**10.期望最大化算法**

**10.1 EM算法的定义**

期望最大化算法是一种迭代算法，用于含有隐含变量（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，简称EM算法。EM算法通过不断求解下界的极大化逼近求解对数似然函数极大化。每次迭代由两步组成：（1）E步，求期望（expectation）；（2）M步，求极大化（maximization）。

**10.2 EM算法的导出**

一个含有隐变量（未观测变量）Z的概率模型，目标是极大化观测数据（不完全数据）Y关于参数θ的对数似然函数

这一目标函数极大化的主要困难是式中包含有未观测数据并有包含和（或积分）的对数。所以EM算法通过迭代逐步近似极大化L(θ)，在迭代过程中我们希望新估计值θ能使L(θ)增加，并逐步达到极大值。假设第i次迭代后θ的估计值是θ(i)，对L(θ)和L(θ(i))作差

利用Jensen不等式（Jensen inequality）

其中λi ≥ 0，。可得

令

则有

即函数B(θ，θ(i))是L(θ)的一个下界。由B(θ，θ(i))的定义式可知

L(θ(i)) = B(θ(i)，θ(i))

因此任何可以使B(θ，θ(i))增大的θ也可以使L(θ)增大。为使L(θ)尽可能的增大，更新时选择θ(i+1)使B(θ，θ(i))达到极大

代入可得θ(i+1)的表达式（省去对θ的极大化而言是常数的项）

上式等价于EM算法的一次迭代，即求Q函数及其极大化。

Q函数的定义：

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和当前参数θ(i)下对未观测数据Z的条件概率分步P(Z | Y，θ(i))的期望

Q函数的第一个变元表示要极大化的参数，第二个变元表示参数当前的估计值。

**10.3 EM算法流程**

输入：观测变量数据Y，隐变量数据Z，联合概率分布P(Y，Z | θ)，条件概率分布P(Z | Y，θ)；

输出：模型参数θ。

（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

其中P(Y，Z | θ(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Q(θ，θ(i))极大化的θ，确定第i+1次迭代的参数估计θ(i+1)

（4）重复第（2）步和第（3）步，直到收敛。

EM算法说明：

（a）参数初值可以任意选择，但是需要注意的是EM算法对初值是敏感的；

（b）每次迭代使似然函数增大或达到局部极值。

（c）步骤（4）给出停止迭代的条件，一般是对较小的正数ε1和ε2，若满足||θ(i+1) -θ(i)|| < ε1或||Q(θ(i+1)，θ(i)) - Q(θ(i)，θ(i))|| < ε2则停止迭代。

（d）EM算法可用于生成模型的非监督学习。

（e）EM算法最大的优点是简单性和普适性。

（f）EM算法是收敛的，但是损失函数L(θ)不是光滑的凸函数，所以EM算法不能保证找到全局最优解。

**10.4 EM算法的收敛性**

EM算法的收敛性证明分为两步：a）观测数据的似然函数L(θ)=P(Y | θ)，θ(i)为EM算法得到的参数估计序列，P(Y | θ(i))则单调递增；b）观测数据的对数似然函数L(θ)=P(Y | θ)有界。

P(Y | θ)有界很好证明，因为P(Y | θ)肯定是小于或等于1的。要证明P(Y | θ(i))的单调性，证明P(Y | θ(i+1)) ≥P(Y | θ(i))即可。

对数似然函数为

令

因为同时有Q函数

所以对数似然函数可以写成

比较大小一般是作差或作商，上式θ分别取θ(i)和θ(i+1)并相减，有

根据EM算法的定义，θ(i+1)使Q函数达到极大，所以有

因此只要证明第二项小于等于即可

根据Jensen不等式可得

所以

综上可知

所以P(Y | θ)单调递增，且P(Y | θ)显然有界，因此P(Y | θ)收敛。

在Q(θ，θ’)和L(θ)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θ\*是L(θ)的稳定点。所以参数估计序列收敛到对数似然函数序列的稳定点，但不能保证收敛到极大值点。初值敏感使初值选择变得非常重要，常用的办法是选取几个不同的初值进行迭代，然后比较得到的各个估计值从中选取最好的。

**10.5 EM算法在高斯混合模型学习中的应用**

高斯混合模型是指具有如下形式的概率分布模型：

其中αk是系数，αk ≥ 0，；是高斯分布密度，θk = (μk，σk2)，

称为第k个分模型。由任意概率分布密度代替高斯分布密度就是一般混合模型。

假设由高斯混合模型生成的观测数据yj，j = 1，2，…，N， 用EM算法估计高斯混合模型的参数，需要确定以下三个方面：

（1）明确隐变量，确定完全数据的对数似然函数。

假设这样产生观测数据：首先依概率αk选择第k个高斯分布模型，然后依第k个分模型的概率分布生成观测数据yj。这时观测数据是已知的，生成观测数据的第k个分模型是未知的，以隐变量γjk表示

其中j = 1，2，…，N，k = 1，2，…，K。所以完全数据是

(yj，γj1，γj2，…，γjK)，j = 1，2，…，N

于是就可以写出完全数据的似然函数

其中，。可得完全数据的对数似然函数为

（2）EM算法的E步确定Q函数。

这里需要计算Eγjk = E(γjk | y，θ)，记其为，表示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的对观测数据yj的响应度

同时，将它们代入Q函数表达式

（3）确定EM算法的M步。

M步就是求Q函数对θ的极大值，即求新一轮迭代的模型参数：

θ(i+1)的各参数由Q函数求得的、及表示。和可以由E步时Q函数分别对求偏导并令其为0求得，是在约束条件下求偏导并令其为0得到

综上所述，高斯混合模型参数估计的EM算法如下：

输入：观测数据y1，y2，…，yN；高斯混合模型；

输出：高斯混合模型参数。

（1）参数初始化；

（2）E步：依据当前的模型参数，计算分模型k对观测数据yj的响应度

（3）M步：计算新一轮迭代的模型参数

（4）重复第（2）步和第（3）步直到收敛。

**10.6 EM算法的推广**

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

（1）F函数的定义。

假设隐变量数据Z的概率分布为，定义分布与参数θ的函数如下：

以上函数称为F函数，是分布的熵。通常假设P(Y，Z | θ)是θ的连续函数，因此F(，θ)是和θ的连续函数。

（2）F函数的重要性质。

1）对于固定的θ，存在唯一的分布极大化F(，θ)，这时由下式给出：

并且随θ联系变化。

2）若则。

3）设为观测数据的对数似然函数，θ(i)（i = 1，2，…，）为EM算法得到的参数估计序列。如果F(，θ)在和θ\*有局部极大值，那么L(θ)也在θ\*有局部极大值。类似的，如果F(，θ)在和θ\*达到全局最大值，那么L(θ)也在θ\*达到全局最大值。

4）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θ的估计， (i)为第i次迭代函数的估计。在第i+1次迭代两步为：

a）对固定的θ(i)，求 (i+1)使F(，θ(i))极大化；

b）对固定的 (i+1)，求θ(i+1)使F( (i+1)，θ)极大化。

（3）GEM算法。

1）GEM算法一。

输入：观测数据，F函数；

输出：模型参数。

（a）初始化参数θ(0)，开始迭代；

（b）第i+1次迭代，第1步：记θ(i)为参数θ的估计值， (i)为函数的估计。求 (i+1)使极大化F(，θ(i))；

（c）第2步：求θ(i+1)使F( (i+1)，θ)极大化；

（d）重复（b）和（c）直到收敛。

2）GEM算法二。

输入：观测数据，Q函数；

输出：模型参数。

（a）初始化参数θ(0)，开始迭代；

（b）第i+1次迭代，第1步：记θ(i)为参数θ的估计值，计算

（c）第2步：求θ(i+1)使Q(θ(i+1)，θ(i)) > Q(θ(i)，θ(i))；

（d）重复（b）和（c）直到收敛。

3）GEM算法三。

输入：观测数据，Q函数；

输出：模型参数。

（a）初始化参数θ(0) = ( θ1(0)，θ2(0)，…，θd(0))，开始迭代；

（b）第i+1次迭代，第1步：记θ(i) = ( θ1(i)，θ2(i)，…，θd(i))为参数θ=( θ1，θ2，…，θd)的估计值，计算

（c）第2步：进行d次条件极大化：

首先在θ2(i)，…，θk(i)保持不变的条件下求使Q(θ，θ(i))达到极大的θ1(i+1)；然后在θ1=θ1(i+1)，θj=θj(i+1)，j = 3，4，…，k的条件下求使Q(θ，θ(i))达到极大的θ2(i+1)；如此迭代经过d次条件极大化得到θ(i+1) = ( θ1(i+1)，θ2(i+1)，…，θd(i+1))使得Q(θ(i+1)，θ(i)) > Q(θ(i)，θ(i))；

（d）重复（b）和（c）直到收敛。