EM 算法

cloud

2017.1.6

1 概述

EM 算法是一种迭代算法，用于含有隐变量的概率模型参数的极大似然估计或极大后验概率估计。EM 算法分为求期望的 E 步和求极大的 M 步，所以这一算法称之为期望极大算法。

2 基本原理

2.1 算法流程

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；

解释：参数的初值可以任意选择，算法对初值时敏感的。

（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 $Q(\theta,\theta^{(i)})$。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；

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

解释：Z 是未观测数据，Y 是观测数据，$Q(\theta,\theta^{(i)})$ 的第一个变元表示要极大化的参数，第二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。

（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 $\theta^{(i+1)}$：

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

解释：M 步求 Q 函数的极大化，得到 $\theta^{(i+1)}$，完成一次 $\theta^{(i)} \to \theta^{(i+1)}$ 的迭代。

（4）重复第 (2)(3) 步直至收敛。

解释：这里的收敛指的是迭代前后参数值之差或者 Q 函数之差小到一定的限定值。

2.2 Q 函数

上节中的 Q 函数指的是完全数据的对数似然函数 $logP(Y,Z|\theta)$ 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。即

$$Q(\theta,\theta^{(i)}) = E_Z[logP(Y,Z|\theta)|Y,\theta^{(i)}]$$

那么 Q 函数是怎么来的呢？为何这样定义呢？下面给出推导过程。

2.3 算法推导

EM 算法能近似实现对观测数据的极大似然估计，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。

当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

$$\begin{aligned} L(\theta) = logP(Y|\theta) &= log\sum_Z P(Y,Z|\theta) \\ &= log(\sum_Z P(Y|Z,\theta)P(Z|\theta)) \end{aligned} \tag{1}$$

这一极大化的困难是 (1) 式中有未观测数据并有和的对数。EM 算法就是解决这一困难的算法，它通过迭代逐步近似极大化 $L(\theta)$，即使得新估计值 θ 的 $L(\theta)$ 大于上次估计值 $\theta^{(i)}$ 的 $L(\theta^{(i)})$。为此需要考虑两者的差：

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

上式最后一步由等式转化为不等式利用了 $Jensen$ 不等式如下的性质：

$$log\sum_j \lambda_j y_j \geqslant \sum_j \lambda_j log y_j$$

$$\lambda_j \geqslant 0, \sum_j \lambda_j = 1$$

我们令

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

就有 $L(\theta) \geqslant B(\theta, \theta^{(i)})$，即函数 $B(\theta, \theta^{(i)})$ 是 $L(\theta)$ 的一个下界，而且由式 (2) 可知 $L(\theta^{(i)}) = B(\theta^{(i)}, \theta^{(i)})$。任何使得 $B(\theta, \theta^{(i)})$ 增大的 θ，也可以使 $L(\theta)$ 增大。选择 $\theta^{(i+1)}$ 使 $B(\theta, \theta^{(i)})$ 达到极大，即

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

由式 (2) 进一步可以写成如下的式子，其中省略对 θ 而言是常数的项。

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

这里将 $\sum_Z logP(Y, Z|\theta)P(Z|Y, \theta^{(i)})$ 写成期望的形式即 $E_Z[logP(Y, Z|\theta)|Y, \theta^{(i)}]$，由此得到了 EM 算法 E 步中 Q 函数的形式，同时也得到了 M 步的极大化目标。这里其实相当于将参数估计所使用的的极大似然方法转化成求极大似然函数下界的极大。EM 算法就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。

2.4 算法解释

在下图中，上方的曲线为 $L(\theta)$，下方曲线为 $B(\theta, \theta^{(i)})$，由前面 $L(\theta^{(i)}) = B(\theta^{(i)}, \theta^{(i)})$ 可知两个函数在点 $\theta = \theta^{(i)}$ 处相等。在进行迭代的时候，EM 算法找到下一个点 $\theta^{(i+1)}$ 使函数 $B(\theta, \theta^{(i)})$ 极大化，也会使得函数 $Q(\theta, \theta^{(i)})$ 极大。找到使极大化的点 $\theta^{(i+1)}$ 后重新计算 Q 函数的值，进入下一次迭代。从图中可知 EM 算法不能保证找到全局最优值。

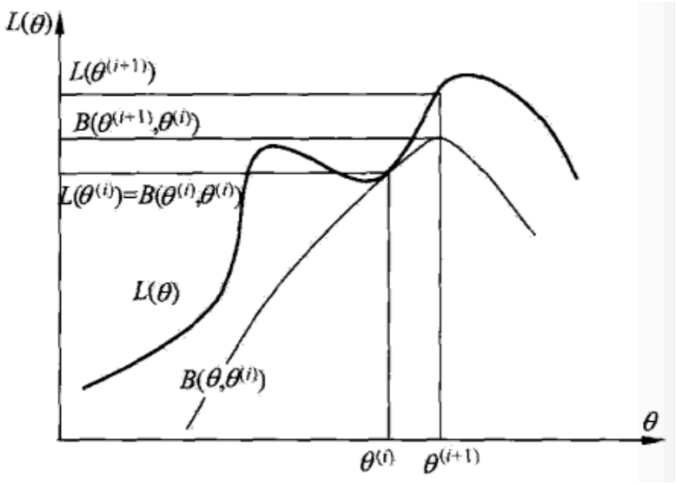

Figure 1: 算法解释

3 高斯混合模型中的应用

3.1 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 $\alpha_k \geqslant 0, \sum_{k=1}^{K} \alpha_k = 1$，$\phi(y|\theta_k)$ 是高斯分布密度称为第 k 个模型，$\phi(y|\theta_k) = \frac{1}{\sqrt{2\pi}\sigma_k} exp(-\frac{(y-\mu_k)^2}{2\sigma^2})$，$\theta_k = (\mu_k, \sigma_k^2)$。

3.2 高斯混合模型使用 EM 算法进行参数估计

假设观测数据 $y^{(1)}, y^{(2)}, ..., y^{(N)}$ 由高斯混合模型生成

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

其中 $\theta = (\alpha_1, \alpha_2, ..., \alpha_k; \theta_1, \theta_2, ..., \theta_k)$，我们要用 EM 算法估计高斯混合模型的参数 θ。

（1）明确隐变量，写出完全数据的对数似然函数

对于每一个观测数据 $y^{(j)}, j = 1, 2, ..., N$，它是这样产生的。首先依概率 α_k 选择第 k 个高斯分布模型 $\phi(y|\theta_k)$，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 $y^{(j)}$ 是已知的，反映观测数据 $y^{(j)}$ 来自第 k 个分模型的数据时未知的，$k = 1, 2, ..., K$，以隐变量 γ_{jk} 为例其定义如下:

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

那么通过观测数据 $y^{(j)}$ 和未观测数据 γ_{jk} 就可得到完全数据时 $(y^{(j)}, \gamma_{j1}, \gamma_{j2}, ..., \gamma_{jk})$，于是可以写出完全数据的似然函数：

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

其中 $n_k = \sum_{j=1}^{N} \gamma_{jk}$，$\sum_{k=1}^{K} n_k = N$，这里的 n_k 指的是由第 k 个分模型产生的观测数据数量。那么完全数据的对数似然函数为：

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

（2）EM 算法的 E 步：确定 Q 函数

$$\begin{aligned} Q(\theta, \theta^{(i)}) &= E[logP(y, \gamma|\theta)|y, \theta^{(i)}] \\ &= E\{\sum_{k=1}^{K} \{n_k log\alpha_k + \sum_{j=1}^{N} \gamma_{jk} [log(\frac{1}{\sqrt{2\pi}}) - log\sigma_k - \frac{1}{2\sigma_k^2}(y^{(j)} - \mu_k)^2]\}\} \\ &= E\{\sum_{k=1}^{K} \{\sum_{j=1}^{N} \gamma_{jk} log\alpha_k + \sum_{j=1}^{N} \gamma_{jk} [log(\frac{1}{\sqrt{2\pi}}) - log\sigma_k - \frac{1}{2\sigma_k^2}(y^{(j)} - \mu_k)^2]\}\} \\ &= \sum_{k=1}^{K} \{\sum_{j=1}^{N} (E\gamma_{jk}) log\alpha_k + \sum_{j=1}^{N} (E\gamma_{jk}) [log(\frac{1}{\sqrt{2\pi}}) - log\sigma_k - \frac{1}{2\sigma_k^2}(y^{(j)} - \mu_k)^2]\} \end{aligned} \tag{3}$$

这里的 $E\gamma_{jk}$ 也即 $E(\gamma_{jk}|y, \theta)$ 记为 $\hat{\gamma}_{jk}$ 进一步可以写成：

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

由上式的结果可知 $\hat{\gamma}_{jk}$ 等于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 $\hat{\gamma}_{jk} = E\gamma_{jk}$，$n_k = \sum_{j=1}^{N} E\gamma_{jk}$ 代入式 (3) 得：

$$\begin{aligned} Q(\theta, \theta^{(i)}) &= \sum_{k=1}^{K}\{\sum_{j=1}^{N}(E\gamma_{jk})log\alpha_k + \sum_{j=1}^{N}(E\gamma_{jk})[log(\frac{1}{\sqrt{2\pi}}) - log\sigma_k - \frac{1}{2\sigma_k^2}(y^{(j)} - \mu_k)^2]\} \\ &= \sum_{k=1}^{K}\{n_k log\alpha_k + \sum_{j=1}^{N}\hat{\gamma}_{jk}[log(\frac{1}{\sqrt{2\pi}}) - log\sigma_k - \frac{1}{2\sigma_k^2}(y^{(j)} - \mu_k)^2]\} \end{aligned} \tag{4}$$

（3）EM 算法的 M 步

迭代的 M 步是求函数 $Q(\theta, \theta^{(i)})$ 对 θ 的极大值，即求新一轮迭代的模型参数：

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

$\theta^{(i+1)}$ 的参数用 $\hat{\mu}_k, \hat{\sigma}_k^2, \hat{\alpha}_k$ 表示。求 $\hat{\mu}_k, \hat{\sigma}_k^2$ 只需将式 (4) 分别对 μ_k, σ_k^2 求偏导令其为 0 即可得到，求 $\hat{\alpha}_k$ 是在 $\sum_{k=1}^{K}\alpha_k = 1$ 条件下求偏导数并令其为 0 得到，结果如下：

$$\hat{\mu}_k = \frac{\sum_{j=1}^{N}\hat{\gamma}_{jk}y^{(j)}}{\sum_{j=1}^{N}\hat{\gamma}_{jk}}$$

$$\hat{\sigma}_k^2 = \frac{\sum_{j=1}^{N}\hat{\gamma}_{jk}(y^{(j)} - \mu_k)^2}{\sum_{j=1}^{N}\hat{\gamma}_{jk}}$$

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

重复以上计算，直到对数似然函数值不再有明显的变化为止。

4 参考文献

1. 李航博士的统计学习方法
2. 统计学习方法勘误表
http://www.hangli-hl.com/uploads/3/4/4/6/34465961/errata.pdf