

日期： /

Discrete problem

为什么使用现有的模型建模离散问题很困难？

① 现有的模型高度依赖微积分，粗暴地离散化通常会导致不可微

② 离散空间通常难以具有与连续空间一样的表现力

③ 将离散值嵌入连续空间会造成稀疏区域

因此目前只有一个真正适用于离散值的生成模型，即 Transformer

这种自回归模式的优点有：

① 高拓展性 ② 模型强大 ③ 对于 NLP 有合理的归纳偏置

缺点有：

① 自回归过程易导致错误积累

② 对于非语言问题（如 DNA），归纳偏置并不合理

③ 结构受限

④ 采样顺序进行，低效

能否将分步匹配用于离散任务？

日期： /

将分数推广至离散空间，首先考虑梯度的推广：

$$\nabla f(x) = [f(y) - f(x)]_{y \text{ neighbor of } x}$$

那么对于分数函数：

$$\nabla_x \log p(x) = \frac{\nabla p(x)}{p(x)} = \left[\underbrace{\frac{p(y)}{p(x)}}_{\substack{\text{归一化取值} \\ \rightarrow \text{梯度}}}\right]_{y=1}$$

如果对 x 的所有邻居 y 时，复杂度为 $O(N^{d-1})$
所以只选择一个邻居建模： $\frac{p(x^1 \dots x^{d-1}, y)}{p(x^1 \dots x^{d-1})}$
此时复杂度为 $O(Nd)$

$x^1 \quad x^2 \quad \dots \quad x^d$

↓ Seq-to-Seq Neural Network

$$\frac{p(x^1, x^2 \dots x^d)}{p(x^1, x^2 \dots x^d)} \quad \frac{p(x^1, 1 \dots x^d)}{p(x^1, x^2 \dots x^d)} \quad \dots \quad \frac{p(x^1, x^2 \dots 1)}{p(x^1, x^2 \dots x^d)}$$

$$\frac{p(x^1, x^2 \dots x^d)}{p(x^1, x^2 \dots x^d)} \quad \frac{p(x^1, 2 \dots x^d)}{p(x^1, x^2 \dots x^d)} \quad \dots \quad \frac{p(x^1, x^2 \dots 2)}{p(x^1, x^2 \dots x^d)}$$

⋮ ⋮ ⋮

$$\frac{p(x^1, x^2 \dots x^d)}{p(x^1, x^2 \dots x^d)} \quad \frac{p(x^1, N \dots x^d)}{p(x^1, x^2 \dots x^d)} \quad \dots \quad \frac{p(x^1, x^2 \dots N)}{p(x^1, x^2 \dots x^d)}$$

如何学习推广的分数？

学习目标： $S_\theta(x)$ s.t. $S_\theta(x)_y \approx \frac{p(y)}{p(x)}$

$$\min_{\theta} \sum_{y \neq x} S_\theta(x)_y - \frac{p(y)}{p(x)} \log S_\theta(x)_y \quad (\text{评分熵})$$

无法取得，使用隐式评分或去噪评分熵

日期： /

当 $S - \frac{p(y)}{p(x)} \log S$ 最小时，

$$(S - \frac{p(y)}{p(x)} \log S)'_S = 0$$

$$1 - \frac{p(y)}{p(x)} \cdot \frac{1}{S} = 0$$

$S = p(y)/p(x)$ 因此损失函数合理

[Denoising Score Entropy]

假设 $p(x) = \sum_{x_0} p(x|x_0) p_0(x_0)$

$$\begin{aligned} \text{则 } E_{x \sim p} \sum_{y \neq x} \frac{p(y)}{p(x)} \log S_\theta(x)_y &= \sum_x \sum_{y \neq x} \log S_\theta(x)_y p(y) \\ &= \sum_x \sum_{y \neq x} \log S_\theta(x)_y \sum_{x_0} p(y|x_0) p_0(x_0) \\ &= \sum_{x_0} \sum_{y \neq x} \log S_\theta(x)_y \frac{p(y|x_0)}{p(x|x_0) p_0(x_0)} \\ &= E_{x_0 \sim p_0, x \sim p(\cdot|x_0)} \sum_{y \neq x} \frac{p(y|x_0)}{p(x|x_0)} \log S_\theta(x)_y \end{aligned}$$

那么原损失函数可以写作：

$$E_{x_0 \sim p_0, x \sim p(\cdot|x_0)} \sum_{y \neq x} S_\theta(x)_y - \underbrace{\frac{p(y|x_0)}{p(x|x_0)} \log S_\theta(x)_y}_{\text{可计算}}$$

如何采样？

扩散是一个沿时间的进化过程 $P_t \in \mathbb{R}^{l \times l}$

写作微分方程的形式： $dP_t = Q_t P_t$

$$\Rightarrow \log P_t / P_0 = Q_t \Rightarrow P_t = e^{Q_t} \cdot P_0$$

日期： /

矩阵 Q_t 满足：

1. Q_t 的列求和为 0

2. 非对角元不小于 0

Q_t 控制状态人间转换的频率。

$$P(X_{t+\Delta t} = j | X_t = i) = \delta_{i,j} + Q_t(j,i) \Delta t + O(\Delta t^2)$$

高散化时间，类似 Euler-Maruyama 系统

例：有 $Q_t = \begin{bmatrix} -2 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & -2 \end{bmatrix}$, $P_0 = \begin{bmatrix} 0.5 \\ 0.2 \\ 0.3 \end{bmatrix}$

$$P_t = \exp(t Q_t) P_0 = \exp(t \begin{bmatrix} -2 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & -2 \end{bmatrix}) \begin{bmatrix} 0.5 \\ 0.2 \\ 0.3 \end{bmatrix}$$

通常我们取 $Q_t = \sigma(t) Q$

$$P_t = \exp(\sum(t) Q) P_0$$

$$\text{或 } P(X_{t+j} = j | X_0 = i) = \exp(\sum(t) Q)(j, i)$$

且 $t \rightarrow \infty$, $P_t \rightarrow P_{\text{base}}$

对于损失函数：

$$E_{t,x_0 \sim P_0, x_t \sim P_t(\cdot | x_0)} \sum_{y \neq x} S_0(x_t, t) y - \frac{P_t(y|x_0)}{P_t(x|x_0)} \log S_0(x_t, t) y$$

日期： /

对于去噪过程 $dP_{T-t} = \bar{Q}_{T-t} P_{T-t}$

$$\bar{Q}_t(i, j) = \frac{P_t(i)}{P_t(i, j)} Q_t(i, j) \quad (i \neq j)$$

$$\approx S_0(i, j); Q_t(i, j)$$

由于每次只试图对一个位置进行翻转直接以这种方式采样是低效的。

整体来看，模型的工作流程是：

1. 从数据分布中取得样本
2. 定义前向扩散过程
3. 以评分熵损失学习 S_0
4. 反向去噪

如何评估？

自回归常用困惑度： $PPL(x) = e^{-\frac{1}{d} \log p_\theta(x^1 \dots x^d)}$

对于 diffusion：

$$-\log p_\theta(x_0) \leq \int_0^t \sum_{y \neq x_t} Q_t(x_t, y) \left(S_0(x_t, t) y - \frac{p_{\theta|0}(y|x_0)}{p_{\theta|0}(x_t|x_0)} \log S_0(x_t, t) y \right) dt + C$$

$$PPL(x) \leq e^{-\frac{1}{d} DSE(x)}$$