第3章　概率密度函数的参数估计

刘家锋

哈尔滨工业大学

第3章　概率密度函数的参数估计

3.0 引言

问题的提出

- **贝叶斯分类器的学习**
  - 根据贝叶斯决策理论构造判别函数，需要知道每个类别的先验概率$P(\omega_i)$和类条件概率密度$p(\mathbf{x}|\omega_i)$；
  - 实际应用中，往往并不能直接得到这些信息；
  - 贝叶斯分类器的学习，主要是利用训练数据集来估计类条件概率密度$p(\mathbf{x}|\omega_i)$；
  - 先验概率$P(\omega_i)$一般可以根据先验知识得到，也可以统计训练集中各类样本所占比例来做出估计；

问题的表示

- **学习问题的表示**
  - 给定有监督训练样本集$D=\{(\mathbf{x}_1,y_1),\cdots,(\mathbf{x}_n,y_n)\}$，类别标签$y_i=j$表示$\mathbf{x}_i\in\omega_j$；
  - 根据类别标签可以将训练集划分为c个子集：

$$D=D_1\cup D_2\cup\cdots\cup D_c$$

  - 贝叶斯分类器学习的任务就是根据每个类别的训练子集D_i，来估计这个类别的条件概率密度$p(\mathbf{x}|\omega_i)$；
- **概率密度的估计**
  - 这是一个典型的统计学中的概率密度估计问题；
  - 每个类别的估计是单独完成的，为了表述方便省略类别下标，将估计问题表示为：

$$D=\{\mathbf{x}_1,\cdots,\mathbf{x}_n\}\longrightarrow p(\mathbf{x})$$

概率密度的估计方法

- **独立同分布假设**
  - 训练集D中的样本抽样自同一分布，$\{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \sim p(\mathbf{x})$；
  - 每个样本的抽样过程是相互独立的；
  - 独立同分布假设是统计学的一个基本假设；
- **估计方法的分类**
  - 概率密度函数的估计方法可以分为两大类：参数估计方法和非参数估计方法

参数估计和非参数估计

- **参数估计方法（第3章）**
  - 除了训练样本集D之外，需要已知分布$p(\mathbf{x})$的其他信息；
  - 一般假设每个类别概率密度函数的形式是已知的，例如正态分布等等；
  - 但是分布的具体参数$\boldsymbol{\theta}$是未知的，例如正态分布的$\boldsymbol{\mu}, \Sigma$；
  - 学习的任务是根据训练集D估计分布的参数$\boldsymbol{\theta}$；
- **非参数估计方法（第4章）**
  - 非参数估计方法不需要知道分布的形式，直接根据训练集D估计分布$p(\mathbf{x})$；

参数估计和非参数估计

- **参数估计方法的特点**
  - 参数估计方法需要关于分布形式的先验知识，估计的准确程度依赖于先验知识是否准确；
  - 由于有先验知识的存在，参数估计方法使用比较少的训练数据就可以得到较好的估计结果；
- **非参数估计方法的特点**
  - 不需要任何关于分布的先验知识，适用性好；
  - 取得准确的估计结果，需要的训练样本数量远多于参数估计方法；

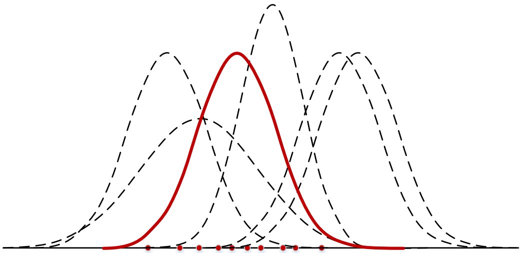
似然估计和贝叶斯估计

- **参数估计方法的分类**
  - 参数估计包括两种方法：最大似然估计和贝叶斯估计
  - 两种方法的目的都是要对分布的参数$\boldsymbol{\theta}$做出估计，但是观点存在差异；
  - 最大似然估计(Maximum Likelihood Estimation)
    - 认为参数$\boldsymbol{\theta}$是一个确定的未知量；
    - 估计的是参数的值：$\hat{\boldsymbol{\theta}}$
  - 贝叶斯估计(Bayesian Estimation)
    - 认为参数$\boldsymbol{\theta}$是一个随机变量；
    - 估计的是参数的分布：$p(\boldsymbol{\theta}|D)$

3.1 最大似然估计

似然函数

- **分布参数的似然**
  - 样本集$D = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$独立抽样自分布$p(\mathbf{x}|\boldsymbol{\theta})$；
  - 不同的分布参数$\boldsymbol{\theta}$决定了不同的分布；
  - 所有不同的分布中，哪一个产生出样本集D的可能性最大？



似然函数

- **似然函数**
  - $p(\mathbf{x}|\boldsymbol{\theta})$是给定参数$\boldsymbol{\theta}$条件下，样本$\mathbf{x}$的分布密度；
  - 给定参数$\boldsymbol{\theta}$条件下，样本集D的分布密度是n个随机事件发生的联合密度；
  - 根据独立同分布假设：

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

  - $L(\boldsymbol{\theta}) = p(D|\boldsymbol{\theta})$是参数$\boldsymbol{\theta}$的函数，称为似然函数；

最大似然估计

- **最大似然估计**
  - 最有可能产生样本集D的分布参数$\boldsymbol{\theta}$，是优化问题的解：

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

- **对数似然**
  - 最大似然估计只关心似然函数在哪个参数$\hat{\boldsymbol{\theta}}$上取得最大值，而不关心函数的具体数值；
  - 构造对数似然函数：

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

  - 对数函数是单调上升函数，因此最大似然估计可以等价地求解优化问题：

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

最大似然估计

- **似然函数的优化解**
  - 简单问题的最大似然估计，可以通过求解似然函数极值点的必要条件方程得到：

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \mathbf{0}$$

  - 其中$\nabla_{\boldsymbol{\theta}}$称为梯度算子，是多元函数的一阶导数；
- **梯度算子**
  - 令$\boldsymbol{\theta}$为p维参数矢量，梯度定义为：

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial l}{\partial \theta_1} \\ \vdots \\ \frac{\partial l}{\partial \theta_p} \end{pmatrix}$$

  - 函数$l(\boldsymbol{\theta})$的梯度是一个p维的矢量；

常用的微分公式

- **线性函数**
  - $f(\mathbf{x}) = \mathbf{x}^t\mathbf{y} = \sum_{i=1}^{d} x_i y_i$

$$\frac{\partial f}{\partial x_j} = y_j, \quad \frac{\partial f}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_d \end{pmatrix} = \mathbf{y}$$

- **二次函数**
  - $f(\mathbf{x}) = \mathbf{x}^t A\mathbf{x}$

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = (A + A^t)\mathbf{x}$$

当A为对称矩阵时，$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = 2A\mathbf{x}$

- 其它的微分公式参见“补充材料”

正态分布参数的最大似然估计

- **一维正态分布的最大似然估计**
  - 一维正态分布的参数$\boldsymbol{\theta} = (\mu, \sigma^2)^t$；
  - 构造对数似然函数：

$$l(\mu, \sigma^2) = \sum_{i=1}^{n} \ln p(x_i|\mu, \sigma^2) = \sum_{i=1}^{n} -\frac{1}{2}\left[\ln 2\pi + \ln \sigma^2 + \frac{(x_i-\mu)^2}{\sigma^2}\right]$$

  - 计算偏导数：

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{n} \frac{1}{\sigma^2}(x_i - \mu)$$

$$\frac{\partial l}{\partial \sigma^2} = \sum_{i=1}^{n} \left[-\frac{1}{2\sigma^2} + \frac{(x_i-\mu)^2}{2\sigma^4}\right]$$

正态分布参数的最大似然估计

- **一维正态分布的最大似然估计**
  - 极值点方程：

$$\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(x_i-\hat{\mu})=0$$

$$-\frac{n}{\hat{\sigma}^2}+\frac{1}{\hat{\sigma}^4}\sum_{i=1}^{n}(x_i-\hat{\mu})^2=0$$

  - 参数的最大似然估计：

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

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$$

正态分布参数的最大似然估计

- **多维正态分布的最大似然估计**
  - 假设多维正态分布参数Σ已知，估计未知参数$\boldsymbol{\mu}$；
  - 构造对数似然函数：

$$
\begin{aligned}
l(\boldsymbol{\mu}) &= \sum_{i=1}^{n} \ln p(\mathbf{x}_i|\boldsymbol{\mu}) \\
&= -\frac{1}{2}\sum_{i=1}^{n}\left[\ln (2\pi)^d|\Sigma| + (\mathbf{x}_i-\boldsymbol{\mu})^t\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right]
\end{aligned}
$$

  - 计算偏导数：

$$
\frac{\partial l}{\partial \boldsymbol{\mu}} = \sum_{i=1}^{n}\left[\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right]
$$

正态分布参数的最大似然估计

- **多维正态分布的最大似然估计**
  - 极值点方程：

$$\Sigma^{-1}\left[\sum_{i=1}^{n}(\mathbf{x}_i-\hat{\boldsymbol{\mu}})\right]=\mathbf{0}$$

  - 左乘矩阵Σ，得到参数$\boldsymbol{\mu}$的最大似然估计：

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

  - 参数Σ的似然估计需要较多的数学推导，参见“补充材料”

3.2 期望最大化算法

问题的提出

- **似然函数优化总是可以由极值点方程求解吗？**
  - 简单分布的参数可以求解极值点方程得到最大似然估计；
  - 很多复杂分布参数的极值点方程难于求解，需要迭代优化：
    - 梯度法：通用的迭代优化求解方法；
    - EM算法：专门用于迭代优化（对数）似然函数；
- **训练数据中存在缺失时，如何估计模型参数？**
  - 训练数据可能是不完整的，例如样本的某些特征是缺失的；
  - 不完整训练数据构造的似然函数中，除了需要估计的分布参数之外，还存在一些未知变量（缺失数据）；
  - 缺失数据集的似然函数无法直接优化，需要采用EM算法迭代优化；

Gauss Mixture Model

- **混合密度模型**
  - 复杂的概率密度函数可以由简单密度函数线性组合构成：

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^{M} \alpha_k p_k(\mathbf{x}|\boldsymbol{\theta}_k)$$

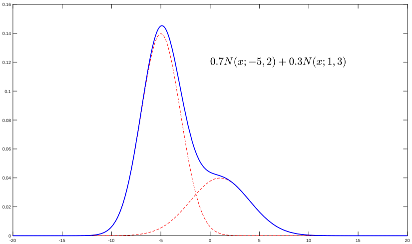
其中，$\sum_{k=1}^{M} \alpha_k = 1, \alpha_k > 0$

- **高斯混合模型**
  - GMM是混合密度模型的一个特例，由多个高斯（正态分布）函数的组合构成：

$$p(\mathbf{x}) = \sum_{k=1}^{M} \alpha_k N(\mathbf{x}; \boldsymbol{\mu}_k, \Sigma_k)$$

Gauss Mixture Model

- **“通用”的概率密度函数**
  - 最大似然估计需要训练数据符合何种分布的先验知识；
  - 实际应用中，往往缺乏这样的先验知识；
  - GMM可以看作是一种"通用"的概率密度函数
    - 数量M足够大，GMM可以任意精度逼近任意分布密度函数；



GMM的参数估计

- **GMM的最大似然估计**
  - GMM需要估计的参数：

$$\boldsymbol{\theta} = (\alpha_1, \cdots, \alpha_M, \boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_M, \Sigma_1, \cdots, \Sigma_M)$$

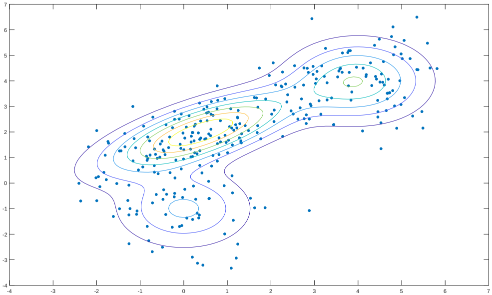
  - 对数似然函数：

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \left[\sum_{k=1}^{M} \alpha_k N(\mathbf{x}_i; \boldsymbol{\mu}_k, \Sigma_k) \right]$$

  - 极值点方程是复杂的超越方程组，很难直接求解；
  - 常用的GMM参数估计方法是EM算法；

样本的产生过程

- **GMM样本的产生过程**
  - 依据概率$\{\alpha_1,\cdots,\alpha_M\}$，选择一个成份高斯分布；
  - 依据选择的成份高斯分布参数，产生具体的样本特征矢量；



GMM的参数估计问题

- **GMM的参数估计问题**
  - 训练数据集：$D_X = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$
  - 学习参数：$\boldsymbol{\theta} = \{\alpha_k, \boldsymbol{\mu}_k, \Sigma_k\}_{k=1,\cdots,M}$
- **存在的问题**
  - 只有样本$\mathbf{x}_i$，但不知道是由哪一个成份高斯产生的；
  - 令$y_i = k$表示$\mathbf{x}_i$是由第k个成份高斯产生，构造集合：

$$D_Y = \{y_1, \cdots, y_n\}$$

  - 完整的数据集$D = D_X \cup D_Y$，其中D_Y为缺失的数据；

GMM的参数估计问题

- **已知数据集D_Y的条件下**
  - 可以很容易地估计GMM的参数；
  - 定义示性函数：

$$I(t) = \begin{cases} 1, & t \text{ is true} \\ 0, & t \text{ is false} \end{cases}$$

  - $\{\alpha_k\}$是选择成份高斯的概率，用高斯被选择的频度估计：

$$\hat{\alpha}_k = \frac{1}{n} \sum_{i=1}^{n} I(y_i = k)$$

  - 每个高斯的参数用该高斯产生的样本估计：

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

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

GMM的参数估计问题

- **已知数据集GMM参数$\boldsymbol{\theta}$的条件下**
  - 可以很容易地估计数据集D_Y；
  - 数据集中的n个样本，按照抽样高斯的不同分成M个子集；
  - 数据集D_Y的估计相当于M个类别的分类问题，应用贝叶斯判别准则：

$$\hat{y}_i = \arg\max_{1\leq k\leq M} \alpha_k N(\mathbf{x}_i; \boldsymbol{\mu}_k, \Sigma_k)$$

- **D_Y和$\boldsymbol{\theta}$均未知时**
  - 可以采用交替的方式，迭代优化；
  - 固定参数$\boldsymbol{\theta}$，优化D_Y；
  - 固定D_Y，优化参数$\boldsymbol{\theta}$；

GMM与聚类分析

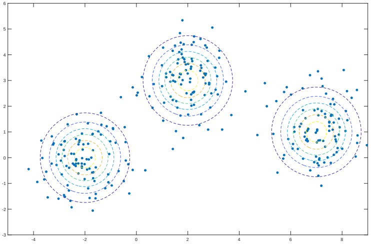
- **聚类分析**
  - 聚类分析属于无监督学习（第10章）；
  - 已知样本集$D_X = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$属于不同的聚类（子集），寻找对样本集的合理划分；
- **GMM与聚类分析**
  - GMM的参数估计与聚类分析之间存在内在的联系；
  - 如果假设数据集D_X来自于K个聚类，每个聚类服从正态分布，聚类的先验概率为$\{\alpha_k\}$，则样本集D_X服从GMM分布；
  - 对数据集D_Y的估计，实质上就是对D_X的聚类划分；

K-means聚类

- **K-means聚类算法**
  - 如果进一步假设，每个聚类的先验概率相同，协方差矩阵均为$\sigma^2 I$，则D_X的分布：

$$p(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} N(\mathbf{x}; \boldsymbol{\mu}_k, \sigma^2 I)$$

  - 上述GMM参数的迭代估计过程，就演变为了著名K-means聚类算法；



K-means聚类算法

Algorithm 1 K-means聚类算法

Input: 数据集$D_X = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$，聚类数$K$

Output: 数据集的聚类标签$D_Y = \{y_1, \cdots, y_n\}$

1: 随机初始化聚类中心$\{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K\}$;

2: **repeat**

3: 更新聚类标签：

$$y_i = \arg \min_{1 \le k \le K} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$$

4: 更新聚类中心：

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

5: **until** D_Y不再改变

GMM的参数估计方法

- **隐变量的概率估计**
  - 迭代优化过程中，参数$\boldsymbol{\theta}$和D_Y都是不准确的中间推断结果；
  - 依据不准确参数$\boldsymbol{\theta}$断定样本$\mathbf{x}_i$由某个高斯产生，过于武断；
  - 合理的方式是推断样本$\mathbf{x}_i$由每一个高斯产生的概率：

$$P(y_i = k) = \frac{\alpha_k N(\mathbf{x}_i; \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{M} \alpha_j N(\mathbf{x}_i; \boldsymbol{\mu}_j, \Sigma_j)} \tag{1}$$

  - 这就是EM算法中的E步，估计隐变量（缺失数据）的概率；

GMM的参数估计方法

- **GMM参数的估计**
  - E步中估计了样本由不同高斯产生的概率；
  - 每个高斯分布参数也需要由所有样本参与估计，同时需要考虑样本由不同高斯产生的概率：

$$\hat{\alpha}_k = \frac{1}{n}\sum_{i=1}^{n} P(y_i = k) \tag{2}$$

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^{n} P(y_i = k)\mathbf{x}_i}{\sum_{i=1}^{n} P(y_i = k)} \tag{3}$$

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{n} P(y_i = k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^t}{\sum_{i=1}^{n} P(y_i = k)} \tag{4}$$

  - 这就是EM算法中的M步，最大化模型的参数；

GMM参数估计的EM算法

Algorithm 2 GMM参数估计的EM算法

Input: 训练数据集$D_X = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$，高斯数$M$

Output: GMM的模型参数$\boldsymbol{\theta}$

1: 随机初始化参数$\boldsymbol{\theta}$；

2: **repeat**

3: **E步**，公式(1)估计样本由不同高斯产生的概率；

4: **M步**，公式(2-4)重新估计模型的参数$\boldsymbol{\theta}$；

5: **until** 达到收敛精度为止

（一般采用似然函数在两轮迭代之间的变化量作为收敛条件。）

EM算法

- **EM算法是含有隐变量或缺失数据的最大似然估计方法**
  - 假设X是观察到的数据集，Y是缺失数据集，$D = X \cup Y$；
  - $\boldsymbol{\theta}$是我们要估计的分布参数，对数似然函数：

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

  - 参数估计存在的问题是Y是未知的，对数似然既是$\boldsymbol{\theta}$的函数，也是Y的函数，无法直接优化；

EM算法

- **对数似然的期望**
  - 将Y视为随机变量，在平均意义下考察对数似然函数：

$$Q(\boldsymbol{\theta}) = \mathbb{E}_Y[\ln p(X, Y|\boldsymbol{\theta})] = \int \ln p(X, Y|\boldsymbol{\theta}) \cdot p(Y) dY$$

  - $Q(\boldsymbol{\theta})$只是$\boldsymbol{\theta}$的函数，用来代替对数似然$l(\boldsymbol{\theta})$的优化；

EM算法

- **E步迭代**
  - $Q(\boldsymbol{\theta})$的计算中需要的$p(Y)$未知，仍然无法直接优化；
  - 给定一个$\boldsymbol{\theta}$的初步估计$\boldsymbol{\theta}^0$，以此来估计Y的分布：

$$p(Y|X,\boldsymbol{\theta}^0) = \frac{p(X,Y|\boldsymbol{\theta}^0)}{p(X|\boldsymbol{\theta}^0)} = \frac{p(X,Y|\boldsymbol{\theta}^0)}{\int p(X,Y|\boldsymbol{\theta}^0)dY}$$

  - 以此来代替$p(Y)$，得到对数似然期望的近似估计：

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^0) = \int \ln p(X,Y|\boldsymbol{\theta}) \cdot p(Y|X,\boldsymbol{\theta}^0)dY$$

EM算法

- **M步迭代**
  - 优化$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^0)$，求解$\boldsymbol{\theta}$近似最优解：

$$\boldsymbol{\theta}' = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^0)$$

  - $\boldsymbol{\theta}'$改进了初步的估计$\boldsymbol{\theta}^0$；
- **EM迭代**
  - 迭代E步：$\boldsymbol{\theta}'$代替$\boldsymbol{\theta}^0$，重新估计$p(Y|X,\boldsymbol{\theta}')$
  - 迭代M步：重新优化$Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$
  - 直到收敛为止；

EM算法

Algorithm 3 形式化的EM算法

Input: 训练数据集X，收敛精度T

Output: 分布的参数估计$\hat{\boldsymbol{\theta}}$

1: 随机初始化参数$\boldsymbol{\theta}^0$，$t \leftarrow -1$;

2: **repeat**

3: $t = t + 1$

4: **E步**，估计分布$p(Y|X, \boldsymbol{\theta}^{t-1})$，计算$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{t-1})$;

5: **M步**，优化分布参数

$$\boldsymbol{\theta}^t = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{t-1})$$

6: **until** $Q(\boldsymbol{\theta}^t|\boldsymbol{\theta}^{t-1}) - Q(\boldsymbol{\theta}^{t-1}|\boldsymbol{\theta}^{t-2}) < T$

7: $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^t$

EM算法

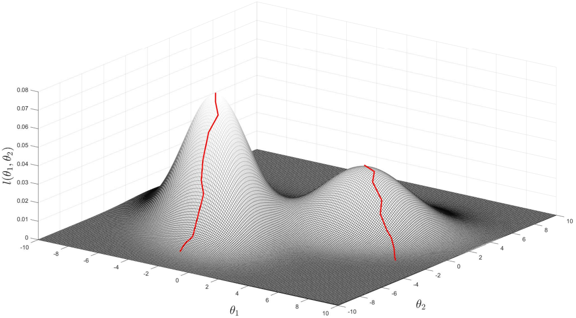
- **EM算法的性质**
  - EM算法是一个形式化的算法，需要根据具体的分布来推导E步和M步的迭代公式；
    - GMM参数估计EM迭代公式的推导参见“补充材料”；
  - 收敛性：EM算法具有收敛性，可以证明

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

    - K-means算法同样具有收敛性；
  - 最优性：EM算法只能保证收敛于似然函数的局部最大值点（极值点），不能保证收敛于全局的最大值点

EM算法

- **EM算法的极值点**
  - 对数似然函数存在多个极值点；
  - EM算法收敛于哪一个极值点，受到初始值设置的影响；
    - K-means算法的聚类结果同样受到初始设置的影响；



3.3 隐含马尔可夫模型

时间序列的识别

- **时间序列**
  - 某些应用中，识别对象是一个动态的过程，无法用静态的矢量描述；
  - 这些动态过程往往是与时间相关的，一般采用时间序列来描述，例如声音、动作、手势的识别；

观察值与观察序列

- **观察序列**
  - 识别对象（样本）需要用一个长度为T的观察序列来描述：

$$V^T = \mathbf{v}_1\mathbf{v}_2\cdots\mathbf{v}_T$$

  - 不同对象，观察序列的长度T是不同的；
- **观察值**
  - 观察序列中的元素$\mathbf{v}_t$对应时刻t的观察值，$t = 1, \cdots, T$；
  - 观察值描述了时刻t的特征，一般用一个特征矢量描述
    - 语音信号，一般用一个采样帧的频谱特征作为观察矢量；
    - 视频信号，可以用每一帧图像的特征作为观察矢量；
  - 为了便于理解，先将观察值简化为1维的离散值；

1阶Markov模型

- **Markov Model**
  - Markov模型包含M个状态：

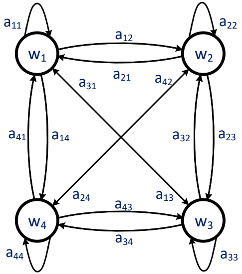
$$w_1, \cdots, w_M$$

  - 时刻t，模型处于状态$w(t)$，并且发生一次状态转移：

$$w(t) \rightarrow w(t+1)$$

  - 经过T个时刻，可以得到状态转移序列：

$$W^T = w(1) \cdots w(T)$$



1阶Markov模型

- **(1阶)Markov性**
  - 模型在时刻t处于状态$w(t) = w_j$的概率，完全由$t-1$时刻的状态$w(t-1) = w_i$决定，与之前以及之后的其它状态无关：

$$P(w(t)|W^T) = P(w(t)|w(1), \cdots, w(T)) = P(w(t)|w(t-1))$$

  - 由状态w_i转移到w_j的概率，与时刻t无关，表示为：

$$a_{ij} = P(w(t) = w_j | w(t-1) = w_i)$$

1阶Markov模型

- **一阶Markov模型的参数：** $\boldsymbol{\theta} = (\boldsymbol{\pi}, A)$
  - 初始状态概率：

$$\boldsymbol{\pi} = (\pi_1, \cdots, \pi_M)^t$$

    其中，$\pi_i = P(w(1) = w_i)$，模型初始于状态w_i的概率；
  - 状态转移概率矩阵：

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

1阶Markov模型

- **输出状态序列的概率**
  - 给定参数$\boldsymbol{\theta} = (\boldsymbol{\pi}, A)$，可以很容易地计算一阶Markov模型输出特定状态转移序列的概率；
  - 例如，模型输出序列$W^5 = w_1 w_1 w_3 w_1 w_2$的概率：

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

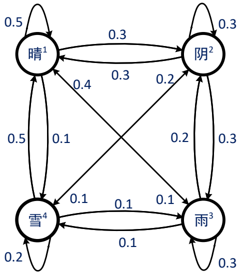
Markov模型实例

- **应用实例**
  - 某个城市天气的变化可以采用一阶马尔科夫模型描述，每天的天气有4种状态{晴, 阴, 雨, 雪}，模型参数：

$$\boldsymbol{\pi} = (0.5, 0.3, 0.1, 0.1)^t$$

$$P(7\text{晴天}) = \pi_1 \times a_{11}^6 = 0.0078125$$

$$P(3\text{晴}4\text{雨}) = \pi_1 \times a_{11}^2 \times a_{13} \times a_{33}^3 = 0.00016875$$



隐含Markov模型

- **HMM：Hidden Markov Model**
  - HMM包含一个一阶Markov模型，但模型的状态以及状态转移过程是隐含的，不可见的；
  - HMM在时刻t由隐状态$w(t)$输出观察值$v(t) \in \{v_1, \cdots, v_K\}$；
  - 经过T个时刻之后，可以观察到HMM输出一个观察序列：

$$V^T = v(1)v(2)\cdots v(T)$$

隐含Markov模型

- **HMM的模型参数：$\boldsymbol{\theta} = (\boldsymbol{\pi}, A, B)$**
  - 初始状态概率：$\boldsymbol{\pi}$，M维矢量
  - 状态转移概率矩阵：A，$M \times M$维矩阵
  - 输出概率矩阵：B，$M \times K$维矩阵
  - 其中：
    - M为状态数，K为观察值数；
    - B的元素b_{ik}表示状态w_i输出观察值v_k的概率，与时刻t无关

$$b_{ik} = P(v(t) = v_k | w(t) = w_i)$$

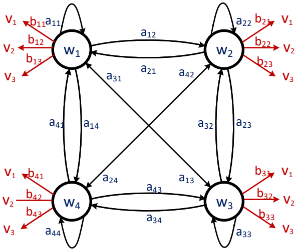
HMM实例

- **应用实例**
  - 我们不知道某城市的天气情况，但知道当地某人每天的活动情况{散步, 购物, 做家务}：

$$\boldsymbol{\pi} = (0.5, 0.3, 0.1, 0.1)^t$$

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

$$B = \begin{bmatrix} 0.5 & 0.4 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.1 & 0.3 & 0.6 \\ 0.1 & 0.2 & 0.7 \end{bmatrix}$$



HMM的核心问题

- **估值问题**
  - 已知HMM的模型参数$\boldsymbol{\theta}$，计算模型输出特定观察序列V^T的概率$P(V^T|\boldsymbol{\theta})$；
  - 识别过程中，需要计算每个类别的HMM模型输出待识别序列V^T的类条件概率(密度)；
- **解码问题**
  - 已知HMM的模型参数$\boldsymbol{\theta}$，计算最有可能输出的特定观察序列V^T的隐状态转移序列W^T；
  - 解码多用于时间序列的分割，如语音分割，动作视频分割；
- **学习问题**
  - 已知HMM模型的结构，根据一组训练序列学习参数$\boldsymbol{\theta}$；

估值问题

- **直接求解**
  - 给定模型参数$\boldsymbol{\theta}$，计算HMM输出观察序列V^T的概率；
  - 隐状态转移序列W^T未知，根据全概公式；

$$
\begin{aligned}
P(V^T|\boldsymbol{\theta}) &= \sum_{r=1}^{r_{max}} P(V^T, W_r^T|\boldsymbol{\theta}) \\
&= \sum_{r=1}^{r_{max}} P(V^T|W_r^T, \boldsymbol{\theta}) P(W_r^T|\boldsymbol{\theta})
\end{aligned}
$$

  - 其中，$r_{max} = M^T$为HMM所有可能的状态转移序列数
    - $P(W_r^T|\boldsymbol{\theta})$可以由参数$\boldsymbol{\pi}$和$A$直接计算；
    - $P(V^T|W_r^T, \boldsymbol{\theta})$可以由参数$B$直接计算；

例题3.1

计算$M=4, K=3$的HMM，输出$V^5=v_2v_2v_1v_3v_1$的概率；

列出所有可能的状态转移序列：

$$W_1^5=w_1w_1w_1w_1w_1 \quad W_2^5=w_1w_1w_1w_1w_2$$

$$W_3^5=w_1w_1w_1w_1w_3 \quad W_4^5=w_1w_1w_1w_4$$

$$\cdots\cdots$$

$$W_{1021}^5=w_4w_4w_4w_4w_1 \quad W_{1022}^5=w_4w_4w_4w_4w_2$$

$$W_{1023}^5=w_4w_4w_4w_4w_3 \quad W_{1024}^5=w_4w_4w_4w_4w_4$$

例题3.1

计算隐状态序列与观察序列同时发生的概率：

$$
\begin{aligned}
P(V^5, W_1^5) &= \pi_1 b_{12} a_{11} b_{12} a_{11} b_{11} a_{11} b_{13} a_{11} b_{11} \\
P(V^5, W_2^5) &= \pi_1 b_{12} a_{11} b_{12} a_{11} b_{11} a_{11} b_{13} a_{12} b_{21} \\
P(V^5, W_3^5) &= \pi_1 b_{12} a_{11} b_{12} a_{11} b_{11} a_{11} b_{13} a_{13} b_{31} \\
P(V^5, W_4^5) &= \pi_1 b_{12} a_{11} b_{12} a_{11} b_{11} a_{11} b_{13} a_{14} b_{41} \\
P(V^5, W_5^5) &= \pi_1 b_{12} a_{11} b_{12} a_{11} b_{11} a_{12} b_{23} a_{21} b_{11} \\
P(V^5, W_6^5) &= \pi_1 b_{12} a_{11} b_{12} a_{11} b_{11} a_{12} b_{23} a_{22} b_{21} \\
P(V^5, W_7^5) &= \pi_1 b_{12} a_{11} b_{12} a_{11} b_{11} a_{12} b_{23} a_{23} b_{31} \\
P(V^5, W_8^5) &= \pi_1 b_{12} a_{11} b_{12} a_{11} b_{11} a_{12} b_{23} a_{24} b_{41} \\
&\cdots\cdots
\end{aligned}
$$

HMM输出V^5的概率：

$$P(V^5) = \sum_{r=1}^{1024} P(V^5, W_r^5)$$

估值问题

- **简化计算**
  - 估值问题直接计算的复杂度高：$O(M^T T)$
  - 计算过程是可以简化的；

$$P(V^5, W_1^5) = \pi_1 b_{12} a_{11} b_{12} a_{11} b_{11} a_{11} b_{13} a_{11} b_{11}$$
$$P(V^5, W_2^5) = \pi_1 b_{12} a_{11} b_{12} a_{11} b_{11} a_{11} b_{13} a_{12} b_{21}$$
$$P(V^5, W_3^5) = \pi_1 b_{12} a_{11} b_{12} a_{11} b_{11} a_{11} b_{13} a_{13} b_{31}$$
$$P(V^5, W_4^5) = \pi_1 b_{12} a_{11} b_{12} a_{11} b_{11} a_{11} b_{13} a_{14} b_{41}$$

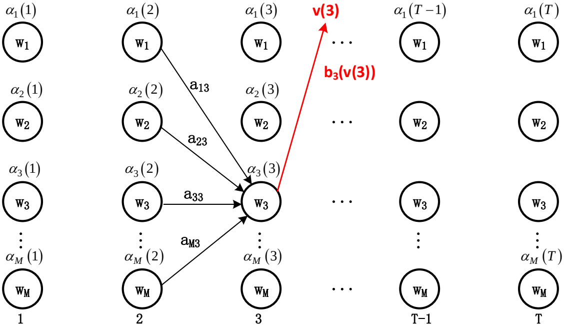
$$P(V^5, W_5^5) = \pi_1 b_{12} a_{11} b_{12} a_{11} b_{11} a_{12} b_{23} a_{21} b_{11}$$
$$P(V^5, W_6^5) = \pi_1 b_{12} a_{11} b_{12} a_{11} b_{11} a_{12} b_{23} a_{22} b_{21}$$
$$P(V^5, W_7^5) = \pi_1 b_{12} a_{11} b_{12} a_{11} b_{11} a_{12} b_{23} a_{23} b_{31}$$
$$P(V^5, W_8^5) = \pi_1 b_{12} a_{11} b_{12} a_{11} b_{11} a_{12} b_{23} a_{24} b_{41}$$

$$\cdots\cdots$$

估值问题

- **前向迭代**
  - 定义变量$\alpha_i(t)$：时刻t，模型处于隐状态w_i，并且输出观察序列$V^{1\to t}=v(1)\cdots v(t)$的概率

$$\alpha_i(t)=P(V^{1\to t},w(t)=w_i|\boldsymbol{\theta})$$



α的迭代计算

$$
\begin{aligned}
\alpha_i(t+1) &= P(V^{1\to t+1}, w(t+1)=w_i) \\
&= P(V^{1\to t}, v(t+1), w(t+1)=w_i) \\
&= \sum_{j=1}^{M} P(V^{1\to t}, v(t+1), w(t)=w_j, w(t+1)=w_i) \\
&= \sum_{j=1}^{M} \left[\begin{array}{c} P(v(t+1)|V^{1\to t}, w(t)=w_j, w(t+1)=w_i)\times \\ P(w(t+1)=w_i|V^{1\to t}, w(t)=w_j) \times P(V^{1\to t}, w(t)=w_j) \end{array} \right] \\
&= \sum_{j=1}^{M} \left[\begin{array}{c} P(v(t+1)|w(t+1)=w_i)\times \\ P(w(t+1)=w_i|w(t)=w_j) \times P(V^{1\to t}, w(t)=w_j) \end{array} \right] \\
&= \sum_{j=1}^{M} \alpha_j(t) a_{ji} b_i(v(t+1))
\end{aligned}
$$

$$
P(V^T|\boldsymbol{\theta}) = \sum_{i=1}^{M} P(V^T, w(T)=w_i|\boldsymbol{\theta}) = \sum_{i=1}^{M} \alpha_i(T)
$$

前向算法

Algorithm 4 前向算法(计算复杂度$O(M^2T)$)

Input: HMM模型参数$\boldsymbol{\theta}$，观察序列V^T

Output: 模型输出观察序列的概率$P(V^T|\boldsymbol{\theta})$

1: 初始化：$i = 1, \cdots, M$

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

2: 迭代递推：$i = 1, \cdots, M, t = 1, \cdots, T-1$

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

3: 结束输出：

$$P(V^T|\boldsymbol{\theta}) = \sum_{i=1}^{M} P(V^T, w(T) = w_i|\boldsymbol{\theta}) = \sum_{i=1}^{M} \alpha_i(T)$$

解码问题

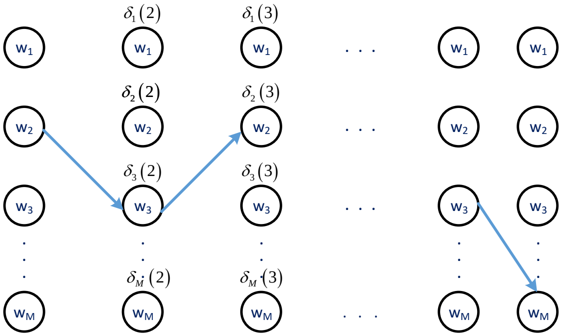
- **解码问题**
  - 已知模型参数$\boldsymbol{\theta}$，计算最有可能输出观察序列V^T的状态转移序列，类似于估值问题：

$$\begin{aligned} W_*^T &= \arg\max_{W^T} P(V^T, W^T|\boldsymbol{\theta}) \\ &= \arg\max_{W^T} P(V^T|W^T, \boldsymbol{\theta})P(W^T|\boldsymbol{\theta}) \end{aligned}$$

  - 直接计算的复杂度为$O(M^T T)$；
  - 存在复杂度为$O(M^2 T)$的简化算法：Viterbi算法

解码问题

- **定义变量**
  - $\delta_i(t)$：时刻t模型处于状态w_i，输出$V^{1\to t}$的最优路径概率；
  - $\phi_i(t)$：最优路径上$t-1$时刻的状态，用于路径回溯；



Viterbi算法

Algorithm 5 Viterbi算法(计算复杂度$O(M^2T)$)

Input: HMM模型参数$\boldsymbol{\theta}$，观察序列V^T

Output: 模型输出观察序列的最优隐状态序列W^*

1: 初始化：$i = 1, \cdots, M$

$$\delta_i(1) = \pi_i b_i(v(1)), \quad \phi_i(1) = 0$$

2: 迭代递推：$i = 1, \cdots, M, t = 1, \cdots, T-1$

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

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

3: 结束输出：

$$P^*(V^T|\boldsymbol{\theta}) = \max_{1 \le j \le M} \delta_i(T), \quad w^*(T) = \arg \max_{1 \le j \le M} \delta_j(T)$$

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

学习问题

- **HMM的学习问题**
  - 已知一组观察序列(训练样本集合)：

$$\mathcal{V} = \{V_1^{T_1}, \cdots, V_n^{T_n}\}$$

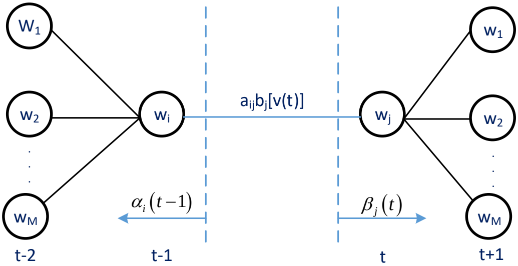
  - 计算参数$\boldsymbol{\theta}$的最大似然估计，使得产生训练集$\mathcal{V}$的概率最大：

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} P(\mathcal{V}|\boldsymbol{\theta})$$

  - 模型在产生每一个观察序列时的状态转移序列是隐含的、未知的，需要使用EM算法来估计参数$\boldsymbol{\theta}$；
  - 下面给出输入单个观察序列V^T时，EM算法的迭代公式；

Baum-Welch算法

- 定义变量
  - 前向变量$\alpha_i(t-1)$：表示$t-1$时刻HMM处于状态w_i，并且在$1 \to t-1$时刻之间产生观察序列$V^{1 \to t-1}$的概率；
  - 后向变量$\beta_j(t)$：表示t时刻HMM处于状态w_j的条件下，$t+1 \to T$时刻之间产生观察序列$V^{t+1 \to T}$的概率；



Baum-Welch算法

- **前向迭代**
  - 初始化于时刻1：

$$\alpha_i(1) = \pi_1 b_i(v(1)), \quad i = 1, \cdots, M$$

  - 由前向后迭代：

$$\alpha_i(t-1) = \left[\sum_{k=1}^{M} \alpha_k(t-2) a_{ki} \right] b_i(v(t-1))$$

- **后向迭代**
  - 初始化于时刻T：

$$\beta_j(T) = 1, \quad j = 1, \cdots, M$$

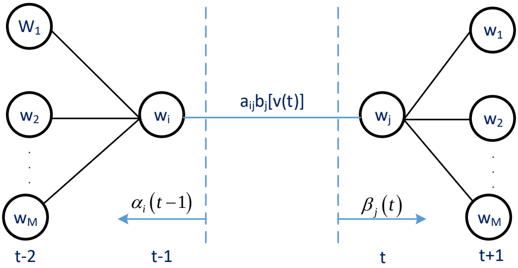
  - 由后向前迭代：

$$\beta_j(t) = \sum_{k=1}^{M} \beta_k(t+1) a_{jk} b_k(v(t+1))$$

Baum-Welch算法

- **定义变量**
  - $\gamma_{ij}(t)$：表示模型$t-1$时刻处于状态w_i，t时刻处于状态w_j，并且输出观察序列V^T的概率

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



Baum-Welch算法

- **HMM参数更新**
  - 初始概率

$$\hat{\pi}_i = \gamma_i(1) = \frac{\pi_i b_i(v(1))\beta_i(1)}{P(V^T|\boldsymbol{\theta})}$$

  - 状态转移概率

$$\hat{a}_{ij} = \frac{\sum_{t=2}^{T} \gamma_{ij}(t)}{\sum_{t=2}^{T}\sum_{k=1}^{M} \gamma_{ik}(t)}$$

  - 输出概率

$$\hat{b}_i(v_k) = \frac{\sum_{t=1,v(t)=v_k}^{T}\sum_{l=1}^{M} \gamma_{il}(t)}{\sum_{t=1}^{T}\sum_{l=1}^{M} \gamma_{il}(t)}$$

  - 学习多个训练样本时，分子和分母分别对样本进行累加；

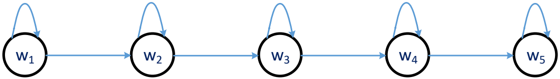
连续型HMM

- **离散型HMM**
  - 观察值为有限的离散值；
  - $\boldsymbol{\theta}$中的输出概率以矩阵B的形式描述，每一行表示相应状态输出观察值的概率分布；
- **连续型HMM**
  - 观察值为连续的特征矢量；
  - $\boldsymbol{\theta}$中的输出概率以M个概率密度函数描述；
  - $b_i(\mathbf{v})$为状态w_i输出观察矢量$\mathbf{v}$的概率密度，通常采用GMM的形式描述；

HMM的拓扑结构

HMM模型的拓扑结构可以根据实际问题的需要来设计;

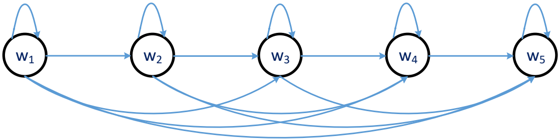
“左-右”模型结构:



$$\boldsymbol{\pi} = (1, 0, \cdots, 0)^t, \quad A = \begin{bmatrix} a_{11} & a_{12} & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{MM} \end{bmatrix}$$

HMM的拓扑结构

带跨越的“左-右”模型结构：



$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1M} \\ 0 & a_{22} & \cdots & a_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{MM} \end{bmatrix}$$

3.4 贝叶斯估计

贝叶斯估计与最大似然估计

- **最大似然估计**
  - 认为分布的参数$\boldsymbol{\theta}$是一个确定的未知矢量；
  - 参数估计的目标就是要寻找最有可能产生训练集D的参数：

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

- **贝叶斯估计**
  - 认为分布的参数$\boldsymbol{\theta}$是一个随机矢量；
  - 参数估计的目标是在已知训练集D的条件下，参数$\boldsymbol{\theta}$的分布：$p(\boldsymbol{\theta}|D)$

贝叶斯学习

- **学习过程**
  - 参数的后验分布：

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

- **识别过程**
  - 计算已知训练集D的条件下，待识模式$\mathbf{x}$的后验概率密度：

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

  已知参数$\boldsymbol{\theta}$的条件下，样本$\mathbf{x}$的分布与训练集D无关，因此有$p(\mathbf{x}|\boldsymbol{\theta}, D) = p(\mathbf{x}|\boldsymbol{\theta})$；

贝叶斯估计和最大似然估计

- **学习和识别过程的差异**
  - 最大似然估计：

$$D \overset{\text{学习}}{\Longrightarrow} \hat{\boldsymbol{\theta}} \overset{\text{识别}}{\Longrightarrow} p(\mathbf{x}|D) \approx p(\mathbf{x}|\hat{\boldsymbol{\theta}})$$

  - 贝叶斯估计

$$\{D, p(\boldsymbol{\theta})\} \overset{\text{学习}}{\Longrightarrow} p(\boldsymbol{\theta}|D) \overset{\text{识别}}{\Longrightarrow} p(\mathbf{x}|D) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|D)d\boldsymbol{\theta}$$

- **优缺点**
  - 最大似然估计计算简单，贝叶斯估计计算过程复杂；
  - 贝叶斯估计能够提高小样本训练集的概率密度估计准确性，但需要知道参数的先验分布$p(\boldsymbol{\theta})$；

正态分布密度参数的贝叶斯估计

- **一维正态分布均值的贝叶斯估计**
  - 数据集$D=\{x_1,\cdots,x_n\}$服从一维正态分布$N(\mu,\sigma^2)$；
  - 假设方差σ^2已知，均值μ的先验为正态分布$N(\mu_0,\sigma_0^2)$：
  - 即样本x和参数μ分布服从两个不同的正态分布：

$$p(x|\mu)=N(x;\mu,\sigma^2)$$

$$p(\mu)=N(\mu;\mu_0,\sigma_0^2)$$

正态分布密度参数的贝叶斯估计

参数μ的后验分布：

$$
\begin{aligned}
p(\mu|D) &= \frac{p(D|\mu)p(\mu)}{\int p(D|\mu)p(\mu)d\mu} = \alpha \prod_{i=1}^{n} p(x_i|\mu)p(\mu) \\
&= \alpha \left[\prod_{i=1}^{n} N(x_i; \mu, \sigma^2) \right] N(\mu; \mu_0, \sigma_0^2) \\
&= \alpha' \exp \left\{ -\frac{1}{2} \left[\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) \mu^2 - 2 \left(\frac{1}{\sigma^2} \sum_{i=1}^{n} x_i + \frac{\mu_0}{\sigma_0^2} \right) \mu \right] \right\} \\
&= \frac{1}{\sqrt{2\pi}\sigma_n} \exp \left[-\frac{1}{2} \left(\frac{\mu - \mu_n}{\sigma_n} \right)^2 \right] \\
&= N(\mu_n, \sigma_n^2)
\end{aligned}
$$

正态分布密度参数的贝叶斯估计

- 参数μ的后验分布
  - 推导过程中，由于exp的指数部分是μ的标准二次型，因此可以断言μ的后验服从正态分布(更详细的过程参见“补充材料”)；
  - 对比μ的二次型系数，可以得到后验分布的均值和方差：

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

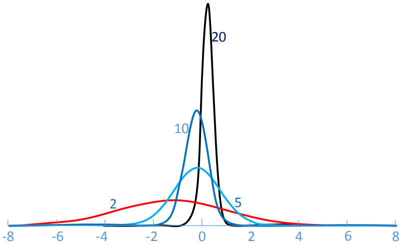
$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}$$

其中，$\hat{\mu}_n$是μ的最大似然估计：

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

正态分布密度参数的贝叶斯估计

- μ的分布与训练样本数n
  - $n=0$时：$\mu_n=\mu_0, \sigma_n^2=\sigma_0^2$
  - 随着n的增大：μ_n由μ_0趋近$\hat{\mu}_n$，σ_n^2减小，估计的确信度增大
  - $n\rightarrow\infty$时：$\mu_n\rightarrow\hat{\mu}_n, \sigma_n^2\rightarrow 0$



$$p(\mu|D)=N(\mu;\mu_n,\sigma_n^2)$$

正态分布密度参数的贝叶斯估计

样本x的后验分布：

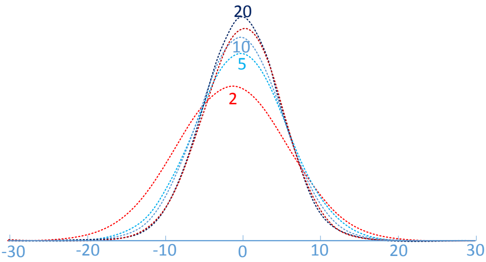
$$
\begin{aligned}
p(x|D) &= \int p(x|\mu)p(\mu|D)d\mu \\
&= \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] d\mu \\
&= \frac{f(\sigma,\sigma_n)}{2\pi\sigma\sigma_n}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

其中：

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

正态分布密度参数的贝叶斯估计

- **x的分布与训练样本数n**
  - $n \to \infty$时：$\mu_n \to \hat{\mu}_n, \sigma_n^2 \to 0$
  - 贝叶斯估计与最大似然估计趋向一致：$p(x|D) \to N(\hat{\mu}_n, \sigma^2)$



$$p(x|D) = N(x; \mu_n, \sigma^2 + \sigma_n^2)$$

共轭先验分布

- 参数的先验分布$p(\boldsymbol{\theta})$
  - 一般将参数的先验分布假设为其共轭分布；
  - 这样可以使得参数的后验分布与先验分布属于同一分布族；
  - 样本的后验分布$p(\mathbf{x}|D)$通常也非常容易计算；
- **GMM参数的共轭先验**
  - $\boldsymbol{\mu}$的共轭先验：Gauss分布
  - Σ的共轭先验：inverse-Wishart分布
  - $\boldsymbol{\alpha}$的共轭先验：Dirichlet分布

End