Statistical Computing
Chap. 2.1: Cases for EM

LIU, Ran

Department of Statistics,
Beijing Normal University (Zhuhai Campus)

March 10, 2024

北京師範大學
BEIJING NORMAL UNIVERSITY

目录

两枚硬币出现正面概率的 EM 算法

假设有 A, B 两枚硬币，其中正面朝上的概率分别为 p_A, p_B，这两个参数是感兴趣的待估参数。

设计 6 组试验，每次实验投掷 5 次硬币，第 i 组试验结果为 $X_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})(i = 1, \cdots, 6)$，$x_{ij} = 1$ 表示硬币出现正面，$x_{ij} = 0$ 表示硬币出现反面。

如果知道每一组实验结果 X_i 是 A 硬币投的结果还是 B 硬币投的结果，也就是观测到 Z_i，$Z_i = 1$ 表示 A 硬币投的结果，$Z_i = 0$ 表示 B 硬币投的结果。

两枚硬币出现正面概率的 EM 算法

基于完整数据 $Y_i = (X_i, Z_i), i = 1, 2, \cdots, 6$ 的对数似然函数为

$$\begin{aligned} l(Y; p_A, p_B) &= \sum_{k=1}^{6} \{Z_k (n_k \log p_A + (5 - n_k) \log(1 - p_A)) \\ &\quad + (1 - Z_k)(n_k \log p_B + (5 - n_k) \log(1 - p_B))\}. \end{aligned}$$

这里 $n_k = \sum_{j=1}^{5} x_{kj}$.

如果 $Z_i(i = 1, 2, \cdots, 6)$ 可观测，不难得到

$$\hat{p}_A = \frac{\sum_{k=1}^{6} n_k Z_k}{5 \sum_{k=1}^{6} Z_k}, \quad \hat{p}_B = \frac{\sum_{k=1}^{6} n_k (1 - Z_k)}{5 \sum_{k=1}^{6} (1 - Z_k)}.$$

也就是参数 p_A, p_B 的估计，只需分别统计 A, B 硬币投的结果出现正面的次数，然后除以分别投的总次数。

由于 $Z = (Z_1, \cdots, Z_6)$ 没有观测到，基于观测数据 $X = (X_1, \cdots, X_6)$ 和 $p_A^{(i-1)}, p_B^{(i-1)}$，条件期望为

$$
\begin{aligned}
& Q(p_A, p_B; p_A^{(i-1)}, p_B^{(i-1)}) = E(l(Y; p_A, p_B)|X; p_A^{(i-1)}, p_B^{(i-1)}) \\
= & \sum_{k=1}^{6} \{E(Z_k|X; p_A^{(i-1)}, p_B^{(i-1)})(n_k \log p_A + (5 - n_k) \log(1 - p_A)) \\
& + (1 - E(Z_k|X; p_A^{(i-1)}, p_B^{(i-1)})(n_k \log p_B + (5 - n_k) \log(1 - p_B))\}.
\end{aligned}
$$

关于 p_A, p_B, 极大化 $Q(p_A, p_B; p_A^{(i-1)}, p_B^{(i-1)})$, 可得

$$
\begin{aligned}
p_A^{(i)} &= \frac{\sum_{k=1}^{6} n_k E(Z_k|X; p_A^{(i-1)}, p_B^{(i-1)})}{5\sum_{k=1}^{5} E(Z_k|X; p_A^{(i-1)}, p_B^{(i-1)})} \\
p_B^{(i)} &= \frac{\sum_{k=1}^{6} n_k (1 - E(Z_k|X; p_A^{(i-1)}, p_B^{(i-1)}))}{5\sum_{k=1}^{5} (1 - E(Z_k|X; p_A^{(i-1)}, p_B^{(i-1)}))}.
\end{aligned}
$$

对 $E(Z_k|X;p_A^{(i-1)},p_B^{(i-1)})$, 不难推导可得

$$E(Z_k|X;p_A^{(i-1)},p_B^{(i-1)})=\frac{(p_A^{(i-1)})^{n_k}(1-p_A^{(i-1)})^{5-n_k}}{(p_A^{(i-1)})^{n_k}(1-p_A^{(i-1)})^{5-n_k}+(p_B^{(i-1)})^{n_k}(1-p_B^{(i-1)})^{5-n_k}}$$

若是将每组硬币正面出现的次数当作 $X=\{n_1,...,n_6\}$:

$$E(Z_k|X;p_A^{(i-1)},p_B^{(i-1)})=\{\frac{C_5^{n_k}}{n_k!(5-n_k)!}(p_A^{(i-1)})^{n_k}(1-p_A^{(i-1)})^{5-n_k}\}$$
$$/\{\frac{C_5^{n_k}}{n_k!(5-n_k)!}(p_A^{(i-1)})^{n_k}(1-p_A^{(i-1)})^{5-n_k}$$
$$+\frac{C_5^{n_k}}{n_k!(5-n_k)!}(p_B^{(i-1)})^{n_k}(1-p_B^{(i-1)})^{5-n_k}\}.$$

估算是一样的。

假设观测到的数据为 $X_1 = (1, 1, 1, 0, 1)$, $X_2 = (0, 1, 0, 1, 1)$, $X_3 = (1, 0, 1, 1, 0)$, $X_4 = (0, 0, 1, 1, 1)$, $X_5 = (1, 0, 1, 0, 1)$ and $X_6 = (1, 1, 0, 0, 0)$, 用 EM 算法估计 p_A 和 p_B.

```
E <- function(nk, a, b){
 A <- (choose(5, nk)/(factorial(nk)*factorial(5 - nk)))
 * a^nk * (1 - a)^(5 - nk)
 B <- (choose(5, nk)/(factorial(nk)*factorial(5 - nk)))
 * b^nk * (1 - b)^(5 - nk)
 E <- A/(A + B)
 return(E)
}
```

```
n <- c(4,3,3,3,3,2)
max.iter <- 100     #初值与最大循环数
pa <- c();pa[1] <- 0.6
pb <- c();pb[1] <- 0.4
for (i in 1:max.iter){#迭代
 EZ <- E(n, pa[i], pb[i])
 pa[i+1] <- sum(n*EZ)/(5*sum(EZ))
 pb[i+1] <- sum(n*(1-EZ))/(5*sum(1-EZ))
 if(abs(pa[i+1] - pa[i]) < 1e-8 &
 abs(pb[i+1] - pb[i]) < 1e-8) break
}
```

```
import math
def E(nk, a, b):
    A = (a ** nk) * ((1 - a) ** (5 - nk))
    B = (b ** nk) * ((1 - b) ** (5 - nk))
    E = A / (A + B)
    return E

n = [4, 3, 3, 3, 3, 2]
max_iter = 100
pa, pb = [0.6], [0.4]
for i in range(max_iter):
    EZ = E(n, pa[i], pb[i])
    pa_next = sum([nk * EZ_i for nk, EZ_i in zip(n, EZ)]) / (5 * sum(EZ))
    pb_next = sum([(nk * (1 - EZ_i)) for nk, EZ_i in zip(n, EZ)]) / (5 * sum(1 - EZ))

    if abs(pa_next - pa[i]) < 1e-8 and abs(pb_next - pb[i]) < 1e-8:
        break
    pa.append(pa_next)
    pb.append(pb_next)
```

目录

多项分布参数的 EM 算法

假设 $x = (x_1, \cdots, x_m)$ 服从多项分布，也就是

$$p(x|p_1, \cdots, p_m) = \frac{n!}{x_1! \cdots x_m!} p_1^{x_1} \cdots p_m^{x_m}.$$

如果 $m = 4$，$(p_1, \cdots, p_4) = (0.5 - \theta/2, \theta/4, \theta/4, 0.5)$. 观测数据的数据为 $x = (x_1, x_2, x_{34})^\tau$，这里 $x_{34} = x_3 + x_4$，感兴趣参数 θ 的估计。

潜变量 x_3, x_4 没有观测到，基于完整数据 $y = (x, z)$，即 $(x_1, x_2, x_3, x_4)^\tau$，对数似然函数为

$$\begin{aligned} l(y;\theta) &= \log L(y;\theta) \\ &= x_1 \log(\frac{1}{2} - \frac{\theta}{2}) + x_2 \log(\frac{\theta}{4}) + x_3 \log(\frac{\theta}{4}) + const. \end{aligned}$$

基于观测到的数据和上次迭代估计 $\theta^{(i-1)}$，计算对数似然的条件期望，

$$\begin{aligned} & E(l(y;\theta)|x;\theta^{(i-1)}) = E(\log L(y;\theta)|x,\theta^{(i-1)}) \\ = \quad & x_1 \log(\frac{1}{2} - \frac{\theta}{2}) + x_2 \log(\frac{\theta}{4}) + E(x_3|x,\theta^{(i-1)}) \log(\frac{\theta}{4}) + const. \end{aligned}$$

下面计算 $E(x_3|x,\theta^{(i-1)})$，由于

$$E(x_3|x;\theta^{(i-1)}) = \sum_{z=0}^{x_{34}} z\, p(z|x;\theta^{(i-1)}) = \sum_{z=0}^{x_{34}} z \frac{p(x,z;\theta^{(i-1)})}{p(x;\theta^{(i-1)})}, \tag{1}$$

这里利用了多项式分布中，其中两项合并也是多项式分布，我们有

$$
\begin{aligned}
& p(x, z; \theta^{(i-1)})) = p(x_1, x_2, z, x_{34} - z; \theta^{(i-1)}) \\
= & \frac{n!}{x_1! x_2! z! (x_{34} - z)!} (\frac{1}{2} - \frac{\theta^{(i-1)}}{2})^{x_1} (\frac{\theta^{(i-1)}}{4})^{x_2} (\frac{\theta^{(i-1)}}{4})^{z} (\frac{1}{2})^{x_{34}-z}; \\
& p(x; \theta^{(i-1)}) = p(x_1, x_2, x_{34}; \theta^{(i-1)}) \\
= & \frac{n!}{x_1! x_2! (x_{34})!} (\frac{1}{2} - \frac{\theta^{(i-1)}}{2})^{x_1} (\frac{\theta^{(i-1)}}{4})^{x_2} (\frac{\theta^{(i-1)}}{4} + \frac{1}{2})^{x_{34}}.
\end{aligned}
$$

基于 $p(x, z; \theta^{(i-1)}))$ 和 $p(x; \theta^{(i-1)})$ 的表达式，可得

$$
\frac{p(x, z; \theta^{(i-1)})}{p(x; \theta^{(i-1)})} = \frac{(x_{34})!}{z!(x_{34} - z)!} (\frac{\theta^{(i-1)}}{4})^{z} (\frac{1}{2})^{x_{34}-z} / (\frac{\theta^{(i-1)}}{4} + \frac{1}{2})^{x_{34}},
$$

以及

$$
\begin{aligned}
E(x_3|x;\theta^{(i-1)}) &= \sum_{z=0}^{x_{34}} z\frac{(x_{34})!}{z!(x_{34}-z)!}(\frac{\theta^{(i-1)}}{4})^z(\frac{1}{2})^{x_{34}-z} \\
&\quad /(\frac{\theta^{(i-1)}}{4}+\frac{1}{2})^{x_{34}} \\
&= x_{34}\frac{\theta^{(i-1)}}{4}/(\frac{\theta^{(i-1)}}{4}+\frac{1}{2}) = x_{34}\frac{\theta^{(i-1)}}{\theta^{(i-1)}+2}.
\end{aligned}
$$

值得一提的是，除了上面一步一步推导，也可以直接这样看待：x_3 在 x 已知的条件下，肯定为二项式分布，此时属于第三类别和第四类别的概率成比例于联合概率中二者的概率，也就是 $\theta/4$ 和 0.5，这个比例除以他们之和，即为条件概率。

其中关于式子：

$$\sum_{z=0}^{x_{34}} z \frac{(x_{34})!}{z!(x_{34}-z)!}(\frac{\theta^{(i-1)}}{4})^z(\frac{1}{2})^{x_{34}-z}/(\frac{\theta^{(i-1)}}{4}+\frac{1}{2})^{x_{34}} = x_{34}\frac{\theta^{(i-1)}}{\theta^{(i-1)}+2}$$

是基于二项分布的期望为 np 来推出的，我们先有对一个二项分布：

$$P(X=k) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

$$\sum_{k=0}^{n} k\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} = np$$

$$(1-p)^n\sum_{k=0}^{n}\frac{n!}{k!(n-k)!}k(\frac{p}{1-p})^k = np$$

把 k 的合在一起

$$\sum_{k=0}^{n} k\frac{n!}{k!(n-k)!}(\frac{p}{1-p})^k = \frac{np}{(1-p)^n}$$

再看看原式

$$\sum_{z=0}^{x_{34}} z \frac{(x_{34})!}{z!(x_{34}-z)!}(\frac{\theta^{(i-1)}}{4})^z(\frac{1}{2})^{x_{34}-z}/(\frac{\theta^{(i-1)}}{4}+\frac{1}{2})^{x_{34}}$$
$$=\sum_{z=0}^{x_{34}} z \frac{(x_{34})!}{z!(x_{34}-z)!}(\frac{\theta^{(i-1)}}{2})^z(\frac{1}{2})^{x_{34}}/(\frac{\theta^{(i-1)}}{4}+\frac{1}{2})^{x_{34}}$$

只关注与 z 相关的部分，看作是一个以 z 为参数，$n = x_{34}$ 的二项式分布，我们令参数对应相等：

$$\frac{\theta^{(i-1)}}{2} = \frac{p}{1-p},$$

则我们有

$$p = \frac{\theta^{(i-1)}}{2+\theta^{(i-1)}}, \quad 1-p = \frac{\theta^{(i-1)}}{2+\theta^{(i-1)}}.$$

所以

$$\frac{np}{(1-p)^n} = \frac{(2+\theta^{(i-1)})^{n-1} n\theta^{(i-1)}}{2^n} = \frac{(2+\theta^{(i-1)})^{x_{34}-1} n\theta^{(i-1)}}{2^{x_{34}}}$$

带回到原式，我们有

$$\begin{aligned} & E(x_3|x;\theta^{(i-1)}) \\ &= \sum_{z=0}^{x_{34}} z \frac{(x_{34})!}{z!(x_{34}-z)!} (\frac{\theta^{(i-1)}}{2})^z (\frac{1}{2})^{x_{34}} / (\frac{\theta^{(i-1)}}{4} + \frac{1}{2})^{x_{34}} \\ &= x_{34} \frac{\theta^{(i-1)}}{\theta^{(i-1)}+2} \end{aligned}$$

式 (1) 关于对数似然的条件期望进一步写为

$$
\begin{aligned}
& Q(\theta, \theta^{(i-1)}) = E(l(y;\theta)|x;\theta^{(i-1)}) \\
= \quad & x_1 \log(\frac{1}{2} - \frac{\theta}{2}) + x_2 \log(\frac{\theta}{4}) + (x_{34}) \frac{\theta^{(i-1)}}{\theta^{(i-1)} + 2} \log(\frac{\theta}{4}) + const.
\end{aligned}
$$

关于 θ 极大化 $Q(\theta, \theta^{(i-1)})$ 得到下次迭代 $\theta^{(i)}$，对 $Q(\theta, \theta^{(i-1)})$ 关于 θ 求导可得

$$\begin{aligned} Q'(\theta, \theta^{(i-1)}) &= -\frac{x_1}{1-\theta} + \frac{x_2}{\theta} + \frac{E(x_3|x;\theta^{(i-1)})}{\theta} \\ &= -\frac{x_1}{1-\theta} + \frac{x_2}{\theta} + \frac{x_{34}}{\theta} \times \frac{\theta^{(i-1)}}{\theta^{(i-1)}+2} = 0. \end{aligned}$$

因此

$$\theta^{(i)} = \frac{x_2 + E(x_3|x;\theta^{(i-1)})}{x_1 + x_2 + E(x_3|x;\theta^{(i-1)})}.$$

通过不断迭代，最终得到 θ 的估计.

如果数据完全观测到，也就是 x_3 被观测，则参数 θ 的极大似然估计为

$$\hat{\theta} = \frac{x_2 + x_3}{x_1 + x_2 + x_3}.$$

对比完全数据的极大似然估计 $\hat{\theta}$ 和 EM 算法第 i 次迭代估计 $\theta^{(i)}$，EM 算法本质上相当于把基于完整数据得到的极大似然估计量中没有观测到的数据，采用观测到数据和上次迭代估计 $\theta^{(i-1)}$ 预测。

如果观测到的数据为 $x_{obs} = (x_1, x_2, x_{34})^\tau = (38, 34, 125)^\tau$，求参数的 EM 估计。

```
#观测数据
  x1 <- 38
  x2 <- 34
  x34 <- 125
  #初值与最大循环数
  max.iter <- 100
  htheta <- rep(0, max.iter); htheta[1] <- 0
  #迭代
  for (i in 1:max.iter){
  hx3 <- 0.25*x34*htheta[i]/(0.25*htheta[i] + 0.5)
  htheta[i+1] <- (x2 + hx3)/(x1 + x2 + hx3)
  if (abs(htheta[i+1] - htheta[i]) < 1e-8)
    break
  }
```

```
# 观测数据
x1 = 38
x2 = 34
x34 = 125

# 初值与最大循环数
max_iter = 100
htheta = [0] * max_iter
htheta[0] = 0

# 迭代
for i in range(max_iter):
    hx3 = 0.25 * x34 * htheta[i] / (0.25 * htheta[i] + 0.5)
    htheta[i+1] = (x2 + hx3) / (x1 + x2 + hx3)

    if abs(htheta[i+1] - htheta[i]) < 1e-8:
        break
```

目录

多项式分布的特点

- 多项式分布其中两个类别合在一起，新分布也是多项式分布。
- 多项式分布的其中两个类别，他们加起来等于一个常数的话，那么它在这个他们两个在这个常数下的条件概率分布成比例于他们在原有的多项式分布中各自的概率，然后他们在这个二项式分布真实的概率为它们在联合概率中的这个概率除以它们的求和。

多项式分布其中两个类别合在一起，新分布也是多项式分布。

假设 $x = (x_1, \cdots, x_m)$ 服从多项分布，也就是

$$p(x|p_1, \cdots, p_m) = \frac{n!}{x_1! \cdots x_m!} p_1^{x_1} \cdots p_m^{x_m}.$$

那么 $x^* = (x_1 + x_2, x_3, ..., x_m)$ 也服从多项式分布：

$$p(x^*|p_1, \cdots, p_m) = \frac{n!}{(x_1 + x_2)! \cdots x_m!} (p_1 + p_2)^{x_1 + x_2} \cdots p_m^{x_m}.$$

证明：

$$\begin{aligned} p(x_1 + x_2 = y, x_3, ..., x_m| \sim) &= \sum_{x_2=0}^{y} p(x_1 = y - x_2, x_2 = x_2, ...| \sim) \\ &= \sum_{x_2=0}^{y} \frac{n!}{(y - x_2)! x_2! ... x_m!} p_1^{y-x_2} p_2^{x_2} \cdots \\ &= \frac{n!}{y! x_3! \cdots x_m!} (p_1 + p_2)^y p_3^{x_3} \cdots \end{aligned}$$

多项式分布的其中两个类别，他们加起来等于一个常数的话，那么它在这个他们两个在这个常数下的条件概率分布成比例于他们在原有的多项式分布中各自的概率，然后他们在这个二项式分布真实的概率为它们在联合概率中的这个概率除以它们的求和。

假设 $x = (x_1, \cdots, x_m)$ 服从多项分布，也就是

$$p(x|p_1, \cdots, p_m) = \frac{n!}{x_1! \cdots x_m!} p_1^{x_1} \cdots p_m^{x_m}.$$

那么 $x_1, x_2|x_1 + x_2 = y$ 也服从多项式分布：

$$p(x_1, x_2|x_1 + x_2 = y, p) = C_y^{x_1} (\frac{p_1}{p_1 + p_2})^{x_1} (\frac{p_2}{p_1 + p_2})^{y - x_1}, \quad x_1 = 0, 1, ..., y.$$

证明：

$$\begin{aligned} p(x_1, x_2|x_1 + x_2 = y, p) &= \sum_{x_3, \cdots, x_m} p(x_1, x_2, ..., x_m) \\ &\propto \frac{1}{x_1! x_2!} p_1^{x_1} p_2^{x_2} \end{aligned}$$

若没有 $x_1 + x_2 = y$，那么 x_1, x_2 不服从多项式分布。简单来说，他俩的和可以为 0 到 n 的所有值。真要算，可以这么算：

$$p(x_1, x_2) = \sum_{y=0}^{n} p(x_1, x_2 | x_1 + x_2 = y) p(x_1 + x_2 = y)$$

更多有趣的性质，可查阅：
https://online.stat.psu.edu/stat504/book/export/html/667

目录

正态分布参数 EM 估计

对来自正态总体 $N(\mu, \sigma^2)$ 的完整数据 $y = (x_1, x_2, \cdots, x_n)^\tau$，假设观测到数据 $x = (x_1, \cdots, x_m)$，隐含数据为 $z = (x_{m+1}, \cdots, x_n)$. 则基于完整数据 y 的的极大似然函数为

$$L(y; \mu, \sigma^2) = (\frac{1}{2\pi\sigma^2})^{n/2} \exp\{-\sum_{j=1}^{n} \frac{(x_j - \mu)^2}{2\sigma^2}\}.$$

相应的对数极大似然函数为

$$\begin{aligned} & l(y; \mu, \sigma^2) = \log L(x; \mu, \sigma^2) \\ = \quad & -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{j=1}^{n} x_j^2 + \frac{\mu}{\sigma^2}\sum_{j=1}^{n} x_j - \frac{n\mu^2}{2\sigma^2}. \end{aligned}$$

这个例子比较简单，隐含数据和观测数据在参数已知下独立，互不干扰。

基于观测数据的条件对数似然为

$$
\begin{aligned}
& Q(\mu, \sigma^2; \mu^{(i-1)}, \sigma^{2(i-1)}) = E(l(y; \mu, \sigma^2)|x; \mu^{(i-1)}, \sigma^{2(i-1)}) \\
= \quad & -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{j=1}^{m} x_j^2 + \frac{\mu}{\sigma^2}\sum_{j=1}^{m} x_j - \frac{n\mu^2}{2\sigma^2} \\
& -\frac{1}{2\sigma^2}E\{\sum_{j=m+1}^{n} x_j^2|x; \mu^{(i-1)}, \sigma^{2(i-1)}\} \\
& +\frac{\mu}{\sigma^2}E\{\sum_{j=m+1}^{n} x_j|x; \mu^{(i-1)}, \sigma^{2(i-1)}\}.
\end{aligned}
$$

对 $Q(\mu,\sigma^2;\mu^{(i-1)},\sigma^{2(i-1)})$ 关于 μ,σ^2 求导可得

$$\frac{\partial Q(\mu,\sigma^2;\mu^{(i-1)},\sigma^{2(i-1)})}{\partial \mu}$$

$$=\frac{1}{\sigma^2}\sum_{j=1}^{n}x_j+\frac{1}{\sigma^2}E\{\sum_{j=m+1}^{n}x_j|x;\mu^{(i-1)},\sigma^{2(i-1)}\}-\frac{n\mu}{\sigma^2}=0$$

$$\frac{\partial Q(\mu,\sigma^2;\mu^{(i-1)},\sigma^{2(i-1)})}{\partial \sigma^2}$$

$$=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{j=1}^{n}x_j^2-\frac{\mu}{\sigma^4}\sum_{j=1}^{n}x_j+\frac{n\mu^2}{2\sigma^4}$$

$$+\frac{1}{2\sigma^4}E\{\sum_{j=m+1}^{n}x_j^2|x;\mu^{(i-1)},\sigma^{2(i-1)}\}-\frac{\mu}{\sigma^4}E\{\sum_{j=m+1}^{n}x_j|x;\mu^{(i-1)},\sigma^{2(i-1)}\}$$

$$=0.$$

所以第 i 次迭代估计 $\theta^{(i)}$ 为

$$\begin{aligned}\mu^{(i)} &= \frac{1}{n}\{\sum_{j=1}^{m} x_j + E(\sum_{j=m+1}^{n} x_j | x; \mu^{(i-1)}, \sigma^{2(i-1)})\} \\ \sigma^{2(i)} &= \frac{1}{n}\{\sum_{j=1}^{m} x_j^2 + E(\sum_{j=m+1}^{n} x_j^2 | x; \mu^{(i-1)}, \sigma^{2(i-1)})\} - (\mu^{(i)})^2.\end{aligned}$$

给定第 $i-1$ 步的 $(\mu^{(i-1)}, \sigma^{2(i-1)})$，且未观测到的隐含数据和观测到的数据独立，则

$$E(\sum_{j=m+1}^{n} x_j | x; \mu^{(i-1)}, \sigma^{2(i-1)}) = (n-m)\hat{\mu}^{(i-1)},$$

$$E(\sum_{j=m+1}^{n} x_j^2 | x; \mu^{(i-1)}, \sigma^{2(i-1)}) = (n-m)(\hat{\mu}^{(i-1)2} + \sigma^{2(i-1)}).$$

若未观测的隐含数据和观测数据相关，则根据变量相关性计算 $E(\sum_{j=m+1}^{n} x_j | x; \mu^{(i-1)}, \sigma^{2(i-1)})$ 和 $E(\sum_{j=m+1}^{n} x_j^2 | x; \mu^{(i-1)}, \sigma^{2(i-1)})$.

设 $y = (x_1, x_2, \cdots, x_{1000})^\tau$ 来自于正态总体 $N(2, 4)$，其中观测到的数据为 $x = (x_1, \cdots, x_{600})$，未观测到的数据为 $z = (x_{601}, \cdots, x_{1000})$，通过 EM 算法给出参数 μ, σ^2 的估计。

```
set.seed(1) #生成数据
n <- 1000;m <- 600
x <- rnorm(n, mean = 2, sd = 2)
sx <- sum(x[1:m]); sx2 <- sum(x[1:m]^2)
max.iter <- 100  #最大循环数
hmu <- rep(0, max.iter)
hsigma2 <- rep(0, max.iter)
hmu[1] <- 0;hsigma2[1] <- 1  #初值
for (i in 1:max.iter){
s1 <- sx + (n - m)*hmu[i]
s2 <- sx2 + (n - m)*(hmu[i]^2 + hsigma2[i])
hmu[i+1] <- s1/n
hsigma2[i+1] <- s2/n - hmu[i+1]^2
if(abs(hmu[i+1] - hmu[i]) < 1e-8
& abs(hsigma2[i+1] - hsigma2[i]) < 1e-8) break
}
```

```
import numpy as np
np.random.seed(1)  # 生成数据
n = 1000
m = 600
x = np.random.normal(loc=2, scale=2, size=n)
sx = np.sum(x[:m])
sx2 = np.sum(x[:m]**2)
max_iter = 100  # 最大循环数
hmu = np.zeros(max_iter)
hsigma2 = np.zeros(max_iter)
hmu[0] = 0
hsigma2[0] = 1  # 初值

for i in range(max_iter):
    s1 = sx + (n - m) * hmu[i]
    s2 = sx2 + (n - m) * (hmu[i]**2 + hsigma2[i])
    hmu[i+1] = s1 / n
    hsigma2[i+1] = s2 / n - hmu[i+1]**2
    if abs(hmu[i+1] - hmu[i]) < 1e-8 and abs(hsigma2[i+1] -
    hsigma2[i]) < 1e-8:
        break
```

目录

二项泊松混合模型的 EM 估计

我们观测 n 个人的 B 站用户等级，等级为 $i(i=0,1,\cdots,6)$ 的人数是 n_i，则 $n=\sum_{i=0}^{6} n_i$. 观测数据如下表所示

用户等级	0	1	2	3	4	5	6
人数	n_0	n_1	n_2	n_3	n_4	n_5	n_6

假定**已注册的人的 B 站等级服从参数为 λ 的泊松分布**，感兴趣的参数是参数 λ 和没有注册的概率 ξ。这里例子中，设定了刚注册的和没有注册的用户等级都为 0。

零膨胀模型（Zero-Inflated Model）常用于处理具有过度零值的数据，其中数据集中包含大量的零值观测。

令 n_A 表示没有注册的人数，则 $n_B = n_0 - n_A$ 表示刚注册等级为 0 的用户. 观测数据为 $x = (n_0, n_1, \cdots, n_6)$，如果隐含数据 $z = n_A$ 观测到，很容易得到参数 (ξ, λ) 的极大似然估计.

以下根据 EM 算法估计 (ξ, λ)，基于完全数据 $y = (x, z)$ 的似然函数为

$$
\begin{aligned}
& L(y; \xi, \lambda) \\
= & \frac{(n_A + n_B + n_1 + \cdots + n_6)!}{n_A! n_B! n_1! \cdots n_6!} \xi^{n_A} [e^{-\lambda}(1-\xi)]^{n_B} \Pi_{k=1}^{6} [\frac{\lambda^k e^{-\lambda}}{k!}(1-\xi)]^{n_k}.
\end{aligned}
$$

相当于先随机分组，再乘以每个人出现在这个组别的概率（多项式分布）。

这里其实有一个问题是，看作 8 个类别的多项式分布时，其中概率为泊松分布概率的话，总体概率求和不为 1。但我们可以假设是无穷类别的多项式分布，与此同时有更多的观测变量，$n_7, ..., n_\infty = 0$，这样代入式子中没有任何影响。

这是从整体考虑的似然函数，而非从个体出发。

二项泊松混合模型的 EM 估计

对数似然函数为

$$\begin{aligned} l(y;\xi,\lambda) &= n_A \log\xi - n_B\lambda + n_B\log(1-\xi) \\ &\quad + \sum_{k=1}^{6} n_k[-\lambda + k\log\lambda + \log(1-\xi)] + cost. \end{aligned}$$

基于观测数据 x 和上次迭代估计 $\xi^{(i-1)}, \lambda^{(i-1)}$, 对数似然的条件期望为

$$\begin{aligned} &Q(\xi,\lambda;\xi^{(i-1)},\lambda^{(i-1)}) = E(l(y;\xi,\lambda)|x;\xi^{(i-1)},\lambda^{(i-1)}) \\ &= E(n_A|x;\xi^{(i-1)},\lambda^{(i-1)})\log\xi - E(n_B|x;\xi^{(i-1)},\lambda^{(i-1)})\lambda \\ &+ E(n_B|x;\xi^{(i-1)},\lambda^{(i-1)})\log(1-\xi) \\ &+ \sum_{x=1}^{6} n_k[-\lambda + k\log\lambda + \log(1-\xi)] + cost. \end{aligned}$$

对 $Q(\xi, \lambda; \xi^{(i-1)}, \lambda^{(i-1)})$ 关于 ξ, λ 求导, 可得

$$\begin{aligned}
& \frac{\partial Q(\xi, \lambda; \xi^{(i-1)}, \lambda^{(i-1)})}{\partial \xi} \\
= & \frac{E(n_A|x; \xi^{(i-1)}, \lambda^{(i-1)})}{\xi} - \frac{E(n_B|x; \xi^{(i-1)}, \lambda^{(i-1)}) + n_1 + \cdots + n_6}{1-\xi} = 0 \\
& \frac{\partial Q(\xi, \lambda; \xi^{(i-1)}, \lambda^{(i-1)})}{\partial \lambda} \\
= & -\{E(n_B|x; \xi^{(i-1)}, \lambda^{(i-1)}) + n_1 + \cdots + n_6\} + \frac{1}{\lambda} \sum_{k=1}^{6} k n_k = 0.
\end{aligned}$$

根据上式不难得出第 i 次迭代结果

$$\xi^{(i)} = \frac{E(n_A|x;\xi^{(i-1)},\lambda^{(i-1)})}{n}, \quad \lambda^{(i)} = \frac{\sum_{k=1}^{6} kn_k}{n - E(n_A|x;\xi^{(i-1)},\lambda^{(i-1)})}.$$

对于上式 $E(n_A|x;\xi^{(i-1)},\lambda^{(i-1)})$，通过推导（$n_0$ 已知，第一和第二类别的比例已知，求和为 1），不难进一步化简得

$$E(n_A|x;\xi^{(i-1)},\lambda^{(i-1)}) = \frac{n_0\xi^{(i-1)}}{\xi^{(i-1)} + (1-\xi^{(i-1)})e^{-\lambda^{(i-1)}}}.$$

按照上述算法反复迭代，得到参数 ξ, λ 的 EM 估计。

目录

实际例子

假设我们有一个大型的商圈，每天每家店各个时间段的营业额数据，如何刻画客户群体？

一天得到的数据是：

1. 小明奶茶店：...，11：30 进账 500 元，11：47 进账 1000 元，12：20 进账 800 元...；
2. 小芳电影院：...，17：30 进账 300 元，19：30 进账 700 元，20：45 进账 900 元...；
3. 小红超市：...，12：30 进账 1000 元，13：20 进账 1270 元，14：50 进账 1150 元...

首先要做的就是数据清洗，将他们合并成同一时间分割。