Flow Model

Chen Gong

27 June 2020

1 Introduction

在上一小节中讲到了 Latent Variable Model（LAM），VAE。其主要思想就是将隐变量扩充为高维连续的分布，来增强模型的表达能力。而 LAM 模型中的核心困难是 $P(X)$ 计算不出来，因为 $P(X) = \int_Z P(X|Z)P(Z)dZ$，而 Z 的维度过高 $P(X)$ 算不出来。而根据 Bayesian 公式：

$$P(Z|X) = \frac{P(Z)P(X|Z)}{P(X)} \tag{1}$$

所以导致 $P(Z|X)$ 无法计算。而 VAE 那章介绍了近似推断的方法，使用一个简单分布 $Q_\phi(Z|X)$ 来近似 $P(Z|X)$，其中还使用重参数化技巧来用一个神经网络来代替分布。

而在 VAE 中通过优化变分下界 ELBO 来达到最终优化的目的，而不是直接对 Log 似然函数进行优化。所以当然会有误差了。那么这将启发我们，可不可以绕过这个 intractable 的 $P(Z)$，使模型变得 tractable。

2 Flow based Model

什么是 flow model 呢？首先用一张图来进行表示：

图 1: Flow 模型基础示意图

可以用一个简单的例子来简单的介绍 Flow model。X 可以代表是当前的自己，人是比较复杂的，所以 $X \to P_X(X)$ 计算非常困难。而一般昨天的我 $Z_k \to P_{Z_k}(Z_k)$，比今天要简单一点，但是很有可能，昨天的我依然很复杂，无法计算。那么，就不但的往前推，到了刚出生的时候 Z_0，这时肯定是非常简单的，$Z_0 \to P_{Z_0}(Z_0)$ 婴儿的世界里是非黑即白的，此时的分布很简单，可以被假设为 $\mathcal{N}(0, I)$。而这个过程：

$$P_{Z_0}(Z_0) \to P_{Z_1}(Z_1) \to P_{Z_2}(Z_2) \cdots \to P_{Z_k}(Z_k) \to P_X(X) \tag{2}$$

就被称为"流"。因为流模型中初始分布是很简单的。极大似然估计中求的是：$\arg\max P(X)$。那么下一个问题就是如何建立 X 和 Z_0 之间的关系，将 $\arg\max P(X)$ 转换成求关于 $P(Z_0)$ 的函数。

3 Change of Variables

假设 $X=f(Z)$，$Z,X\in\mathbb{R}^p$。而 $Z\sim P_Z(Z)$，$X\sim P_X(X)$；f 是一个光滑可逆的函数。那么可以得到：

$$\int_Z P_Z(Z)dZ = 1 = \int_X P_X(X)dX \tag{3}$$

根据不定积分的性质可以得到：

$$|P_Z(Z)dZ| = |P_X(X)dX| \tag{4}$$

$$P_X(X) = \left|\frac{dZ}{dX}P_Z(Z)\right| \tag{5}$$

而 $X=f(Z)$ 且 f 是光滑可逆的，所以 $Z=f^{-1}(X)$，那么有

$$P_X(X) = \left|\frac{\partial f^{-1}(X)}{\partial X}\right| P_Z(Z) \tag{6}$$

但是实际上 Z 和 X 都是高维变量，所以 $\frac{\partial f^{-1}(X)}{\partial X}$ 是一个 Jacobian Matrix。**熟悉矩阵的朋友应该知道，矩阵代表了一个变换，而矩阵行列式的值则代表了变换的尺度**。而在计算中我们关注的是矩阵变换的尺度，所以，

$$P_X(X) = \left|\det\left(\frac{\partial f^{-1}(X)}{\partial X}\right)\right| P_Z(Z) \tag{7}$$

而最终的目的是想将 $P_X(X)$ 完全用一个 Z 为自变量的函数来表达，所以要将 $\left|\frac{\partial f^{-1}(X)}{\partial X}\right|$ 用 Z 来表示。下面先写结论

$$\begin{aligned} P_X(X) &= \left|\det\left(\frac{\partial f^{-1}(X)}{\partial X}\right)\right| P_Z(Z) \\ &= \left|\det\left(\frac{\partial f^{-1}(Z)}{\partial Z}\right)\right|^{-1} P_Z(Z) \end{aligned} \tag{8}$$

这个结论是怎么来的呢？我们来看一个简单的例子，如下图所示：

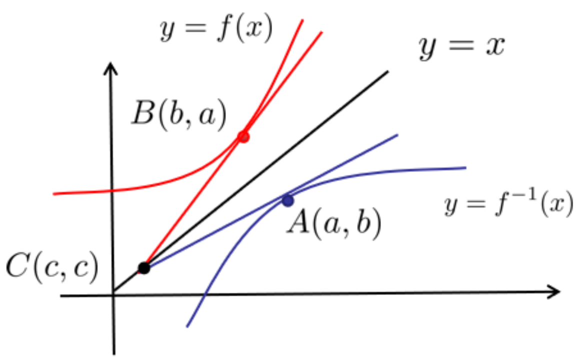

图 2: 实例

如图所示，$y=f(x)$，$x=f^{-1}(y)$。那么有

$$\frac{dy}{dx}=\frac{\partial f(x)}{\partial x}, \frac{dx}{dy}=\frac{\partial f^{-1}(y)}{\partial y} \tag{9}$$

而，

$$\frac{\partial f(x)}{\partial x}\frac{\partial f^{-1}(y)}{\partial y}=1 \tag{10}$$

在本文举的例子中，

$$\begin{aligned}(f^{-1})'(a)&=\frac{b-c}{a-c}\\(f)'(b)&=\frac{a-c}{b-c}\end{aligned} \tag{11}$$

很显然有 $(f^{-1})'(a)f'(b)=1$。这就是 change of variables theorem。我们可以得到两个变量之间关于映射 f 的转换为：

$$P_X(X)=\left|\det\left(\frac{\partial f^{-1}(Z)}{\partial Z}\right)\right|^{-1}P_Z(Z) \tag{12}$$

那么，当训练完成之后，从 $P(Z)$ 中采样比较简单，通过上述公式，就可以得到 $P(X)$，所以 $P(X)$ 是可求解的。如何学习呢？其实并不难，通过极大似然估计可以得到：

$$\log P_X(X)=\log\left|\det\left(\frac{\partial f^{-1}(Z)}{\partial Z}\right)\right|^{-1}+\log P_Z(Z) \tag{13}$$

那么：

$$\begin{aligned}\frac{\partial \log P_X(X)}{\partial X}&=\frac{\partial \log\left|\det\left(\frac{\partial f^{-1}(Z)}{\partial Z}\right)\right|^{-1}+\log P_Z(Z)}{\partial Z}\frac{\partial Z}{\partial X}\\&=\frac{\partial \log\left|\det\left(\frac{\partial f^{-1}(Z)}{\partial Z}\right)\right|^{-1}+\log P_Z(Z)}{\partial Z}\frac{\partial f^{-1}(X)}{\partial X}\end{aligned} \tag{14}$$

由于 f 的逆很要求，上述梯度的计算还是比较简单的。然而，关于大矩阵行列式的计算并不美丽。后续有很多针对这点的改进方法，有兴趣的同学自行查看 flow based 的论文。

[1] ICLR 2015 NICE-Non-linear Independent Components Estimation

[2] ICLR 2017 Density estimation using Real NVP

[3] 2018 Glow: Generative Flow with Invertible 1×1 Convolutions

4 小结

本章主要介绍的是流模型的主要思想，在 Latent Variable Model 经常会遇到后验过于复杂无法求解的问题。流模型绕开了这个部分，对更简单的分布建模，然后建立原分布与简单分布之间的映射关系。个人觉得 Stein 变分梯度下降就有点流模型的影子在里面。在建立映射关系是用到了重要的 change of variables theorem，并之后介绍了变化后的目标函数和梯度求解方法。