第九章

EM 算法及其推广

袁春 清华大学深圳研究生院
李航 华为诺亚方舟实验室

目录

一、EM 算法的引入

- EM 算法
- EM 算法的导出
- EM 算法在非监督学习中的应用

三硬币模型

- 三硬币模型：硬币A、B、C，正面概率π，p，q，
- A正面时选B，反面选C，
- 得到结果：1101001011
- 问题：只能看结果，不能看中间过程，估算π，p，q，
- 解：模型

$$
\begin{aligned}
P(y\mid\theta) &= \sum_z P(y,z\mid\theta) = \sum_z P(z\mid\theta)P(y\mid z,\theta) \\
&= \pi p^y(1-p)^{1-y} + (1-\pi)q^y(1-q)^{1-y}
\end{aligned}
$$

- 随机变量Y是观测变量，表示一次试验观测的结果是1或0，随机变量z是隐变量，表示未观测到的掷硬币A的结果，这一模型是以上数据的生成模型。

三硬币模型

- 观测数据： $Y=(Y_1,Y_2,\cdots,Y_n)^{\mathrm{T}}$
- 未观测数据： $Z=(Z_1,Z_2,\cdots,Z_n)^{\mathrm{T}}$
- 似然函数： $P(Y\mid\theta)=\sum_Z P(Z\mid\theta)P(Y\mid Z,\theta)$
- 即： $P(Y\mid\theta)=\prod_{j=1}^{n}[\pi p^{y_j}(1-p)^{1-y_j}+(1-\pi)q^{y_j}(1-q)^{1-y_j}]$
- 极大似然估计：

$$\hat{\theta}=\arg\max_{\theta}\log P(Y\mid\theta)$$

- 该问题没有解析解，EM迭代法：

EM算法

- 选取初值： $\theta^{(0)}=(\pi^{(0)},p^{(0)},q^{(0)})$
- 第i步的估计值： $\theta^{(i)}=(\pi^{(i)},p^{(i)},q^{(i)})$
- EM算法第i+1次迭代：
- E步：计算在模型参数$\pi^{(i)}$，$p^{(i)}$，$q^{(i)}$下观测数据yi来自掷硬币B的概率：

$$\mu^{(i+1)}=\frac{\pi^{(i)}(p^{(i)})^{y_j}(1-p^{(i)})^{1-y_j}}{\pi^{(i)}(p^{(i)})^{y_j}(1-p^{(i)})^{1-y_j}+(1-\pi^{(i)})(q^{(i)})^{y_j}(1-q^{(i)})^{1-y_j}}$$

- M步: 计算模型参数的新估计值

$$\pi^{(i+1)}=\frac{1}{n}\sum_{j=1}^{n}\mu_j^{(i+1)} \qquad p^{(i+1)}=\frac{\sum_{j=1}^{n}\mu_j^{(i+1)}y_j}{\sum_{j=1}^{n}\mu_j^{(i+1)}} \qquad q^{(i+1)}=\frac{\sum_{j=1}^{n}(1-\mu_j^{(i+1)})y_j}{\sum_{j=1}^{n}(1-\mu_j^{(i+1)})}$$

EM算法

初值：

$$\pi^{(0)}=0.5,\quad p^{(0)}=0.5,\quad q^{(0)}=0.5$$

对 $y_j=1$ 与 $y_j=0$ 均有 $\mu_j^{(1)}=0.5$

利用迭代公式，得： $\pi^{(1)}=0.5,\quad p^{(1)}=0.6,\quad q^{(1)}=0.6$

$$\mu_j^{(2)}=0.5,\quad j=1,2,\cdots,10$$

继续迭代，得：

$$\pi^{(2)}=0.5,\quad p^{(2)}=0.6,\quad q^{(2)}=0.6$$

得到模型参数的极大似然估计：

$$\hat{\pi}=0.5,\quad \hat{p}=0.6,\quad \hat{q}=0.6$$

EM算法

如果取初值：

$$\pi^{(0)}=0.4, \quad p^{(0)}=0.6, \quad q^{(0)}=0.7$$

$$\hat{\pi}=0.4064, \quad \hat{p}=0.5368, \quad \hat{q}=0.6432$$

完全数据 complete-data $P(Y,Z\mid\theta)$

不完全数据 incomplete-data $P(Y\mid\theta)$

EM算法

输入：观测变量数据Y,隐变量数据Z，联合分布P(Y,Z|Θ)

条件分布P(Z|Y, Θ)

输出：模型参数Θ

（1）选择参数的初值$\theta^{(0)}$，开始迭代；

（2）E步：记$\theta^{(i)}$为第i次迭代参数θ的估计值，

在第$i+1$次迭代的E步，计算

$$\begin{aligned} Q(\theta,\theta^{(i)}) &= E_Z[\log P(Y,Z|\theta)|Y,\theta^{(i)}] \\ &= \sum_Z \log P(Y,Z|\theta)P(Z|Y,\theta^{(i)}) \end{aligned}$$

给定观测数据Y和当前参数估计Θ

EM算法

（3）M步：求使 $Q(\theta,\theta^{(i)})$ 极大化的 θ，

确定第 $i+1$ 次迭代的参数的估计值 $\theta^{(i+1)}$

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

Q函数定义：

完全数据的对数似然函数logP(Y,Z|Θ)关于在给定观测数据Y和当前函数Θ^(i)下对未观测数据Z的条件概率分布P(Z|Y, Θ^(i)),的期望称为Q函数，即：

$$Q(\theta,\theta^{(i)}) = E_Z[\log P(Y,Z\mid\theta)\mid Y,\theta^{(i)}]$$

EM算法

- 算法说明：
- 步骤3，完成一次迭代：$\Theta^{(i)}$到$\Theta^{(i+1)}$，将证明每次迭代使似然函数增大或达到局部最大值。
- 步骤4，停止迭代的条件

$$\|\theta^{(i+1)}-\theta^{(i)}\|<\varepsilon_1 \quad \text{或} \quad \|Q(\theta^{(i+1)},\theta^{(i)})-Q(\theta^{(i)},\theta^{(i)})\|<\varepsilon_2$$

EM算法的导出

- 为什么EM算法能近似实现对观测数据的极大似然估计？
- 极大化(不 $$L(\theta)=\log P(Y\mid\theta)=\log\sum_{Z}P(Y,Z\mid\theta)$$ 函数：

$$=\log\left(\sum_{Z}P(Y\mid Z,\theta)P(Z\mid\theta)\right)$$

- 难点：有未观测数据，包含和的对数。
- EM通过迭代逐步近似极大化L(Θ),希望

$$L(\theta)>L(\theta^{(i)})$$

EM算法的导出

- 考虑二者的差：

$$L(\theta)-L(\theta^{(i)})=\log\left(\sum_Z P(Y|Z,\theta)P(Z|\theta)\right)-\log P(Y|\theta^{(i)})$$

- Jason不等式：

$$\begin{aligned}L(\theta)-L(\theta^{(i)})&=\log\left(\sum_Z P(Y|Z,\theta^{(i)})\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Y|Z,\theta^{(i)})}\right)-\log P(Y|\theta^{(i)})\\&\geqslant\sum_Z P(Z|Y,\theta^{(i)})\log\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})}-\log P(Y|\theta^{(i)})\\&=\sum_Z P(Z|Y,\theta^{(i)})\log\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})P(Y|\theta^{(i)})}\end{aligned}$$

EM算法的导出

- 令：

$$B(\theta,\theta^{(i)}) \triangleq L(\theta^{(i)}) + \sum_Z P(Z \mid Y,\theta^{(i)}) \log \frac{P(Y \mid Z,\theta)P(Z \mid \theta)}{P(Z \mid Y,\theta^{(i)})P(Y \mid \theta^{(i)})}$$

- 则：

$$L(\theta) \geqslant B(\theta,\theta^{(i)})$$

$$L(\theta^{(i)}) = B(\theta^{(i)},\theta^{(i)})$$

任何可以使 $B(\theta,\theta^{(i)})$ 增大的 θ，也可以使 $L(\theta)$ 增大

- 选择：

$$\theta^{(i+1)} = \arg\max_{\theta} B(\theta,\theta^{(i)})$$

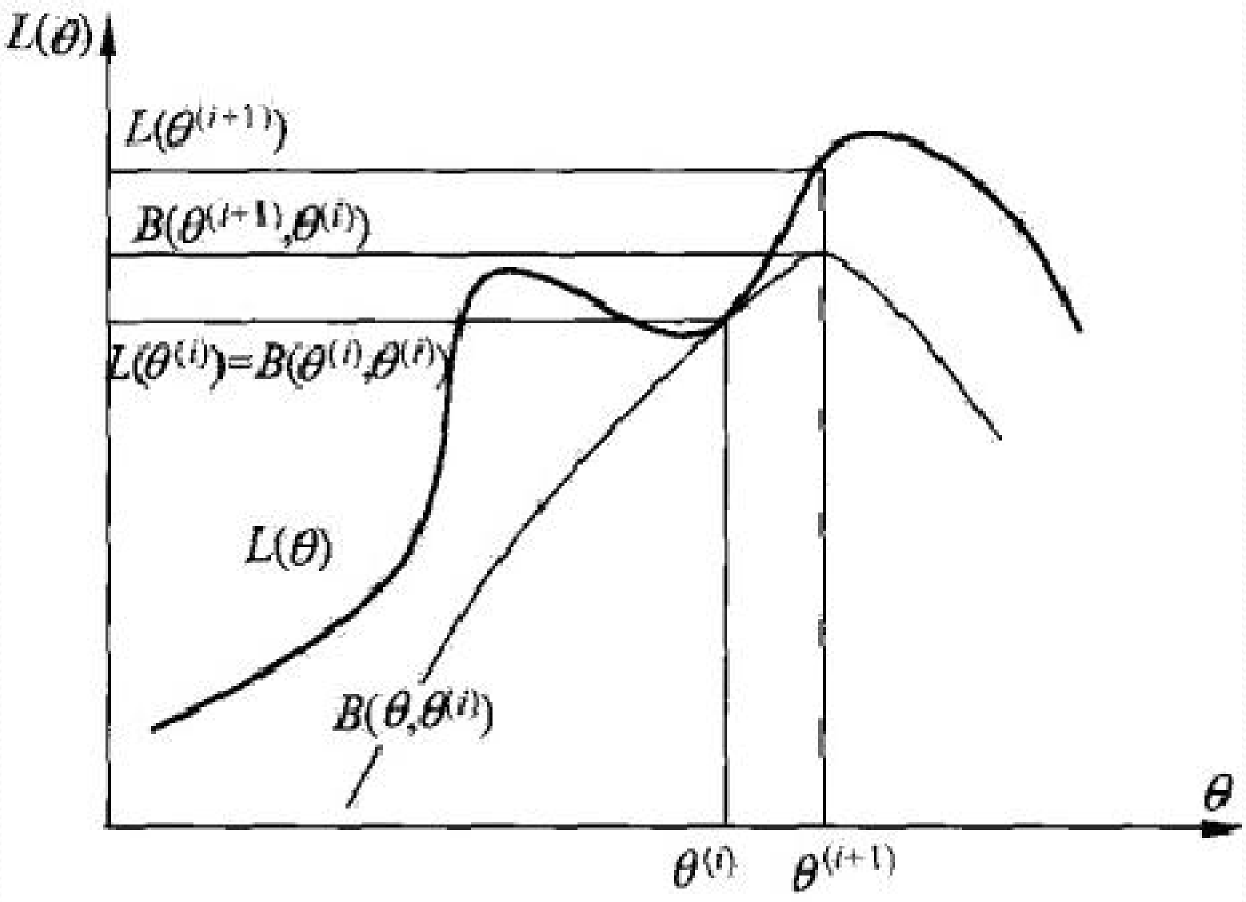
EM算法的导出

- 省去和 Θ 无关的项：

$$
\begin{aligned}
\theta^{(i+1)} &= \arg\max_{\theta}\left(L(\theta^{(i)}) + \sum_{Z} P(Z \mid Y,\theta^{(i)}) \log \frac{P(Y \mid Z,\theta)P(Z \mid \theta)}{P(Z \mid Y,\theta^{(i)})P(Y \mid \theta^{(i)})}\right) \\
&= \arg\max_{\theta}\left(\sum_{Z} P(Z \mid Y,\theta^{(i)}) \log(P(Y \mid Z,\theta)P(Z \mid \theta))\right) \\
&= \arg\max_{\theta}\left(\sum_{Z} P(Z \mid Y,\theta^{(i)}) \log P(Y,Z \mid \theta)\right) \\
&= \arg\max_{\theta} Q(\theta,\theta^{(i)})
\end{aligned}
$$

EM算法的解释

L(Θ)开始



EM在非监督学习中的应用

- 生成模型由联合概率分布P(X,Y)表示，可以认为非监督学习训练数据是联合概率分布产生的数据，X为观测数据，Y为未观测数据。

二、EM算法的收敛性

- EM，提供一种近似计算含有隐变量概率模型的极大似然估计的方法，
- EM，最大优点：简单性和普适性；
- 疑问：
- 1、EM算法得到的估计序列是否收敛？
- 2、如果收敛，是否是全局极大值或局部极大值？

EM算法的收敛性

- 两个收敛定理：
- 定理9.1：设P(Y|Θ)为观测数据的似然函数，$\Theta^{(i)}$(i=1,2..）为EM参数估计序列，$P(Y\mid\theta^{(i)})\ (i=1,2,\cdots)$为对应的似然函数序列，则$P(Y|\Theta^{(i)})$是单调递增的，即：

$$P(Y\mid\theta^{(i+1)})\geq P(Y\mid\theta^{(i)})$$

- 证明：由

$$P(Y\mid\theta)=\frac{P(Y,Z\mid\theta)}{P(Z\mid Y,\theta)}$$

$$\log P(Y\mid\theta)=\log P(Y,Z\mid\theta)-\log P(Z\mid Y,\theta)$$

- 由：

$$Q(\theta,\theta^{(i)})=\sum_Z \log P(Y,Z\mid\theta)P(Z\mid Y,\theta^{(i)})$$

EM算法的收敛性

- 令：

$$H(\theta,\theta^{(i)})=\sum_{Z}\log P(Z\mid Y,\theta)P(Z\mid Y,\theta^{(i)})$$

- 则：

$$\log P(Y\mid\theta)=Q(\theta,\theta^{(i)})-H(\theta,\theta^{(i)})$$

- 得：

$$\log P(Y\mid\theta^{(i+1)})-\log P(Y\mid\theta^{(i)})$$
$$=[Q(\theta^{(i+1)},\theta^{(i)})-Q(\theta^{(i)},\theta^{(i)})]-[H(\theta^{(i+1)},\theta^{(i)})-H(\theta^{(i)},\theta^{(i)})]$$

- 只需证右端非负

EM算法的收敛性

- 前半部分，Θ$^{(i+1)}$为极大值，所以

$$Q(\theta^{(i+1)},\theta^{(i)})-Q(\theta^{(i)},\theta^{(i)})\geqslant 0$$

- 后半部分：

$$\begin{aligned}
H(\theta^{(i+1)},\theta^{(i)})-H(\theta^{(i)},\theta^{(i)})
&=\sum_Z\left(\log\frac{P(Z\mid Y,\theta^{(i+1)})}{P(Z\mid Y,\theta^{(i)})}\right)P(Z\mid Y,\theta^{(i)})\\
&\leqslant\log\left(\sum_Z\frac{P(Z\mid Y,\theta^{(i+1)})}{P(Z\mid Y,\theta^{(i)})}P(Z\mid Y,\theta^{(i)})\right)\\
&=\log P(Z\mid Y,\theta^{(i+1)})=0
\end{aligned}$$

EM算法的收敛性

- 定理9.2:
- 设$L(\Theta)=\log P(Y|\Theta)$,为观测数据的对数似然函数，$\Theta^{(i)}$(i=1,2..）为EM算法得到的参数估计序列，$L(\Theta^{(i)})$为对应的对数似然函数序列，
- 1、如果$P(Y|\Theta)$有上界，则$L(\Theta^{(i)})=\log P(Y|\Theta^{(i)})$收敛到某一值$L^*$；
- 2、在函数$Q(\Theta, \Theta')$与$L(\Theta)$满足一定条件下，由EM算法得到的参数估计序列$\Theta^{(i)}$的收敛值Θ^*是$L(\Theta)$的稳定点。

三、EM算法在高斯混合模型学习中的应用

- 高斯混合模型：
- 概率分布模型：$P(y|\theta)=\sum_{k=1}^{K}\alpha_k\phi(y|\theta_k)$
- 系数：$\alpha_k \geqslant 0$，$\sum_{k=1}^{K}\alpha_k=1$
- 高斯分布密度：$\phi(y|\theta_k)$ $\theta_k=(\mu_k,\sigma_k^2)$
- 第K个分模型：$\phi(y|\theta_k)=\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$

可任意高斯模型

高斯混合模型参数估计的EM算法

- 假设观测数据$y_1,y_2,....y_N$由高斯混合模型生成：

$$P(y|\theta)=\sum_{k=1}^{K}\alpha_k\phi(y|\theta_k)$$

$$\theta=(\alpha_1,\alpha_2,\cdots,\alpha_K;\theta_1,\theta_2,\cdots,\theta_K)$$

- 用EM算法估计参数；
- 1、明确隐变量，写出完全数据的对数似然函数：
  - 设想观测数据yi是依概率a_k选择第k个高斯分模型 $\phi(y|\theta_k)$ 生成，隐变量

$$\gamma_{jk}=\begin{cases}1, & \text{第 } j \text{ 个观测来自第 } k \text{ 个分模型}\\ 0, & \text{否则}\end{cases}$$

EM算法在高斯混合模型学习中的应用

- 1、明确隐变量，写出完全数据的对数似然函数：
- 完全数据：$(y_j, \gamma_{j1}, \gamma_{j2}, \cdots, \gamma_{jK})$， $j=1,2,\cdots,N$
- 似然函数：

$$P(y,\gamma|\theta)=\prod_{j=1}^{N}P(y_j,\gamma_{j1},\gamma_{j2},\cdots,\gamma_{jK}|\theta)$$

$$=\prod_{k=1}^{K}\prod_{j=1}^{N}\left[\alpha_k\phi(y_j|\theta_k)\right]^{\gamma_{jk}}$$

$$=\prod_{k=1}^{K}\alpha_k^{n_k}\prod_{j=1}^{N}\left[\phi(y_j|\theta_k)\right]^{\gamma_{jk}}$$

$$=\prod_{k=1}^{K}\alpha_k^{n_k}\prod_{j=1}^{N}\left[\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y_j-\mu_k)^2}{2\sigma_k^2}\right)\right]^{\gamma_{jk}}$$

$$n_k=\sum_{j=1}^{N}\gamma_{jk}$$

$$\sum_{k=1}^{K}n_k=N$$

EM算法在高斯混合模型学习中的应用

- 1、明确隐变量，写出完全数据的对数似然函数：

$$\log P(y,\gamma|\theta)=\sum_{k=1}^{K}n_k\log\alpha_k+\sum_{j=1}^{N}\gamma_{jk}\left[\log\left(\frac{1}{\sqrt{2\pi}}\right)-\log\sigma_k-\frac{1}{2\sigma_k^2}(y_j-\mu_k)^2\right]$$

EM算法在高斯混合模型学习中的应用

- 2、EM算法的E步，确定Q函数

$$
\begin{aligned}
Q(\theta,\theta^{(i)}) &= E[\log P(y,\gamma|\theta)|y,\theta^{(i)}] \\
&= E\left\{\sum_{k=1}^{K} n_k \log\alpha_k + \sum_{j=1}^{N}\gamma_{jk}\left[\log\left(\frac{1}{\sqrt{2\pi}}\right) - \log\sigma_k - \frac{1}{2\sigma_k^2}(y_j-\mu_k)^2\right]\right\} \\
&= \sum_{k=1}^{K}\left\{\sum_{j=1}^{N}(E\gamma_{jk})\log\alpha_k + \sum_{j=1}^{N}(E\gamma_{jk})\left[\log\left(\frac{1}{\sqrt{2\pi}}\right) - \log\sigma_k - \frac{1}{2\sigma_k^2}(y_j-\mu_k)^2\right]\right\}
\end{aligned}
$$

需要计算 $E(\gamma_{jk}|y,\theta)$，记为 $\hat{\gamma}_{jk}$

- 第j个观测数据来自第k个分模型的概率，称为分模型k对观测数据y_j的响应度。

EM算法在高斯混合模型学习中的应用

- 2、EM算法的E步，确定Q函数

$$\hat{\gamma}_{jk} = E(\gamma_{jk} \mid y, \theta) = P(\gamma_{jk} = 1 \mid y, \theta)$$

$$= \frac{P(\gamma_{jk} = 1, y_j \mid \theta)}{\sum_{k=1}^{K} P(\gamma_{jk} = 1, y_j \mid \theta)}$$

$$= \frac{P(y_j \mid \gamma_{jk} = 1, \theta) P(\gamma_{jk} = 1 \mid \theta)}{\sum_{k=1}^{K} P(y_j \mid \gamma_{jk} = 1, \theta) P(\gamma_{jk} = 1 \mid \theta)}$$

$$= \frac{\alpha_k \phi(y_j \mid \theta_k)}{\sum_{k=1}^{K} \alpha_k \phi(y_j \mid \theta_k)}, \quad j = 1, 2, \cdots, N; \quad k = 1, 2, \cdots, K$$

EM算法在高斯混合模型学习中的应用

- 2、EM算法的E步，确定Q函数

将 $\hat{\gamma}_{jk} = E\gamma_{jk}$ 及 $n_k = \sum_{j=1}^{N} E\gamma_{jk}$ 代入

$$Q(\theta,\theta^{(i)}) = \sum_{k=1}^{K} n_k \log\alpha_k + \sum_{k=1}^{N} \hat{\gamma}_{jk}\left[\log\left(\frac{1}{\sqrt{2\pi}}\right) - \log\sigma_k - \frac{1}{2\sigma_k^2}(y_j-\mu_k)^2\right]$$

EM算法在高斯混合模型学习中的应用

- 3、确定EM算法的M步：
- 求：$$\theta^{(i+1)} = \arg\max_{\theta} Q(\theta, \theta^{(i)})$$

用 $\hat{\mu}_k$ ，$\hat{\sigma}_k^2$ 及 $\hat{\alpha}_k$ ，$k=1,2,\cdots,K$ ，表示 $\theta^{(i+1)}$

- 采用求导的方法：

$$\hat{\mu}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} y_j}{\sum_{j=1}^{N} \hat{\gamma}_{jk}}$$

$$\hat{\sigma}_k^2 = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} (y_j - \mu_k)^2}{\sum_{j=1}^{N} \hat{\gamma}_{jk}}$$

$$\hat{\alpha}_k = \frac{n_k}{N} = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk}}{N}$$

高斯混合模型参数估计的EM算法

- 输入：观测数据$y_1,y_2,\ldots y_N$, 高斯混合模型
- 输出：高斯混合模型参数
- 1、设定初始值开始迭代
- 2、E步，响应度计算

$$\hat{\gamma}_{jk}=\frac{\alpha_k\phi(y_j\mid\theta_k)}{\sum_{k=1}^{K}\alpha_k\phi(y_j\mid\theta_k)}$$

高斯混合模型参数估计的EM算法

- 输入：观测数据$y_1,y_2,\ldots y_N$, 高斯混合模型
- 输出：高斯混合模型参数
- 3、M步，计算新一轮迭代的模型参数：

$$\hat{\mu}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} y_j}{\sum_{j=1}^{N} \hat{\gamma}_{jk}} \qquad \hat{\sigma}_k^2 = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} (y_j - \mu_k)^2}{\sum_{j=1}^{N} \hat{\gamma}_{jk}} \qquad \hat{\alpha}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk}}{N}$$

- 4、重复2，3步直到收敛

四、EM算法的推广

- EM算法可以解释为:
- F函数的极大---极大算法（maximization –maximization algorithm)
- 广义期望极大(Generalization Expectation Maximization. GEM)

F函数的极大—极大算法

- F函数：
- 假设隐变量数据Z的概率分布为$\tilde{P}(Z)$，定义分布$\tilde{P}$ 与参数θ的函数$F(\tilde{P},\theta)$ ：

$$F(\tilde{P},\theta)=E_{\tilde{P}}[\log P(Y,Z\mid\theta)]+H(\tilde{P})$$

- 熵：

$$H(\tilde{P})=-E_{\tilde{P}}\log\tilde{P}(Z)$$

- F函数是θ的连续函数，重要性质：
- 引理9.1：对于固定的θ，存在唯一的分布$\tilde{P}_\theta$ 极大化$F(\tilde{P},\theta)$
- 这时的$\tilde{P}_\theta$

$$\tilde{P}_\theta(Z)=P(Z\mid Y,\theta)$$

- 并且 $\tilde{P}_\theta$ 随θ连续变化。

F函数的极大—极大算法

- 证明：对于固定的θ，拉格朗日函数方法对最优化问题求$\tilde{P}(Z)$，
$$L = E_{\tilde{P}} \log P(Y,Z|\theta) - E_{\tilde{P}} \log \tilde{P}(Z) + \lambda\left(1 - \sum_Z \tilde{P}(Z)\right)$$
- 对$\tilde{P}(Z)$求偏导：$\dfrac{\partial L}{\partial \tilde{P}(Z)} = \log P(Y,Z|\theta) - \log \tilde{P}(Z) - 1 - \lambda$
- 令偏导为0：$\lambda = \log P(Y,Z|\theta) - \log \tilde{P}_\theta(Z) - 1$
- 得：分子分母成比例，$\dfrac{P(Y,Z|\theta)}{\tilde{P}_\theta(Z)} = e^{1+\lambda}$
- 由：$\sum_Z \tilde{P}_\theta(Z) = 1$　得：$\tilde{P}_\theta(Z) = P(Z|Y,\theta)$

由假设$P(Y,Z|\theta)$是θ的连续函数，得到$\tilde{P}_\theta$是θ的连续函数.

F函数的极大—极大算法

- 引理9.2：

  若 $\tilde{P}_{\theta}(Z)=P(Z\mid Y,\theta)$，则 $F(\tilde{P},\theta)=\log P(Y\mid\theta)$

- 定理9.3：
- 设 $L(\theta)=\log P(Y\mid\theta)$ 为观测数据的对数似然数 $\theta^{(i)}$，$i=1,2,\cdots$ 为EM算法得到的参数估计序列，F函数 $F(\tilde{P},\theta)$，如果 $F(\tilde{P},\theta)$ 在 $\tilde{P}^{*}$ 和 θ^{*} 有局部极大值，那么L(θ)也在 θ^{*} 有局部极大值，类似地，如果 $F(\tilde{P},\theta)$ 在 $\tilde{P}^{*}$ 和 θ^{*} 达到全局最大值，那么L(θ)也在 θ^{*} 达到全局最大值。

F函数的极大—极大算法

证明：由定理9.1,9.2

$L(\theta)=\log P(Y|\theta)=F(\tilde{P}_\theta,\theta)$ 对任意 θ 成立；特别的：

对于使 $F(\tilde{P},\theta)$ 达到极大的参数 θ^*，

$$L(\theta^*)=F(\tilde{P}_{\theta^*},\theta^*)=F(\tilde{P}^*,\theta^*)$$

为了证明 θ^* 是 $L(\theta)$ 的极大点

需要证明不存在接近 θ^* 的点 θ^{**}，使 $L(\theta^{**})>L(\theta^*)$.

假如存在这样的点 θ^{**}，那么应有 $F(\tilde{P}^{**},\theta^{**})>F(\tilde{P}^*,\theta^*)$，

这里 $\tilde{P}^{**}=\tilde{P}_{\theta^{**}}$ 但因 $\tilde{P}_\theta$ 是随 θ 连续变化的，$\tilde{P}^{**}$ 应接近 $\tilde{P}^*$，

这与 $\tilde{P}^*$ 和 θ^* 是 $F(\tilde{P},\theta)$ 的局部极大点的假设

类似可以证明关于全局最大值的结论

F函数的极大—极大算法

定理9.4:

EM算法的一次迭代可由F函数的极大----极大算法实现。

设 $\theta^{(i)}$ 为第 i 次迭代参数 θ 的估计，$\tilde{P}^{(i)}$ 为第 i 次迭代函数 $\tilde{P}$ 的估计

在第 $i+1$ 次迭代的两步为

（1）对固定的 $\theta^{(i)}$，求 $\tilde{P}^{(i+1)}$ 使 $F(\tilde{P},\theta^{(i)})$ 极大化；

（2）对固定的 $\tilde{P}^{(i+1)}$，求 $\theta^{(i+1)}$ 使 $F(\tilde{P}^{(i+1)},\theta)$ 极大化.

F函数的极大—极大算法

定理9.4：

EM算法的一次迭代可由F函数的极大----极大算法实现。

证明：

（1）由引理 9.1，对于固定的 $\theta^{(i)}$，

$\tilde{P}^{(i+1)}(Z)=\tilde{P}_{\theta^{(i)}}(Z)=P(Z|Y,\theta^{(i)})$ 使 $F(\tilde{P},\theta^{(i)})$ 极大化．此时

$$
\begin{aligned}
F(\tilde{P}^{(i+1)},\theta) &= E_{\tilde{P}^{(i+1)}}[\log P(Y,Z|\theta)]+H(\tilde{P}^{(i+1)}) \\
&= \sum_{Z}\log P(Y,Z|\theta)P(Z|Y,\theta^{(i)})+H(\tilde{P}^{(i+1)})
\end{aligned}
$$

由 $Q(\theta,\theta^{(i)})$ 的定义

$$
F(\tilde{P}^{(i+1)},\theta)=Q(\theta,\theta^{(i)})+H(\tilde{P}^{(i+1)})
$$

F函数的极大–极大算法

定理9.4：

EM算法的一次迭代可由F函数的极大----极大算法实现。

证明：

（2）固定 $\tilde{P}^{(i+1)}$，求 $\theta^{(i+1)}$ 使 $F(\tilde{P}^{(i+1)},\theta)$ 极大化．得到

$$\theta^{(i+1)}=\arg\max_{\theta} F(\tilde{P}^{(i+1)},\theta)=\arg\max_{\theta} Q(\theta,\theta^{(i)})$$

通过以上两步完成了EM算法的一次迭代，由EM算法与F函数的极大---极大算法得到的参数估计序列 $\theta^{(i)}$， $i=1,2,\cdots$，是一致的。

F函数的极大—极大算法

算法 9.3（GEM 算法 1）

输入：观测数据，F 函数；

输出：模型参数.

（1）初始化参数 $\theta^{(0)}$，开始迭代

（2）第 $i+1$ 次迭代，第 1 步：记 $\theta^{(i)}$ 为参数 θ 的估计值，$\tilde{P}^{(i)}$ 为函数 $\tilde{P}$ 的估计．求 $\tilde{P}^{(i+1)}$ 使 $\tilde{P}$ 极大化 $F(\tilde{P},\theta^{(i)})$

（3）第 2 步：求 $\theta^{(i+1)}$ 使 $F(\tilde{P}^{(i+1)},\theta)$ 极大化

（4）重复 (2) 和 (3)，直到收敛.

问题和方法：有时求 $Q(\theta,\theta^{(i)})$ 的极大化是很困难的

通过：找 $\theta^{(i+1)}$ 使得 $Q(\theta^{(i+1)},\theta^{(i)}) > Q(\theta^{(i)},\theta^{(i)})$

F函数的极大—极大算法

算法 9.4（GEM 算法 2）

输入：观测数据，Q 函数；

输出：模型参数.

（1）初始化参数 $\theta^{(0)}$，开始迭代

（2）第 $i+1$ 次迭代，第 1 步：记 $\theta^{(i)}$ 为参数 θ 的估计值，

计算

$$
\begin{aligned}
Q(\theta,\theta^{(i)}) &= E_Z[\log P(Y,Z\mid\theta)\mid Y,\theta^{(i)}] \\
&= \sum_Z P(Z\mid Y,\theta^{(i)})\log P(Y,Z\mid\theta)
\end{aligned}
$$

（3）第 2 步：求 $\theta^{(i+1)}$ 使

$$Q(\theta^{(i+1)},\theta^{(i)}) > Q(\theta^{(i)},\theta^{(i)})$$

（4）重复 (2) 和 (3)，直到收敛.

F函数的极大—极大算法

- 当参数θ的维数为d大于等于2时，可采用一种特殊的GEM算法，算法的M步分解为d次条件极大化，每次只改变参数向量的一个分量，其余分量不改变。

F函数的极大—极大算法

算法 9.5（GEM 算法 3）

输入：观测数据，Q 函数；

输出：模型参数.

（1）初始化参数 $\theta^{(0)}=(\theta_1^{(0)},\theta_2^{(0)},\cdots,\theta_d^{(0)})$，开始迭代

（2）第 $i+1$ 次迭代，第 1 步：记 $\theta^{(i)}=(\theta_1^{(i)},\theta_2^{(i)},\cdots,\theta_d^{(i)})$

为参数 $\theta=(\theta_1,\theta_2,\cdots,\theta_d)$ 的估计值，计算

$$
\begin{aligned}
Q(\theta,\theta^{(i)}) &= E_Z[\log P(Y,Z\mid\theta)\mid Y,\theta^{(i)}] \\
&= \sum_Z P(Z\mid y,\theta^{(i)})\log P(Y,Z\mid\theta)
\end{aligned}
$$

（3）第 2 步：进行 d 次条件极大化：

首先，在 $\theta_2^{(i)},\cdots,\theta_k^{(i)}$ 保持不变的条件下求使 $Q(\theta,\theta^{(i)})$ 达到极大的 $\theta_1^{(i+1)}$；

F函数的极大—极大算法

然后，在 $\theta_1=\theta_1^{(i+1)}$， $\theta_j=\theta_j^{(i)}$， $j=3,4,\cdots,k$ 的条件下

求使 $Q(\theta,\theta^{(i)})$ 达到极大的 $\theta_2^{(i+1)}$

如此继续，经过 d 次条件极大化，得到 $\theta^{(i+1)}=(\theta_1^{(i+1)},\theta_2^{(i+1)},\cdots,\theta_d^{(i+1)})$

使得 $Q(\theta^{(i+1)},\theta^{(i)})>Q(\theta^{(i)},\theta^{(i)})$

（4）重复 (2) 和 (3)，直到收敛

Q & A