9 Dec

变分自编码器（八）：估计样本概率密度

By 苏剑林 | 2021-12-09 | 61655位读者 引用

在本系列的前面几篇文章中，我们已经从多个角度来理解了VAE，一般来说，用VAE是为了得到一个生成模型，或者是做更好的编码模型，这都是VAE的常规用途。但除了这些常规应用外，还有一些“小众需求”，比如用来估计x的概率密度，这在做压缩的时候通常会用到。

本文就从估计概率密度的角度来了解和推导一下VAE模型。

两个问题

所谓估计概率密度，就是在已知样本$x_1, x_2, \cdots, x_N \sim \tilde{p}(x)$的情况下，用一个待定的概率密度簇$q_\theta(x)$去拟合这批样本，拟合的目标一般是最小化负对数似然：

$$\mathbb{E}_{x\sim\tilde{p}(x)}[-\log q_\theta(x)] = -\frac{1}{N}\sum_{i=1}^{N}\log q_\theta(x_i) \tag{1}$$

但这纯粹都只是理论形式，还有诸多问题没有解决，主要可以归为两个大问题：

> 1、用什么样的$q_\theta(x)$去拟合；
>
> 2、用什么方法去求解上述目标。

混合模型

第一个问题，我们自然是希望$q_\theta(x)$的拟合能力越强越好，最好它有能力拟合所有概率分布。然而很遗憾的是，神经网络虽然理论上有万能拟合能力，但那只是拟合函数的能力，并不是拟合概率分布的能力，概率分布需要满足$q_\theta(x)\geq 0$且$\int q_\theta(x)dx = 1$，后者通常难以保证。

直接的做法做不到，那么我们就往间接的角度想，构建混合模型：

$$q_\theta(x) = \int q_\theta(x|z)q(z)dz = \mathbb{E}_{z\sim q(z)}[q_\theta(x|z)] \tag{2}$$

其中$q(z)$通常被选择为无参数的简单分布，比如标准正态分布；而$q_\theta(x|z)$则是带参数的、以z为条件的简单分布，比如均值、方差跟z相关的标准正态分布。

从生成模型的角度来看，上述模型被解释为先从$q(z)$中采样z，然后传入$q_\theta(x|z)$中生成x的两步操作。但本文的焦点是估计概率密度，我们之所以选择这样的$q_\theta(x|z)$，是因为它有足够的拟合复杂分布的能力，最后的$q_\theta(x)$表示为了多个简单分布$q_\theta(x|z)$的平均，了解高斯混合模型的读者应该知道，这样的模型能够起到非常强的拟合能力，甚至理论上能拟合任意分布，所以分布的拟合能力有保证了。

重要采样

但式(2)是无法简单积分出来的，或者说只有这种无法简单显式地表达出来的分布，才具有足够强的拟合能力，所以我们要估计它的话，都要按照$\mathbb{E}_{z\sim q(z)}[q_\theta(x|z)]$的方式进行采样估计。然而，实际的场景下，$z$和$x$的维度比较高，而高维空间是有“维度灾难”的，这意思是说在高维空间中，我们哪怕采样百万、千万个样本，都很难充分地覆盖高维空间，也就是说很难准确地估计$\mathbb{E}_{z\sim q(z)}[q_\theta(x|z)]$。

为此，我们要想办法缩小一下采样空间。首先，我们通常会将$q_\theta(x|z)$的方差控制得比较小，这样一来，对于给定x，能够使得$q_\theta(x|z)$比较大的z就不会太多，大多数z算出来的$q_\theta(x|z)$都非常接近于零。于是我们只需要想办法采样出使得$q_\theta(x|z)$比较大的z，就可以对$\mathbb{E}_{z\sim q(z)}[q_\theta(x|z)]$进行一个比较好的估计了。

具体来说，我们引入一个新的分布$p_\theta(z|x)$，假设使得$q_\theta(x|z)$比较大的z服从该分布，于是我们有

$$q_\theta(x) = \int q_\theta(x|z)q(z)dz = \int q_\theta(x|z)\frac{q(z)}{p_\theta(z|x)}p_\theta(z|x)dz = \mathbb{E}_{z\sim p_\theta(z|x)}\left[q_\theta(x|z)\frac{q(z)}{p_\theta(z|x)}\right]$$

这样一来我们将从$q(z)$“漫无目的”的采样，转化为从$p_\theta(z|x)$的更有针对性的采样。由

于$q_{\theta}(x|z)$的方差控制得比较小，所以$p_{\theta}(z|x)$的方差自然也不会大，采样效率是变高了。注意在生成模型视角下，$p_{\theta}(z|x)$被视为后验分布的近似，但是从估计概率密度的视角下，它其实就是一个纯粹的重要性加权函数罢了，不需要特别诠释它的含义。

训练目标

至此，我们解决了第一个问题：用什么分布，以及怎么去更好地计算这个分布。剩下的问题就是如何训练了。

其实有了重要性采样的概念后，我们就不用考虑什么ELBO之类的了，直接使用目标(1)就好，代入$q_{\theta}(x)$的表达式得到

$$\mathbb{E}_{x\sim\tilde{p}(x)}\left[-\log\mathbb{E}_{z\sim p_{\theta}(z|x)}\left[q_{\theta}(x|z)\frac{q(z)}{p_{\theta}(z|x)}\right]\right]\tag{4}$$

事实上，如果$\mathbb{E}_{z\sim p_{\theta}(z|x)}$这一步我们通过重参数只采样一个$z$，那么训练目标就变成

$$\mathbb{E}_{x\sim\tilde{p}(x)}\left[-\log q_{\theta}(x|z)\frac{q(z)}{p_{\theta}(z|x)}\right],\quad z\sim p_{\theta}(z|x)\tag{5}$$

这其实已经就是常规VAE的训练目标了。如果采样$M>1$个，那么就是

$$\mathbb{E}_{x\sim\tilde{p}(x)}\left[-\log\left(\frac{1}{M}\sum_{i=1}^{M}q_{\theta}(x|z_i)\frac{q(z_i)}{p_{\theta}(z_i|x)}\right)\right],\quad z_1,z_2,\cdots,z_M\sim p_{\theta}(z|x)\tag{6}$$

这就是“重要性加权自编码器”了，出自《Importance Weighted Autoencoders》，它被视为VAE的加强。总的来说，通过重要性采样的角度，我们可以绕过传统VAE的ELBO等繁琐推导，也可以不用《变分自编码器（二）：从贝叶斯观点出发》所介绍的联合分布视角，直接得到VAE模型甚至其改进版。

文章小结

本文从估计样本的概率密度这一出发点介绍了变分自编码器VAE，结合重要性采样，我们可以得到VAE的一个快速推导，完全避开ELBO等诸多繁琐细节。

如果您需要引用本文，请参考：

苏剑林. (Dec. 09, 2021). 《变分自编码器（八）：估计样本概率密度 》[Blog post]. Retrieved from https://spaces.ac.cn/archives/8791

```
@online{kexuefm-8791,
        title={变分自编码器（八）：估计样本概率密度},
        author={苏剑林},
        year={2021},
        month={Dec},
        url={\url{https://spaces.ac.cn/archives/8791}},
}
```