15 Mar

从最大似然到EM算法：一致的理解方式

By 苏剑林 | 2018-03-15 | 41564位读者 引用

最近在思考NLP的无监督学习和概率图相关的一些内容，于是重新把一些参数估计方法理了一遍。在深度学习中，参数估计是最基本的步骤之一了，也就是我们所说的模型训练过程。为了训练模型就得有个损失函数，而如果没有系统学习过概率论的读者，能想到的最自然的损失函数估计是平均平方误差，它也就是对应于我们所说的欧式距离。而理论上来讲，概率模型的最佳搭配应该是“交叉熵”函数，它来源于概率论中的最大似然函数。

最大似然

合理的存在

何为最大似然？哲学上有句话叫做“存在就是合理的”，最大似然的意思是“存在就是最合理的”。具体来说，如果事件X的概率分布为$p(X)$，如果一次观测中具体观测到的值分别为$X_1, X_2, \dots, X_n$，并假设它们是相互独立，那么

$$\mathcal{P} = \prod_{i=1}^{n} p(X_i) \tag{1}$$

是最大的。如果$p(X)$是一个带有参数θ的概率分布式$p_\theta(X)$，那么我们应当想办法选择θ，使得$\mathcal{L}$最大化，即

$$\theta = \mathop{\arg\max}_{\theta} \mathcal{P}(\theta) = \mathop{\arg\max}_{\theta} \prod_{i=1}^{n} p_\theta(X_i) \tag{2}$$

对概率取对数，就得到等价形式

$$\theta = \mathop{\arg\max}_{\theta} \sum_{i=1}^{n} \log p_\theta(X_i) \tag{3}$$

如果右端再除以n，我们就得到更精炼的表达形式

$$\theta = \mathop{\arg\max}_{\theta} \mathcal{L}(\theta) = \mathop{\arg\max}_{\theta} \mathbb{E}\big[\log p_\theta(X_i)\big] \tag{4}$$

其中我们将$-\mathcal{L}(\theta)$就称为交叉熵。

理论形式

理论上，根据已有的数据，我们可以得到每个X的统计频率$\tilde{p}(X)$，那么可以得到上式的等价形式

$$\theta = \mathop{\arg\max}_{\theta} \mathcal{L}(\theta) = \mathop{\arg\max}_{\theta} \sum_{X} \tilde{p}(X) \log p_\theta(X) \tag{5}$$

但实际上我们几乎都不可能得到$\tilde{p}(X)$（尤其是对于连续分布），我们能直接算的是关于它的数学期望，也就是(4)式，因为求期望只需要把每个样本的值算出来，然后求和并除以n就行了。所以(5)式只有理论价值，它能方便后面的推导。

要注意的是，上面的描述是非常一般的，其中X可以是任意对象，它也有可能是连续的实数，这时候就要把求和换成积分，把$p(X)$变成概率密度函数。当然，这并没有什么本质困难。

更广泛的KL散度

从KL散度出发也可以导出最大似然的形式来。假如两个分布$\tilde{p}(X)$和$p(X)$，我们用KL散度来衡量它们的距离：

$$\begin{aligned}KL\Big(\tilde{p}(X)\Big\Vert p(X)\Big) =& \sum_X \tilde{p}(X)\ln \frac{\tilde{p}(X)}{p(X)}\\ =&\mathbb{E}\left[\ln \frac{\tilde{p}(X)}{p(X)}\right]\end{aligned}\tag{6}$$

当两个分布相同时，KL散度为0，当两个分布不同时，KL散度大于0，假设读者已经知道这些性质。

接着假设X的样本已经给出来了，这就意味着$\tilde{p}(X)$可以视为已知了，这时候：

$$\begin{aligned}\theta =&\mathop{\arg\min}_{\theta} KL\Big(\tilde{p}(X)\Big\Vert p_{\theta}(X)\Big)\\ =&\mathop{\arg\max}_{\theta} \sum_X \tilde{p}(X)\log p_{\theta}(X)\\ =&\mathop{\arg\max}_{\theta} \mathbb{E}\big[\log p_{\theta}(X_i)\big]\end{aligned}\tag{7}$$

这就重新导出了(4)和(5)。事实上KL散度要比简单的最大似然含义更为丰富，因为最大似然相当于假设了$\tilde{p}(X)$是已知的（已知X的样本），这并不总是能实现的（比如EM算法的场景），很多时候我们只知道X的部分信息，这时候就要回归到KL散度中来。

注：如果读者不能很好地理解采样计算，请阅读《变分自编码器（二）：从贝叶斯观点出发》中的《数值计算vs采样计算》一节。

有监督模型

现在我们来观察有监督学习中是如何应用上述内容的。假设输入为X，标签为Y，那么(X,Y)就构成了一个事件，于是我们根据(4)就有

$$\theta = \mathop{\arg\max}_{\theta} \mathbb{E}_{X,Y}\big[\log p_{\theta}(X,Y)\big]\tag{8}$$

这里已经注明了是对X,Y整体求数学期望，然而该式却是不够实用的。

分类问题

以分类问题为例，我们通常建模的是$p(Y|X)$而不是$p(X,Y)$，也就是我们要根据输入确定输出的分布，而不是它们的联合分布。所以我们还是要从(5)式出发，利用$p(X,Y)=p(X)p(Y|X)$，先得到

$$\theta = \mathop{\arg\max}_{\theta} \sum_{X,Y} \tilde{p}(X,Y)\log\big[p_\theta(X)p_\theta(Y|X)\big] \tag{9}$$

因为我们只对$p(Y|X)$建模，因此$p_\theta(X)$我们认为就是$\tilde{p}(X)$，那么这相当于让优化目标多了一个常数项，因此(9)等价于

$$\theta = \mathop{\arg\max}_{\theta} \sum_{X,Y} \tilde{p}(X,Y)\log p_\theta(Y|X) \tag{10}$$

然后，我们还有$\tilde{p}(X,Y)=\tilde{p}(X)\tilde{p}(Y|X)$，于是(8)式还可以再变化成

$$\theta = \mathop{\arg\max}_{\theta} \sum_{X} \tilde{p}(X)\sum_{Y}\tilde{p}(Y|X)\log p_\theta(Y|X) \tag{11}$$

最后别忘了，我们是处理有监督学习中的分类问题，一般而言在训练数据中对于确定的输入X就只有一个类别，所以$\tilde{p}(Y_t|X)=1$，其余为0，Y_t就是X的目标标签，所以

$$\theta = \mathop{\arg\max}_{\theta} \sum_{X} \tilde{p}(X)\log p_\theta(Y_t|X) \tag{12}$$

这就是最常见的分类问题的最大似然函数了：

$$\theta = \mathop{\arg\max}_{\theta} \mathbb{E}_X\big[\log p_\theta(Y_t|X)\big] \tag{13}$$

变变变

事实上，上述的内容只是一些恒等变换，应该说没有特别重要的价值，而它的结果（也就是分类问题的交叉熵损失）也早就被我们用得滚瓜烂熟了。因此，这一节仅仅是展示了如何将最大似然函数从最原始的形式出发，最终落实到一个具体的问题中，让读者熟悉一下这种逐步推进的变换过程。

隐变量

现在就是展示它的价值的时候了，我们要将**用它来给出一个EM算法的直接推导**（本博客还提供了另外一个理解角度，参考《梯度下降和EM算法：系出同源，一脉相承》）。对于EM算法，一般将它分为M步和E步，应当说，M步是比较好理解的，难就难在E步的那个Q函数为什么要这样构造。很多教程并没有给出这个Q函数的解释，有一些教程给出了基于詹森不等式的理解，但我认为这些做法都没有很好凸显出EM算法的精髓。

一般来说，EM算法用于存在隐变量的概率问题优化。什么是隐变量？很简单，还是以刚才的分类问题为例，分类问题要建模的是$p(Y|X)$，当然也等价于$p(X,Y)$，我们说过要用最大似然函数为目标，得到(8)式

$$\theta = \mathop{\arg\max}_{\theta} \mathbb{E}_{X,Y}\big[\log p_\theta(X,Y)\big] \tag{8}$$

如果给出(X,Y)的标签数据对，那就是一个普通的有监督学习问题了，然而如果只给出X不给出Y呢？这时候

Y就称为隐变量，它存在，但我们看不见，所以“隐”。

GMM模型

等等，没有标签数据你也想做分类问题？当然有可能，GMM模型不就是这样的一个模型了吗？在GMM中假设了

$$p_\theta(X,Y)=p_\theta(Y)p_\theta(X|Y) \tag{14}$$

注意，是$p_\theta(Y)p_\theta(X|Y)$而不是$p_\theta(X)p_\theta(Y|X)$，两者区别在于我们难以直接估计$p(X)$，也比较难直接猜测$p(Y|X)$的形式。而$p(Y)$和$p(X|Y)$就相对容易了，因为我们通常假设Y的意义是类别，所以$p(Y)$只是一个有限向量，而$p(X|Y)$表示每个类内的对象的分布，既然这些对象都属于同一个类，同一个类应该都长得差不多吧，所以GMM假设它为正态分布，这时候做的假设就有依据了，不然将所有数据混合在一起，谁知道假设什么分布好呢？

这种情况下，我们完整的数据应该是(X,Y)，但我们并没有这种成对的样本$(X_1,Y_1),\dots,(X_n,Y_n)$（不然就退化为有监督学习了），我们只知道$X$的样本$X_1,\dots,X_n$，这就对应了我们在KL散度这一节描述的情形了。

pLSA模型

当然，并不只有无监督学习才有隐变量，有监督学习也可以有，比如我们可以设

$$p(Y|X)=\sum_Z p_\theta(Y|Z)p_\theta(Z|X) \tag{15}$$

这时候多出了一个变量Z，就算给出(X,Y)这样的标签数据对，但Z仍然是没有数据的，是我们假想的一个变量，它也就是隐变量，pLSA就是这样的一个问题。也就是说，这时候完整的数据对应该是(X,Y,Z)的形式，但我们只知道$(X_1,Y_1),\dots,(X_n,Y_n)$这样的部分样本。

贝叶斯学派

可能有读者“异想天开”：那么参数θ是不是也可以看作一个隐变量呢？恭喜你，如果你有这层领悟，那你已经进入贝叶斯学派的思维范畴了。贝叶斯学派认为，一切都是随机的，一切都服从某个概率分布，参数θ也不例外。不过很遗憾，贝叶斯学派的概率理论很艰深，我们这里还没法派上用场。（其实更重要的是，笔者也还不懂～～）

EM算法

好了，不再废话了，还是正式进入对EM算法的讨论吧。

联合KL散度

我们先来看一下，对于含有隐变量的问题求解，一般教程的处理方案是这样的：由于隐变量不可观测，因此一般改用边缘分布（也就是显变量的分布）的最大似然为目标函数，即

$$\theta=\mathop{\arg\max}_\theta \sum_X \tilde{p}(X)\log\sum_Z p_\theta(X|Z)p_\theta(Z) \tag{16}$$

为最大化的目标。

这种做法不是不行，而是这样一来为了得到EM算法就需要引入比较多的数学知识，而且严格证明还需要比较冗长的推导。事实上可以从KL散度出发，通过分析联合概率分布的KL散度来极大简化EM算法的推导。而如果采用边缘分布最大似然的做法，我们就无法直观地理解那个Q函数的来源了。

以GMM为例，首先我们来算$\tilde{p}(X,Y)$和$p_\theta(X,Y)$的KL散度：

$$\begin{aligned}&KL\Big(\tilde{p}(X,Y)\Big\Vert p_\theta(X,Y)\Big)\\=&\sum_{X,Y}\tilde{p}(X,Y)\log\frac{\tilde{p}(X,Y)}{p_\theta(X,Y)}\\=&\sum_X\tilde{p}(X)\sum_Y\tilde{p}(Y|X)\log\frac{\tilde{p}(Y|X)\tilde{p}(X)}{p_\theta(X|Y)p_\theta(Y)}\\=&\mathbb{E}_X\left[\sum_Y\tilde{p}(Y|X)\log\frac{\tilde{p}(Y|X)\tilde{p}(X)}{p_\theta(X|Y)p_\theta(Y)}\right]\\=&\mathbb{E}_X\left[\sum_Y\tilde{p}(Y|X)\log\frac{\tilde{p}(Y|X)}{p_\theta(X|Y)p_\theta(Y)}+\sum_Y\tilde{p}(Y|X)\log\tilde{p}(X)\right]\\=&\mathbb{E}_X\left[\sum_Y\tilde{p}(Y|X)\log\frac{\tilde{p}(Y|X)}{p_\theta(X|Y)p_\theta(Y)}\right]+\mathbb{E}\left[\log\tilde{p}(X)\right]\\=&\mathbb{E}_X\left[\sum_Y\tilde{p}(Y|X)\log\frac{\tilde{p}(Y|X)}{p_\theta(X|Y)p_\theta(Y)}\right]+\text{常数}\end{aligned}\tag{17}$$

这个过程虽然比较长，但并没有什么迂回的变换，是比较容易接受的。

EM大佬来了

再次回顾(17)式的来源，我们希望找到一组分布的参数θ，使得$KL\Big(\tilde{p}(X,Y)\Big\Vert p_\theta(X,Y)\Big)$越小越好，$p_\theta(X,Y)$我们已经给出为$p_\theta(X|Y)p_\theta(Y)$的形式，只有参数$\theta$是未知的。但是在$(17)$式中，$\tilde{p}(Y|X)$也是未知的，包括形式。

这时候，大佬就发话了：先当它已知的吧，这时候$\tilde{p}(Y|X)$可以视为常数，那么我们就可以算参数θ了：

$$\begin{aligned}\theta^{(r)}=&\mathop{\arg\min}_{\theta}\mathbb{E}_X\left[\sum_Y\tilde{p}^{(r-1)}(Y|X)\log\frac{\tilde{p}^{(r-1)}(Y|X)}{p_\theta(X|Y)p_\theta(Y)}\right]\\=&\mathop{\arg\max}_{\theta}\mathbb{E}_X\left[\sum_Y\tilde{p}^{(r-1)}(Y|X)\log p_\theta(Y)p_\theta(X|Y)\right]\end{aligned}\tag{18}$$

然后这时候算出了新的$\theta^{(r)}$，我们把$p_\theta(X|Y)$当成已知的，来求$\tilde{p}(Y|X)$，

$$\tilde{p}^{(r)}(Y|X)=\mathop{\arg\min}_{\tilde{p}(Y|X)}\mathbb{E}_X\left[\sum_Y\tilde{p}(Y|X)\log\frac{\tilde{p}(Y|X)}{p_{\theta^{(r)}}(X|Y)p_{\theta^{(r)}}(Y)}\right]\tag{19}$$

事实上(19)式是可以直接写出解析解的，答案是：

$$\tilde{p}^{(r)}(Y|X) = \frac{p_{\theta^{(r)}}(Y)p_{\theta^{(r)}}(X|Y)}{\sum\limits_Y p_{\theta^{(r)}}(Y)p_{\theta^{(r)}}(X|Y)} \tag{20}$$

补充推导：(19)式方括号内的部分，可以改写为

$$\begin{aligned}&\sum_Y \tilde{p}(Y|X)\log\frac{\tilde{p}(Y|X)}{p_{\theta^{(r)}}(X,Y)}\\ =&\sum_Y \tilde{p}(Y|X)\log\frac{\tilde{p}(Y|X)}{p_{\theta^{(r)}}(Y|X)} - \sum_Y \tilde{p}(Y|X)\log p_{\theta^{(r)}}(X)\\ =&KL\Big(\tilde{p}(Y|X)\Big\Vert p_{\theta^{(r)}}(Y|X)\Big) - \text{常数}\end{aligned}$$

所以最小化(19)式也就相当于最小化$KL\Big(\tilde{p}(Y|X)\Big\Vert p_{\theta^{(r)}}(Y|X)\Big)$，根据KL散度的性质，显然最优解就是两个分布完全一致，即

$$\tilde{p}(Y|X) = p_{\theta^{(r)}}(Y|X) = \frac{p_{\theta^{(r)}}(Y)p_{\theta^{(r)}}(X|Y)}{\sum\limits_Y p_{\theta^{(r)}}(Y)p_{\theta^{(r)}}(X|Y)}$$

这就得到了(20)式。

因为我们没法一步到位求(17)的最小值，所以现在就将它交替地训练：先固定一部分，最大化另外一部分，然后交换过来。**EM算法就是对复杂目标函数的交替训练方法**！

联合(18)式和(20)式，就构成了整个求解算法。现在来看看(18)式，**有个E（求期望），又有个M（$\arg\max$），就叫它EM算法吧，那个被E的式子，我们就叫它Q函数好了**。于是EM大佬就这样出来了，Q函数也出来了，就这么任性...

当然，EM算法中的E的本意是将$\sum\limits_Y \tilde{p}^{(r-1)}(Y|X)\log p_\theta(Y)p_\theta(X|Y)$看成是对隐变量$Y$求期望，这里我们就随意一点的，结论没错就行～

是不是感觉很突然？感觉啥也没做，EM算法就这么两句话说清楚了？还包括了推导？

究竟在做啥

对于pLSA或者其他含有隐变量的模型的EM算法，也可以类似地推导。对比目前我能找到的EM算法的推导，我相信上面的过程已经是相当简洁了。尽管前面很多铺垫，但其实都是基础知识而已。

那这是如何实现的呢？回顾整个过程，其实我们也没做什么，只是纯粹地使用KL散度作为联合分布的差异性度量，然后对KL散度交替最小化罢了～这样子得到的推导，比从边缘分布的最大自然出发，居然直接快捷了很多，也是个惊喜。

一致的理解

本文是作者对最大似然原理的一翻思考，整体思路是从最大似然的原理和形式出发，来诱导出有监督/无监督学习的一些结果，希望能用一个统一的思想将各种相关内容都串起来。最后发现结果也挺让人满意的，尤其是EM算法部分，以后只需要记住一切的根本都是（联合）分布的最大似然或KL散度，再也不用死记EM算法中的Q函数形式了。

当然，文章有些观点都是“我认为”的，因此可能有不当之处，请读者甄别。不过可以保证的是结果跟现有的都是一样的。欢迎读者继续交流～

如果您需要引用本文，请参考：

苏剑林. (2018, Mar 15). 《从最大似然到EM算法：一致的理解方式 》[Blog post]. Retrieved from https://kexue.fm/archives/5239