从最大似然到EM算法

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

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 $X1, X2, \dots, XN$，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 L 最大化，即：

$$\theta = \arg\max_\theta \mathcal{P}(\theta) = \arg\max_\theta \prod_{i=1}^{N} p_\theta(X_i) \tag{2}$$

对概率取对数，就得到等价形式：

$$\theta = \arg\max_\theta \sum_{i=1}^{N} \log p_\theta(X_i) \tag{3}$$

如果右端再除以 N，我们就得到更精炼的表达形式：

$$\theta = \arg\max_\theta \mathcal{L}(\theta) = \arg\max_\theta \mathbb{E}\left[\log p_\theta(X_i)\right] \tag{4}$$

其中我们将 $-L(\theta)$ 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 p（X)，那么可以得到上式的等价形式：

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

但实际上，我们几乎都不可能得到 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。

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：

$$\theta = \arg\max_{\theta} \mathbb{E}_{X,Y}\Big[\log p_{\theta}(X,Y)\Big] \tag{6}$$

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：

$$\theta = \arg\max_{\theta} \sum_{X,Y} \tilde{p}(X,Y) \log \left[p_\theta(X) p_\theta(Y|X)\right] \tag{7}$$

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 (7) 等价于：

$$\theta = \arg\max_{\theta} \sum_{X,Y} \tilde{p}(X,Y) \log p_\theta(Y|X) \tag{8}$$

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：

$$\theta = \arg\max_{\theta} \sum_{X} \tilde{p}(X) \sum_{Y} \tilde{p}(Y|X) \log p_\theta(Y|X) \tag{9}$$

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：

$$\theta = \arg\max_{\theta} \sum_{X} \tilde{p}(X) \log p_\theta(Y_t|X) \tag{10}$$

这就是最常见的分类问题的最大似然函数了：

$$\theta = \arg\max_{\theta} \mathbb{E}_X \left[\log p_\theta(Y_t|X)\right] \tag{11}$$

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。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 (6) 式：

$$\theta = \arg\max_{\theta} \mathbb{E}_{X,Y}\left[\log p_{\theta}(X,Y)\right] \tag{6}$$

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：

$$p_{\theta}(X,Y) = p_{\theta}(Y)p_{\theta}(X|Y) \tag{12}$$

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？

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：

$$p(Y|X) = \sum_Z p_\theta(Y|Z)p_\theta(Z|X) \tag{13}$$

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。这时候它的最大似然估计是：

$$\theta = \arg\max_\theta \mathbb{E}_{X,Y,Z}\left[\log p_\theta(Y|Z)p_\theta(Z|X)p_\theta(X)\right] \tag{14}$$

联合最大似然

再等等，你这个好像跟我之前看到的 pLSA 的目标函数不大一样呀？还有 (6) 式也跟 GMM 的目标函数不一样呀？你是不是弄错了？

我觉得并没有弄错，**最大似然函数应该要考虑的是整体联合分布，也就是得把 Z 也考虑进去**。而教程一般是这样处理的：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：

$$\theta = \arg\max_\theta \sum_X \tilde{p}(X) \log \sum_Z p_\theta(X|Z)p_\theta(Z) \tag{15}$$

为最大化的目标。

事实上这种做法我认为是不大妥当的，**隐变量虽然“隐”了，但既然我们假设它存在，那么它就是真的存在了，既然真的存在，最大似然函数当然要考虑上它，这才是真正的“存在就是最合理的”，是连同隐变量一起最合理才对**：

$$
\begin{aligned}
\theta =& \arg\max_{\theta} \sum_{X,Y} \tilde{p}(X,Y) \log p_\theta(X,Y) \\
=& \arg\max_{\theta} \sum_{X,Y} \tilde{p}(X,Y) \log p_\theta(X|Y) p_\theta(Y)
\end{aligned} \tag{16}
$$

而事实上这种处理不仅具有理论意义，它还极大简化了 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。

最后，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。

EM算法

好了，不再废话了，还是正式进入对 EM 算法的讨论吧。

再变变变

以式 (6) 的模型为例，假设我们只有 X 的数据，没有对应的标签 Y，这时候 Y 是隐变量，但我们还是要算整体的最大似然，也就是前面的 (16) 式：

$$
\theta = \arg\max_{\theta} \sum_{X,Y} \tilde{p}(X,Y) \log p_\theta(X|Y) p_\theta(Y) \tag{16}
$$

这时候我们依然没有解决的问题是：我们不知道 p（X,Y)，甚至 p（X) 我们也可能不知道（但我们可以算关于它的期望）。那好吧，将式子做一下变换：

$$
\begin{aligned}
\theta =&\arg\max_{\theta} \sum_X \tilde{p}(X) \sum_Y \tilde{p}(Y|X) \log p_\theta(X) p_\theta(Y|X) \\
=&\arg\max_{\theta} \mathbb{E}\left[\sum_Y \tilde{p}(Y|X) \log p_\theta(Y) p_\theta(X|Y)\right]
\end{aligned} \tag{17}
$$

这里的 $\mathbb{E}$ 是对 X 求的期望。现在好像有点意思了，然而并没有什么用，因为 p（$Y|X$）还是未知的。

EM大佬来了

这时候，大佬就发话了：**先当它已知的吧**，那么我们就可以算参数 θ 了：

$$
\theta^{(r)} = \arg\max_{\theta} \mathbb{E}\left[\sum_Y \tilde{p}^{(r-1)}(Y|X) \log p_\theta(Y) p_\theta(X|Y)\right] \tag{18}
$$

然后根据算出来的结果再去更新 p（$Y|X$）就是了，根据定义：

$$
p(Y|X) = \frac{p(X,Y)}{p(X)} = \frac{p(Y)p(X|Y)}{\sum\limits_Y p(Y)p(X|Y)} \tag{19}
$$

所以：

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

就让它们交替更新吧。现在来看看 (18) 式，**有个 E（求期望），又有个 M（argmax），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：

$$\sum_{Y} \tilde{p}^{(r-1)}(Y|X) \log p_\theta(Y) p_\theta(X|Y)$$

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 或者其他含有因变量的模型的 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**坚持使用联合隐变量的整体分布的最大似然，然后该化简的就化简，最终关于隐变量部分没法化简，那就迭代吧，迭着迭着就出来了。**

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。