EM算法

我们的目标

请考虑一下这样一个例子。

假设有3枚硬币, 分别记做A, B, C.现在我们来进行如下试验: 首先投掷硬币C, 如果结果是正面我们选择硬币A, 如果是背面我们硬币B. 然后投掷刚才选择的硬币(A或者B), 并记录下他们的结果. 正面记录为1, 背面记录为0. 独立重复地进行这个实验10次. 我们得到了如下的观测结果:

$$1, 1, 0, 1, 0, 0, 1, 0, 1, 1$$

假设我们最后拿到的结果只有上面这条观测结果(也就是我们只知道我们记录下的结果，却不知道这些结果出自A还是B). 现在我们想要在这样的条件下, 估计三枚硬币各自出现正反面的概率.

整理一下现在已有的信息吧

我们现在的条件其实非常有限, 能够看到的只有试验的结果(A或者B的结果). 我们在这里将试验的结果称作观测变量(observed variable), 并记做X. 那么就有

$$X = \{X_1, X_2, \dots, X_{10}\} = \{1, 1, 0, 1, 0, 0, 1, 0, 1, 1\}$$

同时我们知道在每一个结果诞生前, 我们需要投掷硬币C来决定接下来将要投掷硬币A还是硬币B. 那么我们将C的结果这样我们不能直接获得的数据称作隐含变量(latent variable), 并记做Z. 那么就有

$$Z = \{Z_1, Z_2, \dots, Z_{10}\}, Z_i = A_{或者} B$$

A, B, C 出现正面的概率, 也就是我们要估计的参数, 我们分别将他们记做$\{\alpha, \beta, \gamma\}$. 为了简洁表达, 我们在这里用$\theta$来表示. 也就是

$$\theta = \{\alpha, \beta, \gamma\}$$

策略

要想估计θ, 我们可以想到的一种策略便是枚举每一个θ, 使得最后出现观测结果X的可能性最大(也就是极大似然法(MLE)的思想). 用数学语言表示这种可能性的话就是$p(X|\theta)$, 也就是在θ的条件下, 出现X这个结果序列的可能性. 我们的目标就是求一个$\hat{\theta}$. 这个$\hat{\theta}$满足

$$\hat{\theta} = \arg\max_{\theta} p(X|\theta)$$

从数学到算法

数学

我们从$p(X|\theta)$入手. 因为问题中有隐含变量Z, 那么我们可以考虑将含有Z的表达式纳入计算中. 由条件概率公式$p(AB) = p(A|B)p(B)$, 我们可以得到$p(X, Z|\theta) = p(Z|X, \theta)p(X|\theta)$.也就是

$$p(X|\theta) = \frac{p(X, Z|\theta)}{p(Z|X, \theta)}$$

为了运算简便, 对两边同时取对数,并且由对数的运算性质可得

$$\log p(X|\theta) = \log p(X, Z|\theta) - \log p(Z|X, \theta)$$

接下来我们做出一个构造

$$\begin{aligned}\log p(X|\theta) &= [\log p(X, Z|\theta) - \log q(Z)] - [\log p(Z|X, \theta) - \log q(Z)] \\ &= \log \frac{p(X, Z|\theta)}{q(Z)} - \log \frac{p(Z|X, \theta)}{q(Z)}\end{aligned}$$

在这个构造中, $q(Z)$是一个以Z为未知数的某一个分布, 引入$q(Z)$只是为了进行化简的一种数学技巧, 所以不必过于在意$q(Z)$是什么.

得到上面这个式子以后, 我们对等式两边同时关于$q(Z)$求期望(P.S. 个人认为这是这个算法的整个数学推导中非常精彩的一步, 很快我们就能知道为什么).

$$\text{左边} = \int_Z \log p(X|\theta) q(Z)\, dZ$$

$$\text{由于 } \log p(X|\theta) \text{中没有 } Z, \text{在这里我们可以将它看做一个常数, 那么}$$

$$\text{左边} = \log p(X|\theta) \int_Z q(Z)\, dZ$$

$$\text{观察 } \int_Z q(Z)\, dZ, \text{可以发现这个式子就是在求累积概率所以在整个 } Z \text{的空间的累积概率应该等于} 1$$

$$\text{所以, 左边} = \log p(X|\theta)$$

所以左边通过求期望的步骤后没有发生变化. 而对于右边, 有

$$\begin{aligned}\text{右边} &= \int_Z \log \frac{p(X, Z|\theta)}{q(Z)} \cdot q(Z)\, dZ - \int_Z q(Z) \cdot \log \frac{p(Z|X, \theta)}{q(Z)}\, dZ \\ &= \int_Z \log \frac{p(X, Z|\theta)}{q(Z)} \cdot q(Z)\, dZ + \int_Z q(Z) \cdot \log \frac{q(Z)}{p(Z|X, \theta)}\, dZ\end{aligned}$$

其中$\int_Z q(Z) \cdot \log \frac{q(Z)}{p(Z|X,\theta)}\, dZ$这一项其实可以看做是KL散度(K-L Divergence), 在这里我们只需要知道它具有非负性且$q(Z) = p(Z|X, \theta)$时等于0即可. 也就是

$$\int_Z q(Z) \cdot \log \frac{q(Z)}{p(Z|X, \theta)}\, dZ \geq 0$$

因此

$$\text{原式} \geq \int_Z \log \frac{p(X, Z|\theta)}{q(Z)} \cdot q(Z)\, dZ$$

且等号仅在$q(Z) = p(Z|X, \theta)$时等于成立. 替换$q(Z) = p(Z|X, \theta)$有

$$\int_Z \log \frac{p(X, Z|\theta)}{p(Z|X, \theta)} \cdot p(Z|X, \theta)\, dZ$$

我们称这个式子为下界ELBO(evidence lower bound). 接下来的思想将是整个算法的**核心**所在.

既然我们已经得到ELBO, 且我们知道要最大化的$P(X|\theta)$一定大于等于ELBO. 那么如果$p(X|\theta)$ 的最大值是确定的, 是一个存在的常数, 且我们已知一组$\theta^{(t)}$，注意这里是$\theta^{(t)}$而不是$\hat{\theta}$. 那么如果我们能保证我们利用$\theta^{(t)}$算出的最大化的$p(X|\theta^{(t)})$一定是小于另一组$\theta^{(t+1)}$的下界ELBO的, 那么我们就可以通过迭代的方式逐步获得新的$\theta^{(t)}$, 从而利用$\theta^{(t)}$逐步逼近$p(X|\theta)$的最大值.

算法

上面关于思想的这一段很绕也有很多地方需要解释, 为了解释更方便, 所以我在这里先直接描述算法, 稍后再来讨论这一个思想.

$$
\begin{aligned}
\text{input: } & X, \theta^{(0)} \\
\text{E-Step: } & 1) \text{利用} \theta^{(0)} \text{列出近似的表达式} q^{(0)}(Z) = p^{(0)}(Z|X, \theta^{(0)}) \\
& 2) \text{将近似的} q^{(0)}(Z) \text{带入} \mathrm{ELBO} \text{中，得到} \mathrm{ELBO} \text{的表达式} \\
\text{M-Step: } & 1) \text{求出使得} p^{(1)}(Z|X, \theta^{(1)}) \text{最大的} \theta^{(1)} \\
\text{repeat: } & \text{回到} \mathrm{E\text{-}Step}. \text{利用刚得到的} \theta^{(1)} \text{表示} q^{(1)}, \text{然后重复接下来的步骤，直到达到指定迭代次数}
\end{aligned}
$$

关于这个算法, 首先解释一下为什么用$\theta^{(t)}$ 求ELBO, 而不是$\theta^{(t+1)}$. 按理来说, 确实$\mathrm{ELBO}^{(t)}$中$p(X, Z|\theta)$和$p(Z|X, \theta)$中的θ应该是相同的θ, 但由于如果将$p(Z|X, \theta)$中的θ 作为一个未知数的话计算会非常困难, 所以选择用上一步得到的θ来近似表示.

第二个问题是如果$p(Z|X, \theta^{(t)})$ 就算利用$\theta^{(t)}$来近似表示也很难表示出来应该怎么办呢？事实上我们在分析介绍ELBO的时候隐含了我们能准确得到$p(Z|X, \theta^{(t)})$的假设，也就是我们可以使KL散度等于0. 但如果我们不能求得$p(Z|X, \theta^{(t)})$的表达式时, 我们也就无法保证KL散度等于0了. 在这种情况下我们选择求使得KL散度最小的$q(Z)$来近似. 这也就是广义上的EM算法, 而我们正在介绍的这种被称作狭义上的EM算法. 注意我们这里说的只是$p(Z|X, \theta^{(t)})$的表达式我们无法求出, 而非这是一个未知的量.

现在我们回来解释一下关于上文所说的算法的**核心思想**.

从迭代的角度考虑, 我们给$p(X|\theta^{(t+1)})$的ELBO起个名字叫做$Q(\theta^{(t+1)}, \theta^{(t)})$, 其中$\theta^{(t+1)}$暂时称作未知量, 而$\theta^{(t)}$则是一个已知量.

$$
Q(\theta^{(t+1)}, \theta^{(t)}) = \int_Z \log \frac{p(X, Z|\theta^{(t+1)})}{p(Z|X, \theta^{(t)})} \cdot p(Z|X, \theta^{(t)})\, dZ
$$

从这个角度来看, ELBO既然是一个关于θ的一个函数. 那么我们完全可以让第$t+1$步的ELBO刚好大于等于$p(X|\theta)$第t步的最大值, 当然前提是ELBO *可以*等于$p(X|\theta)$第t步的最大值. 请记住这段话, 因为稍后你将惊叹于这个构思的精巧.

那么什么样的$\theta^{(t+1)}$恰好让ELBO等于$p(X|\theta)$第t步的最大值呢？回忆一下, $\theta^{(t)}$是使得$p(X|\theta)$在第t步取得最大值的一组参数, 所以答案是$\theta^{(t)}$. 那么我们就可以得到这样的关系
$\mathrm{ELBO\ of\ } p(X|\theta^{(t+1)}) = Q(\theta, \theta^{(t)}) \geq p(X|\theta^{(t)})$, 且等号在$\theta^{(t+1)} = \theta^{(t)}$时成立. 那么既然如此, 我们直接假设第$t+1$步的$p(X|\theta^{(t+1)}) = Q(\theta, \theta^{(t)})$. 因为$Q(\theta, \theta^{(t)}) \geq p(X|\theta^{(t)})$. 那么我们就变得可以很主动，如果我们能找到一个$\theta$使得$Q(\theta, \theta^{(t)})$比$Q(\theta^{(t)}, \theta^{(t)})$ 更大, 那么我们用新的θ替换掉$\theta^{(t)}$, 反之则$\theta^{(t+1)} = \theta^{(t)}$, 而这时候也说明迭代已经收敛了. 所以关键的问题变成了 第$t+1$步的ELBO 是否**可以**等于$p(X|\theta)$第t步的最大值, 其实也就是我们是否**总能**找到一个$\theta^{(t+1)}$使得$p(X|\theta^{(t+1)}) \geq p(X|\theta^{(t)})$. 这个证明又被称为EM的收敛性证明. 因为这个证明其实还挺复杂的, 所以我会在后面补充进来, 但这个命题是成立的.

到这里EM算法的核心部分就已经结束了, 整个算法的逻辑很绕, 但当最后想通以后却发现又是那么简单. 真的可以称得上是一个非常优雅, 精致并令人着迷的算法了. 在接下来的一部分我们将运用到目前为止的推导解答开篇的时候提出的投掷硬币的问题.

三硬币模型

正如上面描述的那样, 我们假设已知$\theta^{(0)} = \{\alpha = 0.5, \beta = 0.5, \gamma = 0.5\}$, 我们列出我们第1步的ELBO.

$$
\begin{aligned}
Q(\theta,\theta^{(0)}) &= \sum_Z \log\frac{p(X,Z|\theta)}{p(Z|X,\theta^{(0)})}\cdot p(Z|X,\theta^{(0)})\,dZ \\
&= \sum_Z \log p(X,Z|\theta)\cdot p(Z|X,\theta^{(0)})\,dZ - \sum_Z \log p(Z|X,\theta^{(0)})\cdot p(Z|X,\theta^{(0)})\,dZ
\end{aligned}
$$

由于$\sum_Z \log p(Z|X,\theta^{(0)})\cdot p(Z|X,\theta^{(0)})\,dZ$中不包含$\theta$, 也就是说对于估计$\theta$的值没有什么影响. 所以在这里看成一个任意常数, 比如说0. 加上每一次试验都是独立的. 所以$p(X,Z|\theta)=\prod p(X_i,Z_i|\theta)$, $p(Z|X,\theta^{(0)})=\prod p(Z_i|X_i,\theta^{(0)})$.

$$
\begin{aligned}
Q(\theta,\theta^{(0)}) &= \sum_Z \log p(X,Z|\theta)\cdot p(Z|X,\theta^{(0)})\,dZ \\
&= \sum_Z \log[\prod_{i=1}^N p(X_i,Z_i|\theta)]\cdot[\prod_{i=1}^N p(Z_i|X_i,\theta^{(0)})]\,dZ \\
&= \sum_Z [\log p(X_1,Z_1|\theta)+\log p(X_2,Z_2|\theta)+\cdots]\cdot[\prod_{i=1}^N p(Z_i|X_i,\theta^{(0)})]\,dZ \\
&= \sum_Z \log p(X_1,Z_1|\theta)\cdot[\prod_{i=1}^N p(Z_i|X_i,\theta^{(0)})]\,dZ+\cdots
\end{aligned}
$$

为了方便我们现在只讨论X_1这一项. 注意到$\sum_Z$可以看成是一个对$Z_1,\ldots,Z_N$的多元函数的积分

$$
\begin{aligned}
&\sum_Z \log p(X_1,Z_1|\theta)\cdot[\prod_{i=1}^N p(Z_i|X_i,\theta^{(0)})]\,dZ \\
=&\sum_{Z_1} \log p(X_1,Z_1|\theta)\cdot p(Z_1|X_1,\theta^{(0)})\sum_{Z_2} p(Z_2|X_2,\theta^{(0)})\cdots\sum_{Z_N} p(Z_N|X_N,\theta^{(0)})
\end{aligned}
$$

观察到$\sum_{Z_i} p(Z_i|X_i,\theta^{(0)})$也可以看做是在整个概率空间中的累积概率. 所以这个式子等于1. 就有

$$
\begin{aligned}
&\sum_Z \log p(X_1,Z_1|\theta)\cdot[\prod_{i=1}^N p(Z_i|X_i,\theta^{(0)})]\,dZ \\
=&\sum_{Z_1} \log p(X_1,Z_1|\theta)\cdot p(Z_1|X_1,\theta^{(0)}) \\
&Q(\theta,\theta^{(0)}) \\
=&\sum_{Z_1} \log p(X_1,Z_1|\theta)\cdot p(Z_1|X_1,\theta^{(0)})+\cdots+\sum_{Z_N} \log p(X_N,Z_N|\theta)\cdot p(Z_N|X_N,\theta^{(0)})
\end{aligned}
$$

看到这里其实我们已经成功了一大半了, 现在不如坐下来喝杯咖啡, 稍微休息下将问题解决吧.

现在, 我们试图将已有的数据代入表示一下这个式子, 我们发现好像还是表示不出来. 因为我们不知道应该如何表示$p(X_i,Z_i|\theta)$和$p(Z_i|X_i,\theta)$.

整理一下我们的已知信息, 我们现在知道

$$
\begin{aligned}
p(Z_i=A|\theta^{(0)}) &= \gamma = 0.5 \\
p(Z_i=B|\theta^{(0)}) &= 1-\gamma = 0.5 \\
p(X_i=1|Z_i=A,\theta^{(0)}) &= \alpha = 0.5 \\
p(X_i=1|Z_i=B,\theta^{(0)}) &= \beta = 0.5 \\
p(X_i=0|Z_i=A,\theta^{(0)}) &= 1-\alpha = 0.5 \\
p(X_i=0|Z_i=B,\theta^{(0)}) &= 1-\beta = 0.5
\end{aligned}
$$

由$p(X,Z|\theta)=p(X|Z,\theta)p(Z|\theta)$和全概率公式, 我们可以得到

$$
\begin{cases}
p(X_i,Z_i|\theta^{(0)}) = p(X_i|Z_i,\theta^{(0)})p(Z_i|\theta^{(0)}) \\
p(Z_i|X_i,\theta^{(0)}) = \dfrac{p(X_i|Z_i,\theta^{(0)})p(Z_i|\theta^{(0)})}{p(X_i|Z=A,\theta^{(0)})p(Z=A|\theta^{(0)})+p(X_i|Z=B,\theta^{(0)})p(Z=B|\theta^{(0)})}
\end{cases}
$$

为了更具体, 我们现在只看第一次试验的结果X_1.

$$
\begin{aligned}
&\sum_{Z_1} \log p(X_1, Z_1|\theta) p(Z_1|X_1, \theta^{(0)}) \\
=&\log \frac{p(1|A,\theta)}{p(A|\theta)} \frac{p(1|A,\theta^{(0)})p(A|\theta^{(0)})}{p(1|A,\theta^{(0)})p(A|\theta^{(0)}) + p(1|B,\theta^{(0)})p(B|\theta^{(t)})} \\
+&\log \frac{p(1|B,\theta)}{p(B|\theta)} \frac{p(1|B,\theta^{(0)})p(B|\theta^{(0)})}{p(1|A,\theta^{(0)})p(A|\theta^{(0)}) + p(1|B,\theta^{(0)})p(B|\theta^{(t)})} \\
=&0.5 \cdot [\log p(1|A,\theta) - \log p(A|\theta) + \log p(1|B,\theta) - \log p(B|\theta)]
\end{aligned}
$$

这里稍微解释一下, 因为$Z_1 = A \ or \ B$. 所以$\sum_{Z_1}$就是$Z_1 = A$的情况和$Z_1 = B$的情况. 这个式子最好看做是对Z_1以Z_1中每种情况为未知数的多元函数积分, 有很多的构造都是这么来的. 当然这里情况比较简单, 所以也就直接列出来了.

继续观察我们还可以发现, 其实对于求解这个问题, X的顺序是没关系的, 也就是说, 我们只需要知道试验结果中$X = 1$的个数和$X = 0$的个数就可以完整表达出ELBO了.

$$
\begin{aligned}
L(\theta^{(1)}) &= \log p(X|\theta^{(1)}) \\
&= 0.5(6 \cdot [\log \frac{\alpha}{\gamma} + \log \frac{\beta}{1-\gamma}] + 4 \cdot [\log \frac{1-\alpha}{\gamma} + \log \frac{1-\beta}{1-\gamma}]) \\
&= 3 \cdot [\log \frac{\alpha}{\gamma} + \log \frac{\beta}{1-\gamma}] + 2 \cdot [\log \frac{1-\alpha}{\gamma} + \log \frac{1-\beta}{1-\gamma}] \\
&= 3 \cdot [\log \alpha + \log \beta - \log \gamma - \log(1-\gamma)] + 2 \cdot [\log(1-\alpha) + \log(1-\beta) - \log \gamma - \log(1-\gamma)] \\
&= 3\log \alpha + 2\log(1-\alpha) + 3\log \beta + 2\log(1-\beta) - 5\log \gamma - 5\log(1-\gamma)
\end{aligned}
$$

通过求偏导的方法, 我们得到一组θ, 使$L(\theta^{(1)})$最大.

$$
\begin{cases}
\gamma = 0.5 \\
\alpha = 0.6 \\
\beta = 0.6
\end{cases}
$$

由于在这个问题中为了运算简便, 我们假设的$\theta^{(0)}$数值比较特殊, 所以有些细节并没有展示出来. 我写了一个简单的代码用来解决这个问题. 更多细节结合参考代码来看.