

日期： /

Generative adversarial networks

GAN 提供了一种比较相似性的新方法，而不依赖于极大似然。

Why maximum likelihood?

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^M \log P_{\theta}(x_i) , x_1, x_2, \dots, x_m \sim P_{\text{data}}(x)$$

- 通常有最快的优化速度
- 高似然通常意味着低压缩损失，这是一个合理的训练目标

然而： 极大似然不一定对应高样本生成质量

对于采样自两个分布 P, Q 的样本 $S_1 = \{x \sim P\}$,
 $S_2 = \{x \sim Q\}$, GAN 提供了 KL 散度之外的另一种方式。
判断 P, Q 是否一致。与比较一般统计量的方式不同，我们使用神经网络区分样本，即 GAN 中的判别器。

日期： /

对于判别器 (Discriminator)，我们的训练目标为：

$$\max_{D\phi} V(p_\theta, D\phi) = \mathbb{E}_{x \sim P_{\text{data}}} [\log D\phi(x)] + \mathbb{E}_{x \sim p_\theta} [\log (1 - D\phi(x))]$$
$$\approx \sum_{x \in S_1} \log \phi(x) + \sum_{x \in S_2} \log (1 - \phi(x))$$

即交叉熵损失

最优判别器有以下形式： $D\phi^*(x) = \frac{P_{\text{data}}(x)}{P_{\text{data}}(x) + p_\theta(x)}$

所以，大体上，GAN 由以下两个部分组成：

① Generator (生成器)

$$z \xrightarrow{G\theta} x$$

G^θ 由一个简单的先验分布中采样，通过一系列映射产生输出。 G^θ 的形式可以是任意的，且不具有可逆性的要求。此外，我们不再关心 $x \rightarrow z$ 的编码过程。

对于生成器，训练目标是：

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim P_{\text{data}}} [\log D(x)] + \mathbb{E}_{x \sim p_\theta} [\log (1 - D(x))]$$

日期： /

② 判别器 (Discriminator)

对于最优判别器：

$$\begin{aligned} V(G, D_G^*(x)) &= \mathbb{E}_{x \sim P_{\text{data}}} \left[\log \frac{P_{\text{data}}(x)}{P_{\text{data}}(x) + P_G(x)} \right] + \mathbb{E}_{x \sim P_G} \left[\log \frac{P_G(x)}{P_G(x) + P_{\text{data}}(x)} \right] \\ &= \mathbb{E}_{x \sim P_{\text{data}}} \left[\log \frac{P_{\text{data}}(x)}{[P_{\text{data}}(x) + P_G(x)]/2} \right] + \mathbb{E}_{x \sim P_G} \left[\log \frac{P_G(x)}{[P_G(x) + P_{\text{data}}(x)]/2} \right] - \log 4 \\ &= \underline{D_{KL}[P_{\text{data}}, (P_{\text{data}} + P_G)/2]} + \underline{D_{KL}[P_G, (P_{\text{data}} + P_G)/2]} - \log 4 \\ &= 2D_{\text{JSD}}[P_{\text{data}}, P_G] \quad \text{Jensen-Shannon 散度} - \log 4 \end{aligned}$$

$D_{\text{JSD}}[P, q]$ 有一系列优良性质：

① $D_{\text{JSD}}[P, q] \geq 0$

② $D_{\text{JSD}}[P, q] = 0 \Leftrightarrow P = q$

③ $D_{\text{JSD}}[P, q] = D_{\text{JSD}}[q, p]$

这种设计为网络引入了更多的灵活性，同时使得采样过程更高效

日期： /

训练过程：

① 从数据集 D 中采样 $x^{(1)}, x^{(2)}, \dots, x^{(m)}$

② 从先验分布中采样 $z^{(1)}, z^{(2)}, \dots, z^{(m)}$

③ 更新判别器参数 ϕ

$$\nabla_{\phi} V(G_{\theta}, D_{\phi}) = \frac{1}{m} \nabla_{\phi} \sum_{i=1}^m [\log D_{\phi}(x^{(i)}) + \log (1 - D_{\phi}(G_{\theta}(z^{(i)})))]$$

④ 更新生成器参数 θ

$$\nabla_{\theta} V(G_{\theta}, D_{\phi}) = \frac{1}{m} \nabla_{\theta} \sum_{i=1}^m \log (1 - D_{\phi}(G_{\theta}(z^{(i)})))$$

GAN有以下问题：

① 难以训练，甚至不确定收敛

② 模式崩塌 (mode collapse)：GAN可能仅会学习到真实数据分布的一部分

[f-GAN]

实践中，可以以优化 f 散度 代替 J-S 散度作为目标：

$$D_f(p, q) = \mathbb{E}_{x \sim q} [f(\frac{p(x)}{q(x)})]$$

f 为一个凸的、下半连续的函数且 $f(1)=0$ 当 $f=u \log u$ 时，即 KL 散度类似 VAE 中的处理。 $\mathbb{E}_{x \sim q} [f(\frac{p(x)}{q(x)})] \geq f(\mathbb{E}_{x \sim q} [p(x)/q(x)]) = 0$

| Name | $D_f(P Q)$ | Generator $f(u)$ |
|---|---|---|
| Total variation | $\frac{1}{2} \int p(x) - q(x) dx$ | $\frac{1}{2} u - 1 $ |
| Kullback-Leibler | $\int p(x) \log \frac{p(x)}{q(x)} dx$ | $u \log u$ |
| Reverse Kullback-Leibler | $\int q(x) \log \frac{q(x)}{p(x)} dx$ | $-\log u$ |
| Pearson χ^2 | $\int \frac{(q(x) - p(x))^2}{p(x)} dx$ | $(u - 1)^2$ |
| Neyman χ^2 | $\int \frac{(p(x) - q(x))^2}{q(x)} dx$ | $\frac{(1-u)^2}{u}$ |
| Squared Hellinger | $\int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx$ | $(\sqrt{u} - 1)^2$ |
| Jeffrey | $\int (p(x) - q(x)) \log \left(\frac{p(x)}{q(x)} \right) dx$ | $(u - 1) \log u$ |
| Jensen-Shannon | $\frac{1}{2} \int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} dx$ | $-(u+1) \log \frac{1+u}{2} + u \log u$ |
| Jensen-Shannon-weighted | $\int p(x) \pi \log \frac{p(x)}{\pi p(x)+(1-\pi)q(x)} + (1-\pi)q(x) \log \frac{q(x)}{\pi p(x)+(1-\pi)q(x)} dx$ | $\pi u \log u - (1-\pi+\pi u) \log(1-\pi+\pi u)$ |
| GAN | $\int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} dx - \log(4)$ | $u \log u - (u+1) \log(u+1)$ |
| α -divergence ($\alpha \notin \{0, 1\}$) | $\frac{1}{\alpha(\alpha-1)} \int \left(p(x) \left[\left(\frac{q(x)}{p(x)} \right)^\alpha - 1 \right] - \alpha(q(x) - p(x)) \right) dx$ | $\frac{1}{\alpha(\alpha-1)} (u^\alpha - 1 - \alpha(u-1))$ |

对边缘概率的需求
模型自回归或可逆

$$D_f(P_\theta, P_{\text{data}}) = \mathbb{E}_{x \sim P_{\text{data}}} \left[f\left(\frac{P_\theta(x)}{P_{\text{data}}(x)}\right) \right]$$

外部期望可以采样近似

无法解析计算,
此届由神经网络给出

[Fenchel conjugate]

对于任意 $f(\cdot)$, 其凸共轭为 $f^*(t) = \sup_{u \in \text{dom}_f} (ut - f(u))$ 上确界

dom_f 是 $f(\cdot)$ 定义域

- f^* 为凸函数, 且为下半连续函数
- $f^{**} \leq f$ Prof: $f^*(u) \geq ut - f(t) \Rightarrow f(u) \geq ut - f^*(t) = f^{**}(u)$
- 当 f 为凸函数时, 且下半连续时, $f^{**} = f$

日期： /

通过 Fenchel conjugate 可以实现以下近似：

$$\begin{aligned} D_f(p, q) &= \mathbb{E}_{x \sim q} \left[f\left(\frac{p(x)}{q(x)}\right) \right] \\ &= \mathbb{E}_{x \sim q} \left[f^{**}\left(\frac{p(x)}{q(x)}\right) \right] \\ &= \mathbb{E}_{x \sim q} \left[\sup_{t \in \text{dom } f^*} \left(t \frac{p(x)}{q(x)} - f^*(t) \right) \right] \\ &= \mathbb{E}_{x \sim q} \left[T^*(x) \frac{p(x)}{q(x)} - f^*(T^*(x)) \right] \\ &= \int_X q(x) \left[T^*(x) \frac{p(x)}{q(x)} - f^*(T^*(x)) q(x) \right] dx \\ &= \int_X \left[T^*(x) p(x) - f^*(T^*(x)) q(x) \right] dx \\ &\stackrel{\substack{\text{假定有一个映射} \\ \text{可以将 } x \\ \text{映射为上确界} \\ \text{处 } x \text{ 的取值}}}{=} \sup_{T \in \mathcal{T}} \int_X [T(x) p(x) - f^*(T(x)) q(x)] dx \\ &\geq \sup_{T \in \mathcal{T}} \int_X [T(x) p(x) - f^*(T(x)) q(x)] dx \\ &= \sup_{T \in \mathcal{T}} [\mathbb{E}_{x \sim p_{\text{data}}} [T(x)] - \mathbb{E}_{x \sim p_{\text{gen}}} [f^*(T(x))]] \end{aligned}$$

模型族

那么我们对 f-GAN 的优化目标就可以写为：

$$\min_{\phi} \max_{\theta} F(\theta, \phi) = \mathbb{E}_{x \sim p_{\text{data}}} [T_\phi(x)] - \mathbb{E}_{x \sim p_{\phi}} [f^*(T_\phi(x))]$$

日期： /

[Wasserstein - GAN]

传统的散度指标在生成样本与真实样本相差很大时
难以提供有意义的梯度信息。

Wasserstein (Earth-Mover) 距离的定义是

$$D_w(p, q) = \inf_{r \in \Pi(p, q)} E_{(x, y) \sim r} [\|x - y\|_1]$$

$\Pi(p, q)$ 是所有可能的概率测度集合，描述所有可能的 p, q 联合概率分布，其中元素 $r(x, y)$ 满足 $\int r(x, y) dy = p(x)$
且 $\int r(x, y) dx = q(y)$
直观上， $D_w(p, q)$ 可以理解为将 p 分布下的质量转换至 q 分布所需的“运输成本”。

这个问题有一个变分表征，可以写作以下优化问题：

$$D_w(p, q) = \sup_{\|f\|_{L^1} \leq 1} E_{x \sim p}[f(x)] - E_{x \sim q}[f(x)]$$

$\|f\|_{L^1}$ 意味着 $f(x)$ 的 Lipschitz 常数不超过 1，即

$$\forall x, y : |f(x) - f(y)| \leq \|x - y\|_1$$

日期： /

那么对于 WGAN，优化目标是：

$$\min_{\theta} \max_{\phi} \mathbb{E}_{x \sim P_{\text{data}}} [D\phi(x)] - \mathbb{E}_{z \sim p(z)} [D\phi(G_{\theta}(z))]$$

$\|f\|_{L_1}$ 通过裁剪网络权重的方式实现

如何使用 GAN 获得隐变量？

① 在不同任务上对判别器进行微调

② 添加一个编码器网络，且同时关注样本的
隐变量表示。

[BiGAN]

