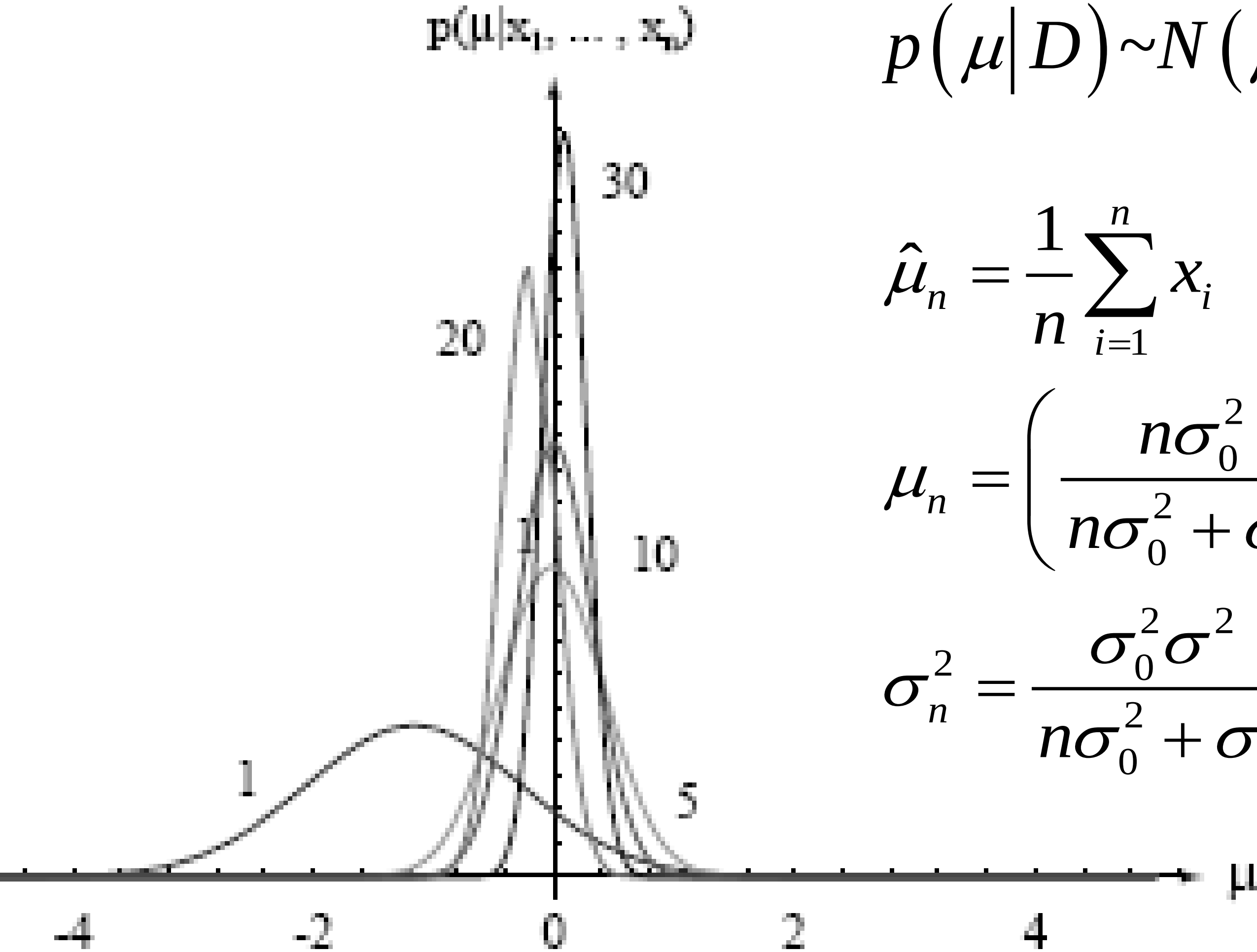
根据对应项相等，可求得μ_n，σ_n。在样本数n增加时，$p(\mu|D)$仍然保持正态分布称为复制密度函数，$p(\mu)$称为共轭先验

根据对应项相等，求得$p\left(\mu\middle|D\right)\sim N\left(\mu_n,\sigma_n\right)$：

$$\begin{cases} p\left(\mu\middle|D\right)=\alpha''\exp\left[-\dfrac{1}{2}\left[\left(\dfrac{n}{\sigma^2}+\dfrac{1}{\sigma_0^2}\right)\mu^2-2\left(\dfrac{1}{\sigma^2}\sum_{i=1}^{n}x_i+\dfrac{\mu_0}{\sigma_0^2}\right)\mu\right]\right] \\ p\left(\mu\middle|D\right)=\beta\exp\left[-\dfrac{1}{2}\left(\dfrac{\mu-\mu_n}{\sigma_n}\right)^2\right] \\ \hat{\mu}_n=\dfrac{1}{n}\sum_{i=1}^{n}x_i \end{cases}$$

$$\Rightarrow\begin{cases} \mu_n=\left(\dfrac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\hat{\mu}_n+\dfrac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0 \\ \sigma_n^2=\dfrac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2} \end{cases}$$

均值分布的变化



$$p(\mu|D) \sim N(\mu_n, \sigma_n)$$

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

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

类条件概率密度的计算

$$p(x|D)=\int p(x|\mu)\,p(\mu|D)\,d\mu$$

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

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

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

$p(x|D)$是正态分布，均值为μ_n，方差为$\sigma^2+\sigma_n^2$

$$\hat{\mu}_n=\frac{1}{n}\sum_{i=1}^{n}x_i \qquad \mu_n=\left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\hat{\mu}_n+\frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0 \qquad \sigma_n^2=\frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}$$

递归贝叶斯学习

$D^n = \{x_1, \cdots x_n\}$表示有$n$个样本的样本集

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

$$p\left(\boldsymbol{\theta} \middle| D^n\right) = \frac{p\left(x_n \middle| \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \middle| D^{n-1}\right)}{\int p\left(x_n \middle| \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \middle| D^{n-1}\right) d\boldsymbol{\theta}}$$

随着训练样本的增加，能够产生一系列概率密度函数

$p\left(\boldsymbol{\theta}\right), p\left(\boldsymbol{\theta} \middle| \mathbf{x}_1\right), p\left(\boldsymbol{\theta} \middle| \mathbf{x}_1, \mathbf{x}_2\right), \cdots$

举例：均匀分布的递归贝叶斯学习

一维样本服从均匀分布如下，参数θ值未知

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

假设θ服从均匀分布$U(0,10)$,根据训练样本集$D=\{4,7,7,8\}$

估计θ及概率密度$p(x)$

$$D^0: \text{无样本} \quad p(\theta|D^0) = p(\theta) = U(0,10)$$

$$D^1: x_1 = 4 \quad p(\theta|D^1) \propto p(x|\theta)\, p(\theta|D^0) = \begin{cases} 1/\theta & 4 \le \theta \le 10 \\ 0 & \text{其他} \end{cases}$$

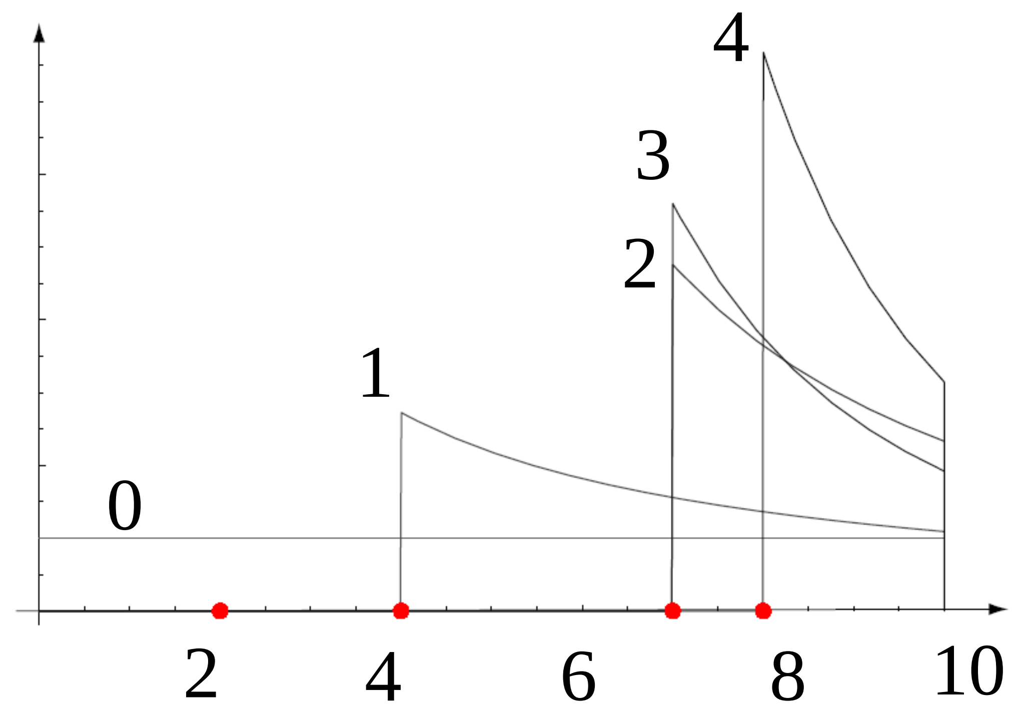
$$D^2: x_2 = 7 \quad p(\theta|D^2) \propto p(x|\theta)\, p(\theta|D^1) = \begin{cases} 1/\theta^2 & 7 \le \theta \le 10 \\ 0 & \text{其他} \end{cases}$$

$$D^1: x_1 = 4 \quad p\left(\theta \middle| D^1\right) \propto p\left(x \middle| \theta\right) p\left(\theta \middle| D^0\right) = \begin{cases} 1/\theta & 4 \le \theta \le 10 \\ 0 & \text{其他} \end{cases}$$

$$D^2: x_1 = 7 \quad p\left(\theta \middle| D^2\right) \propto p\left(x \middle| \theta\right) p\left(\theta \middle| D^1\right) = \begin{cases} 1/\theta^2 & 7 \le \theta \le 10 \\ 0 & \text{其他} \end{cases}$$

每次递归引入系数$1/\theta$，分布仅对大于最大值的区间非0，即：

$$p\left(\theta \middle| D^n\right) \propto \begin{cases} 1/\theta^n & \max_x\left[D^n\right] \le \theta \le 10 \\ 0 & else \end{cases}$$



$$p(x|\theta)=\begin{cases}1/\theta & 0\le x\le\theta\\ 0 & \text{其他}\end{cases} \qquad p(\theta|D^n)\propto\begin{cases}1/\theta^n & 8\le\theta\le 10\\ 0 & else\end{cases}$$

根据 $p(\theta|D^n)$，$p(x|\theta)$ 类条件件概率密度的贝叶斯估计：

$$p(\mathbf{x}|D)=\int p(\mathbf{x}|\boldsymbol{\theta})\,p(\boldsymbol{\theta}|D)\,d\boldsymbol{\theta}=\int_8^{10} p(\mathbf{x}|\boldsymbol{\theta})\,p(\boldsymbol{\theta}|D)\,d\boldsymbol{\theta}$$

当$0\le x\le 8$时，$p(x|\theta)=1/\theta$

$$p(\mathbf{x}|D)\propto\int_8^{10}\frac{1}{\theta}\frac{1}{\theta^n}d\boldsymbol{\theta}=\frac{n+1}{8^{n+2}}-\frac{n+1}{10^{n+2}}$$

当$8\le x\le 10$时, $p(x|\theta)=\begin{cases}1/\theta & 0\le x\le\theta\\ 0 & \text{其他}\end{cases}$

$$p(\mathbf{x}|D)\propto\int_8^{x}0\frac{1}{\theta^n}d\boldsymbol{\theta}+\int_x^{10}\frac{1}{\theta}\frac{1}{\theta^n}d\boldsymbol{\theta}=\frac{n+1}{x^{n+2}}-\frac{n+1}{10^{n+2}}$$

当$x<0$或$x>10$时$p(x|\theta)=0$

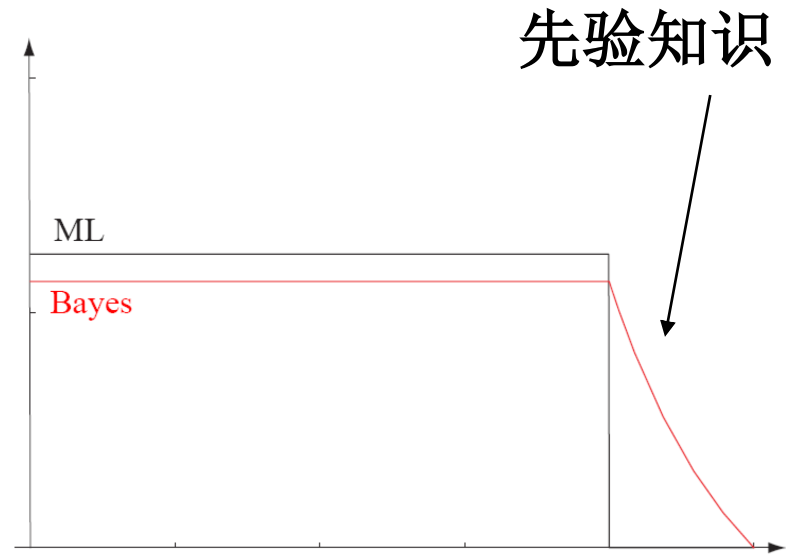
$$p(\mathbf{x}|D)=0$$

$$p(x|\theta)=\begin{cases}1/\theta & 0\le x\le\theta\\ 0 & \text{其他}\end{cases}$$

$$p(\theta|D^n)\propto\begin{cases}1/\theta^n & 8\le\theta\le 10\\ 0 & else\end{cases}$$

根据 $p(\theta|D^n)$，$p(x|\theta)$ 类条件概率密度的贝叶斯估计：

$$p(\mathbf{x}|D)\propto\begin{cases}\dfrac{n+1}{8^{n+2}}-\dfrac{n+1}{10^{n+2}} & 0\le x\le 8\\[2ex] \dfrac{n+1}{x^{n+2}}-\dfrac{n+1}{10^{n+2}} & 8\le x\le 10\end{cases}$$



$p(\mathbf{x}|D)$的最大似然估计：

似然函数$p(D|\boldsymbol{\theta})=\prod_{i=1}^{n}p(x_i|\boldsymbol{\theta})=\dfrac{1}{\theta^n}$

$$\max:\frac{1}{\theta^n}\qquad\Rightarrow\qquad \theta=\max\left[D^n\right]$$

$$\text{s.t.}:\quad \theta\ge\max\left[D^n\right]$$

本节主要内容

- 非参数估计
- 参数估计
- 高斯混合模型
- 隐马尔科夫模型

哈尔滨工业大学 计算机学院
模式识别与智能系统研究中心
ISBN 978-7-5603-4763-9

高斯混合模型

复杂的概率密度函数：可以由多个简单的密度函数混合构成：

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} a_k p_k(\mathbf{x}|\boldsymbol{\theta}_k), \qquad a_k > 0, \quad \sum_{k=1}^{K} a_k = 1$$

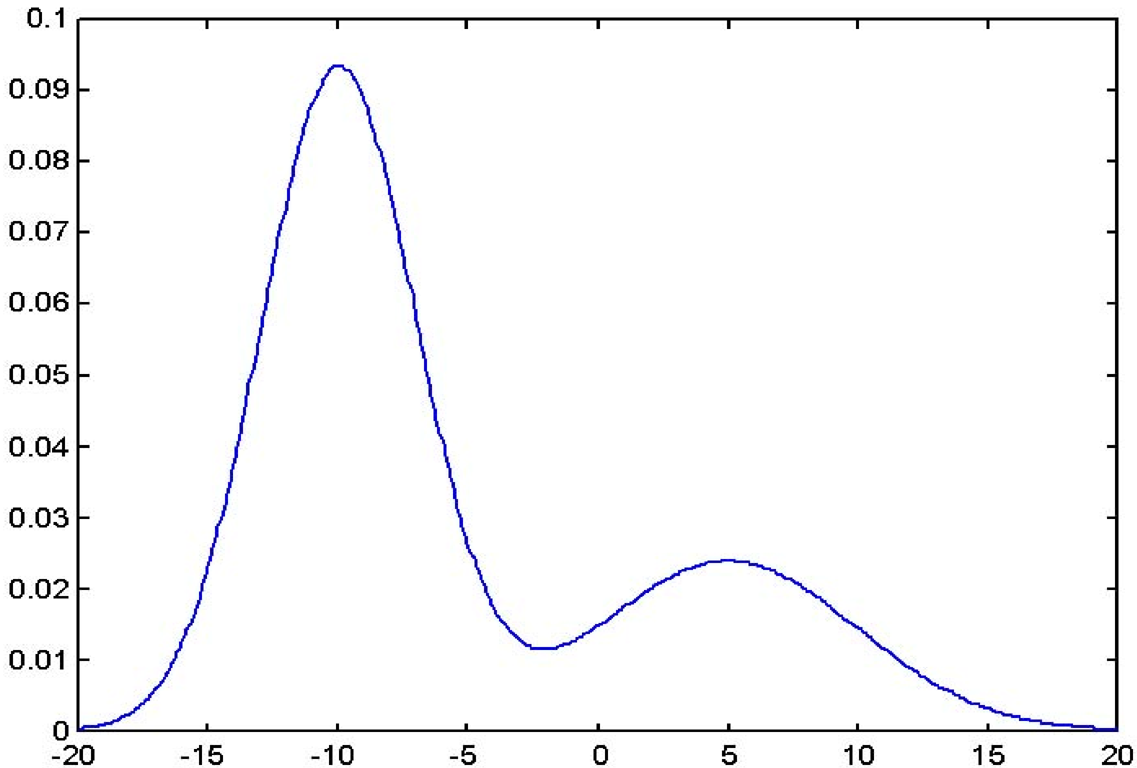
高斯混合模型(GMM，Gauss Mixture Model)：由多个高斯密度函数混合，用于逼近复杂的概率密度

$$p(\mathbf{x}) = \sum_{k=1}^{K} a_k N(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

一定条件下，GMM能够以任意精度逼近任意概率密度函数！

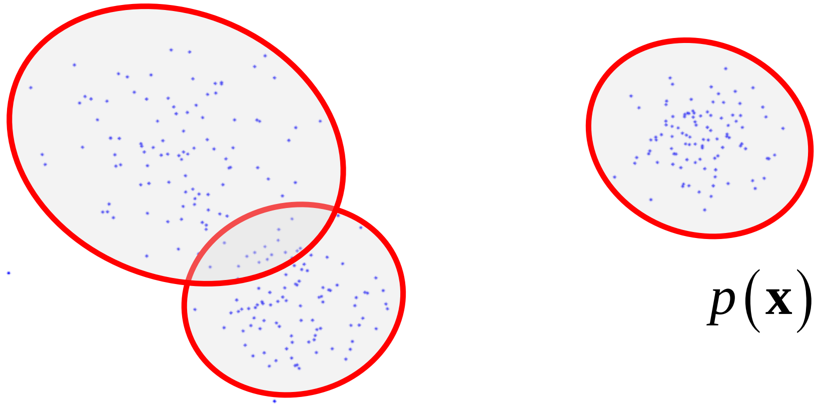
两个高斯函数的混合

$$p(x) = 0.7N(-10,2) + 0.3N(5,3)$$



ISBN 978-7-5603-4763-9

GMM模型的样本产生、参数估计



$$p(\mathbf{x}) = \sum_{k=1}^{3} a_k N(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

样本产生:

1）以组合系数 a_k 作为先验概率，随机选择一个高斯分量k

2）利用该分量 $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ 产生样本

参数估计:

1）样本 $\mathbf{x}_i$ 是由哪个分量产生的，记为 $y_i \in \{1, 2, \cdots, K\}$

2）高斯分量的先验概率 及参数 $\boldsymbol{\theta}_k = \{a_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$

参数估计：1）样本 $\mathbf{x}_i$ 是由哪个分量产生的，记为 $y_i \in \{1, 2, \cdots, K\}$

2）高斯分量的先验概率及参数 $\boldsymbol{\theta}_k = \{a_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$

已知 $\boldsymbol{\theta}_k$，估计 y_i：

计算分量 k 产生 $\mathbf{x}_n$ 的概率

$$p(\mathbf{x}_i|\boldsymbol{\theta}_k) = N(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

根据最大后验概率估计 y_i

$$y_i = \arg\max_k \alpha_k p(\mathbf{x}_i|\boldsymbol{\theta}_k)$$

已知 y_i，估计 $\boldsymbol{\theta}_k$：单个高斯分量参数估计

$$\alpha_k = \frac{1}{n}\sum_{i=1}^{n} \underline{I(y_i = k)}$$

示性函数 $= \begin{cases} 1, & y_i = k \\ 0, & y_i \neq k \end{cases}$

$$\boldsymbol{\mu}_k = \sum_{i=1}^{n} I(y_i = k)\mathbf{x}_i \Big/ \sum_{i=1}^{n} I(y_i = k)$$

$$\boldsymbol{\Sigma}_k = \sum_{i=1}^{n} I(y_i = k)(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^t \Big/ \sum_{i=1}^{n} I(y_i = k)$$

两者都未知时：先随机初始化高斯分量参数 $\boldsymbol{\theta}^0$，再迭代优化

根据第t次迭代得到的参数 $\boldsymbol{\theta}^t$，估计样本集的产生分量 $\mathbf{y}^t$

根据 $\mathbf{y}^t$ 估计新参数 $\boldsymbol{\theta}^{t+1}$

ISBN 978-7-5603-4763-9

参数估计：1）样本 $\mathbf{x}_i$ 是由哪个分量产生的，记为 $y_i \in \{1,2,\cdots,K\}$

2）高斯分量的先验概率及参数 $\boldsymbol{\theta}_k = \{a_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$

两者都未知时：先随机初始化高斯分量参数 $\boldsymbol{\theta}^0$，再迭代优化

根据第t次迭代得到的参数$\boldsymbol{\theta}^t$， 估计样本集的产生分量$\mathbf{y}^t$

根据 $\mathbf{y}^t$ 估计新参数$\boldsymbol{\theta}^{t+1}$

当GMM的混合系数相等、协方差矩阵为相同对角阵时——K均值聚类

示性函数改进： $\mathbf{y}^t$ 是一个不准确估计，采用概率替代示性函数更恰当：

$$I(y_i = k) = P(y_i = k)$$

$$= a_k N(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \Big/ \sum_{j=1}^{K} a_j N(\mathbf{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

ISBN 978-7-5603-4763-9

参数估计： 1）样本 $\mathbf{x}_i$ 是由哪个分量产生的，记为 $y_i \in \{1, 2, \cdots, K\}$

2）高斯分量的先验概率及参数 $\boldsymbol{\theta}_k = \{a_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$

GMM参数估计：先随机初始化高斯分量参数 $\boldsymbol{\theta}^0$，再迭代优化

根据第t次迭代得到的参数 $\boldsymbol{\theta}^t$， 估计样本集的产生分量 $\mathbf{y}^t$

根据 $\mathbf{y}^t$ 估计新参数 $\boldsymbol{\theta}^{t+1}$

已知 $\boldsymbol{\theta}_k$，估计 y_i：

计算 y_i 由分量 k 产生的概率

$$P(y_i = k) = \frac{a_k N(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} a_j N(\mathbf{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

已知 y_i，估计 $\boldsymbol{\theta}_k$：单个高斯分量参数估计

$$\alpha_k = \frac{1}{n} \sum_{i=1}^{n} P(y_i = k)$$

$$\boldsymbol{\mu}_k = \sum_{i=1}^{n} P(y_i = k) \mathbf{x}_i \Big/ \sum_{i=1}^{n} P(y_i = k)$$

$$\boldsymbol{\Sigma}_k = \sum_{i=1}^{n} P(y_i = k)(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^t \Big/ \sum_{i=1}^{n} P(y_i = k)$$

ISBN 978-7-5603-4763-9

参数估计： 1）样本 $\mathbf{x}_i$ 是由哪个分量产生的，记为 $y_i \in \{1, 2, \cdots, K\}$

2）高斯分量的先验概率及参数 $\boldsymbol{\theta}_k = \{a_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$

GMM学习算法

- 设置模型中高斯分量的个数 K，随机初始参数 $\boldsymbol{\theta}$，

  设置收敛精度 η；

- 循环： $t \leftarrow t+1$

  - □ 计算训练样本由各分量产生的概率 $P(y_i = k)$；
  - □ 重新估计参数 $\boldsymbol{\theta}$；
  - □ 计算似然函数值 $L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i | \boldsymbol{\theta})$；

- 直到满足收敛条件： $L_t(\boldsymbol{\theta}) - L_{t-1}(\boldsymbol{\theta}) < \eta$

K 的选择：越大拟合能力越强，结构风险越大

参数初始化：

$$a_k > 0, \quad \sum_{k=1}^{K} a_k = 1$$

$\boldsymbol{\Sigma}_k$ 为对称正定矩阵

收敛条件：似然函数变化小于阈值 η

计算稳定性：

对数阈上计算，克服概率过小

样本过少时，约束为对角阵

当协方差矩阵为奇异阵时，叠加小对角阵

期望最大化算法（Expectation Maximization EM）

样本集合由两部分构成D={X,Y}，其中X已知，Y未知

$$l(\boldsymbol{\theta}) = \ln p(D|\boldsymbol{\theta}) = \ln p(X,Y|\boldsymbol{\theta})$$

Y未知，无法优化；考虑Y所有可能情况下的对数似然函数：

$$Q(\boldsymbol{\theta}) = E_Y\left[\ln p(X,Y|\boldsymbol{\theta})\right] = \int \ln p(X,Y|\boldsymbol{\theta})\, p(Y)\, dY$$

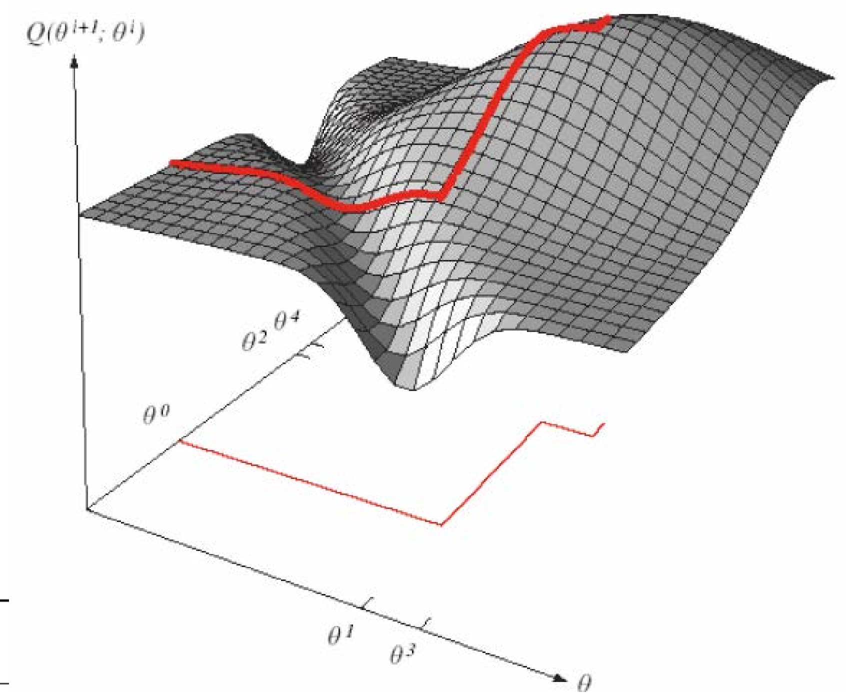
$p(Y)$ 仍然未知，首先设置 $\boldsymbol{\theta}$ 的猜测值 $\boldsymbol{\theta}^g$ ，在已知 X ，$\boldsymbol{\theta}^g$ 的条件下估计 $p(Y|X,\boldsymbol{\theta}^g)$ 替代 $p(Y)$：

E步： $$Q(\boldsymbol{\theta};\boldsymbol{\theta}^g) = \int \ln p(X,Y|\boldsymbol{\theta})\, p(Y|X,\boldsymbol{\theta}^g)\, dY$$

用 $Q(\boldsymbol{\theta};\boldsymbol{\theta}^g)$ 替代对数似然函数进行优化

M步： $$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta};\boldsymbol{\theta}^g)$$

EM算法流程



收敛于对数似然函数极大值，不能保证最大值，

收敛结果与初始值设置有关

EM 算法

- 初始化参数 $\boldsymbol{\theta}^1$，设置收敛精度 η；
- 循环：$t \leftarrow t+1$
  - □ E 步：计算 $Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^t\right)$；
  - □ M 步：$\boldsymbol{\theta}^{t+1}=\arg\max\limits_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^t\right)$；
- 直到满足收敛条件：$Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{t+1}\right)-Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^t\right)<\eta$

广义EM算法：

$$Q\left(\boldsymbol{\theta}^i;\boldsymbol{\theta}^{i-1}\right)>Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{i-1}\right)$$

EM算法应用

- EM算法的应用可以分为两个方面：

  1. 训练样本中某些特征丢失情况下，分布参数的最大似然估计；
  2. 对某些复杂分布模型假设，最大似然估计很难得到解析解时的迭代算法。

ISBN 978-7-5603-4763-9

部分特征丢失

设$\mathbf{x}=\left[\mathbf{x}_{\mathrm{g}},\mathbf{x}_{b}\right]$, $\mathbf{x}_{\mathrm{g}}$为完好特征，$\mathbf{x}_b$为损坏特征，

在给定$\mathbf{x}_{\mathrm{g}}$的前提下的贝叶斯规则：寻找$\mathbf{x}_{\mathrm{g}}$的最大后验概率

$$p\left(\omega_i\middle|\mathbf{x}_{\mathrm{g}}\right)=\frac{p\left(\omega_i,\mathbf{x}_{\mathrm{g}}\right)}{p\left(\mathbf{x}_{\mathrm{g}}\right)}=\frac{\int p\left(\omega_i,\mathbf{x}_{\mathrm{g}},\mathbf{x}_b\right)d\mathbf{x}_b}{p\left(\mathbf{x}_{\mathrm{g}}\right)}$$

$$=\frac{\int P\left(\omega_i\middle|\mathbf{x}_{\mathrm{g}},\mathbf{x}_b\right)p\left(\mathbf{x}_{\mathrm{g}},\mathbf{x}_b\right)d\mathbf{x}_b}{p(\mathbf{x}_{\mathrm{g}})}=\frac{\int P\left(\omega_i\middle|\mathbf{x}\right)p\left(\mathbf{x}\right)d\mathbf{x}_b}{\int p(\mathbf{x})d\mathbf{x}_b}$$

首先，在整个损坏的空间中，对后验概率进行积分。

然后，将贝叶斯判别规则用于后验概率：

$$i=\arg\max_{i}\ p(\omega_i\,|\,\mathbf{x}_g)$$

EM算法举例：

二维空间中，训练样本D包含4个样本点，其中$\mathbf{x}_{41}$丢失

$$D=\{x_1,x_2,x_3,x_4\}=\left\{\begin{pmatrix}0\\2\end{pmatrix},\begin{pmatrix}1\\0\end{pmatrix},\begin{pmatrix}2\\2\end{pmatrix},\begin{pmatrix}*\\4\end{pmatrix}\right\}$$

假设样本服从二维正态分布且协方差矩阵为对角阵，待估计参数为

$\theta=(\mu_1,\mu_2,\sigma_1,\sigma_2)^t$，初始估计为 $\theta^0=(0,0,1,1)^t$，有：

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

$$=\int_{-\infty}^{\infty}\left[\sum_{k=1}^{3}\ln p(x_k\,|\,\theta)+\ln p(x_4\,|\,\theta)\right]p(x_{41}\,|\,\theta^0;x_{42}=4)dx_{41}$$

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

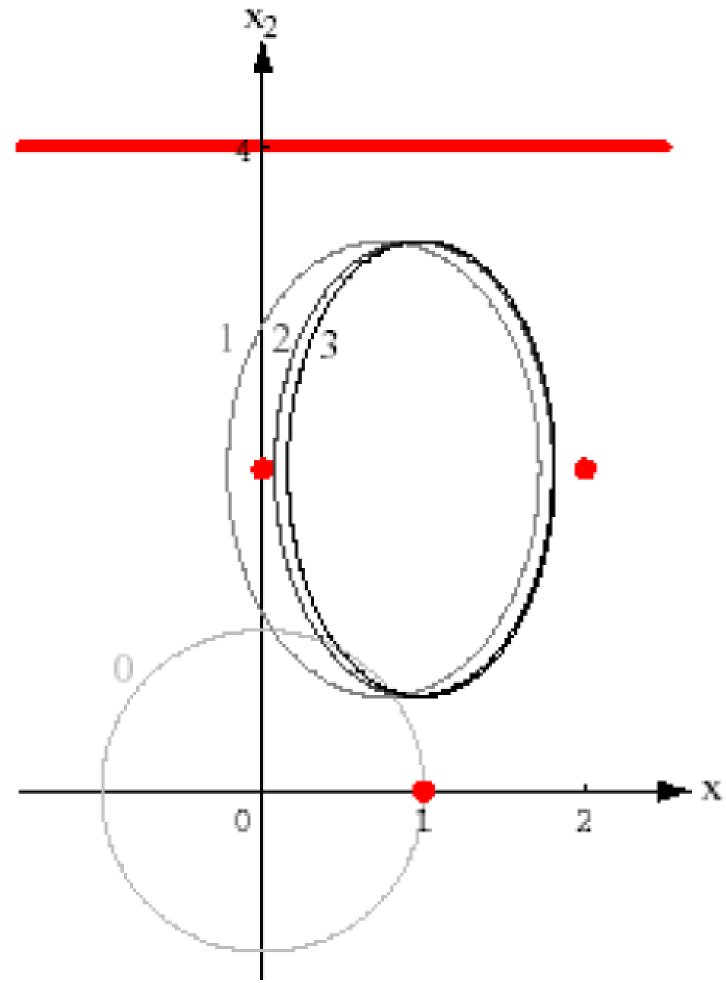
$$Q(\theta;\theta^0)=\sum_{k=1}^{3}\left[\ln p(x_k\mid\theta)\right]-\frac{1+\mu_1^2}{2\sigma_1^2}-\frac{(1-\mu_2)^2}{2\sigma_2^2}-\ln(2\pi\sigma_1\sigma_2)$$

通过直接计算，求出最大化$Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^0\right)$之后的$\boldsymbol{\theta}$：

$$\boldsymbol{\theta}^1=\left(0.75,2.0,0.938,2.0\right)$$

3次迭代后，算法收敛：

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



本节主要内容

- 非参数估计
- 参数估计
- 高斯混合模型
- 隐马尔科夫模型

ISBN 978-7-5603-4763-9

隐Markov模型 (Hidden Markov Model, HMM)

- 有一些模式识别系统处理的是与时间相关的问题，如语音识别，手势识别，唇读系统等；

- 对这类问题采用一个特征矢量序列描述比较方便，这类问题的识别HMM取得了很好的效果。

观察序列

- 信号的特征需要用一个特征矢量的序列来表示：

$$V^T = v_1, v_2, \cdots, v_T$$

- 其中的v_i为一个特征矢量，称为一个观察值。

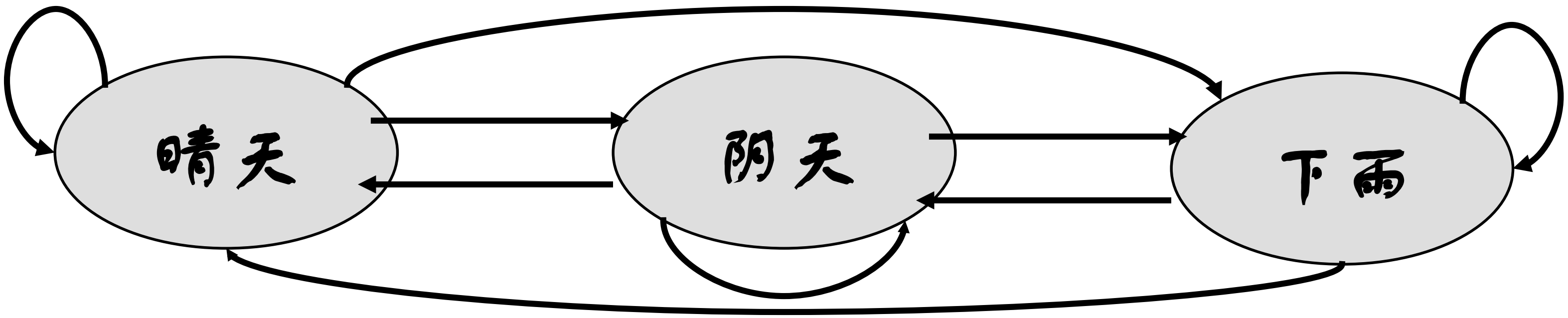
ISBN 978-7-5603-4763-9

马尔可夫性

- 1870年，俄国有机化学家Vladimir V. Markovnikov第一次提出马尔科夫模型
- 如果一个过程的“将来”仅依赖“现在”而不依赖“过去”，则此过程具有马尔可夫性,或称此过程为马尔可夫过程

$$X(t+1) = f(X(t))$$

转移概率



	晴天	阴天	下雨
晴天	0.50	0.25	0.25
阴天	0.375	0.25	0.375
下雨	0.25	0.125	0.625

马尔科夫链

- 时间和状态都离散的马尔科夫过程称为马尔科夫链

  记作$\{X_n = X(n), n = 0,1,2,\ldots\}$

  - 在时间集$T_1 = \{0,1,2,\ldots\}$上对离散状态的过程相继观察的结果

- 链的状态空间记做$I = \{a_1, a_2,\ldots\}, a_i \in R$.

- 条件概率$P_{ij}(m, m+n)=P\{X_{m+n} = a_j | X_m = a_i\}$为马氏链在时刻m处于状态$a_i$条件下，在时刻m+n转移到状态$a_j$的转移概率。

ISBN 978-7-5603-4763-9

转移概率矩阵(续)

- 由于链在时刻m从任何一个状态a_i出发，到另一时刻m+n，必然转移到a_1，a_2…，诸状态中的某一个，所以有

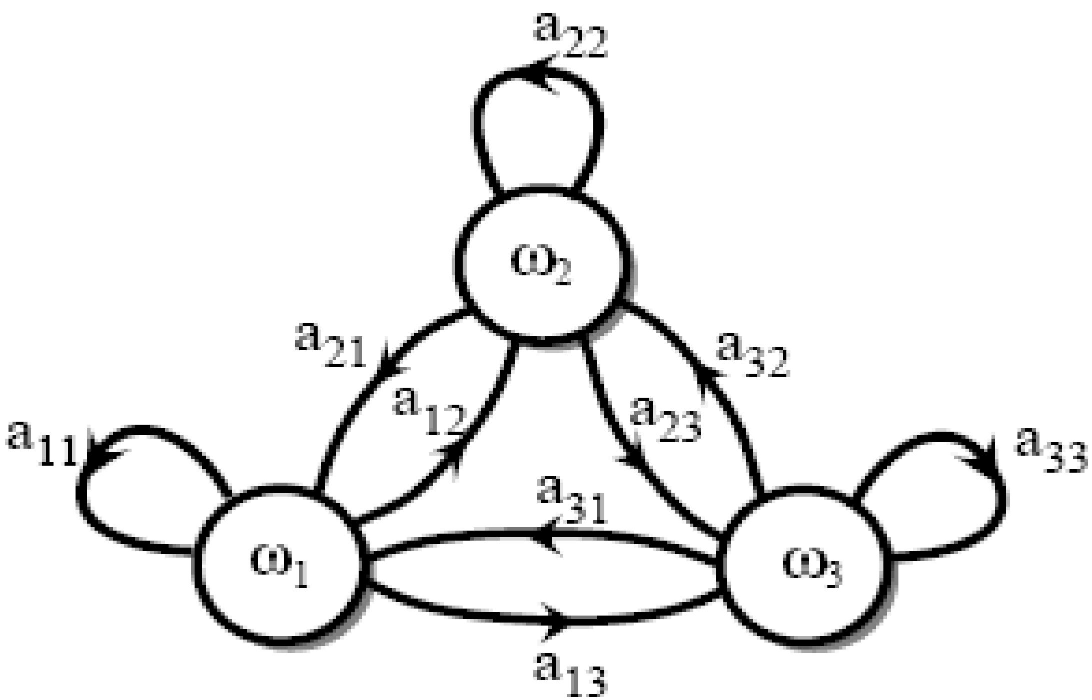
$$\sum_{j=1}^{\infty} P_{ij}(m, m+n) = 1, i = 1, 2, \ldots$$

- 当$P_{ij}(m,m+n)$与m无关时，称马尔科夫链为齐次马尔科夫链，通常说的马尔科夫链都是指齐次马尔科夫链。

ISBN 978-7-5603-4763-9

一阶Markov模型

- 一阶Markov模型由M个状态构成，在每个时刻t，模型处于某个状态w(t)，经过T个时刻，产生出一个长度为T的状态序列W^T=w(1),…,w(T)。



ISBN 978-7-5603-4763-9

一阶Markov模型的状态转移

- 模型在时刻 t 处于状态 w_j 的概率完全由 t-1时刻的状态 w_i 决定，而且与时刻 t 无关，即：

$$P\left(w(t) \middle| W^T\right) = P\left(w(t) \middle| w(t-1)\right)$$

$$P\left(w(t)=\omega_j \middle| w(t-1)=\omega_i\right) = a_{ij}$$

Markov模型的初始状态概率

- 模型初始于状态w_i的概率用 π_i 表示。
- 完整的一阶Markov模型可以用参数 $\boldsymbol{\theta}=(\boldsymbol{\pi},\mathbf{A})$ 表示，其中：

$$\boldsymbol{\pi}=\left(\pi_1,\cdots,\pi_M\right)$$

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1M} \\ a_{21} & a_{22} & \cdots & a_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MM} \end{bmatrix}$$

一阶Markov模型输出状态序列的概率

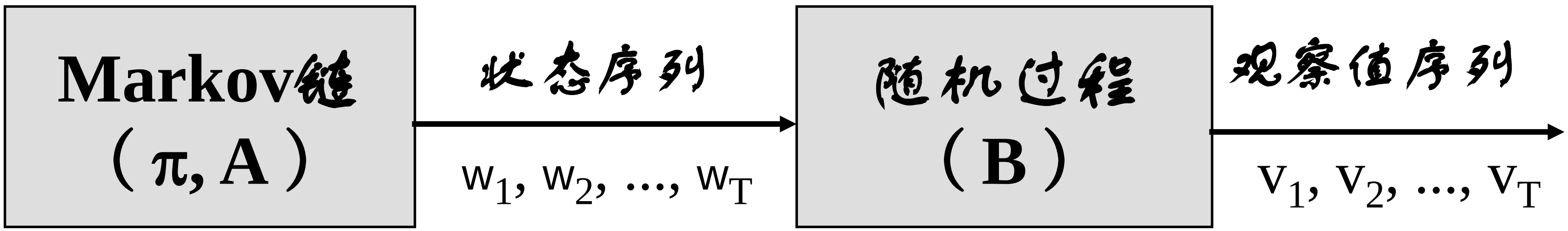
- 模型输出状态序列的概率可以由初始状态概率与各次状态转移概率相乘得到。
- 例如：$W^5=w_1, w_1, w_3, w_1, w_2$，则模型输出该序列的概率为：

$$P\left(W^5\right)=\pi_1 a_{11} a_{13} a_{31} a_{12}$$

HMM概念

- HMM的状态是不确定或不可见的，只有通过观测序列的随机过程才能表现出来
- 观察到的事件与状态并不是一一对应，而是通过一组概率分布相联系
- HMM是一个双重随机过程，两个组成部分：
  - 马尔可夫链：描述状态的转移，用转移概率描述。
  - 一般随机过程：描述状态与观察序列间的关系，用观察值概率描述。

HMM组成



HMM的组成示意图

ISBN 978-7-5603-4763-9

一阶隐含Markov模型

- 隐含Markov模型中，状态是不可见的，在每一个时刻t，模型当前的隐状态可以输出一个观察值。
- 隐状态输出的观察值可以是离散值，连续值，也可以是一个矢量。

ISBN 978-7-5603-4763-9

HMM的工作原理

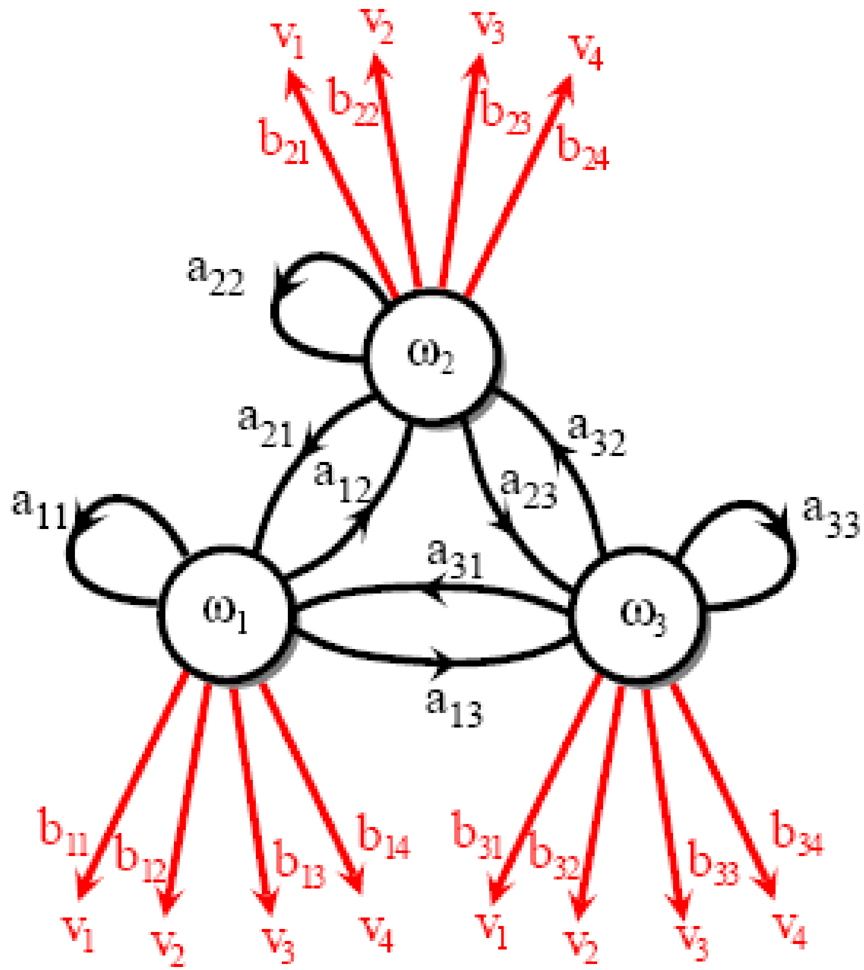
- HMM的内部状态转移过程同Markov模型相同，在每次状态转移之后，由该状态输出一个观察值，只是状态转移过程无法观察到，只能观察到输出的观察值序列。

- 以离散的HMM为例，隐状态可能输出的观察值集合为$\{v_1, v_2, \ldots, v_K\}$，第i个隐状态输出第k个观察值的概率为$b_{ik}$。

- 例如：T=5时，可能的观察序列$V^5=v_3v_2v_3v_4v_1$

ISBN 978-7-5603-4763-9

HMM的工作过程



ISBN 978-7-5603-4763-9

HMM假设

对于一个随机事件，有一个观察值序列：$v_1,\dots,v_T$

该事件隐含着一个状态序列：$w_1,\dots,w_T$

假设1：马尔可夫假设（状态构成一阶马尔可夫链）

$$p(w_i \mid w_{i-1} \dots w_1) = p(w_i \mid w_{i-1})$$

假设2：不动性假设（状态与具体时间无关）

$$p(w_i \mid w_{i-1}) = p(w_j \mid w_{j-1}) \qquad \text{对任意} i, j \text{成立}$$

假设3：输出独立性假设（输出仅与当前状态有关）

$$p(v_1,\cdots,v_T \mid w_1,\cdots,w_T) = \prod_{t=1}^{T} p(v_t \mid w_t)$$

HMM的参数表示

$$\boldsymbol{\theta} = \left(\boldsymbol{\pi}, \mathbf{A}, \mathbf{B}\right)$$

M个状态，K个可能的输出值。

初始概率：$\boldsymbol{\pi}$，包括M个元素。

$$\boldsymbol{\pi} = \left(\pi_1, \cdots, \pi_M\right)^T \qquad \pi_i: \text{第一个时刻处于状态 } w_i \text{的概率}$$

状态转移矩阵：$\mathbf{A}$，M*M方阵；

$$\mathbf{A} = \left(a_{ij}\right)_{M\times M} \qquad a_{ij} = P\left(w(t) = w_j \middle| w(t-1) = w_i\right)$$

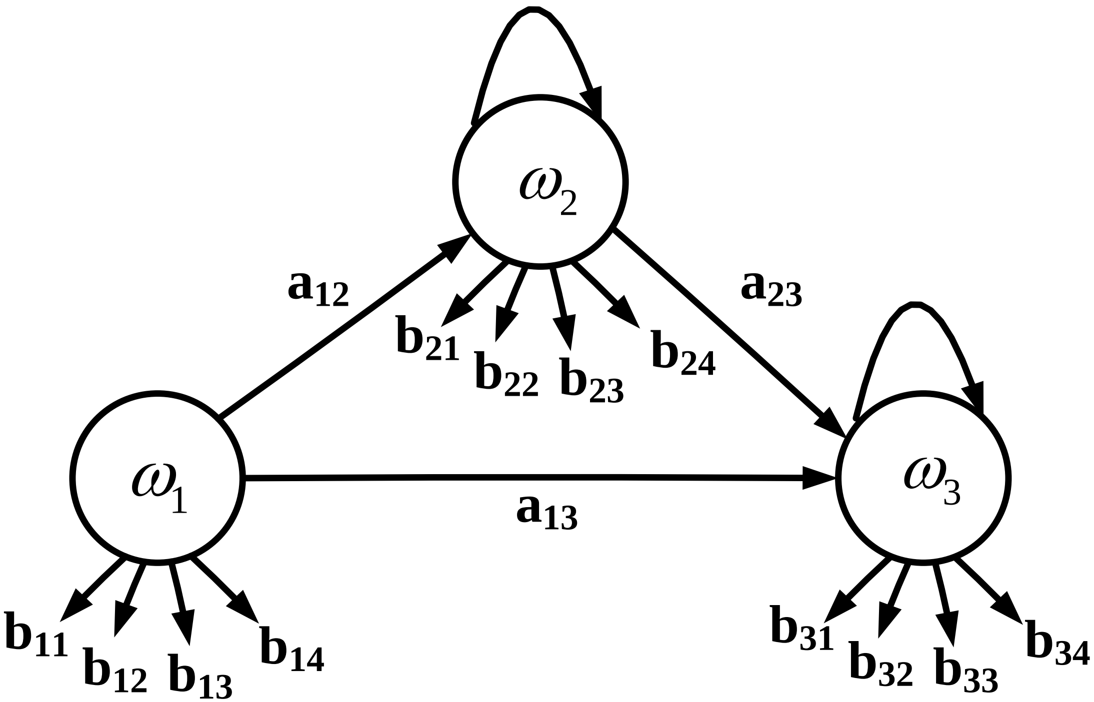
状态输出概率：B，M*K矩阵；

$$\mathbf{B} = \left(b_{ij}\right)_{M\times K} \qquad b_{ij} = P(v_j | w_i)$$

ISBN 978-7-5603-4763-9

HMM示例

如图HMM模型，初始概率：$\pi_1=1\ \pi_2=0\ \ \pi_3=0$ 状态转移概率矩阵：



$$\mathbf{A}=\begin{bmatrix}0 & 0.3 & 0.7\\ 0 & 0.6 & 0.4\\ 0 & 0 & 1\end{bmatrix}$$

状态输出概率矩阵：

$$\mathbf{B}=\begin{bmatrix}0.1 & 0.3 & 0.4 & 0.2\\ 0.5 & 0.2 & 0.1 & 0.2\\ 0.3 & 0.2 & 0.2 & 0.3\end{bmatrix}$$

1，请列出所有可能输出序列 $V=v_2v_4v_4v_1$ 的状态转移序列。

2，分别计算由每一个状态转移序列输出观察序列V的概率

3，计算最有可能输出观察序列V的状态转移序列

HMM的三个核心问题

- **估值问题：** 已有一个HMM模型，其参数已知，计算这个模型输出特定的观察序列V^T的概率；

- **解码问题：** 已有一个HMM模型，其参数已知，计算最有可能输出特定的观察序列V^T的隐状态转移序列W^T；

- **学习问题：** 已知一个HMM模型的结构，其参数未知，根据一组训练序列对参数进行训练；

估值问题

HMM模型产生观察序列 V^T 可以由下式计算：

$$P\left(V^T \middle| \boldsymbol{\theta}\right)=\sum_{r=1}^{r_{\max}} P\left(V^T \middle| W_r^T\right) P\left(W_r^T \middle| \boldsymbol{\theta}\right)$$

- $r_{\max}=M^T$ 为HMM所有可能的状态转移序列数；
- $P\left(V^T \middle| W_r^T\right)$ 为状态转移序列 W_r^T 输出观察序列 V^T 的概率；
- $P\left(W_r^T \middle| \boldsymbol{\theta}\right)$ 为状态转移序列 W_r^T 发生的概率。

□计算复杂度：$O\left(M^T \times T\right)$

HMM前向算法

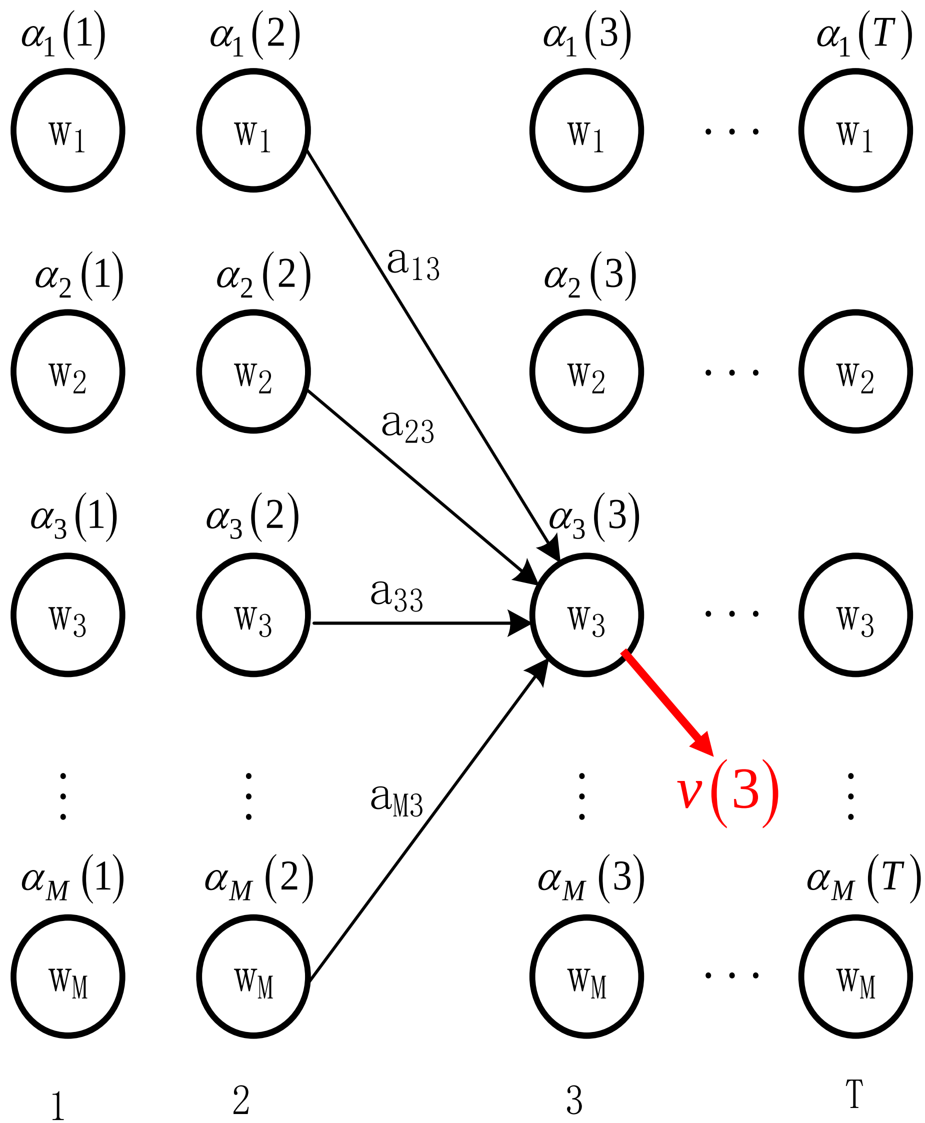
$\alpha_i(t)$： t 时刻，位于隐状态 ω_i 产生前 t 个可见符号的概率

$$\alpha_i(1)=\pi_i b_i\left[v(1)\right]$$

$$\alpha_3(3)=\left[\sum_{j=1}^{M}\alpha_j(2)a_{ji}\right]b_{iv(3)}$$

$$\alpha_i(t+1)=\left[\sum_{j=1}^{M}\alpha_j(t)a_{ji}\right]b_{iv(t+1)}$$

下标 i 表示隐状态



$b_{iv(t)}$:隐状态为i，输出为$v(t)$的概率

HMM的前向算法

前向算法

- 初始化 $t=1$；
- 计算第1列每个节点的 α 值：$\alpha_i(1)=\pi_i b_i\left[v(1)\right]$ ；
- 迭代计算t=2至 T 列每个节点的 α 值：

$$\alpha_i(t+1)=\left[\sum_{j=1}^{M}\alpha_j(t)a_{ji}\right]b_{iv(t+1)} \qquad i=1,\cdots,M$$

- 输出：

$$P\left(V^T\middle|\boldsymbol{\theta}\right)=\sum_{i=1}^{M}\alpha_i(T)$$

计算复杂度：$O\left(M^2T\right)$

HMM的后向算法

$\beta_i(t)$：t 时刻，位于隐状态 ω_i 产生V^T的t时刻值后的T-t个可见符号的概率

初始化：$\beta_i(T)=1, i=1,\cdots M$

迭代计算：

$$\beta_i(t)=\left[\sum_{j=1}^{M}\beta_j(t+1)a_{ji}\right]b_i\big(v(t+1)\big), i=1,\cdots,M$$

结束输出：

$$P\left(V^T\middle|\boldsymbol{\theta}\right)=\sum_{i=1}^{M}\beta_i(T)$$

计算复杂度：$O(M^2T)$

ISBN 978-7-5603-4763-9

前向算法与后向算法的结合

将V^T的在T′时刻分为V1，V2两个序列，有：

$$P\left(V^T\middle|\boldsymbol{\theta}\right)=\sum_{i=1}^{M}\alpha_i\left(T'\right)\beta_i\left(T'\right)$$

$$P\left(V^T\middle|\boldsymbol{\theta}\right)=P\left(V_1,V_2\middle|\boldsymbol{\theta}\right)=\sum_{i=1}^{M}P\left(V_1,V_2,\omega(T')=i\middle|\boldsymbol{\theta}\right)$$

$$=\sum_{i=1}^{M}P\left(V_1,\omega(T')=i\middle|\boldsymbol{\theta}\right)P\left(V_2,\omega(T')=i\middle|\boldsymbol{\theta}\right)$$

$$=\sum_{i=1}^{M}\alpha_i\left(T'\right)\beta_i\left(T'\right)$$

ISBN 978-7-5603-4763-9

HMM的三个核心问题

- **估值问题**：已有一个HMM模型，其参数已知，计算这个模型输出特定的观察序列V^T的概率；

- **解码问题**：已有一个HMM模型，其参数已知，计算最有可能输出特定的观察序列V^T的隐状态转移序列W^T；

- **学习问题**：已知一个HMM模型的结构，其参数未知，根据一组训练序列对参数进行训练；

解码问题

- 解码问题的计算同估值问题的计算类似，最直观的思路是遍历所有的可能状态转移序列，取出最大值，计算复杂度为：$O(M^T T)$。

- 同样存在着优化算法：Viterbi算法。

ISBN 978-7-5603-4763-9

Viterbi算法

$\delta_i(t)$：t 时刻，位于隐状态 ω_i 产生V^T的前t个可见符号的最大概率

$$\delta_i(1)=\pi_i b_i(v(1))$$

$$\delta_i(t+1)=\max_{1\le j\le M}\left[\delta_j(t)a_{ji}\right]b_{iv(t+1)}$$

建立一个矩阵Φ，其元素 $\varphi_i(t)$保存（第t步为第i个状态时）在第t-1步的最优状态。

$$\varphi_1(i)=0 \qquad \varphi_i(t+1)=\arg\max_{1\le j\le M}\left[\delta_j(t)a_{ji}\right]$$

Viterbi算法

1. 初始化： $\delta_i(1)=\pi_i b_i\left(v(1)\right), i=1,\cdots M,\quad \phi_1(i)=0$

2. 迭代计算：

$$\phi_i(t+1)=\arg\max_{1\le j\le M}\left[\delta_j(t)a_{ji}\right]$$

$$\delta_i(t+1)=\max_{1\le j\le M}\left[\delta_j(t)a_{ji}\right]b_i\left(v(t+1)\right), i=1,\cdots,M,$$

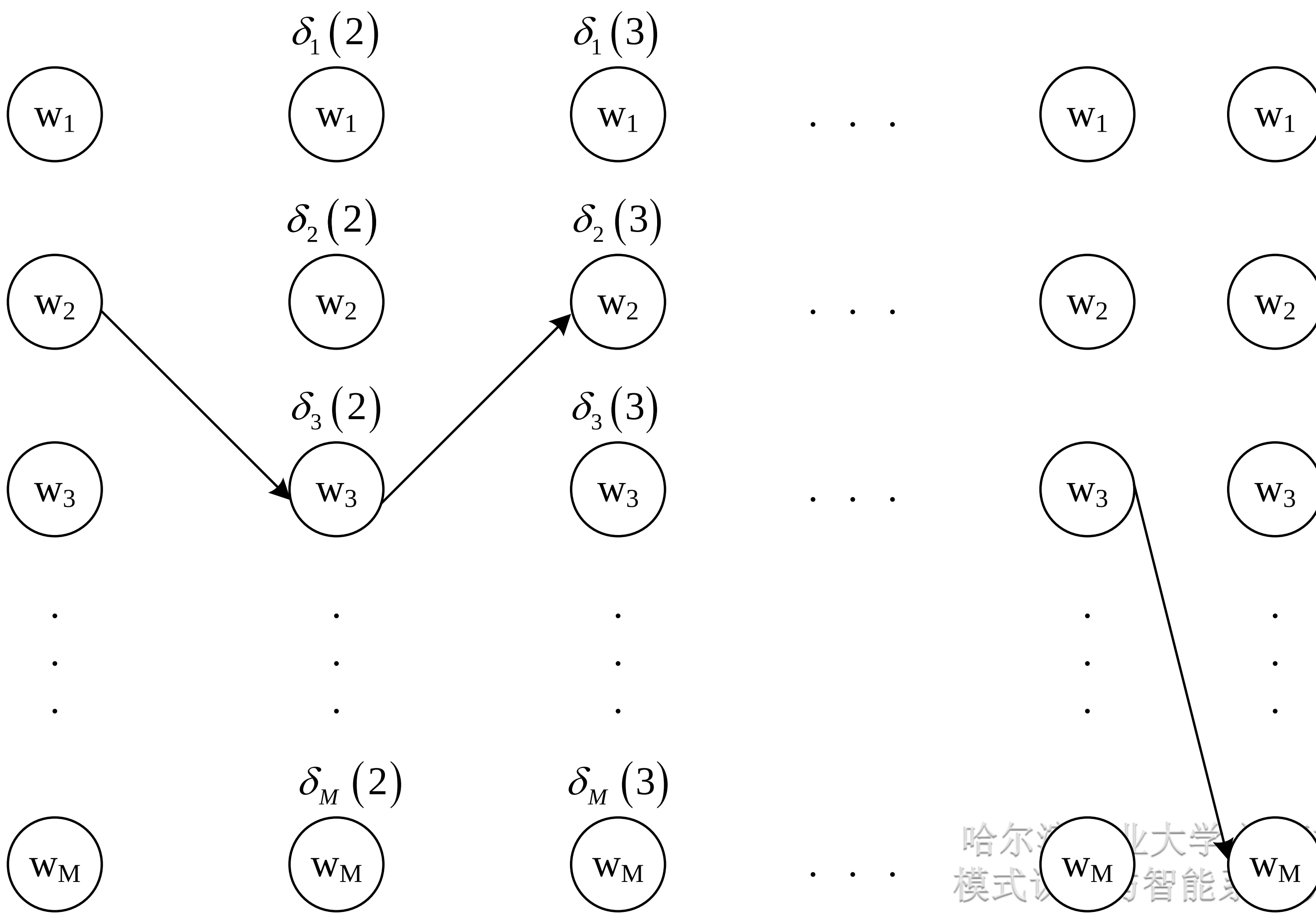
3. 结束：

$$P^*\left(V^T\middle|\boldsymbol{\theta}\right)=\max_{1\le j\le M}\left[\delta_j(T)\right],\ w^*(T)=\arg\max_{1\le j\le M}\left[\delta_j(T)\right]$$

4. 路径回朔： $w^*(t)=\phi_{w^*(t+1)}(t+1)$

哈尔滨工业大学 计算机学院
模式识别与智能系统研究中心
ISBN 978-7-5603-4763-9

Viterbi算法图示



Viterbi算法

$$\delta_i(t+1)=\max_{1\le j\le M}\left[\delta_j(t)a_{ji}\right]b_i\left(v(t+1)\right),i=1,\cdots,M,$$

$$\ln\delta_i(t+1)=\max_{1\le j\le M}\ln\left\{\left[\delta_j(t)a_{ji}\right]b_i\left(v(t+1)\right)\right\}$$

$$=\max_{1\le j\le M}\left\{\underbrace{\ln\delta_j(t)}_{=\rho_j(t)}+\ln a_{ji}+\ln b_i\left(v(t+1)\right)\right\}$$

$$\rho(t+1)=\max_{1\le j\le M}\left\{\rho_j(t)+\ln a_{ji}+\ln b_i\left(v(t+1)\right)\right\}$$

计算复杂度：$O\left(M^2T\right)$

ISBN 978-7-5603-4763-9

HMM应用举例

假设你有一个朋友在外地，每天做三种活动之一——Walk, Shop, Clean。从事活动的概率与天晴、下雨有关，天气与运动的关系及天气间转换的关系如下表：

	Rainy	Sunny
Walk	0.1	0.6
Shop	0.4	0.3
Clean	0.5	0.1

下雨散步的可能性是0.1。

	Rainy	Sunny
Rainy	0.7	0.3
Sunny	0.4	0.6

从行到列：从今天是晴天而明天就开始下雨的可能性是0.4 。

第一天的天气有0.6的概率是Rainy，有0.4概率是Sunny。如果连续三天，你发现你的朋友的活动是：Walk->Shop->Clean；那么，如何判断你朋友那里这几天的天气是怎样的？

ISBN 978-7-5603-4763-9

HMM的三个核心问题

- **估值问题**：已有一个HMM模型，其参数已知，计算这个模型输出特定的观察序列V^T的概率；

- **解码问题**：已有一个HMM模型，其参数已知，计算最有可能输出特定的观察序列V^T的隐状态转移序列W^T；

- **学习问题**：已知一个HMM模型的结构，其参数未知，根据一组训练序列对参数进行训练；

ISBN 978-7-5603-4763-9

HMM的学习问题

已知一组观察序列(训练样本集合)：

$$V=\left\{V_1^{T_1},V_2^{T_2},\cdots,V_n^{T_n}\right\}$$

如何确定最优的模型参数θ，使得模型产生训练集合V的联合概率最大

$$\hat{\boldsymbol{\theta}}=\max_{\boldsymbol{\theta}} P\left(V\middle|\boldsymbol{\theta}\right)$$

这同样是一个最大似然估计问题，需要采用EM算法。

ISBN 978-7-5603-4763-9

HMM学习问题

Baum-Welch算法：先设定一个转移概率的初值；然后获得对该初值的一个修正；反复迭代、直到收敛。

——广义EM算法在HMM中的具体实现

$$a_{ij}=\frac{\text{从状态}i\text{跳转到状态}j\text{的概率}}{\text{从状态}i\text{跳出的概率}} \qquad b_{ik}=\frac{\text{从状态}i\text{输出观测}k\text{的概率}}{\text{跳转到状态}i\text{的概率}}$$

设t时刻，从 i 跳转到 j 的概率为$\gamma_{ij}(t)$，观测长度为T的训练序列，对t求和：

如何计算?

$$a_{ij}=\frac{\sum_{t=2}^{T}\gamma_{ij}(t)}{\sum_{t=2}^{T}\sum_{k=1}^{M}\gamma_{ik}(t)} \qquad b_{ik}=\frac{\sum_{t=1,v(t)=v_k}^{T}\sum_{l=1}^{M}\gamma_{li}(t)}{\sum_{t=1}^{T}\sum_{l=1}^{M}\gamma_{li}(t)}$$

t 时刻，从状态 i 跳转到 j 的概率 $\gamma_{ij}(t)$

模型参数为θ，观测到序列V^T的条件下，$t-1$时刻处于ω_i，t时刻处于ω_j的概率

$$\gamma_{ij}(t)=P\left[\omega(t-1)=\omega_i,\omega(t)=\omega_j\middle|V^T,\boldsymbol{\theta}\right]$$

$$=\frac{P\left[\omega(t-1)=\omega_i,\omega(t)=\omega_j,V^T\middle|\boldsymbol{\theta}\right]}{P\left(V^T\middle|\boldsymbol{\theta}\right)}=\frac{\alpha_i(t-1)a_{ij}b_{jv(t)}\beta_j(t)}{P\left(V^T\middle|\boldsymbol{\theta}\right)}$$

三个独立子事件:

t-1时刻处于状态 ω_i，从1到t-1产生序列 $V^{1\to t-1}$,概率为 $\alpha_i(t-1)$

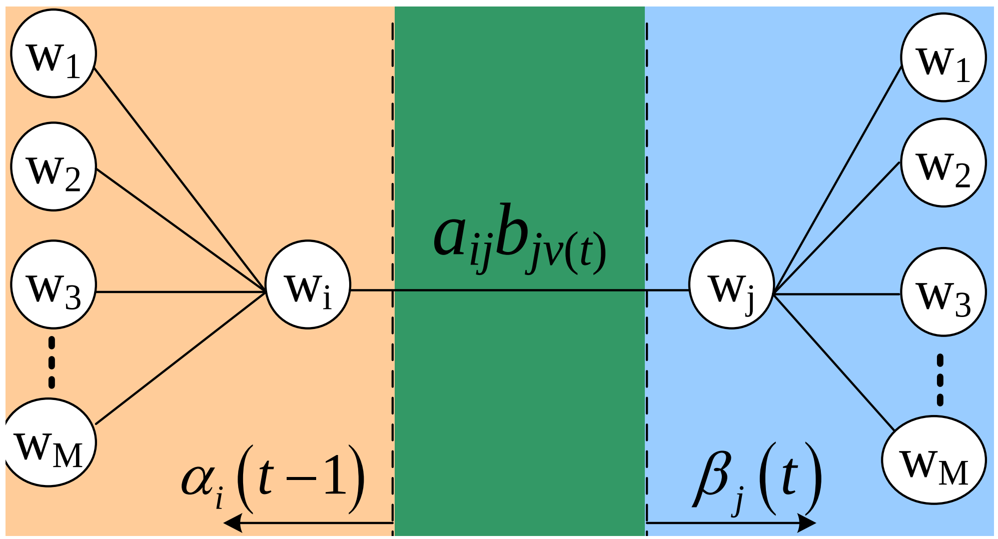
t 时刻处于状态ω_j，从$t+1$ 到 **T** 产生序列$V^{t+1\to T}$ ，概率为 $\beta_j(t)$

在t时刻由状态ω_i转移到ω_j，概率为$a_{ij}b_{jv(t)}$

ISBN 978-7-5603-4763-9

参数为θ,观测到序列V^T的条件下,$t-1$ 时刻处于ω_i,t 时刻处于ω_j 的概率

$$\gamma_{ij}(t)=\frac{P\left[\omega(t-1)=\omega_i,\omega(t)=\omega_j,V^T\middle|\boldsymbol{\theta}\right]}{P\left(V^T\middle|\boldsymbol{\theta}\right)}=\frac{\alpha_i(t-1)a_{ij}b_{jv(t)}\beta_j(t)}{P\left(V^T\middle|\boldsymbol{\theta}\right)}$$



$\alpha_i(t-1)$：在 t-1时刻处于状态 ω_i，从1到t-1时刻之间产生序列的概率；

$$\alpha_i(t-1)=\left[\sum_{j=1}^{M}\alpha_j(t-2)a_{ji}\right]b_{iv(t-1)}$$

$\beta_j(t)$：在 t 时刻处于状态ω_j，从t+1 到 T 时刻之间产生序列的概率；

$$\beta_j(t)=\left[\sum_{i=1}^{M}a_{ji}\beta_i(t+1)\right]b_{jv(t+1)}$$

状态转移概率的估计

输出观察序列V^T时，在t-1时刻HMM处于ω_i状态，在时刻t处于ω_j状态的概率：

$$\gamma_{ij}(t)=\frac{\alpha_i(t-1)a_{ij}b_j(v(t))\beta_j(t)}{P(V^T|\boldsymbol{\theta})}$$

ω_i到ω_j的预期数： $\sum_{t=1}^{T}\gamma_{ij}(t)$

从 ω_i的任何转移的总预期数： $\sum_{t=1}^{T}\sum_{k}\gamma_{ik}(t)$

从ω_i到ω_j的 转移的概率估计： $\hat{a}_{ij}=\frac{\sum_{t=1}^{T}\gamma_{ij}(t)}{\sum_{t=1}^{T}\sum_{k}\gamma_{ik}(t)}$

初始概率的估计

输出观察序列V^T时，在t=1时刻HMM处于ω_j状态的概率：

$$\gamma_j(1) = \frac{\pi_j b_{jv(1)} \beta_j(1)}{P\left(V^T \middle| \boldsymbol{\theta}\right)}$$

π_i 的迭代公式为：

$$\pi_i = P\left[w(1) = w_i \middle| V^T, \boldsymbol{\theta}\right] = \gamma_i(1)$$

ISBN 978-7-5603-4763-9

观察概率的估计

输出观察序列V^T时，在t-1时刻HMM处于ω_i状态，在时刻t处于ω_j状态的概率：

$$\gamma_{ij}(t)=\frac{\alpha_i(t-1)a_{ij}b_j(v(t))\beta_j(t)}{P(V^T|\boldsymbol{\theta})}$$

ω_i 上观察到 v_k 的预期数：
$$\sum_{t=1,v(t)=v_k}^{T}\sum_{l=1}^{M}\gamma_{li}(t)$$

到 ω_i的任何转移的总预期数：
$$\sum_{t=1}^{T}\sum_{l=1}^{M}\gamma_{li}(t)$$

从ω_i到ω_j的 转移的概率估计：
$$\hat{b}_{ik}=\frac{\sum_{t=1,v(t)=v_k}^{T}\sum_{l=1}^{M}\gamma_{li}(t)}{\sum_{t=1}^{T}\sum_{l=1}^{M}\gamma_{li}(t)}$$

ISBN 978-7-5603-4763-9