第6章　贝叶斯分类器

刘家锋

哈尔滨工业大学

第6章 贝叶斯分类器

6.1 贝叶斯决策论

分类与概率

- **概率的角度看分类问题**
  - 将样例$\mathbf{x}$视作随机向量，类别标记y视作有N种取值的离散随机变量：
  $$y \in \mathcal{Y} = \{c_1, \cdots, c_N\}$$
  - 分类可以看作是在已知样例$\mathbf{x}$的条件下，对类别y的决策；
  - y是随机的，因此任何的决策都有可能发生错误，分类问题自然希望发生决策错误的概率越小越好；
- **类别的先验概率**
  - 如果我们不知道样例的属性$\mathbf{x}$，那么只能依据类别的先验概率$P(y)$来决策；
  - 哪个类别的先验概率大，就判别样例属于哪个类别：
  $$y^* = \arg\max_{c \in \mathcal{Y}} P(y = c)$$

最小错误率

- **类别的后验概率**
  - 如果我们知道样例的属性$\mathbf{x}$，就可以依据类别的后验概率$P(y|\mathbf{x})$来决策；
  - 哪个类别的后验概率大，就判别样例属于哪个类别：

$$y^* = \arg\max_{c \in \mathcal{Y}} P(y = c|\mathbf{x})$$

- **最小错误率判别**
  - 依据后验概率的判别，可以取得最小的错误率；
  - 如果决策$y = c_i$，则当真实类别为$c_j, j \neq i$时发生错误，因此决策的错误率为：

$$P_i(\mathsf{error}|\mathbf{x}) = \sum_{j \neq i} P(y = c_j|\mathbf{x}) = 1 - P(y = c_i|\mathbf{x})$$

最小化风险

- 条件风险
  - 最小错误率认为所有的判别错误都是相同的；
  - 如果将一个真实标记为c_j的样本误分类为c_i的损失为λ_{ij}，那么将$\mathbf{x}$判别为c_i类的条件风险为：

$$R(c_i|\mathbf{x}) = \sum_{j=1}^{N} \lambda_{ij} P(y = c_j|\mathbf{x})$$

  - 依据最小化条件风险的准则判别为：

$$y^* = \arg\min_{c \in \mathcal{Y}} R(c|\mathbf{x})$$

  - 最小错误率判别等价于：

$$\lambda_{ij} = \begin{cases} 0, & \text{if} \quad i = j \\ 1, & \text{otherwise} \end{cases}$$

生成式与判别式模型

- 判别式模型(discriminative models)
  - 模型化后验概率$P(y|\mathbf{x})$来判别的方法，称为判别式模型；
  - 线性判别，SVM，神经网络和决策树都属于判别式模型；
- 生成式模型(generative models)
  - 模型化联合概率$P(\mathbf{x},y)$或类条件概率$p(\mathbf{x}|y)$来判别的方法，称为生成式模型；
  - 条件概率公式：

$$P(\mathbf{x},y)=P(y|\mathbf{x})p(\mathbf{x})=p(\mathbf{x}|y)P(y)$$

  - 贝叶斯公式：

$$P(y|\mathbf{x})=\frac{P(y)p(\mathbf{x}|y)}{p(\mathbf{x})}$$

贝叶斯判别

- **联合概率的判别**
  - $p(\mathbf{x})$是一个与类别无关的归一化因子，称为“证据”；
  - 依据联合概率的判别等价于后验概率的判别：

$$\arg\max_{c\in\mathcal{Y}} P(\mathbf{x}, y=c) \quad \Leftrightarrow \quad \arg\max_{c\in\mathcal{Y}} P(y=c|\mathbf{x})$$

- **贝叶斯判别**
  - 同样，依据贝叶斯公式可以得到：

$$\arg\max_{c\in\mathcal{Y}} p(\mathbf{x}|y=c)P(y=c) \quad \Leftrightarrow \quad \arg\max_{c\in\mathcal{Y}} P(y=c|\mathbf{x})$$

  - 先验概率$P(y)$可以利用先验知识得到，也可以用训练集中各个类别样本所占的比例来估计；
  - 贝叶斯判别的学习，主要是估计类条件概率$p(\mathbf{x}|y)$；

6.2 极大似然估计

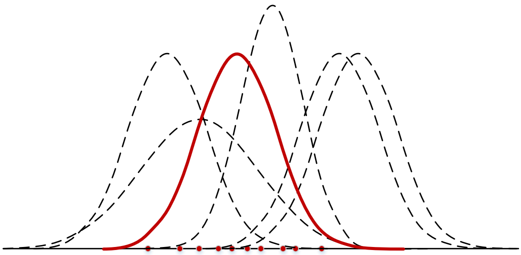
极大似然估计

- **概率分布的参数估计**
  - 假定类条件概率$p(\mathbf{x}|y=c)$具有确定的分布形式，并且被参数$\boldsymbol{\theta}_c$唯一确定；
  - 令D_c表示训练集D中第c类样本组成的集合，并且是独立同分布的样本；
  - 贝叶斯分类器的学习，就是利用数据集D_c来估计参数$\boldsymbol{\theta}_c$，其中$c \in \mathcal{Y} = \{c_1, \cdots, c_N\}$；
- **似然函数**
  - 定义给定参数$\boldsymbol{\theta}_c$条件下，样本集D_c中样本发生的联合概率为似然函数；
  - 似然函数为参数$\boldsymbol{\theta}_c$的函数，根据独立同分布假设有：

$$p(D_c|\boldsymbol{\theta}_c) = \prod_{\mathbf{x} \in D_c} p(\mathbf{x}|\boldsymbol{\theta}_c)$$

极大似然估计

- **极大似然估计**
  - 极大似然估计的思路是在给定的分布形式中，找到一个最有可能产生出训练集D_c的分布；
  - 给定形式的分布由参数$\boldsymbol{\theta}_c$唯一确定，因此以最大化似然函数的参数作为估计结果；



极大似然估计

- **对数似然函数**
  - 概率密度函数的值往往比较小，连乘容易造成计算下溢；
  - 对数函数是单调上升的，一般以对数似然函数代替似然函数作为最大似然估计的优化目标：

$$LL(\boldsymbol{\theta}_c) = \ln p(D_c|\boldsymbol{\theta}_c) = \sum_{\mathbf{x}\in D_c} \ln p(\mathbf{x}|\boldsymbol{\theta}_c)$$

- **极大似然估计**
  - 极大似然估计需要求解如下优化问题：

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

例6.1 正态分布参数估计

类别c的训练集$D_c=\{x_1,\cdots,x_{m_c}\}$，服从参数$\boldsymbol{\theta}_c=(\mu_c,\sigma_c^2)^t$的1维正态分布：

$$p(x|\mu_c,\sigma_c^2)=\frac{1}{\sqrt{2\pi}\sigma_c}\exp\left[-\frac{(x-\mu_c)^2}{2\sigma_c^2}\right]$$

对数似然函数：

$$LL(\mu_c,\sigma_c^2)=\sum_{i=1}^{m_c}\ln p(x_i|\mu_c,\sigma_c^2)=\sum_{i=1}^{m_c}-\frac{1}{2}\left[\ln 2\pi+\ln\sigma_c^2+\frac{(x_i-\mu_c)^2}{\sigma_c^2}\right]$$

计算偏导数，求极值：

$$\frac{\partial LL(\mu_c,\sigma_c^2)}{\partial\mu_c}=\sum_{i=1}^{m_c}\frac{1}{\sigma_c^2}(x_i-\mu_c)=0$$

$$\frac{\partial LL(\mu_c,\sigma_c^2)}{\partial\sigma_c^2}=\sum_{i=1}^{m_c}\left[-\frac{1}{2\sigma_c^2}+\frac{(x_i-\mu_c)^2}{2\sigma_c^4}\right]=0$$

例6.1 正态分布参数估计

求解方程，得到参数的极大似然估计：

$$\hat{\mu}_c = \frac{1}{m_c}\sum_{i=1}^{m_c} x_i, \qquad \hat{\sigma}_c^2 = \frac{1}{m_c}\sum_{i=1}^{m_c}(x_i - \hat{\mu})_c^2$$

样本集D_c服从d维正态分布：

$$p(\mathbf{x}|\boldsymbol{\mu}_c, \Sigma_c) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_c|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_c)^t\Sigma_c^{-1}(\mathbf{x}-\boldsymbol{\mu}_c)\right]$$

同样方法，可以得到多元正态分布参数的极大似然估计：

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{|D_c|}\sum_{\mathbf{x}\in D_c}\mathbf{x}$$

$$\hat{\Sigma}_c = \frac{1}{|D_c|}\sum_{\mathbf{x}\in D_c}(\mathbf{x}-\hat{\boldsymbol{\mu}}_c)(\mathbf{x}-\hat{\boldsymbol{\mu}}_c)^t$$

Naïve Bayes Classifier

- **朴素贝叶斯分类器**
  - 有限训练样本估计高维联合概率(密度)$p(\mathbf{x}|y=c)$存在困难；
  - 朴素贝叶斯对模型进行了简化，假设$\mathbf{x}$的属性之间是相互独立的，即：

$$p(\mathbf{x}|y=c)=\prod_{i=1}^{d}p(x_i|y=c)$$

  - 相应的贝叶斯判别：

$$y^*=\arg\max_{c\in\mathcal{Y}}P(y=c)\prod_{i=1}^{d}p(x_i|y=c)$$

  - 学习时，可以由D_c单独估计每个属性的分布$p(x_i|y=c)$；

例6.2 朴素贝叶斯分类

西瓜数据集

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

例6.2 朴素贝叶斯分类

估计先验概率：

$$P(\text{好瓜}=\text{是})=\frac{8}{17}\approx 0.471, \qquad P(\text{好瓜}=\text{否})=\frac{9}{17}\approx 0.529$$

估计属性“色泽”的条件概率：

$$P_{\text{青绿}|\text{是}}=P(\text{色泽}=\text{青绿}|\text{好瓜}=\text{是})=\frac{3}{8}=0.375$$

$$P_{\text{乌黑}|\text{是}}=P(\text{色泽}=\text{乌黑}|\text{好瓜}=\text{是})=\frac{4}{8}=0.500$$

$$P_{\text{浅白}|\text{是}}=P(\text{色泽}=\text{浅白}|\text{好瓜}=\text{是})=\frac{1}{8}=0.125$$

$$P_{\text{青绿}|\text{否}}=P(\text{色泽}=\text{青绿}|\text{好瓜}=\text{否})=\frac{3}{9}\approx 0.333$$

$$P_{\text{乌黑}|\text{否}}=P(\text{色泽}=\text{乌黑}|\text{好瓜}=\text{否})=\frac{2}{9}\approx 0.222$$

$$P_{\text{浅白}|\text{否}}=P(\text{色泽}=\text{浅白}|\text{好瓜}=\text{否})=\frac{4}{9}\approx 0.444$$

例6.2 朴素贝叶斯分类

估计属性“根蒂”、“敲声”、“纹理”、“脐部”、“触感”的条件概率：

… …

估计属性“密度”的条件概率密度，假设属性服从正态分布：

$$\hat{\mu}_{密度|是} = \frac{1}{8}\sum_{i=1}^{8} x^i_{密度} \approx 0.574, \qquad \hat{\sigma}^2_{密度|是} = \frac{1}{8}\sum_{i=1}^{8} (x^i_{密度} - \hat{\mu}_{密度|是})^2 \approx 0.0166$$

$$\hat{\mu}_{密度|否} = \frac{1}{9}\sum_{i=9}^{17} x^i_{密度} \approx 0.496, \qquad \hat{\sigma}^2_{密度|否} = \frac{1}{9}\sum_{i=9}^{17} (x^i_{密度} - \hat{\mu}_{密度|否})^2 \approx 0.0380$$

估计属性“含糖率”的条件概率密度，假设属性服从正态分布：

$$\hat{\mu}_{含糖|是} = \frac{1}{8}\sum_{i=1}^{8} x^i_{含糖} \approx 0.279, \qquad \hat{\sigma}^2_{含糖|是} = \frac{1}{8}\sum_{i=1}^{8} (x^i_{含糖} - \hat{\mu}_{含糖|是})^2 \approx 0.0102$$

$$\hat{\mu}_{含糖|否} = \frac{1}{9}\sum_{i=9}^{17} x^i_{含糖} \approx 0.154, \qquad \hat{\sigma}^2_{含糖|否} = \frac{1}{9}\sum_{i=9}^{17} (x^i_{含糖} - \hat{\mu}_{含糖|否})^2 \approx 0.0117$$

例6.2 朴素贝叶斯分类

判别下列测试样本$\mathbf{x}$“是/否”好瓜？

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测试1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

$$
\begin{aligned}
p(\mathbf{x}|\text{是})P(\text{是}) =& P(\text{是}) \times [P_{\text{青绿}|\text{是}} \cdot P_{\text{蜷缩}|\text{是}} \cdot P_{\text{浊响}|\text{是}} \cdot P_{\text{清晰}|\text{是}} \cdot P_{\text{凹陷}|\text{是}} \cdot P_{\text{硬滑}|\text{是}} \cdot \\
& p(\text{密度} = 0.697|\text{是}) \cdot p(\text{含糖率} = 0.460|\text{是})] \\
p(\mathbf{x}|\text{否})P(\text{否}) =& P(\text{否}) \times [P_{\text{青绿}|\text{否}} \cdot P_{\text{蜷缩}|\text{否}} \cdot P_{\text{浊响}|\text{是}} \cdot P_{\text{清晰}|\text{是}} \cdot P_{\text{凹陷}|\text{是}} \cdot P_{\text{硬滑}|\text{是}} \cdot \\
& p(\text{密度} = 0.697|\text{否}) \cdot p(\text{含糖率} = 0.460|\text{否})]
\end{aligned}
$$

例6.2 朴素贝叶斯分类

离散属性的概率值可以查表得到，连续属性的概率密度需要计算：

$$p(\text{密度}=0.697|\text{好瓜}=\text{是})=\frac{1}{\sqrt{2\pi}\sigma_{\text{密度}|\text{是}}}\exp\left(-\frac{(0.697-\mu_{\text{密度}|\text{是}})^2}{2\sigma^2_{\text{密度}|\text{是}}}\right)\approx 1.959$$

$$p(\text{密度}=0.697|\text{好瓜}=\text{否})=\frac{1}{\sqrt{2\pi}\sigma_{\text{密度}|\text{否}}}\exp\left(-\frac{(0.697-\mu_{\text{密度}|\text{否}})^2}{2\sigma^2_{\text{密度}|\text{否}}}\right)\approx 1.203$$

$$p(\text{含糖}=0.460|\text{好瓜}=\text{是})=\frac{1}{\sqrt{2\pi}\sigma_{\text{含糖}|\text{是}}}\exp\left(-\frac{(0.460-\mu_{\text{含糖}|\text{是}})^2}{2\sigma^2_{\text{含糖}|\text{是}}}\right)\approx 0.788$$

$$p(\text{含糖}=0.460|\text{好瓜}=\text{否})=\frac{1}{\sqrt{2\pi}\sigma_{\text{含糖}|\text{否}}}\exp\left(-\frac{(0.460-\mu_{\text{含糖}|\text{否}})^2}{2\sigma^2_{\text{含糖}|\text{否}}}\right)\approx 0.066$$

代入，得到：

$$p(\mathbf{x}|\text{是})P(\text{是})\approx 0.052 \quad > \quad p(\mathbf{x}|\text{否})P(\text{否})\approx 6.80\times 10^{-5}$$

判别“测试1”样本“是”好瓜；

拉普拉斯修正

- **未出现的属性值**
  - 某个属性值在训练集没有与某个类别同时出现，则该属性的条件概率为0；
  - 如果在测试数据中该属性值出现，直接将其判别为不属于此类，是不合理的；

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测试2	青绿	蜷缩	清脆	清晰	凹陷	硬滑	0.697	0.460	?

$$p(\mathbf{x}|\text{是})P(\text{是}) = 0 \quad < \quad p(\mathbf{x}|\text{否})P(\text{否}) \approx 2.56 \times 10^{-5}$$

判别“测试2”样本“不是”好瓜；

拉普拉斯修正

- **概率估计的平滑**
  - 拉普拉斯修正可以对概率估计值进行平滑；
  - 令N表示类别数，N_i表示第i个属性的取值数，D_{c,x_i}表示类别c的训练集中属性i取值x_i的样本集合；
  - 类别c的先验概率和条件概率估计的修正为：

$$\hat{P}(y=c) = \frac{|D_c|+1}{|D|+N}, \qquad \hat{P}(x_i|y=c) = \frac{|D_{c,x_i}|+1}{|D_c|+N_i}$$

  - 修正后的概率估计：

$$\hat{P}(\text{好瓜}=\text{是}) = \frac{8+1}{17+2} \approx 0.474, \quad \hat{P}(\text{好瓜}=\text{否}) = \frac{9+1}{17+2} \approx 0.526$$

$$\hat{P}_{\text{清脆}|\text{是}} = \hat{P}(\text{敲声}=\text{清脆}|\text{好瓜}=\text{是}) = \frac{0+1}{8+3} \approx 0.091$$

$$\hat{P}_{\text{清脆}|\text{否}} = \hat{P}(\text{敲声}=\text{清脆}|\text{好瓜}=\text{否}) = \frac{2+1}{9+3} = 0.25$$

6.3 GMM与EM

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由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的线性组合构成：

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{i=1}^{k} \alpha_i p_i(\mathbf{x}|\boldsymbol{\theta}_i)$$

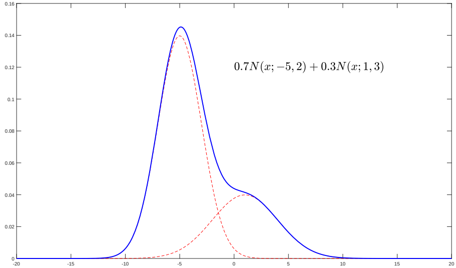
其中，$\sum_{i=1}^{k} \alpha_i = 1, \alpha_i > 0$

- **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：

$$p(\mathbf{x}) = \sum_{i=1}^{k} \alpha_i N(\mathbf{x}; \boldsymbol{\mu}_i, \Sigma_i)$$

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k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：

$$\boldsymbol{\theta} = (\alpha_1, \cdots, \alpha_k, \boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k, \Sigma_1, \cdots, \Sigma_k)$$

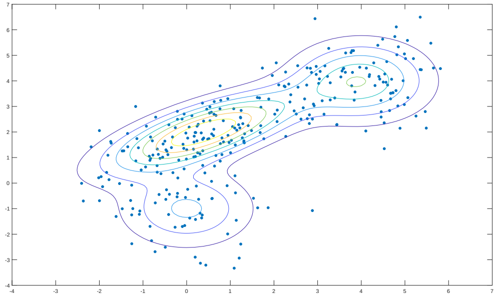
  - 对数似然函数：

$$LL(\boldsymbol{\theta}) = \ln\left[\sum_{i=1}^{k} \alpha_i N(\mathbf{x}; \boldsymbol{\mu}_i, \Sigma_i)\right]$$

  - 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$\{\alpha_1, \cdots, \alpha_k\}$，选择一个成份高斯分布；
  - 依据成份高斯的分布参数，产生具体的样本属性向量；



GMM的参数估计问题

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

$$Z = \{z_1, \cdots, z_m\}$$

  - 完整的数据集$D = X \cup Z$，其中Z为缺失的数据；

GMM的参数估计问题

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

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

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

$$\hat{\alpha}_i=\frac{1}{m}\sum_{j=1}^{m} I(z_j=i)$$

  - 每个高斯的参数用该高斯产生的样本估计：

$$\hat{\boldsymbol{\mu}}_i=\frac{\sum_{j=1}^{m} I(z_j=i)\mathbf{x}_j}{\sum_{j=1}^{m} I(z_j=i)}$$

$$\hat{\Sigma}_i=\frac{\sum_{j=1}^{m} I(z_j=i)(\mathbf{x}_j-\hat{\boldsymbol{\mu}}_i)(\mathbf{x}_j-\hat{\boldsymbol{\mu}}_i)^t}{\sum_{j=1}^{m} I(z_j=i)}$$

GMM的参数估计问题

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

$$\hat{z}_j = \arg\max_{1\le i\le k} \alpha_i N(\mathbf{x}_j; \boldsymbol{\mu}_i, \Sigma_i)$$

- **Z和θ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Z；
  - 固定Z，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8章）；
  - 已知样本集$X = \{\mathbf{x}_1, \cdots, \mathbf{x}_m\}$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着内在的联系；
  - 如果假设数据集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$\{\alpha_i\}$，则样本集X服从GMM分布；
  - 对数据集Z的估计，实质上就是对X的聚类划分；

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Z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$\mathbf{x}_j$由某个高斯产生，过于武断；
  - 合理的方式是推断样本$\mathbf{x}_j$由每一个高斯产生的概率：

$$\gamma_{ji} = P(z_j = i) = \frac{\alpha_i N(\mathbf{x}_j; \boldsymbol{\mu}_i, \Sigma_i)}{\sum_{i=1}^{k} \alpha_i N(\mathbf{x}_j; \boldsymbol{\mu}_i, \Sigma_i)} \tag{1}$$

  - 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：

$$\hat{\alpha}_i = \frac{1}{m}\sum_{j=1}^{m}\gamma_{ji} \tag{2}$$

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{j=1}^{m}\gamma_{ji}\mathbf{x}_j}{\sum_{j=1}^{m}\gamma_{ji}} \tag{3}$$

$$\hat{\Sigma}_i = \frac{\sum_{j=1}^{m}\gamma_{ji}(\mathbf{x}_j-\hat{\boldsymbol{\mu}}_i)(\mathbf{x}_j-\hat{\boldsymbol{\mu}}_i)^t}{\sum_{j=1}^{m}\gamma_{ji}} \tag{4}$$

  - 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 1 GMM参数估计的EM算法

Input: 训练数据集$X = \{\mathbf{x}_1, \cdots, \mathbf{x}_m\}$

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，Z是缺失的数据集，$D = X \cup Z$；
  - $\boldsymbol{\theta}$是我们要估计的分布参数，对数似然函数：

$$LL(\boldsymbol{\theta}) = \ln p(X, Z|\boldsymbol{\theta})$$

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

EM算法

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

$$Q(\boldsymbol{\theta}) = \mathbb{E}_Z[\ln p(X, Z|\boldsymbol{\theta})] = \int \ln p(X, Z|\boldsymbol{\theta}) \cdot p(Z) dZ$$

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

EM算法

- **E步迭代**
  - $Q(\boldsymbol{\theta})$的计算中需要的$p(Z)$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Z的分布：

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

  - 以此来代替$p(Z)$，得到对数似然期望的近似估计：

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

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$p(Z|X,\boldsymbol{\theta}')$；
  - 迭代M步：重新优化$Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$；
  - 直到收敛为止；

EM算法

Algorithm 2 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$p(Z|X, \boldsymbol{\theta}^{t-1})$，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: **return** $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^t$

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；
  - 收敛性：EM算法具有收敛性，可以证明

$$\sum_{j=1}^{m} \ln p(\mathbf{x}_j|\boldsymbol{\theta}^t) \geq \sum_{j=1}^{m} \ln p(\mathbf{x}_j|\boldsymbol{\theta}^{t-1})$$

  - 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；

End