

日期： /

## flow-based models

使用 VAE 的一个问题是边缘分布  $p(x)$ ，流模型的类似于有空间结构的 VAE，使得训练过程大大简化

VAE 的特点是  $p(x|z)$  是易得的而  $p(z|x)$  则需要借助  
示例。flow models 的方案是构建可逆的映射，使得  
 $z \rightarrow x$  的过程可以轻松逆转为  $x \rightarrow z$ 。

[变量替换定理] 当  $X = f(Z)$ , 且  $f(x)$  可逆,  $Z = f^{-1}(X) = h(X)$

$$P_X(X) = P_Z(h(X)) |h'(x)| \quad \nearrow h(x) \text{ 的 Jacobian}$$

$$\text{对于多维情况: } P_X(X) = P_Z(h(X)) \left| \det \left( \frac{\partial h(x)}{\partial x} \right) \right| \quad \nearrow \text{矩阵}$$

类比 VAE 中的隐变量，但现在由一个可逆变换得到  
给定  $Z_1$  和  $Z_2$  的联合分布  $P_{Z_1, Z_2}$ , 有可逆变换  $U: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ ,

记作  $U = (U_1, U_2)$ , 逆变换记作  $V = (V_1, V_2)$ , 令  $X_1, X_2 \sim U_1(Z_1, Z_2)$

$X_2 \sim U_2(Z_1, Z_2)$ , 则  $Z_1 \sim V_1(X_1, X_2)$ ,  $Z_2 \sim V_2(X_1, X_2)$

$$\text{则 } P_{X_1, X_2}(X_1, X_2) = P_{Z_1, Z_2}(V_1(X_1, X_2), V_2(X_1, X_2)) \left| \det \left( \frac{\partial V_1(X_1, X_2)}{\partial X_1}, \frac{\partial V_1(X_1, X_2)}{\partial X_2} \right) \right|$$

为确保可逆,  $Z$  与  $X$  维度应一致

日期： /

在一个归一化的流模型中， $Z$  与  $X$  间的映射由参数化的神经网络  $f_\theta : \mathbb{R}^n \rightarrow \mathbb{R}^n$  给出， $X = f_\theta(Z)$ ,  $Z = f_\theta^{-1}(X)$

边际概率  $P_X(X; \theta) = P_Z(f_\theta^{-1}(X)) \underbrace{|\det(\frac{\partial f_\theta^{-1}(x)}{\partial x})|}_{\text{归一化部分}}$

归一化流模型的基本结构是  $Z_m = f_{\theta_m}^{(m)} \circ \dots \circ f_{\theta_1}^{(1)}(Z_0)$

$P_X(X; \theta) = P_Z(f_\theta^{-1}(X)) \prod_{m=1}^M |\det(\frac{\partial f_\theta^{(m)}(Z_{m-1})}{\partial Z_m})|$

由已知的先验分布  
描述，这是采样的  
基础

关键在于，如何确保  $f_\theta$  为可逆的映射

在数据集中的优化目标： $\max \log P_X(D; \theta) = \sum_{x \in D} [\log P_Z(f_\theta^{-1}(x)) + \log |\det(\frac{\partial f_\theta^{-1}(x)}{\partial x})|]$

另一个问题是如何计算 Jacobian 矩阵（或近似）及行列式。  
做法是使  $f_\theta$  满足一些约束，使其 Jacobian 矩阵为  
三角阵

[NICE - Additive coupling layers]

将  $Z$  分为 2 个部分： $Z_{1:d}$ ,  $Z_{d+1:n}$  且  $1 < d < n$

对于正向过程  $Z \mapsto X$  映射如下：

-  $X_{1:d} = Z_{1:d}$

-  $X_{d+1:n} = Z_{d+1:n} + M_\theta(Z_{1:d})$ ,  $M_\theta(\cdot)$  为一个  $d$  维输入,  $n-d$  维输出的神经网络

日期： /

对于反向过程  $X \rightarrow Z$ :

$$-Z_{1:d} = X_{1:d}$$

$$-Z_{d+1:n} = X_{d+1:n} - M_\theta(X_{1:d})$$

NICE 层的 Jacobian 矩阵有良好的性质:

$$J = \frac{\partial X}{\partial Z} = \begin{pmatrix} I_d & 0 \\ \frac{\partial M_\theta(Z_{1:d})}{Z_{1:d}} & I_{n-d} \end{pmatrix} \det(J) = 1$$

这意味着无须进行额外的归一化

NICE 的最后一层使用了一个重缩放策略:

$$Z \mapsto X: X_i = \underline{s_i} \overset{\text{缩放因子}}{\overbrace{z_i}}$$

$$X \mapsto Z: Z_i = X_i / s_i$$

$$J = \text{diag}(s) \quad \det(J) = \prod_{i=1}^n s_i$$

[Real-NVP]

与 NICE 类似，但在每一层均施加缩放：

前向过程:

$$-X_{1:d} = Z_{1:d}$$

$$-X_{d+1:n} = Z_{d+1:n} \odot \exp(\alpha_\theta(Z_{1:d})) + \mu_\theta(Z_{1:d})$$

日期： /

反向过程  $X \mapsto Z$ :

$$- Z_{1:d} = X_{1:d}$$

$$- Z_{d+1:n} = [X_{d+1:n} - \mu_\theta(Z_{1:d})] / \exp(Z_{1:d})$$

$$J = \frac{\partial \mathbf{x}}{\partial \mathbf{z}} = \begin{pmatrix} I_d & 0 \\ \frac{\partial X_{d+1:n}}{\partial Z_{1:d}} & \text{diag}(\exp(\alpha_\theta(Z_{1:d}))) \end{pmatrix}$$

$$\det(J) = \prod_{i=d+1}^n \exp(\alpha_\theta(Z_{1:d})) = \exp\left(\sum_{i=d+1}^n \alpha_\theta(Z_{1:d})\right)$$

使用 flow model, 如果对  $Z$  进行混合, 例如,

$$Z = \cos \phi (Z^{(1)} \cos \phi' + Z^{(2)} \sin \phi') + \sin \phi (Z^{(3)} \cos \phi' + Z^{(4)} \sin \phi')$$

$Z$  生成的  $X$  将可能混合  $Z^{(1)} Z^{(2)} Z^{(3)} Z^{(4)}$  的风格  
某些自回归模型可以被看做流模型

[Masked autoregression flow (MAF)]

例如 : Sample  $Z_i \sim N(0, 1)$  for  $i=1, 2, \dots, n$

对应  $n$  个随机量

对于训练过程, 所有的 Let  $\hat{X}_1 = \exp(\alpha_1) Z_1 + \mu_1$  计算  $\mu_1(x_1), \alpha_1(x_1)$

$\alpha_1, \mu_1$  可以并行计算, 但 Let  $\hat{X}_2 = \exp(\alpha_2) Z_2 + \mu_2$  计算  $\mu_2(x_1, x_2), \alpha_2(x_1, x_2)$

对于预测过程,  $\alpha_1, \mu_1$  需循环计算,  $\dots$

因此采样是缓慢的 Let  $\hat{X}_n = \exp(\alpha_n) Z_n + \mu_n$  计算  $\mu_n(x_1, \dots, x_{n-1}), \alpha_n(x_1, \dots, x_{n-1})$

日期: /

## [Inverse Autoregressive Flow (IAF)]

如果将过程反转，就可以实现高效的采样  
过程。

$Z \mapsto X$ : Sample  $Z_i \sim N(0, 1)$ ,  $i=1, 2, \dots, n$

并行计算  $\mu_i(Z_{\leq i})$   $\alpha_i(Z_{\leq i})$

Let  $X_1 = \exp(\alpha_1)Z_1 + \mu_1$

Let  $X_2 = \exp(\alpha_2)Z_2 + \mu_2$

...

$X \mapsto Z$ : Let  $Z_1 = (X_1 - \mu_1) / \exp(\alpha_1)$

缺点是仍然的计算代价更大 计算  $\mu_2(Z_1)$   $\alpha_2(Z_1)$

Let  $Z_2 = (X_2 - \mu_2) / \exp(\alpha_2)$

等等  $\mu_3(Z_1, Z_2)$   $\alpha_3(Z_1, Z_2)$

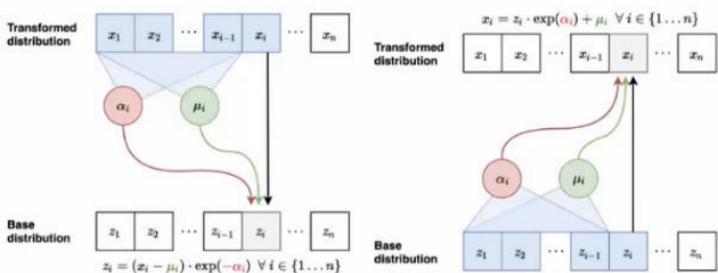


Figure: Inverse pass of MAF (left) vs. Forward pass of IAF (right)

日期： /

## [Porbability density distillation]

概率密度蒸馏是一种能够结合 MAF 的高效训练过程和 IAF 的高效采样过程的模型。在蒸馏过程中，student model 和 teacher model 的概率分布的 KL 散度被最小化。

$$D_{KL}(s, t) = \mathbb{E}_{x \sim s} [\log s(x) - \log t(x)]$$

训练过程是：

- 从 IAF(s) 中采样  $x$
- 以 IAF 模型评估似然（由于  $x$  为模型生成的样本，所以可以高效完成）
- 以 MAF 模型（已训练）评估似然
- 最小化 KL 散度

[MintNet] 提供了一种可逆的卷积方式使其适用于流模型

日期： /

一种思考流模型训练目标的方式：

$$\min D_{KL}(P_{\text{data}} \parallel P_{\theta}(x))$$

$$= \min D_{KL}(P_{\bar{x}} \parallel P_x)$$

$$= \min D_{KL}(P_{f_0^{-1}(\bar{x})} \parallel P_{f_e^{-1}(x)})$$

$$= \min D_{KL}(P_{f_0^{-1}(\bar{x})} \parallel P_z)$$