EM算法简介

EM算法是一种迭代算法，用于含有隐变量(hidden variable)的概率模型参数的极大似然估计，或极大后验概率估计。EM算法每次由两步组成：E步，求期望；M步，求极大。

EM算法的引入

EM算法

一般地，我们用Y表示观测随机变量的数据，Z表示隐随机变量的数据。Y, Z连在一起成为完全数据(conplete data)，观测数据Y又称为不完全数据(incomplete data)。假设给定观测数据Y，其概率分布为$P(Y|\theta)$，那么不完全数据Y的似然函数是$P(Y|\theta)$，对数似然函数$L(\theta)=\log P(Y|\theta)$；假设$Y, Z$的联合概率分布是$P(Y,Z|\theta)$，那么完全数据的对数似然是$\log P(Y,Z|\theta)$。

EM算法通过迭代求$L(\theta)=\log P(Y|\theta)$的极大似然估计。每次迭代包含两步：E步，求期望；M步，求极大化。

算法(EM算法)：

1. 选择参数的初值θ^0，开始迭代；
2. E步：记θ^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

$$Q(\theta,\theta_i)=E_Z[\log P(Y,Z|\theta)|Y,\theta^i]=\sum_Z \log P(Y,Z|\theta)P(Z|Y,\theta^i)$$

   这里，$P(Z|Y,\theta^i)$是在给定观测数据Y和当前参数估计θ^i下隐变量Z的条件概率分布。

3. M步：求使得$Q(\theta,\theta^i)$极大化的θ，即

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

4. 重复2,3步，直到收敛。

EM算法的导出

首先我们需要介绍一个定理。

定理(Jensen 不等式)： f是一个凸函数，X为一个随机变量，则有

$$E[f(X)]\geq f(EX)$$

现在我们想要极大化观测数据Y关于θ的对数似然，即

$$L(\theta)=\log P(Y|\theta)=\log\sum_Z P(Y,Z|\theta)=\log(\sum_Z P(Y|Z,\theta)P(Z|\theta))$$

主要困难是上式中有未观测数据并有包含和的对数。

EM算法通过迭代逐步近似极大化$L(\theta)$。假设在第i次后估计值为θ^i，我们希望新估计值θ使得$L(\theta)$增加，为此，考虑两者的差：

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

令

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

则

$$
L(\theta) \geq B(\theta,\theta^i)
$$

选择θ^{i+1}使得B达到最大，即

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

现在求θ^{i+1}，省去常数项，我们有

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

EM算法的收敛性

定理：设$P(Y|\theta)$为观测数据的似然函数，θ^i为EM算法得到的参数估计序列，$P(Y|\theta^i)$为对应的似然函数序列，则$P(Y|\theta^i)$是单调递增的。

证明：由于

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

取对数有

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

由于

$$
Q(\theta,\theta^i) = \sum_Z \log P(Y,Z|\theta)P(Z|Y,\theta^i)
$$

令

$$
H(\theta,\theta^i) = \sum_Z \log P(Z|Y,\theta)P(Z|Y,\theta^i)
$$

于是对数似然可以写成

$$\log P(Y|\theta) = Q(\theta, \theta^i) - H(\theta, \theta^i)$$

上式分别取θ为θ^i, θ^{i+1}并相减，有

$$\begin{aligned} &\log P(Y|\theta^{i+1}) - \log P(Y|\theta^i) \\ &= [Q(\theta^{i+1}, \theta^i) - Q(\theta^i, \theta^i)] - [H(\theta^{i+1}, \theta^i) - H(\theta^i, \theta^i)] \end{aligned}$$

只需要证右式非负即可。第一项，由于θ^{i+1}使$Q(\theta, \theta^i)$最大，所以非负。第二项，我们有

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

所以第二项也是非负的。

定理：设$L(\theta) = \log P(Y|\theta)$，$\theta^i$为EM算法得到的参数估计序列，$L(\theta^i)$为对应的似然函数序列。

1. 如果$P(Y|\theta)$有上界，则$L(\theta)$收敛。
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θ^i的收敛值是$L(\theta)$的稳定点。

证明从略。

EM算法的收敛性包含关于对数似然函数$L(\theta)$的收敛性和关于参数估计序列的收敛性两层意思，前者不蕴含后者。此外，定理只能保证参数估计序列收敛到对数似然函数序列的稳定，不能保证收敛到极大值点。所以初始值很重要，往往选取几个不同的初值迭代，选择最好的。

高斯混合模型

定义(高斯混合模型)：高斯混合模型是指具有如下形式的概率分布模型：

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

其中$\alpha_k \geq 0, \sum_k \alpha_k = 1$，$\phi(y|\theta_k)$是高斯分布密度，$\theta_k = (\mu_k, \sigma_k^2)$，

$$\phi(y|\theta_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2} \right)$$

称为第k个分模型。

假设观测数据$y_1, \cdots, y_N$由高斯混合模型生成，

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

其中，$\theta = (\alpha_1, \cdots, \alpha_K; \theta_1, \cdots, \theta_K)$。

1. 明确隐变量，写出完全数据的对数似然函数

可以设想观测数据y_j是由这样产生的：首先依概率α_k选择第k个高斯分布分模型，然后依第k个分模型的概率分布$\phi(y|\theta_k)$生产观测数据y_j。这时观测数据是已知的，反映观测数据来自第k个分模型的数据是未知的，以隐变量γ_{jk}表示，定义如下：

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

有了观测数据y_j和未观测数据γ_{jk}，那么完全数据就是

$$(y_j,\gamma_{j1},\cdots,\gamma_{jk}),j=1,\cdots,N$$

于是，可以写出完全数据的对数似然：

$$\begin{aligned}P(y,\gamma|\theta)&=\prod_{j=1}^{N}P(y_j,\gamma_{j1},\cdots,\gamma_{jk}|\theta)\\&=\prod_{k=1}^{K}\prod_{j=1}^{N}[\alpha_k\phi(y_j|\theta_k)]^{\gamma_{jk}}\\&=\prod_{k=1}^{K}\alpha_k^{n_k}\prod_{j=1}^{N}[\phi(y_j|\theta_k)]^{\gamma_{jk}}\\&=\prod_{k=1}^{K}\alpha_k^{n_k}\prod_{j=1}^{N}\left[\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)\right]^{\gamma_{jk}}\end{aligned}$$

其中，$n_k=\sum_j\gamma_{jk},\sum_k n_k=N$。

那么，完全数据的对数似然函数为

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

2. EM算法的E步：确定Q函数

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

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

$$\begin{aligned}\hat{\gamma}_{jk}&=E(\gamma_{jk}|y,\theta)=P(\gamma_{jk}=1|y,\theta)\\&=\frac{P(\gamma_{jk}=1,y_j|\theta)}{\sum_{k=1}^{K}P(\gamma_{jk}=1,y_j|\theta)}\\&=\frac{P(y_j|\gamma_{jk}=1,\theta)P(\gamma_{jk}=1|\theta)}{\sum_{k=1}^{K}P(y_j|\gamma_{jk}=1,\theta)P(\gamma_{jk}=1|\theta)}\\&=\frac{\alpha_k\phi(y_j|\theta_k)}{\sum_{k=1}^{K}\alpha_k\phi(y_j|\theta_k)}\end{aligned}$$

$\hat{\gamma}_{jk}$是在当前模型参数下第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。

将$\hat{\gamma}_{jk}=E\gamma_{jk}$与$n_k=\sum_j E\gamma_{jk}$代回，得到

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

3. 确定EM算法的M步

迭代的M步是求函数Q对θ的极大值，即求新一轮迭代的模型参数：

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

用$\hat{\mu}_k,\hat{\sigma}_k^2,\hat{\alpha}_k$，$k=1,\cdots,K$，表示$\theta^{i+1}$的各参数。求$\hat{\mu}_k,\hat{\sigma}_k^2$只需将上式分别对$\mu_k,\sigma_k^2$求偏导令为0，即可得到；求$\hat{\alpha}_k$是在$\sum_k\alpha_k=1$条件下求偏导并令为0得到的，结果如下：

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

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

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

算法(高斯混合模型参数估计的EM算法)：

1. 取参数的初始值开始迭代
2. E步：根据当前模型参数，计算分模型k对观测数据y_j的响应度

$$\hat{\gamma}_{jk}=\frac{\alpha_k\phi(y_j|\theta_k)}{\sum_{k=1}^{K}\alpha_k\phi(y_j|\theta_k)},j=1,\cdots,N;k=1,\cdots,K$$

3. M步：计算新一轮迭代的模型参数

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

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

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

4. 重复2,3直到收敛

EM算法的推广

EM算法还可以解释为F函数的极大-极大算法，基于这个解释有若干变形与推广，如广义期望极大(generalized expectation maximization, GEM)算法。

F函数的极大-极大算法

定义(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$F(\tilde{P},\theta)$如下：

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

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

通常假设$P(Y,Z|\theta)$是θ的连续函数，因而$F(\tilde{P},\theta)$是$\tilde{P}$和θ的连续函数。F函数有如下重要性质：

引理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$\tilde{P}_\theta$为

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

并且$\tilde{P}_\theta$随着θ连续变化。

证明：对于固定的θ，可以求得使$F(\tilde{P},\theta)$达到极大的分布$\tilde{P}_\theta(Z)$。为此，引进拉格朗日乘子λ，拉格朗日函数为

$$L=E_{\tilde{P}}\log P(Y,Z|\theta)-E_{\tilde{P}}\log\tilde{P}(Z)+\lambda(1-\sum_Z\tilde{P}(Z))$$

对$\tilde{P}$求偏导数：

$$\frac{\partial L}{\partial\tilde{P}(Z)}=\log P(Y,Z|\theta)-\log\tilde{P}(Z)-1-\lambda$$

令为零，得到

$$\lambda=\log P(Y,Z|\theta)-\log\tilde{P}_\theta(Z)-1$$

于是推出$\tilde{P}_\theta(Z)$与$P(Y,Z|\theta)$成比例

$$\frac{P(Y,Z|\theta)}{\tilde{P}_\theta(Z)}=e^{1+\lambda}$$

再从约束条件$\sum_Z\tilde{P}_\theta(Z)=1$就得到了我们的结论。

引理2：若$\tilde{P}_\theta(Z)=P(Z|Y,\theta)$，则

$$F(\tilde{P},\theta)=\log P(Y|\theta)$$

由上面的引理，可以得到关于EM算法用F函数的极大-极大算法的解释。

定理：设$L(\theta)=\log P(Y|\theta)$为观测数据的对数似然函数。如果$F(\tilde{P},\theta)$在$\tilde{P}^*$和$\theta^*$由局部极大值，那么$L(\theta)$在$\theta^*$也有局部最大值。对全局最大值相同。

证明：由上述引理可知，$L(\theta)=\log P(Y|\theta)=F(\tilde{P}_\theta,\theta)$对任意的$\theta$成立，特别地，对于使$F(\tilde{P},\theta)$达到极大值的$\theta^*$，有

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

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，那么应有$F(\tilde{P}',\theta')>F(\tilde{P}^*,\theta^*)$，但因为$\tilde{P}_\theta$是$\theta$的连续函数，$\tilde{P}'$应该接近$\tilde{P}_\theta$，这与$\tilde{P}^*$和$\theta^*$是$F(\tilde{P},\theta)$的局部极大值矛盾。

定理：EM算法的一次迭代可由F函数的极大-极大算法实现。

设θ^i为第i次迭代参数的估计，$\tilde{P}^i$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θ^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θ^{i+1}使$F(\tilde{P}^{i+1},\theta)$极大化。

证明：由引理1，对于固定的θ^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：

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

也即

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

下面固定$\tilde{P}^{i+1}$，求θ^{i+1}使$F(\tilde{P}^{i+1},\theta)$极大，得到

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

由此可知，由EM算法与F函数的极大-极大算法得到的参数估计序列θ^i是一致的。

GEM算法

GEM算法1：

1. 初始化参数θ^o
2. 第$i+1$次迭代，第一步：记θ^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. 第二步：求θ^{i+1}使$F(\tilde{P}^{i+1},\theta)$极大化
4. 重复2,3直到收敛

在上述算法中，有时候求$Q(\theta,\theta^i)$的极大化是困难的，下面的算法2和算法3不是直接求极大化，而是找一个θ^{i+1}使$Q(\theta^{i+1},\theta^i)>Q(\theta^i,\theta^i)$。

GEM算法2：

1. 初始化参数θ^0
2. 第$i+1$次迭代，第一步：计算

$$Q(\theta,\theta^i)=\sum_Z P(Z|Y,\theta^i)\log P(Y,Z|\theta)$$

3. 第二步：求θ^{i+1}使得

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

4.

当参数θ的维数$d\geq 2$时，可采用一种特殊的GEM算法，它将EM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。

GEM算法3：

1. 初始化参数θ^0
2. 第$i+1$次迭代，第一步：计算

$$Q(\theta,\theta^i)=\sum_Z P(Z|Y,\theta^i)\log P(Y,Z|\theta)$$

3. 第二步：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$Q$达到极大的$\theta_1^{i+1}$；

   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$Q$达到极大的$\theta_2^{i+1}$；

   如此，经过d次条件极大化，得到θ^{i+1}使得

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

4. 重复2,3直到收敛

总结

- EM算法是含有隐变量的概率模型极大似然估计或极大后验概率估计的迭代算法，含有隐变量的概率模型的数据表示为$P(Y,Z|\theta)$。EM通过迭代求解观测数据的对数似然函数$L(\theta)=\log P(Y|\theta)$的极大化，实现极大似然估计。每次迭代分为两步：E步求期望，M步求极大。
- EM算法每次迭代后均提高观测数据的似然函数值，即

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

  一般EM是收敛的，但不能保证收敛到全局最优。
- EM算法应用很广，主要用于含有隐变量的概率模型的学习。
- EM还可以解释为F函数的极大-极大算法，可以变形为GEM算法，特点是每次迭代增加F函数值。