机器学习十大算法-EM算法

一.两个例子

例1： Dempster,Laird,Rubin(1977)

多项分布

设$X=(x_1,x_2,x_3,x_4)\sim MN(n,\frac{1}{2}+\frac{\theta}{4},\frac{1-\theta}{4},\frac{1-\theta}{4},\frac{\theta}{4})$。

对数似然函数为：

$$L(\theta|X)=\log f(X;\theta)=\log[\frac{n!}{x_1!x_2!x_3!x_4!}(\frac{2+\theta}{4})^{x_1}(\frac{1-\theta}{4})^{x_2+x_3}(\frac{\theta}{4})^{x_4}]$$

$$=x_1\log(2+\theta)+(x_2+x_3)\log(1-\theta)+x_4\log\theta+\log C$$

其中C为与θ无关的常数。

于是似然方程为：

$$\frac{\partial L(\theta \mid X)}{\partial \theta} = \frac{x_1}{2+\theta} - \frac{x_2 + x_3}{1-\theta} + \frac{x_4}{\theta} = 0$$

解称为极大似然估计。

下面给出另外一种算法

将x_1分为(y_1, y_2),使

$$(y_1, y_2, x_2, x_3, x_4) \sim MN(n, \frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4})。$$

于是上述完全数据的对数似然函数为：

$$L_C(\theta \mid X, Y) = \log[\frac{n!}{y_1! y_2! x_2! x_3! x_4!}(\frac{1}{2})^{y_1}(\frac{\theta}{4})^{y_2+x_4}(\frac{1-\theta}{4})^{x_2+x_3}]$$

$$= (y_2 + x_4)\log\theta + (x_2 + x_3)\log(1-\theta) + \log C'$$

其中C'为与θ无关的常数。

容易看出其MLE为：

$$\hat{\theta}=\frac{y_2+x_4}{y_2+x_2+x_3+x_4}$$，但y_2为缺失数据，不能使用。

EM算法的思路如下：

二项分布

因为$y_2 \mid X,\theta^{(t)} \sim BN(x_1,\frac{\theta^{(t)}/4}{1/2+\theta^{(t)}/4}=\frac{\theta^{(t)}}{2+\theta^{(t)}})$

于是（E步）$E[L_C(\theta \mid X,Y) \mid X,\theta^{(t)}]$

$$=E[(y_2+x_4)\log\theta+(x_2+x_3)\log(1-\theta)+\log C' \mid X,\theta^{(t)}]$$

$$=(E[y_2 \mid X,\theta^{(t)}]+x_4)\log\theta+(x_2+x_3)\log(1-\theta)+\log C'$$

$$=(\frac{x_1\theta^{(t)}}{2+\theta^{(t)}}+x_4)\log\theta+(x_2+x_3)\log(1-\theta)+\log C' \triangleq Q(\theta \mid \theta^{(t)})。$$

然后求解（M步）

$$\max_{\theta} Q(\theta \mid \theta^{(t)})$$

得迭代公式 $\theta^{(t+1)} = \dfrac{\dfrac{x_1\theta^{(t)}}{2+\theta^{(t)}} + x_4}{\dfrac{x_1\theta^{(t)}}{2+\theta^{(t)}} + x_2 + x_3 + x_4}$

注意与 $\hat{\theta} = \dfrac{y_2 + x_4}{y_2 + x_2 + x_3 + x_4}$ 进行比较，可以看出*EM*算法的合理性。

实例计算如下：

若X=(125,18,20,34),不难计算此时极大似然估计为0.6268.
现用EM算法进行计算,可以看出不管初值如何,均收敛到极大似然估计。

```
k: 0   theta: 0.5

k: 1   theta: 0.608247

k: 2   theta: 0.624321

k: 3   theta: 0.626489

k: 4   theta: 0.626777

k: 5   theta: 0.626816
```

```
k: 0   theta: 0.2

k: 1   theta: 0.544166

k: 2   theta: 0.615135

k: 3   theta: 0.625256

k: 4   theta: 0.626613

k: 5   theta: 0.626794

k: 6   theta: 0.626818
```

若X=(1997,906,904,32),不难此时计算极大似然估计为0.0357.现用EM算法计算如下：

k: 0 theta: 0.3

k: 1 theta: 0.139111

k: 2 theta: 0.0820893

k: 3 theta: 0.0576522

k: 4 theta: 0.0463409

k: 5 theta: 0.0409191

k: 6 theta: 0.0382769

k: 7 theta: 0.0369788

k: 8 theta: 0.0363386

k: 9 theta: 0.0360222

k: 10 theta: 0.0358657

k: 11 theta: 0.0357882

k: 12 theta: 0.0357499

k: 13 theta: 0.0357309

k: 0 theta: 0.9

k: 1 theta: 0.264753

k: 2 theta: 0.127901

k: 3 theta: 0.0774875

k: 4 theta: 0.0555629

k: 5 theta: 0.0453485

k: 6 theta: 0.0404376

k: 7 theta: 0.0380408

k: 8 theta: 0.0368625

k: 9 theta: 0.0362811

k: 10 theta: 0.0359938

k: 11 theta: 0.0358516

k: 12 theta: 0.0357813

k: 13 theta: 0.0357465

k: 14 theta: 0.0357292

例2：混合分布

设$X_1, X_2, \cdots, X_n$独立同分布，密度函数为

$$\mathrm{g}(x, p_1, \cdots, p_m) = \sum_{j=1}^{m} p_j \varphi_j(x),$$

其中$p_j \geq 0$，$\mathrm{j} = 1, 2, \cdots, m; \sum_{j=1}^{m} p_j = 1.$

对数似然函数为$\mathrm{L}(p \mid X) = \sum_{i=1}^{n} Log[\sum_{j=1}^{m} p_j \varphi_j(X_i)]$

上述函数的最大似然估计不易求解。

设$Y_1, Y_2, \cdots, Y_n$为缺失数据，独立同分布。

$P(Y_i = e_j = (0, \cdots, 0, 1, 0, \cdots 0)) = p_j;$

且$X_i \mid Y_i = e_j$的条件分布服从$\phi_j(x), j = 1, \cdots, m.$

于是完全对数似然函数为

$$L_C(\mathrm{p} \mid Z) = Log[\prod_{i=1}^{n}\prod_{j=1}^{m} p_j^{Y_{ij}} \phi_j^{Y_{ij}}(X_i)]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m}[Y_{ij}Logp_j + Y_{ij}Log\phi_j(X_i)]$$

$$= \sum_{j=1}^{m}(\sum_{i=1}^{n} Y_{ij})Logp_j + C$$

其中C为与p无关的常数。

容易看出其MLE为：

$$\hat{p}_j = \frac{1}{n}\sum_{i=1}^{n} Y_{ij},\ j = 1,\cdots,m.$$ 但Y_{ij}为缺失数据，无法使用。

EM算法如下：

E步：

$$Q(p|p^{(t)}) = E[L_C(p \mid Z) \mid X, p^{(t)}]$$

$$= \sum_{j=1}^{m} Logp_j [\sum_{i=1}^{n} E(Y_{ij} \mid X, p^{(t)})] + C'$$

$$= \sum_{j=1}^{m} Logp_j [\sum_{i=1}^{n} \frac{p_j^{(t)} \varphi_j(X_i)}{\sum_{j=1}^{m} p_j^{(t)} \varphi_j(X_i)}] + C'$$

上式条件期望的计算利用贝叶斯公式

$$P(Y_i = e_j \mid X_i, p^{(t)}) = \frac{P_{p^{(t)}}(Y_i = e_j) P_{p^{(t)}}(X_i \mid Y_i = e_j)}{\sum_{j=1}^{m} P_{p^{(t)}}(Y_i = e_j) P_{p^{(t)}}(X_i \mid Y_i = e_j)}$$

$$= \frac{p_j^{(t)} \varphi_j(X_i)}{\sum_{j=1}^{m} p_j^{(t)} \varphi_j(X_i)} \triangleq Y_{ij}^{(t)}, \; j = 1, \cdots, m.$$

M步：

$$\max_{p} Q(p \mid p^{(t)})$$

解得迭代公式$p_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} Y_{ij}^{(t)}, \; j = 1, \cdots, m.$

```
p <- 0.7              #mixing parameter
n <- 50               #sample size
mu <- c(3, 10)        #parameters of the normal densities
sigma <- c(1, 1)

          # generate the mixed normal distribution sample

i <- sample(1:2, size=n, replace=TRUE, prob=c(p, 1-p))
x <- rnorm(n, mu[i], sigma[i])

          # mle of the mixed normal distribution

logL <- function(y) {
  -sum( log(y * dnorm(x, mu[1], sigma[1]) +
          (1-y) * dnorm(x, mu[2], sigma[2])) )
}
optimize(logL, lower=0, upper=1)
```

$minimum
[1] 0.7399919

$objective
[1] 102.5042

EM alogrithm

```
    p <- 0.2        #intialize of parameter

    y <- c(1:n)
    k <- 1
while(k <= 5) {
 for(i in 1:n) {
    y[i] <-  ( p * dnorm(x[i], mu[1], sigma[1]) ) /
    ( p * dnorm(x[i], mu[1], sigma[1]) +  ( 1 - p ) * dnorm(x[i], mu[2], sigma[2]) )
    }
 p <- sum( y ) / n
 print(p)
 k <- k+1
}
```

[1] 0.7400001
[1] 0.740003
[1] 0.740003
[1] 0.740003
[1] 0.740003

例3：带参数的混合分布

设$X_1, X_2, \cdots, X_n$独立同分布，密度函数为

$$f(x;\theta)=\sum_{j=1}^{m} p_j f_j(x),$$

其中$p_j \geq 0, \sum_{j=1}^{m} p_j = 1,\ f_j(x)=\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_i)^2}{2\sigma^2}}$,j$=1,2,\cdots,m.$

$$\theta=(p_1,\cdots,p_m,\mu_1,\cdots,\mu_m,\sigma^2).$$

对数似然函数为$\mathrm{L}(p \mid X)=\sum_{i=1}^{n} Log[\sum_{j=1}^{m} p_j f_j(X_i)]$

上述函数的最大似然估计不易求解。

设$Y_1, Y_2, \cdots, Y_n$为缺失数据，独立同分布。

$$P(Y_i = e_j = (0, \cdots, 0, 1, 0, \cdots 0)) = p_j;$$

且$X_i \mid Y_i = e_j$的条件分布服从$f_j(x), j = 1, \cdots, m.$

于是完全对数似然函数为

$$L_C(\theta \mid Z) = Log[\prod_{i=1}^{n}\prod_{j=1}^{m} p_j^{Y_{ij}} f_j^{Y_{ij}}(X_i)]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m}[Y_{ij} Logp_j + Y_{ij} Logf_j(X_i)]$$

$$= \sum_{j=1}^{m}(\sum_{i=1}^{n} Y_{ij}) Logp_j - \frac{1}{2}\sum_{j=1}^{m}\sum_{i=1}^{n} Y_{ij}[\log\sigma^2 + \frac{(X_i - \mu_j)^2}{\sigma^2}] + C$$

其中C为与θ无关的常数。

EM算法如下：

E步：

$$Q(p|p^{(t)}) = E[L_C(\theta \mid Z) \mid X, \theta^{(t)}]$$

$$= \sum_{j=1}^{m} Logp_j [\sum_{i=1}^{n} E(Y_{ij} \mid X, \theta^{(t)})]$$

$$-\frac{1}{2}\sum_{j=1}^{m}\sum_{i=1}^{n} E(Y_{ij} \mid X, \theta^{(t)})[\log \sigma^2 + \frac{(X_i - \mu_j)^2}{\sigma^2}] + C$$

上式条件期望

$$E(Y_{ij} \mid X_i, \theta^{(t)}) = \frac{p_j^{(t)} f_j(X_i)}{\sum_{j=1}^{m} p_j^{(t)} f_j(X_i)} \triangleq Y_{ij}^{(t)},\ j = 1, \cdots, m.$$

M步：

$$\max_{p} Q(\theta \mid \theta^{(t)})$$

解得迭代公式$p_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} Y_{ij}^{(t)},\ j = 1, \cdots, m;$

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} Y_{ij}^{(t)} X_i}{\sum_{i=1}^{n} Y_{ij}^{(t)}},\ j = 1, \cdots, m; (\sigma^2)^{(t+1)} = \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n} Y_{ij}^{(t)} (X_i - \mu_j^{(t+1)})^2$$

```
p <- 0.7                #mixing parameter
n <- 200                #sample size
mu <- c(3, 10)      #parameters of the normal densities
sigma0 <- c(1, 1)

          # generate the mixed normal distribution sample

i <- sample(1:2, size=n, replace=TRUE, prob=c(p, 1-p))
x <- rnorm(n, mu[i], sigma0[i])

                    # EM alogrithm
y <- c(1:n)
k <- 1
for(j in 1:5) {
  for(i in 1:n) {
    y[i] <-  ( p * dnorm(x[i], mu1, sigma) ) /
    ( p * dnorm(x[i], mu1, sigma) +  ( 1 - p ) * dnorm(x[i], mu2, sigma) )
    }
  p <- sum( y ) / n
  mu1 <- sum(y * x) / sum(y)
  mu2 <- sum((1 - y) * x) / sum(1-y)
  sigma <- ( sum(y*(x-mu1)^2) + sum((1-y)*(x-mu2)^2) ) / n
  print(cbind(p,mu1,mu2,sigma))
}
```

```
p <- 0.6        #intialize of parameter
mu1 <- 5
mu2 <- 7
sigma <- 1
```

```
p      mu1     mu2     sigma
[1,] 0.7066418 2.919782 9.94336 0.9731005
      p     mu1     mu2     sigma
[1,] 0.71 2.92178 10.0198 0.8256728
      p     mu1     mu2     sigma
[1,] 0.71 2.92178 10.0198 0.8256686
      p     mu1     mu2     sigma
[1,] 0.71 2.92178 10.0198 0.8256686
      p     mu1     mu2     sigma
[1,] 0.71 2.92178 10.0198 0.8256686
```

```
p <- 0.6        #intialize of parameter
mu1 <- 2
mu2 <- 5
sigma <- 1
```

```
p      mu1     mu2    sigma
[1,] 0.491604 2.627146 7.500634 4.755987
      p     mu1     mu2    sigma
[1,] 0.4950997 4.027033 6.161662 9.552981
      p     mu1     mu2    sigma
[1,] 0.4950526 4.978672 5.228472 10.67643
      p     mu1     mu2   sigma
[1,] 0.495052 5.092976 5.116407 10.6919
      p     mu1     mu2    sigma
[1,] 0.495052 5.103701 5.105893 10.69203
```

二.EM算法及其收敛性

设Z=(X,Y),其中X为观测数据，Y为缺失数据，Z为完全数据。

对数似然函数为：

$$L(\theta \mid X) = \log f(X;\theta),$$

$\max_{\theta \in \Omega} L(\theta \mid X)$可能不易求解，其中$\Omega$为$\theta$的值域。

EM算法思路如下：

E步：在假定$\theta = \theta^{(t)}$的条件下，计算对数完全似然的条件期望，即：

$$Q(\theta \mid \theta^{(t)}) = E[L_C(\theta \mid Z) \mid X, \theta^{(t)}]$$

M步：通过最大化$Q(\theta|\theta^{(t)})$,求出$\theta^{(t+1)}$。即

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)}), \forall \theta \in \Omega.$$

注：若只要求$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)})$，则称为$GEM$算法。

定理：$L(\theta^{(t+1)}|X) \geq L(\theta^{(t)}|X), t = 0,1,2,\cdots$

证明：$L_C(\theta \mid Z) = L(\theta \mid X) + Logf(Y \mid X, \theta)$

两边对$f(Y \mid X, \theta^{(t)})$求期望，得

$$Q(\theta \mid \theta^{(t)}) = L(\theta \mid X) + \int Logf(Y \mid X, \theta) f(Y \mid X, \theta^{(t)}) dY$$

由*Jensen*不等式，有$\forall \theta \in \Omega$

$$\int Logf(Y \mid X, \theta) f(Y \mid X, \theta^{(t)}) dY \leq \int Logf(Y \mid X, \theta^{(t)}) f(Y \mid X, \theta^{(t)}) dY$$

$$\text{又} Q(\theta^{(t+1)} \mid \theta^{(t)}) = L(\theta^{(t+1)} \mid X) + \int Logf(Y \mid X, \theta^{(t+1)}) f(Y \mid X, \theta^{(t)}) dY$$

$$\geq Q(\theta^{(t)} \mid \theta^{(t)}) = L(\theta^{(t)} \mid X) + \int Logf(Y \mid X, \theta^{(t)}) f(Y \mid X, \theta^{(t)}) dY$$

于是有$L(\theta^{(t+1)} \mid X) \geq L(\theta^{(t)} \mid X)$

问题：$\theta^{(t)} \to \hat{\theta}_{MLE}$是否成立？

定义规则条件如下：

1).$\Omega \in R^d$;

2).$\Omega_{\theta_0} = \{\theta \in \Omega \mid L(\theta \mid X) > L(\theta_0 \mid X)\}$为紧集，

$\forall \{\theta_0 \mid L(\theta_0 \mid X) > -\infty\}$;

3).$L(\theta \mid X)$在Ω上关于θ连续，在Ω内部关于θ可导；

4).$\left.\frac{\partial Q(\theta \mid \theta^{(t)})}{\partial \theta}\right|_{\theta^{(t+1)}} = 0.$

定理：

在上述规则条件下，若$Q(\theta|\theta^{(t)})$关于θ，$\theta^{(t)}$为连续函数，则$\{\theta^{(t)}\}$的任意收敛点均为$L(\theta|X)$的驻点。证明：参见Wu(1983).

注:1).$L(\theta|X)$的驻点可能为局部极大点或鞍点。关于收敛到局部极大点的例子参见后面的例题；若收敛到鞍点，只要将初值进行变动，就可以避开鞍点。

2).规则条件中2）可能不一定满足，例如带参数的混合分布：

$$L(\theta|X)=\log\left[p\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(X-\mu_1)^2}{2\sigma_1^2}}+(1-p)\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{(X-\mu_2)^2}{2\sigma_2^2}}\right],$$

对任意θ_0，$\Omega_{\theta_0}=\{\theta|L(\theta|X)>L(\theta_0|X)\}$对$\mu_2$，$\sigma_2$无界。

（令$\theta=(\mu_1,\mu_2,\sigma_1,\sigma_2,p)$中$\mu_1=X$,$\sigma_1$趋于0，则$\mu_2$，$\sigma_2$可以任意大）