《机器学习系列-强填EM算法在理论与工程之间的鸿沟（上）》

原创 夕小瑶 夕小瑶的卖萌屋 2017-03-11

小夕曾经问一位做机器学习理论的学姐：“学姐学姐，EM算法是什么呢？”

学姐回答：“EM算法啊，就是解决包含隐变量的参数估计问题。”

小夕：

然后小夕去问一位做工程的学长：“学长学长，EM算法是什么呢？”

学长回答：“EM算法就那样啊，就是先用有标签的样本训练个分类器，再给未知标签的样本贴标签，然后再拿全部样本再训练分类器，就这样来回倒腾~”

小夕：

于是小夕自己一个人看了一整天的EM算法QAQ

前言

首先说，其实学长和学姐说的都很对。但是对于一个路人来说，很难将学长与学姐的说法联系到同一个东西上。而最终小夕总结出来的就是，**做工程的学长的回答其实是做理论的学姐的回答下的一个简化的特例。**

首先，我们来看一下理论上的期望最大化算法，也就是EM算法（不要想了，对于这个算法，小夕打死也绕不开数学公式了，所以有公式恐惧症的同学请自行用手指盖住它们...

另外，严正声明一下，对于没有微积分与概率统计基础的同学，请直接等下一篇中得出的结论！非要看这一篇的话，请时刻保持理智，请时刻保持理智，请时刻保持理智。

理论家眼中的EM

开门见山，EM算法的目标是**使包含隐变量的数据集的后验概率或似然函数最大化，进而得到最优的参数估计。**

我们知道，通过贝叶斯公式，可以发现后验概率中包含了似然函数和先验概率（忽略分母的那个evidence项），因此求最大后验概率的过程中包含了求极大似然估计的过程。因此虽然EM算法的目标是最大化后验概率或似然函数，而本质上就可以认为是最大化似然函数。因此下面我们直接讨论最大化似然函数。

似然函数设为l(θ)，描述样本可以用多维随机变量（对应于机器学习的多维特征），每一维的随机变量都可以认为服从某种概率分布。因此要描述每一维的样本情况，我们只需要估计出这一维度的概率分布模型的参数就可以啦。而将所有维度的分布模型的参数放在一起，就是似然函数的参数，即θ。因此根据定义，

$$\ell(\theta) = \sum_{i=1}^{m} \log p(x;\theta)$$

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

即似然函数代表着该包含m个样本的样本集存在的合理性（似然函数值越大，该样本集的存在就越合理，即意味着参数取的越正确），描述每个样本的多维随机变量的分布模型的参数即上面的θ，p(x; θ)代表着固定θ的值，求p(x)的概率。

第二行的z则代表隐变量，确切的说是**隐含的随机变量**。哈？看不懂第二步怎么来的？请回去复习微积分...算了，小夕太过善良，还是讲讲吧。

> 显然，这里似然函数讨论的是离散情况（毕竟都是∑符号而不是∫符号呀），因此，在p(x; θ)中加上z这个随机变量后，只能将这个随机变量积分掉才能保证加上z以后的式子依然等于p(x;θ)，当然，z是离散的，所以积分掉的意思是“求和”掉。
>
> （回顾一下，对于任何一个连续随机变量x，∫p(x)dx=1；对于任何一个离散随机变量x，∑p(x)=1）

好，懂了第二步，在继续往下推之前，想一想我们可不可以直接计算第二步呢？当然不行啦，不仅有θ，还有隐变量啊。因此继续往下推。

$$\sum_i \log p(x^{(i)};\theta) = \sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)};\theta)$$

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

诶?又出来个Qi。这个Qi是什么呢?这个Qi是隐变量z的概率分布函数啦。为什么要加上它呢?再好好观察一下最后这一步中的这一部分!

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

有没有发现什么!?对!这就是数学期望呀~别说数学期望都忘了啊,小夕还是再啰嗦一下吧...对于某离散随机变量X来说,其数学期望

$$\mathrm{E}[X] = \sum_i p_i x_i$$

看吧~加上Qi这个概率分布函数后,是不是就出来了一个数学期望啦!但好像还是不能计算,懂数值计算的读者应该知道log(∑...)的计算量是十分恐怖的,而且我们还被我们加上了一个不知道怎么计算的Qi!!!因此要继续变!!!怎么变呢?Jensen不等式来啦!

直接抠了个定义(看不懂没关系):

如果f是一个凸函数，X是一个随机变量，那么：

$$E[f(X)] \geq f(E[X])$$

如果f是严格凸函数，那么只有当$X = E[X]$恒成立时，上式取等号(即X是一个常量)，或者说当$X = E[X]$成立的概率是1时，上式取等号。

通过这个Jensen不等式，我们发现可以进一步往下推了。

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

诶？虽然是往下推了一步，但是我们必须要让等号恒成立才行啊，否则这个推理是不成立的呀。。。那么怎么让等号恒成立呢？

根据Jensen不等式的等号成立条件，E[f(X)]≥f(E[X])中的随机变量X必须恒等于常数！！也就是说：

$$\frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} \equiv c \text{（c为常数）}$$

于是重点来了，将分母的Qi移到右边，将右边的c移到左边！我们发现：

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

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

好，再利用 （概率分布函数的基本性质），发现我们可以继续这样推！

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

推到最后竟然是这个?????

这个不就是每个样本的隐变量z的后验概率吗!!!

也就是说我们只要求出来了每个样本的隐变量的每个取值的后验概率，就能得到这个样本的Qi!!!

就能让Jensen不等式的等号成立!!!

就能让log(∑...)的不可计算成功变成可计算!!!

就能计算我们的目标——似然函数啦!!!

所以，咳咳，总之，我们首先固定一个θ（也就是随便给θ取个初始值），然后我们计算出隐变量z的取值的后验概率，就能让这个包含隐变量的似然函数变成传统意义上的似然函数~也就是只考虑参数θ的似然函数~**（这个过程称为E步）**

而最大化传统意义上的似然函数就不用啰嗦啦~那就用传统的方法最大化呀~最大化了以后就得到了当前的最优θ。**（这个过程称为M步）**

而得到了当前的最优θ以后，我们又可以重新计算出隐变量z的取值的后验概率，就能……~~~总之就又可以E步，然后又M步，然后又E，又M……

就这样一直重复，一直重复，直到似然函数的值不再变化，此时每个样本的Qi就是每个样本的标签~而此时的θ就是最终那个最优的θ啦~

至此，理论上的EM算法完成了，最终得到的就是我们要估计的最优参数θ，顺便得到了每个样本的隐变量的取值。

那么工程上看似是跟分类器打交道，小夕则说其实是理论的特例又是怎么回事呢？敬请期待《机器学习系列-强填EM算法在理论与工程之间的鸿沟（下）》，待小夕华丽丽的填上理论与工程的鸿沟。（下一篇没有这一篇这么恐怖，2333）

虽然您可能没有看懂，但是看在生敲公式后发现微信编辑器不识别然后又一个个截图的份上_(:3」∠)_