EM算法及其推广

演讲者：顾嘉辉

日期：2020.1.13

目录
CONTENTS

PART ONE

EM算法的引入

1EM算法概述

1. EM算法是一种迭代算法

2. 用于含有隐变量的概率模型参数的极大似然估计

3. EM算法由两步组成：E步（求期望），M步（求极大），故称为Expectation Maximization Algorithm

4. 一句话总结：EM算法就是含有隐变量的概率模型参数的极大似然估计法

1三硬币模型

首先介绍一个使用 EM 算法的例子。

例 9.1（三硬币模型） 假设有 3 枚硬币，分别记作 A，B，C。这些硬币正面出现的概率分别是 π，p 和 q。进行如下掷硬币试验：先掷硬币 A，根据其结果选出硬币 B 或硬币 C，正面选硬币 B，反面选硬币 C；然后掷选出的硬币，掷硬币的结果，出现正面记作 1，出现反面记作 0；独立地重复 n 次试验（这里，$n=10$），观测结果如下：

$$1,1,0,1,0,0,1,0,1,1$$

假设只能观测到掷硬币的结果，不能观测掷硬币的过程。问如何估计三硬币正面出现的概率，即三硬币模型的参数。

$$P(y|\theta) = \sum_z P(y,z|\theta) = \sum_z P(z|\theta)P(y|z,\theta)$$
$$= \pi p^y (1-p)^{1-y} + (1-\pi) q^y (1-q)^{1-y}$$

1. y是随机变量0或1（可观测）

2. z是隐变量，表示未观测到的硬币A的投掷结果（不可观测）

3. θ =（π, p, q）是模型的参数ABC硬币正面出现的概率

1三硬币模型

将观测数据表示为 $Y=(Y_1,Y_2,\cdots,Y_n)^{\mathrm{T}}$，未观测数据表示为 $Z=(Z_1,Z_2,\cdots,Z_n)^{\mathrm{T}}$
则观测数据的似然函数为

$$P(Y|\theta)=\sum_Z P(Z|\theta)P(Y|Z,\theta) \tag{9.2}$$

即

$$P(Y|\theta)=\prod_{j=1}^{n}[\pi p^{y_j}(1-p)^{1-y_j}+(1-\pi)q^{y_j}(1-q)^{1-y_j}] \tag{9.3}$$

考虑求模型参数 $\theta=(\pi,p,q)$ 的极大似然估计，即

$$\hat{\theta}=\arg\max_{\theta}\log P(Y|\theta) \tag{9.4}$$

1.观测到数据为Y，有N个

2、未观测到的数据也就是隐变量，和观测到的数据一一对应，也有N个

3、对这个模型求极大似然估计

1三硬币模型

E 步：计算在模型参数 $\pi^{(i)}$，$p^{(i)}$，$q^{(i)}$ 下观测数据 y_j 来自掷硬币 B 的概率

$$\mu_j^{(i+1)} = \frac{\pi^{(i)}(p^{(i)})^{y_j}(1-p^{(i)})^{1-y_j}}{\pi^{(i)}(p^{(i)})^{y_j}(1-p^{(i)})^{1-y_j}+(1-\pi^{(i)})(q^{(i)})^{y_j}(1-q^{(i)})^{1-y_j}} \tag{9.5}$$

M 步：计算模型参数的新估计值

$$\pi^{(i+1)} = \frac{1}{n}\sum_{j=1}^{n}\mu_j^{(i+1)} \tag{9.6}$$

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

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

和常规EM算法一样，分成E步和M步：

1、E步：就是计算观测数据yj来自硬币B的概率

2、M步：找到模型新的参数估计值

1三硬币模型

E 步：计算在模型参数 $\pi^{(i)}$，$p^{(i)}$，$q^{(i)}$ 下观测数据 y_j 来自掷硬币 B 的概率

$$\mu_j^{(i+1)} = \frac{\pi^{(i)}(p^{(i)})^{y_j}(1-p^{(i)})^{1-y_j}}{\pi^{(i)}(p^{(i)})^{y_j}(1-p^{(i)})^{1-y_j} + (1-\pi^{(i)})(q^{(i)})^{y_j}(1-q^{(i)})^{1-y_j}} \tag{9.5}$$

M步:针对Q函数求导，Q函数的表达式是

$$Q(\theta,\theta^i) = \sum_{j=1}^{N}\sum_{z} P(z|y_j,\theta^i)\underline{logP(y_j,z|\theta)} = \sum_{j=1}^{N}[\boxed{\mu_j}\underline{log(\pi p^{y_j}(1-p)^{1-y_j})} + \boxed{(1-\mu_j)}\underline{log((1-\pi)q^{y_j}(1-q)^{1-y_j})}]$$

1、这一步计算的其实是：未观测数据（隐变量）的条件概率分布（未观测到的数据是掷硬币A的结果）

2、图中给出的是来自硬币B的概率，那么相对应的，观测结果yj来自硬币C的概率就是1-μ

1三硬币模型

M步:针对Q函数求导，Q函数的表达式是

$$Q(\theta,\theta^i)=\sum_{j=1}^{N}\sum_{z}P(z|y_j,\theta^i)\underline{logP(y_j,z|\theta)}=\sum_{j=1}^{N}\boxed{\mu_j}\underline{log(\pi p^{y_j}(1-p)^{1-y_j})}+\boxed{(1-\mu_j)}\underline{log((1-\pi)q^{y_j}(1-q)^{1-y_j})}]$$

对Q函数关于π求导，在令其等于0，就可以得到书上的答案：

$$\frac{\partial Q}{\partial \pi}=(\frac{\mu_1}{\pi}-\frac{1-\mu_1}{1-\pi})+\cdots+(\frac{\mu_N}{\pi}-\frac{1-\mu_N}{1-\pi})=\frac{\mu_1-\pi}{\pi(1-\pi)}+\cdots+\frac{\mu_N-\pi}{\pi(1-\pi)}=\frac{\sum_{j=1}^{N}\mu_j-N\pi}{\pi(1-\pi)}$$

M 步：计算模型参数的新估计值

$$\pi^{(i+1)}=\frac{1}{n}\sum_{j=1}^{n}\mu_j^{(i+1)} \tag{9.6}$$

1EM算法

算法 9.1（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 $P(Z|Y,\theta)$；

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

$$\begin{aligned} Q(\theta,\theta^{(i)}) &= E_Z[\log P(Y,Z|\theta)|Y,\theta^{(i)}] \\ &= \sum_Z \log P(Y,Z|\theta)P(Z|Y,\theta^{(i)}) \end{aligned} \tag{9.9}$$

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

$$\theta^{(i+1)} = \arg\max_{\theta} Q(\theta,\theta^{(i)}) \tag{9.10}$$

（4）重复第 (2) 步和第 (3) 步，直到收敛。■

式 (9.9) 的函数 $Q(\theta,\theta^{(i)})$ 是 EM 算法的核心，称为 Q 函数（Q function）。

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

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

1、选择初值

2、根据引进的条件概率分布得出期望

3、求得使上一步得出的期望最大化的参数

4、重复2/3步，直到收敛停止（停止迭代条件就是参数变化较小或者Q函数变化较小的时候）

1EM算法的导出

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

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）[①]得到其$\boxed{\text{下界}}$：

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

令

$$B(\theta,\theta^{(i)}) \doteq L(\theta^{(i)}) + \sum_Z P(Z|Y,\theta^{(i)}) \log \frac{P(Y|Z,\theta)P(Z|\theta)}{P(Z|Y,\theta^{(i)})P(Y|\theta^{(i)})} \tag{9.13}$$

则

$$L(\theta) \geqslant \boxed{B(\theta,\theta^{(i)})} \tag{9.14}$$

① 这里用到的是 $\log \sum_j \lambda_j y_j \geqslant \sum_j \lambda_j \log y_j$，其中 $\lambda_j \geqslant 0$，$\sum_j \lambda_j = 1$。

这里其实就是引入了一个分布：

$$P(Z|Y,\theta^{(i)})$$

然后再利用Jensen Inequality的性质得到这个大于的式子：

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

1EM算法的导出

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 (9.16)、式 (9.13) 及式 (9.10)，有

$$
\begin{aligned}
\theta^{(i+1)} &= \arg\max_{\theta} \left(\boxed{L(\theta^{(i)})} + \sum_Z P(Z|Y,\theta^{(i)}) \log \frac{P(Y|Z,\theta)P(Z|\theta)}{\boxed{P(Z|Y,\theta^{(i)})P(Y|\theta^{(i)})}} \right) \\
&= \arg\max_{\theta} \left(\sum_Z P(Z|Y,\theta^{(i)}) \log (P(Y|Z,\theta)P(Z|\theta)) \right) \\
&= \arg\max_{\theta} \left(\sum_Z P(Z|Y,\theta^{(i)}) \log P(Y,Z|\theta) \right) \\
&= \arg\max_{\theta} Q(\theta,\theta^{(i)}) \qquad (9.17)
\end{aligned}
$$

在这个argmax中，
只有θ是变量，θi是在上一步迭代中确定了的数字，所以可以视为常数，化简后发现其实就是对Q函数求argmax

PART TWO

EM算法的收敛性

2EM算法的收敛性

定理 9.1 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，即

$$P(Y|\theta^{(i+1)}) \geqslant P(Y|\theta^{(i)}) \tag{9.18}$$

EM算法的收敛性其实就是讨论：EM算法得到的估计序列是否收敛，如果收敛，是否收敛到全局最大值或者局部最大值，接下来给出定理：

定理9.1说明了P(Y|θi)是单调递增的

2定理9.1的证明

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

取对数得到这个式子：

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

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

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

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

2EM算法的收敛性

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

（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 $\theta^{(i+1)}$

$$\theta^{(i+1)} = \arg\max_{\theta} Q(\theta, \theta^{(i)}) \tag{9.10}$$

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

这个式子的第一项由于之前M步求得arg max所以Q(θ i+1)天然大于Q(θ i)

2EM算法的收敛性

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

$$\boxed{\leqslant \log \left(\sum_Z \frac{P(Z|Y,\theta^{(i+1)})}{P(Z|Y,\theta^{(i)})} P(Z|Y,\theta^{(i)}) \right)}$$

$$= \log \left(\sum_Z P(Z|Y,\theta^{(i+1)}) \right) = 0 \tag{9.23}$$

这里的不等号由 Jensen 不等式得到。

第二项由Jensen inequality得到不等号：

PART THREE

EM算法在高斯混合模型学习中的应用

3高斯混合模型

定义 9.2（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：

$$P(y|\theta)=\sum_{k=1}^{K}\alpha_k\phi(y|\theta_k) \tag{9.24}$$

其中，α_k 是系数，$\alpha_k \geqslant 0$，$\sum_{k=1}^{K}\alpha_k=1$；$\phi(y|\theta_k)$ 是高斯分布密度，$\theta_k=(\mu_k,\sigma_k^2)$，

$$\phi(y|\theta_k)=\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right) \tag{9.25}$$

称为第 k 个分模型。

一般混合模型可以由任意概率分布密度代替式 (9.25) 中的高斯分布密度，我们只介绍最常用的高斯混合模型。

α 是系数，其实可以理解为权重

Θ 就是参数，包括均值和方差

3高斯混合模型参数估计的EM算法

假设观测数据 $y_1, y_2, \cdots, y_N$ 由高斯混合模型生成，

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

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

观测数据为：y1, y2…yN
其中参数θ由权重和高斯分布密度的均值和方差组成

3高斯混合模型参数估计的EM算法

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，$j=1,2,\cdots,N$，是这样产生的：首先依概率 α_k 选择第 k 个高斯分布分模型 $\phi(y|\theta_k)$，然后依第 k 个分模型的概率分布 $\phi(y|\theta_k)$ 生成观测数据 y_j。这时观测数据 y_j，$j=1,2,\cdots,N$，是已知的；反映观测数据 y_j 来自第 k 个分模型的数据是未知的，$k=1,2,\cdots,K$，以隐变量 γ_{jk} 表示，其定义如下：

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

$$j=1,2,\cdots,N;\quad k=1,2,\cdots,K \tag{9.27}$$

γ_{jk} 是 0-1 随机变量。

白板画图解释

3高斯混合模型参数估计的EM算法

于是，可以写出完全数据的似然函数：

$$
\begin{aligned}
P(y,\gamma|\theta) &= \prod_{j=1}^{N} P(y_j,\gamma_{j1},\gamma_{j2},\cdots,\gamma_{jK}|\theta) \\
&= \prod_{k=1}^{K}\prod_{j=1}^{N}[\alpha_k\phi(y_j|\theta_k)]^{\gamma_{jk}} \\
&= \prod_{k=1}^{K}\alpha_k^{n_k}\prod_{j=1}^{N}[\phi(y_j|\theta_k)]^{\gamma_{jk}} \\
&= \prod_{k=1}^{K}\alpha_k^{n_k}\prod_{j=1}^{N}\left[\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y_j-\mu_k)^2}{2\sigma_k^2}\right)\right]^{\gamma_{jk}}
\end{aligned}
$$

式中，$n_k=\sum_{j=1}^{N}\gamma_{jk}$，$\sum_{k=1}^{K}n_k=N$。

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

对完全数据的似然函数取log即可得到下面的式子

3高斯混合模型参数估计的EM算法

2. EM算法的E步：确定Q函数

$$Q(\theta,\theta^{(i)})=E[\log P(y,\gamma|\theta)|y,\theta^{(i)}]$$
$$=E\left\{\sum_{k=1}^{K}\left\{n_k\log\alpha_k+\sum_{j=1}^{N}\gamma_{jk}\left[\log\left(\frac{1}{\sqrt{2\pi}}\right)-\log\sigma_k-\frac{1}{2\sigma_k^2}(y_j-\mu_k)^2\right]\right\}\right\}$$
$$=\sum_{k=1}^{K}\left\{\sum_{j=1}^{N}(E\gamma_{jk})\log\alpha_k+\sum_{j=1}^{N}(E\gamma_{jk})\left[\log\left(\frac{1}{\sqrt{2\pi}}\right)-\log\sigma_k-\frac{1}{2\sigma_k^2}(y_j-\mu_k)^2\right]\right\} \tag{9.28}$$

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

$$\hat{\gamma}_{jk}=E(\gamma_{jk}|y,\theta)=P(\gamma_{jk}=1|y,\theta)$$
$$=\frac{P(\gamma_{jk}=1,y_j|\theta)}{\sum_{k=1}^{K}P(\gamma_{jk}=1,y_j|\theta)}$$
$$=\frac{P(y_j|\gamma_{jk}=1,\theta)P(\gamma_{jk}=1|\theta)}{\sum_{k=1}^{K}P(y_j|\gamma_{jk}=1,\theta)P(\gamma_{jk}=1|\theta)}$$
$$=\frac{\alpha_k\phi(y_j|\theta_k)}{\sum_{k=1}^{K}\alpha_k\phi(y_j|\theta_k)},\quad j=1,2,\cdots,N;\quad k=1,2,\cdots,K$$

将$\hat{\gamma}_{jk}=E\gamma_{jk}$及$n_k=\sum_{j=1}^{N}E\gamma_{jk}$代入式(9.28)，即得

$$Q(\theta,\theta^{(i)})=\sum_{k=1}^{K}\left\{n_k\log\alpha_k+\sum_{j=1}^{N}\hat{\gamma}_{jk}\left[\log\left(\frac{1}{\sqrt{2\pi}}\right)-\log\sigma_k-\frac{1}{2\sigma_k^2}(y_j-\mu_k)^2\right]\right\} \tag{9.29}$$

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yj的响应度，画图解释。

3高斯混合模型参数估计的EM算法

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

$$\hat{\mu}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} y_j}{\sum_{j=1}^{N} \hat{\gamma}_{jk}}, \quad k = 1, 2, \cdots, K$$

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

$$\hat{\alpha}_k = \frac{n_k}{N} = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk}}{N}, \quad k = 1, 2, \cdots, K$$

1、对Q函数求argmax

2、Q函数对其中的参数求偏导并令其等于零

3高斯混合模型参数估计的EM算法

算法 9.2（高斯混合模型参数估计的EM算法）

输入：观测数据 $y_1, y_2, \cdots, y_N$，高斯混合模型；

输出：高斯混合模型参数。

(1) 取参数的初始值开始迭代；

(2) E 步：依据当前模型参数，计算分模型 k 对观测数据 y_j 的响应度

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

(3) M 步：计算新一轮迭代的模型参数

$$\hat{\mu}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} y_j}{\sum_{j=1}^{N} \hat{\gamma}_{jk}}, \quad k = 1, 2, \cdots, K$$

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

$$\hat{\alpha}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk}}{N}, \quad k = 1, 2, \cdots, K$$

(4) 重复第 (2) 步和第 (3) 步，直到收敛。

1、选择初值开始迭代

2、E步：计算模型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2/3步，直到收敛停止

PART FOUR

EM算法的推广

4F函数的极大极大算法

定义 9.3（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：

$$F(\tilde{P},\theta) = E_{\tilde{P}}[\log P(Y,Z|\theta)] + H(\tilde{P}) \tag{9.33}$$

称为 F 函数。式中 $H(\tilde{P}) = -E_{\tilde{P}} \log \tilde{P}(Z)$ 是分布 $\tilde{P}(Z)$ 的熵。

在定义 9.3 中，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。函数 $F(\tilde{P},\theta)$ 还有以下重要性质。

定义 9.1（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

$$Q(\theta,\theta^{(i)}) = E_Z[\log P(Y,Z|\theta)|Y,\theta^{(i)}] \tag{9.11}$$

1、在很多情况下其实这个无法观测的数据Z的概率分布是无法直接得到的，那么就有了这个通用EM算法的推出

4F函数的极大极大算法

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 $\tilde{P}_\theta$ 由下式给出：

$$\tilde{P}_\theta(Z)=P(Z|Y,\theta) \tag{9.34}$$

并且 $\tilde{P}_\theta$ 随 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

$$L=\boxed{E_{\tilde{P}}\log P(Y,Z|\theta)}\boxed{-E_{\tilde{P}}\log\tilde{P}(Z)}+\boxed{\lambda\left(1-\sum_Z\tilde{P}(Z)\right)} \tag{9.35}$$

将其对 $\tilde{P}$ 求偏导数：

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

令偏导数等于 0，得出

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

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

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

再从约束条件 $\sum_Z\tilde{P}_\theta(Z)=1$ 得式 (9.34)。

4 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。

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 $Q(\theta,\theta^{(i)})$ 的定义式 (9.11) 有

argmax F

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

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

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

4 GEM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。

1、有时对Q(θ，θi)极大化是有困难的

4 GEM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

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

（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。

4 GEM算法

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

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

（3）第 2 步：进行 d 次条件极大化：

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

然后，在 $\theta_1=\theta_1^{(i+1)}$，$\theta_j=\theta_j^{(i)}$，$j=3,4,\cdots,d$ 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。

1、当参数的维度d(>=2)可将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，其余分量不变

PART FIVE

代码演示

THANKS!