

日期： /

## Energy-based model

EBMs 的特点包括：

- ① 非常灵活的网络结构
- ② 稳定的训练过程
- ③ 基模型可以灵活组合

对  $P(x)$  的建模是生成式模型的重要一环，但是参数化的模型  $P_\theta(x)$  需要满足

① 非负，  $P_\theta(x) \geq 0$

② 归一化，  $\int P_\theta(x) dx = 1$  不易满足，特别是对子连续变量

此前解决归一化的方式是可逆变换等，这限制了网络类型的  
选择。而 EBMs 使用如下的方式处理归一化：

$$P_\theta(x) = \frac{1}{Z(\theta)} g_\theta(x) = \frac{1}{\int g_\theta(x) dx} g_\theta(x) = \frac{1}{\text{Volume}(\theta)} g_\theta(x)$$

如果选择简单的  $g_\theta(x)$  形式，则可以封闭计算

$$\cdot g_{(\mu, \sigma)}(x) = e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ 时, } Z(\mu, \sigma) = \int e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sqrt{2\pi\sigma^2}$$

$$\cdot g_\lambda(x) = e^{-\lambda x} \text{ 时, } Z(\lambda) = \int e^{-\lambda x} dx = \frac{1}{\lambda}$$

$$\cdot g_\theta(x) = h(x) \cdot \exp(\theta \cdot T(x)) \text{ 时, } Z(\theta) = \exp(\log \int h(x) \exp(\theta \cdot T(x)) dx)$$

日期： /

在自回归模型中，我们处理归一化的方式类似于

$$\int_x \int_y P_{\theta}(x) P_{\theta'(x)}(y) dx dy = 1$$

其中  $P_{\theta}(x)$  与  $P_{\theta'(x)}(y)$  均为归一化分布。

在 VAE 中 (GMM)，处理归一化的方式类似于

$$\int_x [\alpha P_{\theta}(x) + (1-\alpha) P_{\theta'}(x)] dx = 1$$

其中  $P_{\theta}(x)$  与  $P_{\theta'}(x)$  均为归一化分布。

EBMs 中，我们关注一类特殊的模型形式：

$$P_{\theta}(x) = \frac{1}{\int e^{\exp(f_{\theta}(x))} dx} \exp(f_{\theta}(x)) = \frac{1}{Z(\theta)} \exp(f_{\theta}(x))$$

选择指数的原因在于：

① 可以捕捉大尺度的概率变化

② 是许多概率分布的形式，且满足非负的条件

③ 与统计力学的对偶性

-  $f_{\theta}(x)$  称为能量，能量越低的状态越可能发生

这种形式导致 EBM<sub>s</sub> 有极高的灵活性但同时难以采样和估计似然（因为依赖  $Z(\theta)$ ）。

日期： /

此外，在高维数据建模的情况下， $Z(\theta)$ 难以封闭计算

但是，对于一些任务， $Z(\theta)$ 是不必要的，例如比较可能性，没得这类模型由其擅长一些特殊任务，例如去噪和异常检测。

### [ Restricted Boltzmann machine (RBMs) ]

假设有2种变量：

$x \in \{0,1\}^n$ , 为显变量（例如像素点）

$z \in \{0,1\}^m$ , 为隐变量

$$\begin{aligned}\text{联合概率分布描述为 } P_{w,b,c}(x,z) &= \frac{1}{Z} \exp(x^T W z + b_x + c_z) \\ &= \frac{1}{Z} \exp\left(\sum_{i=1}^n \sum_{j=1}^m x_i z_j w_{ij} + b_x + c_z\right)\end{aligned}$$

- 并不考虑  $z_1, z_2$  之间， $x_1, x_2$  之间的相关性

- RBM 可以堆叠

$$- \text{RBM 中, } Z(w, b, c) = \sum_{x \in \{0,1\}^n} \sum_{z \in \{0,1\}^m} \exp(x^T W z + b_x + c_z)$$

仍极其昂贵，这也使得基于似然的学习难以实现

日期： /

EBMs 的训练目标可以写作： maximize  $\frac{\exp\{f_\theta(x_{train})\}}{Z(\theta)}$   
最大化分子，最小化分母

想法是并不精确计算  $Z(\theta)$ ，而是通过 Monte-Carlo 采样进行估计

[ Contrastive divergence algorithm ]

CD 算法通过梯度近似优化模型参数，避开了对配分函数的计算。

$$\max_\theta \frac{\exp[f_\theta(X_{train})]}{Z(\theta)} \Rightarrow \max_\theta \log \frac{\exp[f_\theta(X_{train})]}{Z(\theta)} \\ \Rightarrow \max_\theta [f_\theta(X_{train}) - \log Z(\theta)]$$

$$\nabla_\theta [f_\theta(X_{train}) - \log Z(\theta)]$$

$$= \nabla_\theta f_\theta(X_{train}) - \nabla_\theta \log Z(\theta)$$

$$= \nabla_\theta f_\theta(X_{train}) - \frac{\nabla_\theta Z(\theta)}{Z(\theta)}$$

$$= \nabla_\theta f_\theta(X_{train}) - \frac{1}{Z(\theta)} \int \nabla_\theta \exp\{f_\theta(x)\} dx$$

$$= \nabla_\theta f_\theta(X_{train}) - \frac{1}{Z(\theta)} \int \exp\{f_\theta(x)\} \cdot \nabla_\theta f_\theta(x) dx$$

$$= \nabla_\theta f_\theta(X_{train}) - \int \frac{\exp\{f_\theta(x)\}}{Z(\theta)} \cdot \nabla_\theta f_\theta(x) dx$$

目前的问题是，如何采样？

$$= \nabla_\theta f_\theta(X_{train}) - E_{X_{sample}} [\nabla_\theta f_\theta(X_{sample})] \\ \approx \nabla_\theta f_\theta(X_{train}) - \nabla_\theta f_\theta(X_{sample})$$

日期： /

## [Markov Chain Monte Carlo (MCMC)]

- Initialize  $x^0$  randomly ( $x^0 \sim \pi(x)$ )

- Let  $x' = x^t + \text{noise}$ :

if  $f_0(x') > f_0(x^t)$ , let  $x^{t+1} = x'$

else  $x^{t+1} = x^t$  with probability  $\exp(f_0(x') - f_0(x^t))$

- Go to step 2

高效探索概率空间，避免局部最优

$x^T$  converges to  $p_0(x)$  when  $T \rightarrow \infty$

## [Langevin MCMC]

- $x^0 \sim \pi(x)$

- Repeat for  $t = 0, 1, 2, \dots, T-1$ :

-  $z^t \sim N(0, 1)$

-  $x^{t+1} = x^t + \epsilon \nabla_x \log p_0(x) |_{x=x^t} + \sqrt{2\epsilon} z^t$  噪声

- $x^T$  converges to a sample from  $p_0(x)$  when  $T \rightarrow \infty$  and  $\epsilon \rightarrow 0$

对于连续的EBMs,  $\nabla_x \log p_0(x) = \nabla_x f_0(x)$

注意是对  $x$  的梯度

日期： /

尽管 MCMC 提供了可行的采样方式，但这些方式仍是昂贵的、特别是对于训练来说。

### [Score function]

对于 EBM  $P_\theta(x) = \frac{\exp[-f_\theta(x)]}{Z(\theta)}$ ,  $\log P_\theta(x) = -f_\theta(x) - \log Z(\theta)$

定义 (Stein) Score function 为

$$S_\theta(x) := \nabla_x \log P_\theta(x) = \nabla_x f_\theta(x) - \nabla_x \log Z(\theta) = \underline{\nabla_x f_\theta(x)}$$

与配分函数无关

对于 Gaussian 分布,  $P_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ ,  $S_\theta(x) = -\frac{x-\mu}{\sigma^2}$

对于 Gamma 分布,  $P_\theta(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ ,  $S_\theta(x) = \frac{\alpha-1}{x} - \beta$

$S_\theta(x)$  以向量场(梯度)的方式描述分布

### [Fisher divergency]

对于概率分布  $p(x), q(x)$ , 二者的 Fisher 散度定义为:

$$D_F(p, q) := \frac{1}{2} \mathbb{E}_{x \sim p} [\|\nabla_x \log p(x) - \nabla_x \log q(x)\|_2^2]$$

日期： /

## [ Score matching ]

分数匹配旨在使用 Fisher 敏度作为优化目标：

$$\frac{1}{2} \mathbb{E}_{x \sim P_{\text{data}}} [\|\nabla_x \log P_{\text{data}}(x) - \nabla_x \log p_\theta(x)\|_2^2]$$

然而  $\nabla_x \log P_{\text{data}}(x)$  是不可计算的，因为我们仅有有限的样本

若  $x$  为标量，对原式作以下处理：

$$\begin{aligned} & \frac{1}{2} \mathbb{E}_{x \sim P_{\text{data}}} [\|\nabla_x \log P_{\text{data}}(x) - \nabla_x \log p_\theta(x)\|_2^2] \\ = & \frac{1}{2} \int_{P_{\text{data}}(x)} [\nabla_x \log P_{\text{data}}(x)]^2 dx + \frac{1}{2} \int_{P_{\text{data}}(x)} [\nabla_x \log p_\theta(x)]^2 dx \\ & - \underbrace{\int_{P_{\text{data}}(x)} \nabla_x \log P_{\text{data}}(x) \nabla_x \log p_\theta(x) dx} \end{aligned}$$

$$\hookrightarrow \int F(x) dx = - \int_{P_{\text{data}}(x)} \frac{1}{P_{\text{data}}(x)} \nabla_x P_{\text{data}}(x) \nabla \log p_\theta(x) dx$$

$$\begin{aligned} (\text{分部积分}) &= - \underbrace{P_{\text{data}}(x) \nabla_x \log p_\theta(x)}_{x \rightarrow -\infty} + \int_{P_{\text{data}}(x)} \nabla_x^2 \log p_\theta(x) dx \\ &= \int_{P_{\text{data}}(x)} P_{\text{data}}(x) \nabla_x^2 \log p_\theta(x) dx \quad \downarrow \lim_{x \rightarrow +\infty} P_{\text{data}}(x) = 0 \end{aligned}$$

综上，损失函数可以写作：

$$\begin{aligned} L(x) &= P_{\text{data}}(x) \nabla_x^2 \log p_\theta(x) dx + \frac{1}{2} \int_{P_{\text{data}}(x)} (P_{\text{data}}(x) \nabla_x \log p_\theta(x))^2 dx + \\ &\quad \text{const.} \end{aligned}$$

$$= \mathbb{E}_{x \sim P_{\text{data}}} [\frac{1}{2} (\nabla_x \log p_\theta(x))^2 + \nabla_x^2 \log p_\theta(x)] + \text{const.}$$

日期： /

若  $x$  为变量，类似地有：

Hessian 矩阵

$$L(x) = \mathbb{E}_{x \sim \text{data}} \left[ \frac{1}{2} \|\nabla_x \log p_\theta(x)\|_2^2 + \text{tr}(\nabla_x^2 \log p_\theta(x)) \right] + \text{const.}$$

- 使用梯度下降优化
- 无需采样
- Hessian 矩阵在高维下极其昂贵

## [Noise contrastive estimation for training EBM's]

这是一种基于GAN的EBM训练方式，我们不从生成器而从已知的随机噪声  $P_n(x)$ 中采样：  
对训练结果影响大

对于传统GAN，最优的判别器有以下形式：

$$D_0(x) = \frac{P(x)}{P_{x \sim G(x)} + P(x)}$$

我们构建如下的判别器

$$D_{\theta, 2}(x) = \frac{\frac{e^{f(x)}}{z}}{\frac{e^{f(x)}}{z} + P_n(x)}$$

EBM, 其中  
z 也为可学习参数  
噪声

那么通过最小化交叉熵的方式训练  $D(x)$  分辨噪声和样本的能力，EBM就会趋近真实分布

日期： /

为了使噪声分布尽可能接近真实分布，我们同样可以对噪声分布参数化： $P_{n,\phi}(x)$ ，这可以是一个基于流的模型。此时，

$$D_{0.2,\phi}(x) = \frac{e^{f_\phi(x)}}{e^{f_\phi(x)} + P_{n,\phi}(x)} Z$$

模型的训练目标改为：

$$\min_{\phi} \max_{0.2} E_{x \sim p_{\text{data}}} [\log D_{0.2,\phi}(x)] + E_{x \sim p_{n,\phi}} [1 - D_{0.2,\phi}(x)]$$

更接近 GAN