博文

EM算法(期望最大化算法)简介

和HMM简介一样（有关HMM，隐马尔科夫模型的简介，请参见我的另一篇博文 http://blog.sciencenet.cn/blog-2970729-1188964.html），我们还是通过一个例子引入EM算法(Expectation Maximization Algorithm)

1. 一个经典例子

我们有两枚硬币(coin A & coin B)，这两枚硬币是用特殊材质做的，硬币A抛出正面 (**H**ead)和反面(**T**ail)的概率为θ_A和1-θ_A，硬币B抛出正面和反面的概率为θ_B 和1-θ_B。我们不知道θ_A和θ_B，因此想通过不断的抛硬币来推测出θ_A和θ_B，为了方便，写成向量形式：

θ= (θ_A,θ_B)。

因为有两枚硬币，我们随机地在硬币A和硬币B中挑一个(概率相等，各为50%)，然后再用选中的硬币独立地抛10次，为了使整个事件更具说服力，我们选硬币抛硬币的整个过程重复做了5次。因此，总的来说选了5次硬币，抛了5×10=50次。

选了5次硬币，每次记为$z_i \in \{A, B\}$，5次合到一起记为z=(z_1, z_2, z_3, z_4, z_5)；每选1次硬币(抛10次)，我们记录其中正面出现的次数$x_i \in \{0,1,...,10\}$，5次合到一起记为x=(x_1, x_2, x_3, x_4, x_5)。于是，很容易我们就可以评估出θ_A：

$$\theta_A = \frac{A\text{硬币抛出正面的次数}}{A\text{硬币抛出的总次数}}$$

这种通过观测来评估模型参数的方法称为**极大似然评估**(Maximum likelihood estimation)。

上面这个例子比较简单，是一个完备问题(complete data case)。如果上面这个例子中，已知x向量，但是不知道z向量，这时再来评估θ_A和θ_B。这类问题属于不完备问题(incomplete data case)，我们称未知的z向量为**隐变量**(hidden variables)或者**潜在因素**(latent factors)。

注：我们可以想象一下，开篇的例子中，我们挑一枚硬币且记录下挑了什么硬币，再抛这枚硬币且记录下硬币的正反面。现在由于意外发生，挑硬币的那部分记录不幸遗失，只有硬币正反向次数的记录，让我们求解两枚硬币抛出正面的概率分别是多少。仔细一想，这个问题其实还是挺难的！聪明的人们想出了一个很直觉的解决办法——EM算法。

首先我们随便给定θ_A和θ_B，如果有先验知识也可以用先验知识。

$$\hat{\theta}^{(t)}=(\hat{\theta}_A^{(t)},\hat{\theta}_B^{(t)})$$

基于这两个参数，计算这5次硬币A/B最可能的出现情况。再固定这5次硬币A/B抽取情况，估计θ，此时记为

$$\hat{\theta}^{(t+1)}$$

再不断重复这一过程直到收敛。

总之，EM算法轮流执行两个步骤：其一，在当前模型的前提下，猜测最可能的概率分布，例如推测硬币A/B的抽取概率和期望，这一步称为**E-step**；其二，在推测的结果基础上重新评估模型的参数，推测θ_A和θ_B，这一步称为**M-step**。

之所以称为E-step，是因为通常不需要在绝对完整的前提下来生成概率分布，而是在当前完整的情况下，计算"期望的(Expected)"的充分统计量。与此类似，之所以称为M-step，是因为模型的重新评估可以认为是使期望的log值"最大化(Maximization)"。

2. 数学推导

在推导之前我们需要一点数学知识：

> Jensen's inequality

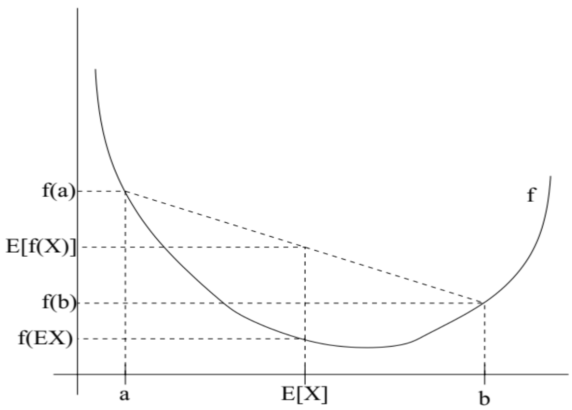
Theorem 0.2 (Jensen's Inequality) *Let $f(x)$ be a convex function defined on an interval I. If $x_1, x_2, \ldots, x_N \in I$ and $\lambda_1, \lambda_2, \ldots, \lambda_N \geq 0$ with $\sum_{i=1}^{N} \lambda_i$,*

$$f\left(\sum_{i=1}^{N} \lambda_i x_i\right) \leq \sum_{i=1}^{N} \lambda_i f(x_i)$$

Alternatively, if $f(x)$ is a convex function and $X \in \{x_i : 1, \ldots, N\}$ is a random variable with probabiliites $P(x_i)$ where $\sum P(x_i) = 1$, then,

$$f(E\{X\}) \leq E\{f(X)\}$$
$$f\left(\sum_{i=1}^{N} x_i P(x_i)\right) \leq \sum_{i=1}^{N} f(x_i) P(x_i)$$

对应的图形如下：



问题：已知一组独立的样本x={x_1, x_2, x_3, ..., x_m}，求模型p(x, z)的参数θ，使p(x, z)最大，其中z是隐变量。

开始推导：

因为log函数是单调递增的，所以求p(x, z)的最大值，即求log(p(x, z))的最大值。完整的写出似然函数如下，即求L(θ)最大时参数θ的值：

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^{m} \log p(x;\theta) \\ &= \sum_{i=1}^{m} \log \sum_{z} p(x,z;\theta).\end{aligned}$$

因为z通常观测不到，这使得问题比较复杂，无法直接求L(θ)的最大值。我们采用EM算法，构建L(θ)的下界(E-step)，再优化下界(M-step)，并不断重复这一过程。

$$\begin{aligned}\sum_i \log p(x^{(i)};\theta) &= \sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)};\theta) &\quad (1)\\ &= \sum_i \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} &\quad (2)\\ &\geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} &\quad (3)\end{aligned}$$

(1)式是上述所求的。上述公式来自Andrew Ng[3]，Andrew Ng喜欢用上标表示index，可能和我们的习惯写法有点不一致，在此不展开，大家可以简单地将$x^{(i)}$看成x_i

(2)式是分子分母同乘了一个$Q_i(z^{(i)})$它是一种分布，而且

$$\sum_z Q_i(z) = 1$$

$$Qi(z) \geqslant 0$$

(3)由第2个式子到第3个式子的推导如下:

令$f(x) = \log(x)$, $j = z^{(i)}$, $\lambda_j = Q_i(j)$，其中$\sum \lambda_j = 1$，且$\lambda_j \geqslant 0$. 而且

$$\beta_i = \frac{p(x^{(i)},\ z^{(i)};\ \theta)}{Q_i(z^{(i)})}$$

那么第2个式子到第3个式子推导如下:

$$\begin{aligned}\sum_i \log\ p(x^{(i)};\theta) &= \sum_i \log \sum_{z(i)} Q_i(z^{(i)}) \frac{p(x^{(i)},\ z^{(i)};\ \theta)}{Q_i(z^{(i)})} \\ &= \sum_i f\left(\sum_j \lambda_j \beta_j\right) \\ &\geqslant \sum_i \sum_j \lambda_j f(\beta_j) \\ &= \sum_i \sum_{z(i)} Q_i(z^{(i)}) \log \frac{p(x^{(i)},\ z^{(i)};\ \theta)}{Q_i(z^{(i)})}\end{aligned}$$

当且仅当以下等式成立时，L(θ)取得极大值:

$$\frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} = c$$

其中c表示常数。也即

$$Q_i(z^{(i)}) \propto p(x^{(i)}, z^{(i)};\theta).$$

又因为$Q_i(z^{(i)})$是概率分布，按照定义

$$\sum_z Q_i(z^{(i)}) = 1$$

所以有

$$\begin{aligned} Q_i(z^{(i)}) &= \frac{p(x^{(i)}, z^{(i)}; \theta)}{\sum_z p(x^{(i)}, z; \theta)} \\ &= \frac{p(x^{(i)}, z^{(i)}; \theta)}{p(x^{(i)}; \theta)} \\ &= p(z^{(i)} | x^{(i)}; \theta) \end{aligned}$$

综上，不断重复以下两个步骤，直到收敛：

(E-step) For each i, set

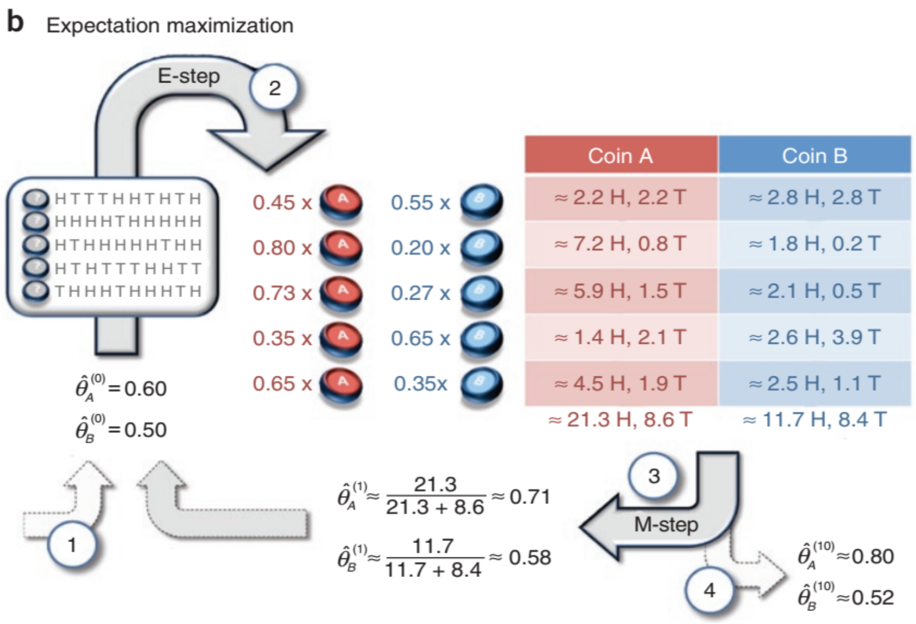
$$Q_i(z^{(i)}) := p(z^{(i)} | x^{(i)}; \theta).$$

(M-step) Set

$$\theta := \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}.$$

其中:=表示当前等于，也即覆盖之前的值

3. 用Chuong[1]文章中的实例详细计算一下



✇ i=0时，令

$$\hat{\theta}_A^{(0)} = 0.60$$
$$\hat{\theta}_B^{(0)} = 0.50$$

E-Step：

(1) 计算第一次选硬币，且投出的结果为H-T-T-T-H-H-T-H-T-H时，选的是硬币A的概率

根据二项分布定义，如果第一次选的是A硬币，投10次有5次为正面的概率为：

$$P(HTTTHHTHTH|A)=C_{10}^{5}(0.6)^{5}(1-0.6)^{5}$$

同理，可计算出如果第一次选的是B硬币，投10次有5次为正面的概率：

$$P(HTTTHHTHTH|B)=C_{10}^{5}(0.5)^{5}(1-0.5)^{5}$$

再由贝叶斯定律可计算出，投币结果为H-T-T-T-H-H-T-H-T-H，且这一结果是硬币A投出的概率为：

$$\begin{aligned}P(A|HTTTHHTHTH)&=\frac{P(HTTTHHTHTH|A)P(A)}{P(HTTTHHTHTH|A)P(A)+P(HTTTHHTHTH|B)P(B)}\\&=\frac{C_{10}^{5}(0.6)^{5}(1-0.6)^{5}\times0.5}{C_{10}^{5}(0.6)^{5}(1-0.6)^{5}\times0.5+C_{10}^{5}(0.5)^{5}(1-0.5)^{5}\times0.5}\\&\approx0.45\end{aligned}$$

注意：其中P(A)和P(B)都为0.5，因为抽取A硬币和抽取B硬币是等可能的。

同理，投币结果为H-T-T-T-H-H-T-H-T-H，且这一结果是硬币B投出的概率为：

$$P(B|HTTTHHTHTH)\approx0.55$$

(2) 再依次计算第二到第五次选硬币时，抽取到A/B硬币的概率。所有结果汇总如下表：

Observation	Coin A	Coin B
HTTTHHTHTH	0.45	0.55
HHHHTHHHHH	0.80	0.20
HTHHHHHTHH	0.73	0.27
HTHTTTHHTT	0.35	0.65
THHHTHHHTH	0.65	0.35
Total	2.98	2.02
Score	21.24	11.76

其中，Coin A列表示投出该行硬币正反面情况下，推测是用A硬币来投的概率。

Total行是计算的5次试验的总和，而Score行的计算如下：

Score = (0.45×5) + (0.8×9) + (0.73×8) + (0.35×4) + (0.65×7)=21.24

仔细回想，这里的Score其实就是5次选币（50次投币）事件的期望值(Expectation)！以上整个过程也正是在计算$P(z^{(i)} | x^{(i)};\theta)$，即

每次观测(Observation)索引 => i

当前θ参数条件 => (θ_A，θ_B)

每次观测，如H-T-T-T-H-H-T-H-T-H => $x^{(i)}$

挑选的硬币（Coin A, Coin B） => $z^{(i)}$

的概率$P(z^{(i)} | x^{(i)};\theta)$

另外，我们可以在文中看到如下表格，计算的也是一样的，以第二行为例Coin A列的2.2H表示第一次选硬币，再抛10次时硬币为H的总和0.45×5；Coin B列的2.8H表示0.55×5

Coin A	Coin B
≈ 2.2 H, 2.2 T	≈ 2.8 H, 2.8 T
≈ 7.2 H, 0.8 T	≈ 1.8 H, 0.2 T
≈ 5.9 H, 1.5 T	≈ 2.1 H, 0.5 T
≈ 1.4 H, 2.1 T	≈ 2.6 H, 3.9 T
≈ 4.5 H, 1.9 T	≈ 2.5 H, 1.1 T
≈ 21.3 H, 8.6 T	≈ 11.7 H, 8.4 T

同理，我们也可以算出观测到是HTTTHHTHTH时Coin A抛出反面的概率得分2.2T

M-Step：

再计算当前假设下，硬币A投正面的概率，以及硬币B投正面的概率，即

$$\hat{\theta}_A^{(1)} \approx \frac{21.3}{21.3+8.6} \approx 0.71$$

$$\hat{\theta}_B^{(1)} \approx \frac{11.7}{11.7+8.4} \approx 0.58$$

从$\theta^{(i)}$到$\theta^{(i+1)}$的过程为什么叫做极大化(Maximization)？其实这里有个认证：每次迭代的过程(从$\theta^{(i)}$到$\theta^{(i+1)}$)，都会使$P(z^{(i)} \mid x^{(i)};\theta)$增大，越来越接近或者达到局部极值

>再将$\theta^{(i+1)}$作为参数，不断重复上述步骤，直到结果收敛。

经过10次重复之后，算出如下结果：

$$\hat{\theta}_A^{(10)} \approx 0.80$$

$$\hat{\theta}_B^{(10)} \approx 0.52$$

最后，需要强调的是EM也有自身的缺陷：

(1) 最终的结果与初始值的选取有关，不同的初始值可能得到不同的参数估计值

(2) 很可能会陷入局部最优解，而无法达到全局最优解

总结：

本文简约地介绍了EM算法，旨在深入浅出，面向的对象是算法小白(如我一般)。留下两个问题，有兴趣的读者可以继续探索，如果有时间后续我也会整理出来：

(1) EM算法的收敛性：E-Step和M-Step不断重复，最后是否会收敛？

(2) EM算法的应用，例如高斯混合模型、HMM中的第三个问题 (这个问题我会尽快整理到博文中，敬请期待)等。

参考材料：

[1] Chuong B Do, Serafim Batzoglou. What is the expectation maximization algorithm?

[2] Konstantinos G. Derpanis . Jensen's Inequality

[3] Andrew Ng. CS229 Lecture notes

[4] https://ynorm.com/blog/expectation-maximization/

上一篇：隐马尔科夫模型简介(三)
下一篇：隐马尔科夫模型简介(四)