EM Algorithm With Applications

曾舸舵 2020302121048

2022-05-16

摘要

本文介绍了 EM 算法的理论模型及其正确性证明，并构造一个实例以应用 EM 算法。

引言

EM 算法最初用于缺失数据模型参数估计，现在已经用在许多优化问题中。

方法介绍

设模型中包含观测数据 X_{obs} 和缺失数据 X_{mis} 两个随机成分，有联合密度函数或概率函数 $f(X_{obs}, X_{mis}|\theta)$, θ 为未知参数。称 $f(X_{obs}, X_{mis}|\theta)$ 为完全数据的密度，一般具有简单的形式。实际上我们只有 X_{obs} 的观测数据 $X_{obs} = x_{obs}$，不能观测得到 X_{mis}，这一部分可能是缺失观测数据，也可能是潜在影响因素。所以实际的似然函数为

$$L(\theta) = (x_{obs}|\theta) = \int f(x_{obs}, x_{mis}|\theta) dx_{mis}$$

这个似然函数通常比完全数据的似然函数复杂得多，所以很难直接从求最大似然估计。

理论模型

EM 算法的想法是，已经有了参数的近似估计值 $\theta^{(t)}$ 后，假设 (X_{obs}, X_{mis}) 近似服从完全密度 $f(X_{obs}, X_{mis}|\theta^{(t)})$，这里 $X_{obs} = x_{obs}$ 已知，所以认为 X_{mis} 近似服从由 $f(x_{obs}, x_{mis}|\theta^{(t)})$ 导出的条件分布

$$f(x_{mis}|x_{obs}, \theta^{(t)}) = \frac{f(x_{obs}, x_{mis}|\theta^{(t)})}{f(x_{obs}|\theta^{(t)})}$$

其中 $f(x_{obs}|\theta^{(t)})$ 是由 $f(x_{obs}, x_{mis}|\theta^{(t)})$ 决定的边缘密度。据此近似条件分布，在完全数据对数似然函数 $\log f(X_{obs}, X_{mis}|\theta)$ 中，把 $X_{obs} = x_{obs}$ 看成已知，关于未知部分 X_{mis} 按密度 $f(x_{mis}|x_{obs}, \theta^{(t)})$ 求期望，得到 $\theta^{(t)}$ 的函数 $Q_t(\theta)$，再求 $Q_t(\theta)$ 的最大值点作为 $\theta^{(t+1)}$。

EM 算法每次迭代有如下的 E 步（期望步）和 M 步（最大化步）：

- E 步: 计算完全数据对数似然函数的期望 $Q_t(\theta) = E\ \log f(x_{obs}, X_{mis}|\theta)$，其中期望针对随机变量 X_{mis}，求期望时假定 X_{mis} 服从条件密度 $f(x_{mis}|x_{obs}, \theta^{(t)})$ 决定的分布。
- M 步: 求 $Q_t(\theta)$ 的最大值点，记为 $\theta^{(t+1)}$，迭代进入下一步。

正确性证明

引入信息熵的概念 $H(X) = EI(X) = E[-\ln(P(X))]$。其中，$P$ 为 X 的概率质量函数，E 为期望函数，而 $I(X)$ 是 X 的信息量。

EM 算法有助于改进 $Q(\theta \mid \theta^{(t)})$ 而不是直接改进 $\log p(\mathbf{X} \mid \theta)$. 这里证明，对前者的改进意味着对后者的改进。

对于任何未知数据 $\mathbf{Z}$ 以及非零概率 $p(\mathbf{Z} \mid \mathbf{X}, \theta)$，我们可以得到

$$\log p(\mathbf{X} \mid \theta) = \log p(\mathbf{X}, \mathbf{Z} \mid \theta) - \log p(\mathbf{Z} \mid \mathbf{X}, \theta)$$

我们对未知数据 $\mathbf{Z}$ 在当前参数估计下 $\theta^{(t)}$ 求期望，即两边乘以 $p(\mathbf{Z} \mid \mathbf{X}, \theta^{(t)})$ 并在 $\mathbf{Z}$ 上求和（积分）. 左边是一个常数的期望，所以我们得到:

$$\begin{aligned}\log p(\mathbf{X} \mid \theta) &= \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \theta^{(t)}) \log p(\mathbf{X}, \mathbf{Z} \mid \theta) - \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \theta^{(t)}) \log p(\mathbf{Z} \mid \mathbf{X}, \theta) \\ &= Q(\theta \mid \theta^{(t)}) + H(\theta \mid \theta^{(t)})\end{aligned}$$

最后一个等式适用于每个 θ。

移项，

$$\log p(\mathbf{X} \mid \theta) - \log p(\mathbf{X} \mid \theta^{(t)}) = Q(\theta \mid \theta^{(t)}) - Q(\theta^{(t)} \mid \theta^{(t)}) + H(\theta \mid \theta^{(t)}) - H(\theta^{(t)} \mid \theta^{(t)}),$$

由吉布斯不等式 $H(\theta \mid \theta^{(t)}) \geq H(\theta^{(t)} \mid \theta^{(t)})$，

$$\log p(\mathbf{X} \mid \theta) - \log p(\mathbf{X} \mid \theta^{(t)}) \geq Q(\theta \mid \theta^{(t)}) - Q(\theta^{(t)} \mid \theta^{(t)}).$$

所以选择 θ 改进 $Q(\theta \mid \theta^{(t)})$ 会带来对 $\log p(\mathbf{X} \mid \theta)$ 同样的改进。

模拟研究

设某种设备的寿命总体 $Y \sim Exp(1/\theta)$，对 n 个这样的设备进行寿命试验，其中 n_1 个观测到失效时间 $t_1, \cdots, t_{n_1}$，另外的 $n_2 = n - n_1$ 个没有观测到失效，仅知道失效时间分别超过 $c_{n_1+1}, \cdots, c_n$。求参数 θ 的估计.

设完全数据为 $(t_1, \cdots, t_n)$ 都是失效时间，完全数据的似然函数为

$$L_c(\theta) = \theta^{-n} \exp(-\frac{1}{\theta}\sum_1^n t_i)$$

取适当初值 $\theta^{(0)}$. 在迭代中设已有 $\theta^{(t)}$, 求 $\ln L_c(\theta)$ 的条件期望。

$$Q_t(\theta) = E(\ln L_c(\theta)|x) = -n\ln\theta - \frac{1}{\theta}\sum_1^{n_1} t_i - \frac{1}{\theta}\sum_{n_1+1}^{n}(c_j + \theta^{(t)})$$

令 $\frac{d}{dt}Q_t(\theta) = 0$, 得

$$\theta^{t+1} = \frac{1}{n}(\sum x_i + n_2\theta^{(t)})$$

结论

在存在缺失数据时，EM 算法可以有效地求解参数。

参考文献

[1] Expectation–maximization algorithm, Wikipedia

[2] 高惠璇. 1995. 统计计算. 北京大学出版社.

[3] Section 8.5 The EM Algorithm, *The Elements of Statistical Learning*, 2016.

附录

```r
set.seed(101)
true.theta <- 2.0
n <- 25
t.comp <- rexp(n, 1/true.theta)
t.cens <- rexp(n, 1/true.theta)
sele <- t.comp <= t.cens
x.comp <- t.comp[sele]
x.cens <- t.cens[!sele]
x <- c(x.comp, x.cens) # 观测数据
n1 <- length(x.comp)
n2 <- n - n1

##EM 算法
eps <- 1E-6
max.iter <- 500
theta <- mean(x)
k <- 0
repeat{
  k <- k+1
  theta0 <- theta # 上一个 $\theta$ 值
  theta <- 1/n*(sum(x) + n2*theta0)
  if(abs(theta - theta0) < eps || k >= max.iter) break
```

```
}
print(theta)
```

```
## [1] 2.555049
```