29 Sep

f-GAN简介：GAN模型的生产车间

By 苏剑林 | 2018-09-29 | 68083位读者

今天介绍一篇比较经典的工作，作者命名为f-GAN，他在文章中给出了通过一般的f散度来构造一般的GAN的方案。可以毫不夸张地说，这论文就是一个GAN模型的“生产车间”，它一般化的囊括了很多GAN变种，并且可以启发我们快速地构建新的GAN变种（当然有没有价值是另一回事，但理论上是这样）。

局部变分

整篇文章对f散度的处理事实上在机器学习中被称为“局部变分方法”，它是一种非常经典且有用的估算技巧。事实上本文将会花大部分篇幅介绍这种估算技巧在f散度中的应用结果。至于GAN，只不过是这个结果的基本应用而已。

f散度

首先我们还是对f散度进行基本的介绍。所谓f散度，是KL散度的一般化：

$$\mathcal{D}_f(P\Vert Q) = \int q(x) f\left(\frac{p(x)}{q(x)}\right)dx \tag{1}$$

注意，按照通用的约定写法，括号内是p/q而不是q/p，大家不要自然而言地根据KL散度的形式以为是q/p。

可以发现，这种形式能覆盖我们见过的很多概率分布之间的度量了，这里直接把论文中的表格搬进来（部分）

距离名称	计算公式	对应的f
总变差	$\frac{1}{2}\int \vert p(x) - q(x)\vert dx$	$\frac{1}{2}\vert u-1\vert$
KL散度	$\int p(x)\log\frac{p(x)}{q(x)}dx$	$u\log u$
逆KL散度	$\int q(x)\log\frac{q(x)}{p(x)}dx$	$-\log u$
Pearson χ^2	$\int\frac{(q(x)-p(x))^2}{p(x)}dx$	$\frac{(1-u)^2}{u}$
Neyman χ^2	$\int\frac{(p(x)-q(x))^2}{q(x)}dx$	$(u-1)^2$
Hellinger距离	$\int\left(\sqrt{p(x)}-\sqrt{q(x)}\right)^2dx$	$(\sqrt{u}-1)^2$
Jeffrey距离	$\int(p(x)-q(x))\log\left(\frac{p(x)}{q(x)}\right)dx$	$(u-1)\log u$
JS散度	$\frac{1}{2}\int p(x)\log\frac{2p(x)}{p(x)+q(x)} + q(x)\log\frac{2q(x)}{p(x)+q(x)}dx$	$-\frac{u+1}{2}\log\frac{1+u}{2} + \frac{u}{2}\log u$

凸函数

上面列举了一堆的分布度量以及对应的f，那么一个很自然的问题是这些f的共同特点是什么呢？

答案是：

1、它们都是非负实数到实数的映射（$\mathbb{R}^* \to \mathbb{R}$）；

2、$f(1)=0$；

3、它们都是凸函数。

第一点是常规的，第二点$f(1)=0$保证了$\mathcal{D}_f(P\Vert P)=0$，那第三点凸函数是怎么理解呢？其实它是凸函数性质的一个最基本的应用，因为凸函数有一个非常重要的性质（詹森不等式）：

$$\mathbb{E}\big[f(x)\big]\geq f\big(\mathbb{E}[x]\big)\tag{2}$$

也就是“函数的平均大于平均的函数”，有些教程会直接将这个性质作为凸函数的定义。而如果$f(u)$是光滑的函数，我们一般会通过二阶导数$f''(u)$是否恒大于等于0来判断是否凸函数。

利用(2)，我们有

$$\begin{aligned}\int q(x)f\left(\frac{p(x)}{q(x)}\right)dx =&\mathbb{E}_{x\sim q(x)}\left[f\left(\frac{p(x)}{q(x)}\right)\right]\\
\geq&f\left(\mathbb{E}_{x\sim q(x)}\left[\frac{p(x)}{q(x)}\right]\right)\\
=&f\left(\int q(x)\frac{p(x)}{q(x)}dx\right)\\
=&f\left(\int p(x)dx\right)\\
=&f(1) = 0
\end{aligned}\tag{3}$$

也就是说，这三个条件保证了f散度是非负，而且当两个分布一模一样时f散度就为0，这使得$\mathcal{D}_f$可以用来简单地度量分布之间的差异性。当然，f散度原则上并没有保证$P\neq Q$时$\mathcal{D}_f(P\Vert Q) > 0$。但通常我们会选择严格凸的$f$（即$f''(u)$恒大于0），那么这时候可以保证$P\neq Q$时$\mathcal{D}_f(P\Vert Q) > 0$，也就是说这时候有$\mathcal{D}_f(P\Vert Q)=0\,\,\Leftrightarrow\,\,P=Q$。（注：即便如此，一般情况下$\mathcal{D}_f(P\Vert Q)$仍然不是满足公理化定义的“距离”，不过这个跟本文主题关系不大，这里只是顺便一提。）

凸共轭

现在从比较数学的角度讨论一下凸函数，一般地，记凸函数的定义域为$\mathbb{D}$（对于本文来说，$\mathbb{D}=\mathbb{R}_+$）。选择任意一个点ξ，我们求$y=f(u)$在$u=\xi$处的切线，结果是

$$y = f(\xi) + f'(\xi)(u - \xi)\tag{4}$$

考虑两者的差函数

$$h(u) = f(u) - f(\xi) - f'(\xi)(u - \xi)\tag{5}$$

所谓凸函数，直观理解，就是它的图像总在它的（任意一条）切线上方，因此对于凸函数来说下式恒成立

$$f(u) - f(\xi) - f'(\xi)(u - \xi)\geq 0\tag{6}$$

整理成

$$f(u) \geq f(\xi) - f'(\xi)\xi + f'(\xi)u\tag{7}$$

因为不等式是恒成立的，并且等号是有可能取到的，因此可以导出

$$f(u) = \max_{\xi\in\mathbb{D}}\left\{f(\xi) - f'(\xi)\xi + f'(\xi)u\right\} \tag{8}$$

换新的记号，记$t = f'(\xi)$，并从中反解出ξ（对于凸函数，这总能做到，读者可以自己尝试证明），然后记

$$g(t) = -f(\xi) + f'(\xi)\xi \tag{9}$$

那么就有

$$f(u) = \max_{t\in f'(\mathbb{D})}\left\{tu - g(t)\right\} \tag{10}$$

这里的$g(t)$就称为$f(u)$的共轭函数。留意花括号里边的式子，给定f后，g也确定了，并且整个式子关于u是线性的。所以总的来说，我们做了这样的一件事情：

> 对一个凸函数给出了线性近似，并且通过最大化里边的参数就可以达到原来的值。

注意给定u，我们都要最大化一次t才能得到尽可能接近$f(u)$的结果，否则随便代入一个t，只能保证得到下界，而不能确保误差大小。所以它称为“局部变分方法”，因为要在每一个点（局部）处都要进行最大化（变分）。这样一来，我们可以理解为t实际上是u的函数，即

$$f(u) = \max_{T\text{是值域为}f'(\mathbb{D})\text{的函数}}\left\{T(u)u - g(T(u))\right\} \tag{11}$$

上述讨论过程实际上已经给出了计算凸共轭的方法，在这里我们直接给出上表对应的凸函数的共轭函数。

$f(u)$	对应的共轭$g(t)$	$f'(\mathbb{D})$	激活函数
$\frac{1}{2}\lvert u-1\rvert$	t	$\left[-\frac{1}{2},\frac{1}{2}\right]$	$\frac{1}{2}\tanh(x)$
$u\log u$	e^{t-1}	$\mathbb{R}$	x
$-\log u$	$-1-\log(-t)$	$\mathbb{R}_-$	$-e^x$
$\frac{(1-u)^2}{u}$	$2-2\sqrt{1-t}$	$(-\infty,1)$	$1-e^x$
$(u-1)^2$	$\frac{1}{4}t^2+t$	$(-2,+\infty)$	e^x-2
$(\sqrt{u}-1)^2$	$\frac{t}{1-t}$	$(-\infty,1)$	$1-e^x$
$(u-1)\log u$	$W(e^{1-t})+\frac{1}{W(e^{1-t})}+t-2$	$\mathbb{R}$	x
$-\frac{u+1}{2}\log\frac{1+u}{2}+\frac{u}{2}\log u$	$-\frac{1}{2}\log(2-e^{2t})$	$\left(-\infty,\frac{\log 2}{2}\right)$	$\frac{\log 2}{2}-\frac{1}{2}\log(1+e^{-x})$

（注：这里的W为朗伯W函数。）

f-GAN

由上述推导，我们就可以给出f散度的估算公式，并且进一步给出f-GAN的一般框架。

f散度估计

计算f散度有什么困难呢？根据定义(1)，我们同时需要知道两个概率分布P,Q才可以计算两者的f散度，但事实上在机器学习中很难做到这一点，有时我们最多只知道其中一个概率分布的解析形式，另外一个分布只有采

样出来的样本，甚至很多情况下我们两个分布都不知道，只有对应的样本（也就是说要比较两批样本之间的相似性），所以就不能直接根据(1)来计算f散度了。

结合(1)和(11)，我们得到

$$\begin{aligned}\mathcal{D}_f(P\|Q) &= \max_T \int q(x)\left[\frac{p(x)}{q(x)}T\left(\frac{p(x)}{q(x)}\right) - g\left(T\left(\frac{p(x)}{q(x)}\right)\right)\right]dx\\ &= \max_T \int \left[p(x)\cdot T\left(\frac{p(x)}{q(x)}\right) - q(x)\cdot g\left(T\left(\frac{p(x)}{q(x)}\right)\right)\right]dx\end{aligned} \tag{12}$$

将$T\left(\frac{p(x)}{q(x)}\right)$记为整体$T(x)$，那么就有

$$\mathcal{D}_f(P\|Q) = \max_T \Big(\mathbb{E}_{x\sim p(x)}[T(x)] - \mathbb{E}_{x\sim q(x)}[g(T(x))]\Big) \tag{13}$$

式(13)就是估计f散度的基础公式了。意思就是说：分别从两个分布中采样，然后分别计算$T(x)$和$g(T(x))$的平均值，优化T，让它们的差尽可能地大，最终的结果就是f散度的近似值了。显然$T(x)$可以用足够复杂的神经网络拟合，我们只需要优化神经网络的参数。

注意在对凸函数的讨论中，我们在最大化目标的时候，对T的值域是有限制的。因此，在T的最后一层，我们必须设计适当的激活函数，使得T满足要求的值域。当然激活函数的选择不是唯一的，参考的激活函数已经列举在前表。注意，尽管理论上激活函数的选取是任意的，但是为了优化上的容易，应该遵循几个原则：

> 1、对应的定义域为$\mathbb{R}$，对应的值域为要求值域（边界点可以忽略）；
>
> 2、最好选择全局光滑的函数，不要简单地截断，例如要求值域为$\mathbb{R}_+$的话，不要直接用$relu(x)$，可以考虑的是e^x或者$\log(1+e^x)$；
>
> 3、注意式(13)的第二项包含了$g(T(x))$，也就是g和T的复合计算，因此选择激活函数时，最好使得它与g的复合运算比较简单。

GAN批发

好了，说了那么久，几乎都已经到文章结尾了，似乎还没有正式说到GAN。事实上，GAN可以算是整篇文章的副产物而已。

GAN希望训练一个生成器，将高斯分布映射到我们所需要的数据集分布，那就需要比较两个分布之间的差异了，经过前面的过程，其实就很简单了，随便找一种f散度都可以了。然后用式(13)对f散度进行估计，估计完之后，我们就有f散度的模型了，这时候生成器不是希望缩小分布的差异吗？最小化f散度就行了。所以写成一个表达式就是

$$\min_G \max_T \Big(\mathbb{E}_{x\sim p(x)}[T(x)] - \mathbb{E}_{x=G(z),z\sim q(z)}[g(T(x))]\Big) \tag{14}$$

或者反过来：

$$\min_G \max_T \Big(\mathbb{E}_{x=G(z),z\sim q(z)}[T(x)] - \mathbb{E}_{x\sim p(x)}[g(T(x))]\Big) \tag{15}$$

就这样完了～

需要举几个例子？好吧，先用JS散度看看。把所有东西式子一步步代进去，你会发现最终结果是（略去了$\log 2$的常数项）

$$\min_G \max_D \Big(\mathbb{E}_{x\sim p(x)}[\log D(x)] + \mathbb{E}_{x=G(z),z\sim q(z)}[\log(1-D(x))]\Big) \tag{16}$$

其中D用$\sigma(x)=1/(1+e^{-x})$激活。这就是最原始版本的GAN了。

用Hellinger距离试试？结果是

$$\min_G \max_D \Big(-\mathbb{E}_{x\sim p(x)}[e^{D(x)}] - \mathbb{E}_{x=G(z),z\sim q(z)}[e^{-D(x)}]\Big) \tag{17}$$

这里的$D(x)$是线性激活。这个貌似还没有命名？不过论文中已经对它做过实验了。

那用KL散度呢？因为KL散度是不对称的，所以有两个结果，分别为

$$\min_G \max_D \Big(\mathbb{E}_{x\sim p(x)}[D(x)] - \mathbb{E}_{x=G(z),z\sim q(z)}[e^{D(x)-1}]\Big) \tag{18}$$

或

$$\min_G \max_D \Big(\mathbb{E}_{x=G(z),z\sim q(z)}[D(x)] - \mathbb{E}_{x\sim p(x)}[e^{D(x)-1}]\Big) \tag{19}$$

这里的$D(x)$也是线性激活。

好吧，不再举例了。其实这些f散度本质上都差不多，看不到效果差别有多大。不过可以注意到，JS散度和Hellinger距离都是对称的、有界的，这是一个非常好的性质，以后我们会用到。

总结

说白了，本文主要目的还是介绍f散度及其局部变分估算而已～所以大部分还是理论文字，GAN只占一小部分。

当然，经过一番折腾，确实可以达到“GAN生产车间”的结果（取决于你有多少种f散度），这些新折腾出来的GAN可能并不像我们想象中的GAN，但它们确实在优化f散度。不过，以往标准GAN（对应JS散度）有的问题，其实f散度照样会有，因此f-GAN这个工作更大的价值在于“统一”，从生成模型的角度，并没有什么突破。

如果您需要引用本文，请参考：

苏剑林. (Sep. 29, 2018). 《f-GAN简介：GAN模型的生产车间 》[Blog post]. Retrieved from https://kexue.fm/archives/6016