Autoregressive Quantile Networks for Generative Modeling

Georg Ostrovski [*1] Will Dabney [*1] Rémi Munos [1]

【机器学习 54】AIQN

张楚珩
清华大学 交叉信息院博士在读

AIQN是Autoregressive Implicit Quantile Network的简称。

原文传送门

Ostrovski, Georg, Will Dabney, and Rémi Munos. "Autoregressive quantile networks for generative modeling." arXiv preprint arXiv:1806.05575 (2018).

特色

这是一个做生成模型（generative model）的最新工作，基本上是本专栏之前讲的IQN的的原班人马做的（一二作位置换了一下）。读IQN的时候就觉得这种网络结构能够替代reparametrization trick，用于生成一个概率分布，结果搜了一下果然有人做掉了。

过程

一、背景

生成模型主要用于拟合数据的分布，然后能够用算法去生成数据，比较常见的应用就是用训练好的模型去生成高质量的图片或者音频。生产模型主要的方法有对抗生成网络（GAN）、变分推断（variational inference）和自回归密度估计（autoregressive density estimation）。

估计数据分布最直接的方式***参数化***，即假定一个数据分布并且做参数化，然后利用真实数据来估计相应的参数。比如假定数据从一个离散分布中得到 $x_1, x_2, \cdots \in \mathcal{X}$ ，并且认为数据取到某个离散值的概率密度为 $p_\theta(x_i) \propto \exp(\theta_i)$ ，接下来就能通过最小化KL散度来求导相应的密度估计 $\theta^* = arg\min_\theta D_{KL}(p_X \| p_\theta)$ 。对于连续分布也可以认为数据从高斯分布中得到，并且最小化KL散度来得到高斯分布的参数。另外文章为了对于高维连续概率空间进行建模，用到了一种特殊情况，这种情况介于各个维度之间完全独立和各个维度可以任意相关之间，即每一维都依赖于前面的维度，但不依赖后面的维度

$$p_X(x) = \prod_{i=1}^{n} p_{X_{\sigma(i)}}(x_{\sigma(i)} | x_{\sigma(1)}, \ldots, x_{\sigma(i-1)}).$$

估计数据分布的第二种方法是***隐变量方法***（latent variable），即大家很熟悉的VAE，假设一个低维隐空间内的一个隐变量 $z \in \mathcal{Z}$ 通过一个decoder决定了最后生成的数据，隐变量服从一些简单的概

率分布，而通过decoder产生的数据则服从数据的分布。VAE的训练过程就是要找到decoder神经网络的参数，但是整个训练过程是需要在encoder神经网络的帮助下进行的。即

$$\log p_\theta(x) \geq -D_{KL}(q_\theta(z|x)\|p(z)) + \mathbb{E}\left[\log p_\theta(x|z)\right].$$

估计数据分布的第三种方法是***对抗生成网络***（GAN）。它利用博弈的思路，同时训练生成器网络（generator, G）和判别器网络（discriminator, D），生成器网络的目标是为了生成更“像”真实数据的样本，对抗器网络的目标是分辨出哪个是真实数据、哪个是人造数据；通过博弈最后得到一个近乎以假乱真的生成器。其优化目标为

$$\arg\min_G \sup_D \left[\mathbb{E}_X \log(D(X)) + \mathbb{E}_Z \log(1 - D(G(Z)))\right],$$

以上三种方法归根结底都是在最小化KL散度（GAN是在最小化Jensen-Shannon散度，它可以看做是一种特殊的KL散度），通过本专栏讲的Wasserstein距离可以看到KL散度不是一个很好的距离衡量标准，而Wasserstein距离会更好，它对于距离的衡量在语义上更为有意义（semantically meaningful)。而从之前本专栏讲的IQN等工作可以知道，quantile regressional是在一维连续空间上最小化1-Wasserstein距离，那么能不能***推广到高维连续空间，并且用它来估计数据分布***呢？

二、Autoregressive Implicit Quantile

开始不太懂啥叫自回归（autoregression），查了一下，大概的意思是回归（regression）就是拟合数据x和数据y之间的关系，与此对应，自回归就是拟合数据x自己的概率分布。

那么如何用IQN来做自回归呢？

1. 考虑到IQN其实是做的回归，输入是状态 s 和分位数 τ ，输出是该分位数下的价值函数 $Z_\tau(s,\cdot)$ 。最简单的做法就是把IQN本来的输入只留下一个分位数 $\tau\in[0,1]$ 。在一维情况下，概率分布的CDF是一个一一映射，其逆函数是定义良好的，但是在高维空间中，其逆就不止一个了。只有当高维概率分布满足comonotonicity性质的时候，才具有一个定义良好的逆CDF，即 $F_X^{-1}(\tau)=(F_{X_1}^{-1}(\tau),F_{X_2}^{-1}(\tau),\cdots)$ 。

2. 另一种做法是输入 $\tau\in[0,1]^n$ ，这种情况其实假设了各个维度之间相互独立。但是这种假设在图像领域肯定不行，因为每个像素点之间肯定不是相互独立的，这样产生的生成模型只会得到噪声。

3. 文章采用的假设介于各个维度间有 full covariance 和各个维度间相互独立之间，即每个维度依赖于在其之前的所有维度，即可以定义 CDF

$$\begin{aligned} F_X(x) &= \mathrm{P}(X_1 \leq x_1, \ldots, X_n \leq x_n), \\ &= \prod_{i=1}^{n} F_{X_i|X_{i-1},\ldots,X_1}(x_i). \end{aligned}$$

这样给定一个分位数 $\tau_{joint}=\prod_{i=1}^{n}\tau_i$ ，有相应的 CDF 逆函数

$$F_X^{-1}(\tau_{joint}) = (F_{X_1}^{-1}(\tau_1), \ldots, F_{X_n|X_{n-1},\ldots}^{-1}(\tau_n)).$$

由此可以把文章中的生成模型看成一个黑盒子，随机采样一个 $\tau\sim U([0,1])^n$ 送入黑盒子中，黑盒子输出上式中的CDF逆函数，再联合对应的数据 $x\sim X$ 得到 quantile loss 并且使用梯度下降算法来学习。

$$\sum_i \rho^{\kappa}_{\tau_i}(x_i - Q_X(\tau_i|x_{i-1},\ldots)).$$

三、网络结构

刚刚我们说到，该黑盒子接受一个输入 $\tau\sim U([0,1])^n$ ，输出 $F_X^{-1}(\tau)$ ，那么它具体网络结构是怎样的呢？它可以看做一个类似RNN的结构，每过一次block新生成一个像素点的数据，对于一张图片来说，需要分别生成 $3n^2$ 个像素点的数据。每个block的输出是一个像素点数据 x_i ，其输入是一个 $\tau_i\sim U([0,1])$ 和之前的数据点 $x_{i-1},x_{i-2},\cdots$ 。结构如下图所示

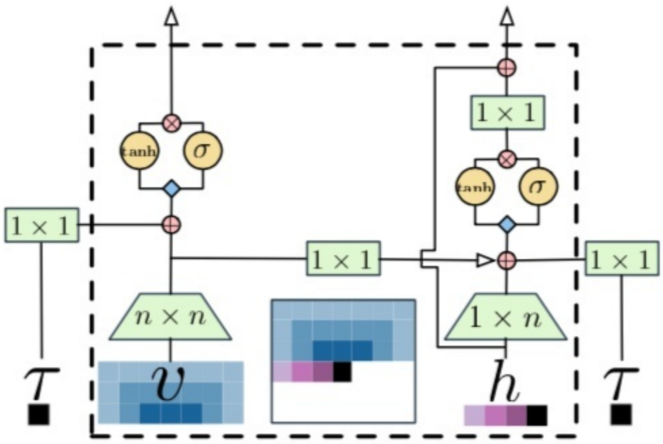

Figure 2. Illustration of Gated PixelCNN layer block for PixelIQN. Dashed line shows boundary of standard Gated PixelCNN, with v the vertical and h the horizontal stack. Conditioning on τ is identical to the location-dependent conditioning in Gated PixelCNN.

四、Quantile Regression的性质

1. CDF逆函数与PDF函数之间的关系

$$\frac{\partial}{\partial \tau} F_X^{-1}(\tau) = \frac{1}{p_X(F_X^{-1}(\tau))}.$$

当CDF逆函数 $F_X^{-1}(\tau)$ 是用神经网络表示出来的时候，可以使用神经网络的求导功能来得到输出点上的PDF。

2. Quantile Regression的梯度及其优化的函数

在前面的工作中（参考本专栏Quantile Regression）如果把quantile representation看做是Dirac函数的加和，那么可以证明最小化quantile loss就等于最小化1-Wasserstein距离。这里考虑更为一般情况，quantile representation是任意的函数，那么最小化quantile loss就等于最小化如下的quantile divergence。

$$q(P,Q) := \int_0^1 \left[\int_{F_P^{-1}(\tau)}^{F_Q^{-1}(\tau)} (F_P(x) - \tau) dx \right] d\tau.$$

有如下定理

Proposition 1. *For any distributions P and Q, define the quantile divergence*

$$q(P,Q) := \int_0^1 \left[\int_{F_P^{-1}(\tau)}^{F_Q^{-1}(\tau)} (F_P(x) - \tau)\, dx \right] d\tau.$$

Then the expected quantile loss of a quantile function $\bar{Q}$ implicitly defining the distribution Q satisfies

$$\mathop{\mathbb{E}}_{\tau \sim \mathcal{U}([0,1])} \mathop{\mathbb{E}}_{X \sim P} \left[\rho_\tau(X - \bar{Q}(\tau)) \right] = q(P,Q) + h(P),$$

where $h(P)$ does not depend on Q.

因此，当我们在最小化quantile loss（也就是等式左边）的时候，其实就是在最小化对应的quantile divergence。

证明

定义

$$\rho_\tau(u) = u(\tau - \mathbb{I}\{u \le 0\}),$$
$$g_\tau(q) = \mathop{\mathbb{E}}_{X \sim P}[\rho_\tau(X - q)].$$

通过把分段函数展开，并且使用分部积分，可以把 $g_\tau(q)$ 写成

$$
\begin{aligned}
g_\tau(q) &= \int_{-\infty}^{q} (x-q)(\tau-1) f_P(x)\, dx \\
&\quad + \int_{q}^{\infty} (x-q)\tau f_P(x)\, dx \\
&= \int_{-\infty}^{q} (q-x) f_P(x) dx + \int_{-\infty}^{\infty} (x-q)\tau f_P(x)\, dx \\
&= qF_P(q) + \int_{-\infty}^{q} F_P(x)\, dx - [xF_P(x)]_{-\infty}^{q} \\
&\quad + \tau \left(\mathop{\mathbb{E}}_{X\sim P}[X] - q \right) \\
&= \int_{-\infty}^{q} F_P(x)\, dx + \tau \left(\mathop{\mathbb{E}}_{X\sim P}[X] - q \right).
\end{aligned}
$$

外面再套一层期望的话， $g_\tau(q)$ 就是quantile loss了；当 $q = F_P^{-1}$ 的时候，该函数值取最小，即即使q完全是P分布的CDF逆函数，也会产生一个constant，但如果把这个constant减去，就能得到定义的quantile divergence。

$$
g_\tau(F_P^{-1}(\tau)) = \int_{-\infty}^{F_P^{-1}(\tau)} F_P(x)\, dx + \tau \left(\mathop{\mathbb{E}}_{X\sim P}[X] - F_P^{-1}(\tau) \right)
$$

可以得到

$$
\begin{aligned}
& g_\tau(q) - g_\tau(F_P^{-1}(\tau)) \\
&= \int_{F_P^{-1}(\tau)}^{q} F_P(x)\, dx + \tau(F_P^{-1}(\tau) - q) \\
&= \int_{F_P^{-1}(\tau)}^{q} (F_P(x) - \tau)\, dx.
\end{aligned}
$$

注意到当 $q=F_Q^{-1}$ 的时候，该式子就等于 $q(P,Q)$ ，由此

$$\mathop{\mathbb{E}}_{\tau\sim\mathcal{U}([0,1])}\left[g_\tau(Q(\tau))\right]=q(P,Q)+\underbrace{\mathop{\mathbb{E}}_{\tau\sim\mathcal{U}([0,1])}\left[g_\tau(F_P^{-1}(\tau))\right]}_{\text{does not depend on } Q}.$$

由此可以知道，在sample上面求到的quantile loss的梯度是quantile divergence的无偏估计。

最后一幅图总结一下，如果认为Q是Dirac函数的组合，那么quantile loss就在最小化1-Wasserstein距离；如果Q是任意函数，那么quantile loss就在最小化quantile散度。

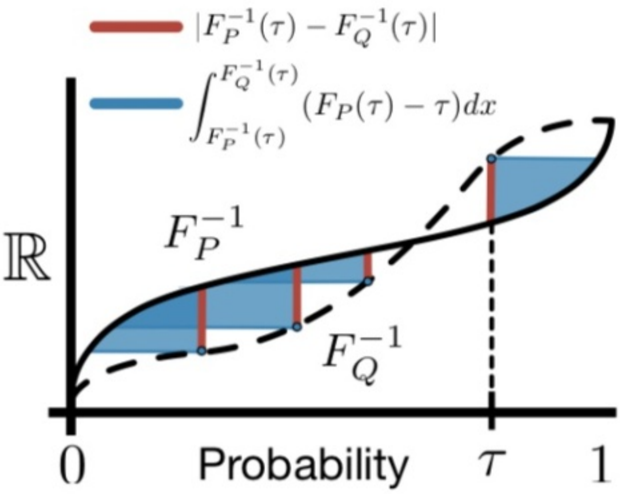

Figure 8. Illustration of the relation between the 1-Wasserstein metric (red) and the quantile divergence (blue).

实验

实验在CIFAR-10和ImageNet 32x32上做的，效果优于所比较的算法，并且能够使用更小的神经网络规模达到更好的效果。

Figure 5. ImageNet 32x32: Real example images (left), samples generated by PixelCNN (center), and samples generated by PixelIQN (right). Neither of the sampled image sets were cherry-picked. More samples by PixelIQN in the Appendix.

发布于 2019-04-22

机器学习

文章被以下专栏收录

强化学习前沿
读呀读paper
进入专栏