

Data Mining and Statistics

Expectation Maximization algorithm

Kong

2020-03-28

目录

EM 算法	1
1.1 EM 算法的说明	1
1.2 EM 算法的导出	3

EM 算法

1.1 EM 算法的说明

EM 算法与初值的选择有关,选择不同的初值可能得到不同的参数估计值。一般地,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 和 Z 连在一起称为完全数据 (complete-data), 观测数据 Y 又被称为不完全数据 (incomplete-data)。假设给定观测数据 Y , 其概率分布为 $P(Y|\theta)$, 其中 θ 是需要估计的模型参数, 那么不完全数据 Y 的似然函数是 $P(Y|\theta)$, 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 $P(Y, Z|\theta)$, 那么完全数据的对数似然函数是 $\log P(Y, Z|\theta)$ 。

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。下面介绍 EM 算法。

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 $P(Z|Y, \theta)$;

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

$$\begin{aligned} Q(\theta, \theta^{(i)}) &= E_Z[\log P(Y, Z|\theta)|Y, \theta^{(i)}] \\ &= \sum_Z \log P(Y, Z|\theta) P(Z|Y, \theta^{(i)}) \end{aligned} \quad (1.1)$$

这里， $P(Z|Y, \theta^{(i)})$ 是在给定观测数据 Y 和当前的参数估计 $\theta^{(i)}$ 下隐变量数据 Z 的条件概率分布；

(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 $\theta^{(i+1)}$

$$\theta^{(i+1)} = \arg \max_{\theta} Q(\theta, \theta^{(i)}) \quad (1.2)$$

(4) 重复第 (2) 步和第 (3) 步，直到收敛。

式 (1.1) 的函数 $Q(\theta, \theta^{(i)})$ 是 EM 算法的核心，称为 Q 函数 (Q function)。

定义 (Q 函数) 完全数据的对数似然函数 $\log P(Y, Z|\theta)$ 关于在给定观测数据 Y 和当前参数 $\theta^{(i)}$ 下对未观测数据 Z 的条件概率分布 $P(Z|Y, \theta^{(i)})$ 的期望称为 Q 函数，即

$$Q(\theta, \theta^{(i)}) = E_Z[\log P(Y, Z|\theta)|Y, \theta^{(i)}] \quad (1.3)$$

下面关于 EM 算法作几点说明：

步骤 (1) 参数的初值可以任意选择，但需要注意 EM 算法对初值是敏感的。

步骤 (2) E 步求 $Q(\theta, \theta^{(i)})$ 。 Q 函数式中 Z 是未观测数据， Y 是观测数据。注意， $Q(\theta, \theta^{(i)})$ 的第 1 个变元表示要极大化的参数，第 2 个变元表示参数的当前估计值。每次迭代实际在求 Q 函数及其极大。

步骤 (3) M 步求 $Q(\theta, \theta^{(i)})$ 的极大化，得到 $\theta^{(i+1)}$ ，完成一次迭代 $\theta^{(i)} \rightarrow \theta^{(i+1)}$ 。后面将证明每次迭代使似然函数增大或达到局部极值。

步骤 (4) 给出停止迭代的条件，一般是对较小的正数 $\varepsilon_1, \varepsilon_2$ ，若满足

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

则停止迭代。

1.2 EM 算法的导出

下面通过近似求解观测数据的对数似然函数的极大化问题来导出 EM 算法。我们面对一个含有隐变量的概率模型，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，即极大化

$$\begin{aligned} L(\theta) &= \log P(Y|\theta) = \log \sum_Z P(Y, Z|\theta) \\ &= \log \left(\sum_Z P(Y|Z, \theta) P(Z|\theta) \right) \end{aligned} \quad (1.4)$$

注意到这一极大化的主要困难是式（1.4）中含有未观测数据以及和（或积分）的对数。

事实上，EM 算法是通过迭代逐步近似极大化 $L(\theta)$ 的。假设在第 i 次迭代后 θ 的估计值是 $\theta^{(i)}$ 。我们希望新估计值 θ 能使 $L(\theta)$ 增加，即 $L(\theta) > L(\theta^{(i)})$ ，并逐步达到极大值。为此，考虑两者的差：

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

利用 Jensen 不等式（Jensen inequality）得到其下界：

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

令

$$B(\theta, \theta^{(i)}) \triangleq L(\theta^{(i)}) + \sum_Z P(Z|Y, \theta^{(i)}) \log \frac{P(Y|Z, \theta) P(Z|\theta)}{P(Z|Y, \theta^{(i)}) P(Y|\theta^{(i)})} \quad (1.5)$$

则

$$L(\theta) \geq B(\theta, \theta^{(i)}) \quad (1.6)$$

即函数 $B(\theta, \theta^{(i)})$ 是 $L(\theta)$ 的一个下界，而且由式（1.5）可知，

$$L(\theta^{(i)}) = B(\theta^{(i)}, \theta^{(i)}) \quad (1.7)$$

因此，任何可以使 $B(\theta, \theta^{(i)})$ 增大的 θ ，也可以使 $L(\theta)$ 增大。为了使 $L(\theta)$ 有尽可能大的增长，选择 $\theta^{(i+1)}$ 使 $B(\theta, \theta^{(i)})$ 达到极大，即

$$\theta^{(i+1)} = \arg \max_{\theta} B(\theta, \theta^{(i)}) \quad (1.8)$$

现在求 $\theta^{(i+1)}$ 的表达式。省去对 θ 的极大化而言是常数的项，由式 (1.8)、式 (1.5) 及式 (1.2)，有

$$\begin{aligned}
\theta^{(i+1)} &= \arg \max_{\theta} \left(L(\theta^{(i)}) + \sum_Z P(Z|Y, \theta^{(i)}) \log \frac{P(Y|Z, \theta)P(Z|\theta)}{P(Z|Y, \theta^{(i)})P(Y|\theta^{(i)})} \right) \\
&= \arg \max_{\theta} \left(\sum_Z P(Z|Y, \theta^{(i)}) \log (P(Y|Z, \theta)P(Z|\theta)) \right) \\
&= \arg \max_{\theta} \left(\sum_Z P(Z|Y, \theta^{(i)}) \log P(Y, Z|\theta) \right) \\
&= \arg \max_{\theta} Q(\theta, \theta^{(i)})
\end{aligned} \tag{1.9}$$

式 (1.9) 等价于 EM 算法的一次迭代，即求 Q 函数及其极大化。故 EM 算法是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的算法。