Noise Contrast Estimation and Generative Adversarial Network

YWJ

2017年7月11日

1 Introduction

这篇文章梳理了概率统计模型的框架，并详细推导了NCE损失的原型。这同样是tensorflow.nn.nce-loss的原型，也是所谓负采样的理论依据。NCE理论有趣的地方是用作干扰的样本分布越接近训练样本的真实分布，这个模型反而越有效，在后文中将会介绍由这个推论如何衍生出的所谓生成式对抗网络（GAN）。

2 概率统计模型的一般套路，问题与解决办法

假设观察到的样本服从未知的真实分布$p_d(.)$, 而$p_d(.)$属于分布函数空间$p_d(.;\alpha)$，α为分布函数空间的参数向量，一个满足约束的α唯一的确定一个分布函数希望找到一个α，使之对应的分布对所有样本的似然最大，即在这个分布下，所有样本出现的概率之积最大，那么求真实分布的问题，转化为了带约束条件的优化问题，优化目标为：

$$l(\alpha) = \sum_{u \in U} \ln p_d(u;\alpha)$$

u表示单个样本，U表示样本空间所有样本的集合，但并非所有的α都能使$p_d(.;\alpha)$成为一个概率密度函数，概率密度函数要求其在所有可能样本上的概率之和为1，所以α需满足约束条件：

$$\sum_{u \in ALL} p_d(u;\alpha) = 1$$

注意这里的ALL是所有可能样本集合，U是训练样本集合，可能存在未被训练集观测到的样本,总之问题简化成了求:

$$\arg\max_\alpha \sum_{u \in U} \ln p_d(u;\alpha)$$
$$s.t. \sum_{u \in ALL} p_d(u;\alpha) = 1$$

如果考虑未被观测到的样本，$s.t.$是无法被求解的，只好假设所有可能的样本均已被训练集观测到，在此之上为了消除$s.t.$，定义规范化因子:

$$Z(\alpha) = \sum_{u \in U} p_d(u;\alpha)$$

而另真实分布属于空间：

$$p_k(u;\alpha) = \frac{p_d(u;\alpha)}{Z(\alpha)}$$

那么一定有(注意到$Z(\alpha)$是常数且不一定为1):

$$\sum_{u\in U} p_k(u;\alpha) = \sum_{u\in U} \frac{p_d(u;\alpha)}{Z(\alpha)} = 1$$

于是原问题简化成了求（限定条件没有了）:

$$\arg\max_{\alpha} \sum_{u\in U} \ln \frac{p_d(u;\alpha)}{Z(\alpha)}$$

然而这样还是有问题，如果训练集样本数量足够多，描述样本的向量维数足够大，不论是求所有样本的非规范化概率和$Z(\alpha)$还是非规范化概率积 $\sum_{u\in U} \ln p_d(u;\alpha)$计算量都是可观的，但若要用梯度下降优化$\alpha$，就不得不反反复复对每次$\alpha$更新后所有样本的非规范化概率求积求和。

解决 $\sum_{u\in U} \ln p_d(u;\alpha)$可以使用批量梯度下降，即每次更新 α 时使用从训练集U中随机选取的少量样本的集合U',直观的看要使所有样本的似然最大，对于任意选取样本子集的似然也应该趋向较大的值，那么每次优化一个随机选取的子集并重复多次貌似可以达到跟直接在整个训练集上反复优化相近的效果（这个替换确实有效果，但不知道哪里有构造性的证明能说明这样的替代一定有效）。

显然，不考虑乘法和加法的开销差异，求$Z(\alpha)$与求 $\sum_{u\in U} \ln p_d(u;\alpha)$的代价是相近的，但每次Z(α)的求值似乎是无法绕开的问题，2010年的这篇文章 [2]给出了一种绕过极大似然估计以绕过求规范化概率和的靠谱的解决办法。

思路是：

规范化因子$Z(\alpha)$本身也当作需要优化的参数，但这样以极大似然为原则的目标函数是没有上界的，因为$Z(\alpha)$更大总能让各个样本非规范化概率更大，从而让似然函数更大。

解决的办法是：

人为地引入分布律已知的干扰样本，对于一个待评估的分布函数来说，希望他跟训练样本符合的更好而跟干扰样本符合的更差，即以在训练集上的似然和在干扰集上的负似然之和作为优化目标，如过$Z(\alpha)$增大使训练集的似然增大，同样会使干扰集的负似然增大，这样定义的目标函数似乎不会随着$Z(\alpha)$增大而没有上界。

3 NCE损失的推导

定义参数向量:

$$\theta = (\alpha, c)$$

其中c是$-\ln(Z(\alpha))$的估计量，样本集合$U(u_1, u_2, \cdots, u_{2T})$,样本类别标记（正类还是负类）$C_T \in (0, 1)$,记标记为1的正类为训练样本$X(x_1, x_2, \cdots, x_T)$,服从带求解的分布律$p_m(x;\theta)$,标记为0的负类样本为$Y(y_1, y_2, , \cdots y_3)$,服从分布律$p_n(y)$,注意$Y$是人为产生的干扰样本，它的分布律$p_n(y)$是已知的。希望所有正负类的似然最大，据此定义损失函数：

$$\ln(\theta) = \sum_{t=1}^{T} \ln P(C=1|u;\theta) + \sum_{t=1}^{T} \ln P(C=0|u;\theta)$$

$$=\sum_{t=1}^{T}\ln P(C=1|u;\theta)+\sum_{t=1}^{T}\ln(1-P(C=1|u;\theta))$$

考察条件概率：

$$P(C=1|u;\theta)=\frac{P(C=1,u;\theta)}{P(u;\theta)}$$

$$=\frac{P(C=1)\cdot P(u|C=1;\theta)}{P(C=1)\cdot P(u|C=1;\theta)+P(C=0)\cdot P(u|C=0;\theta)}$$

$$=\frac{p_m(u;\theta)}{p_m(u;\theta)+p_n(u)}$$

记：

$$h(u;\theta)=\frac{p_m(u;\theta)}{p_m(u;\theta)+p_n(u)}$$

那么：

$$P(C=0|u;\theta)=1-P(C=1|u;\theta)=1-h(u;\theta)$$

$$\ln(\theta)=\sum_{t=1}^{T}\ln P(C=1|u;\theta)+\sum_{t=1}^{T}\ln P(C=0|u;\theta)$$

$$=\sum_{t=1}^{T}\ln h(x_t;\theta)+\sum_{t=1}^{T}\ln(1-h(y_t;\theta))$$

记每个样本的平均损失：

$$J_T(\theta)=\frac{1}{2T}\sum_{t=1}^{T}\ln h(x_t;\theta)+\ln(1-h(y_t;\theta))$$

由弱大数定律，训练集基数趋于无穷时，$J_T(\theta)$依概率收敛于：

$$J(\theta)=\frac{1}{2}E[\ln h(X;\theta)+\ln(1-h(Y;\theta))]$$

$$=\frac{1}{2}E[\ln\frac{1}{1+e^{-(\ln p_m(X;\theta)-\ln p_n(X))}}+\ln(1-\frac{1}{1+e^{-(\ln p_m(Y;\theta)-\ln p_n(Y))}})]$$

作者直接给出的结论是：

如果干扰样本分布律$p_n(.)$满足：

$$\mathrm{p}_d(.)\neq 0\Rightarrow p_n(.)\neq 0$$

（即干扰样本分布律的定义域覆盖真实分布的定义域，实际上在另一篇文章中你会看到这个限定条件几乎弱到等于没有。）

那么$p_m(.;\theta)=p_d(.)$时，$J(\theta)$取得全局唯一极大值。

NCE的作者抛出这个结论而没有证明，但从上下文看这个证明过成似乎与极大似然估计的依概率收敛证明类似，所以作者懒得再写，但我确实想不通这能怎么证明。

4 一个反常识的推论与GAN的关系

NCE损失相当于人为地加入一些不相干的样本，这些干扰样本也能提供一些关于训练样本的信息，因为这些干扰样本被打上了不服从训练样本真实分布的标记，变相缩小了训练样本分布函数的空间。但如果这些人为确定的干扰样本的分布律恰好与真实目标分布函数相近的话会怎么样呢？

作者给出了一个反直觉的结论，干扰样本的分布律与真实分布律越接近，那么目标函数的似然反而越大，学习到的$p_m(.;\theta)$反而越好，同样没有给出构造性的证明，但可以通过下面的方式印证这个结论：

不妨令J(θ)中的干扰分布就等于真实分布：

$$p_n(.) = p_d(.)$$

那么：

$$\begin{aligned} J(\theta) &= \frac{1}{2}E[\ln \frac{1}{1+e^{-(\ln p_m(X;\theta)-\ln p_d(X))}} + \ln(1-\frac{1}{1+e^{-(\ln p_m(Y;\theta)-\ln p_d(Y))}})] \\ &= \frac{1}{2}E[\ln \frac{p_m(X;\theta)}{p_d(\mathrm{X})+p_m(X;\theta)} + \ln \frac{p_n(Y)}{p_d(Y)+p_m(Y;\theta)}] \\ &= \frac{1}{2}E[\ln \frac{1}{\frac{p_d(\mathrm{X})}{p_m(X;\theta)}+1} + \ln \frac{1}{\frac{p_m(Y;\theta)}{p_d(Y)}+1}] \end{aligned}$$

定义函数：

$$f(.;\theta) = \frac{p_d(.)}{p_m(.;\theta)}$$

那么有：

$$J(\theta)=\frac{1}{2}E[\ln \frac{1}{1+f(X;\theta)} + \ln \frac{1}{1+\frac{1}{f(Y;\theta)}}]$$

考虑到X,Y服从同一真实分布：

$$J(\theta)=\frac{1}{2}E[\ln \frac{f(X;\theta)}{(1+f(X;\theta))^2}]$$

考察函数：

$$\mathrm{G}(x) = \frac{x}{(1+x)^2}$$

$$\mathrm{G}'(x) = \frac{(1+x)^2-2x(1+x)}{(1+x)^3} = \frac{1-x^2}{(1+x)^3}$$

因为有$f(\mathrm{X};\theta) > 0$,在$x > 0$的区间上,$G(x)$有唯一极大值$G(1) = 1/2$,这说明当调整θ使$f(.;\theta) \equiv 1$时，即$p_m(.;\theta) = p_d(.)$时，对任意样本X,$\frac{f(X;\theta)}{(1+f(X;\theta))^2}$都取得最大值1/2,此时$J(\theta)$是常数且是全局最大值。

这说明了干扰样本分布等于真实分布的极端条件下似然才可能取到全局最大值，这从侧面印证了作者关于干扰分布越接近真实分布学习到的分布律越好的结论。

这个推论实际上就是GAN的来源，GAN使用判别器D来区分来自发生器G或真实分布的样本 [1]，实际上这两个学习器的任务都是学习真实分布。可以把G看作一个负采样发生器，D看作使用NCE损失的学习器，固定住G优化D,使NCE似然达到一个极大值，这时根据上面的推论，G的分布率越接近真实分布，这个似然就能优化的越大，因此可以固定住D，优化G,继续增大似然，反复固定住一个优化另一个，直到

似然没有明显增长，这时G的分布实际上是接近真实分布的，实际上依据上面的计算，这个极大的似然值也就是每个样本的平均损失的期望是1/2，这个值代表学习器判断样本属于真实分布的概率，当干扰样本分布等于真实分布时，学习器无法判断一个样本究竟来自干扰还是真实采样，只能给出两者各占一半的判断，这个概率的理论值是符合直觉的。

可以看见GAN同其它采用NCE损失的概率学习器的区别在于GAN不直接指定负采样的来源分布，而是通过生成器G学习负采样分布，而且仅当D以真实分布作为判别依据且G以真实分布作为生成依据时，D和G的博弈趋于稳定的均衡，此时G只输出服从真实分布的样本，D对给定的任何样本均以1/2的概率判为真实样本或干扰样本。

参考文献

[1] Ian J. Goodfellow, Jean Pougetabadie, Mehdi Mirza, Bing Xu, David Wardefarley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Advances in Neural Information Processing Systems*, 3:2672–2680, 2014.

[2] Michael Gutmann and Aapo Hyv?rinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *Journal of Machine Learning Research*, 9:297–304, 2010.