EM 算法

舒双林

2025 年 4 月 7 日

目录

1 EM 算法的理论基础

1.1 EM 算法的引入

EM(Expectation Maximization, 期望最大) 算法是一种从不完全数据或含有隐变量的数据中估计参数的方法，由 Dempster 等人于 1977 年总结提出。EM 算法的基本思想是：在每次迭代中，分两步进行，第一步是求期望 (E 步)，即求隐变量的期望，第二步是求极大 (M 步)，即求参数的极大似然估计。

一般地，用 Y 表示观测随机变量的数据，Z 表示隐变量的数据。Y 和 Z 组合一起称为完全数据，而观测数据 Y 称为不完全数据。假设给定观测数据 Y，其概率分布是 $P(Y|\theta)$，其中 θ 是需要估计的模型参数，那么不完全数据 Y 的似然函数为 $P(Y|\theta)$，对数似然函数 $L(\theta) = \log P(Y|\theta)$；假设 Y 和 Z 的联合概率分布是 $P(Y, Z|\theta)$，那么完全数据 Y 和 Z 的对数似然函数为 $L(\theta) = \log P(Y, Z|\theta)$。

1.2 EM 算法的数学推导

对于含有隐变量 Z 的概率模型，直接极大化对数似然函数 $L(\theta) = \log P(Y|\theta)$ 是困难的，因为其含有未观测数据且包含和（或积分）的对数。

假设在第 i 次迭代后 θ 的估计值是 $\theta^{(i)}$。考虑新估计值 θ 能否使 $L(\theta)$ 增加，对两者作差：

$$L(\theta) - L(\theta^{(i)}) = \log\left(\sum_Z P(Y|Z,\theta)P(Z|\theta)\right) - \log P(Y|\theta^{(i)}) \tag{1}$$

利用 Jensen 不等式，得到其下界：

$$L(\theta) - L(\theta^{(i)}) = \log\left(\sum_Z P(Z|Y,\theta^{(i)})\frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})}\right) - \log P(Y|\theta^{(i)}) \tag{2}$$

$$\geq \sum_Z P(Z|Y,\theta^{(i)}) \log \frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})} - \log P(Y|\theta^{(i)}) \tag{3}$$

定义辅助函数 $Q(\theta, \theta^{(i)})$：

$$Q(\theta, \theta^{(i)}) = \sum_Z P(Z|Y,\theta^{(i)}) \log P(Y, Z|\theta) \tag{4}$$

可以证明，最大化 $Q(\theta, \theta^{(i)})$ 可以保证 $L(\theta)$ 不减。因此，EM 算法的迭代公式为：

$$\theta^{(i+1)} = \arg\max_\theta Q(\theta, \theta^{(i)}) \tag{5}$$

这等价于：

1. E 步：计算 $Q(\theta, \theta^{(i)})$

2. M 步：求解 $\theta^{(i+1)} = \arg\max_\theta Q(\theta, \theta^{(i)})$

EM 算法通过不断求解下界的极大化来逼近求解对数似然函数极大化，但不能保证找到全局最优值。

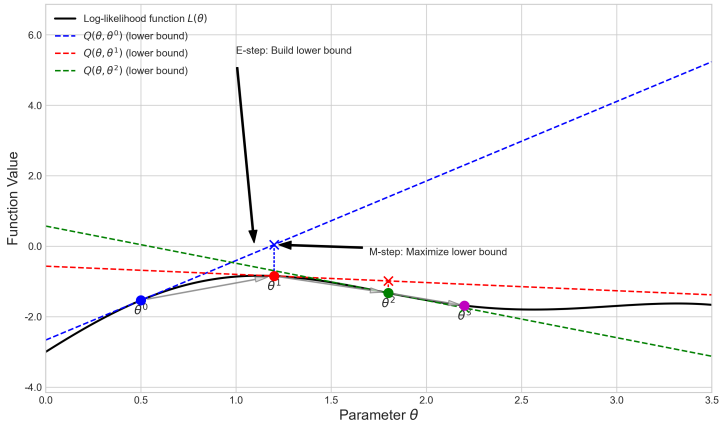

图 1: EM 算法示意图

1.3 EM 算法流程

算法 1 EM 算法

输入: 观测变量数据 Y、隐变量数据 Z、联合分布 $P(Y,Z|\theta)$、条件分布 $P(Z|Y,\theta)$

过程:

1: 选择参数初值 $\theta^{(0)}$、最大迭代次数 N、迭代精度 δ，开始迭代

2: **for** $i=1$ to N **do**

3: 令 $\theta^{(i-1)}$ 为第 $i-1$ 次迭代参数 θ 的估计值

4: E-step: 计算在给定观测数据 Y 和当前参数 $\theta^{(i-1)}$ 下 Z 的条件概率分布期望

$$Q(\theta,\theta^{(i-1)})=\sum_{Z}\log P(Y,Z|\theta)P(Z|Y,\theta^{(i-1)}) \tag{6}$$

5: M-step: 极大化 $Q(\theta,\theta^{(i-1)})$，确定第 i 次迭代的参数估计值 $\theta^{(i)}$

$$\theta^{(i)}=\arg\max_{\theta}Q(\theta,\theta^{(i-1)}) \tag{7}$$

6: 计算 $\theta^{(i-1)}$ 与 $\theta^{(i)}$ 的差值的二范数 $\delta^{(i)}=||\theta^{(i)}-\theta^{(i-1)}||$

7: **if** $\delta^{(i)}<\delta$ **then**

8: 迭代结束，$\hat{\theta}=\theta^{(i)}$ 为参数的极大似然估计值

9: **else**

10: 继续迭代，$i=i+1$，返回第 3 步

11: 迭代结束，$\hat{\theta}=\theta^{(N)}$ 为参数的极大似然估计值

输出: 模型的参数估计值 $\hat{\theta}$

1.4 关于 EM 算法的说明

下面关于 EM 算法作几点说明:

1. 参数的初值可以任意选择，但需注意 EM 算法对初值是敏感的。

2. 迭代停止的条件可以是参数的变化小于一个给定的阈值，也可以是 Q 函数的增益小于一个给定的阈值, 即

$$\|\theta^{(i+1)}-\theta^{(i)}\|<\delta_1 \quad \text{or} \quad \|Q(\theta^{(i+1)},\theta^{(i)})-Q(\theta^{(i)},\theta^{(i)})\|<\delta_2 \tag{8}$$

3. M 步求 $Q(\theta,\theta^{(i)})$ 的极大化，得到 $\theta^{(i+1)}$，完成一次迭代 $\theta^{(i)}\to\theta^{(i+1)}$。后续将给出定理保证 EM 算法的收敛性。

1.5 EM 算法的收敛性

EM 算法的收敛性由以下定理保证:

[单调性] 设 $P(Y|\theta)$ 为观测数据的似然函数，$\theta^{(i)}(i=1,2,\cdots)$ 为 EM 算法得到的参数估计序列，$P(Y|\theta^{(i)})(i=1,2,\cdots)$ 为对应的似然函数序列，则 $P(Y|\theta^{(i)})$ 是单调递增的。

$$P(Y|\theta^{(i+1)}) \geq P(Y|\theta^{(i)}) \tag{9}$$

[收敛性] 设 $L(\theta)=\log P(Y|\theta)$ 为观测数据的对数似然函数，$\theta^{(i)}(i=1,2,\cdots)$ 为 EM 算法得到的参数估计序列，$L(\theta^{(i)})(i=1,2,\cdots)$ 为对应的对数似然函数序列。

(1) 收敛性: 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) 稳定点性质: 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。

1.6 EM 算法的应用场景

EM 算法广泛应用于统计学、机器学习和数据挖掘等领域，尤其在处理缺失数据和隐变量模型时表现出色。以下是一些常见的应用场景:

- 高斯混合模型 (GMM)
- 隐马尔可夫模型 (HMM)
- 潜在语义分析 (LSA)
- 概率主成分分析 (PPCA)

EM 算法的优势在于:

- 能够从不完全数据中有效估计模型参数
- 实现简单，计算效率较高
- 理论基础良好，易于理解和分析
- 可与其他算法结合使用 (如变分推断、MCMC 等)