GAN, VAE and Semi-Supervised Learning: Part I

Shell Xu Hu

January 19, 2018

Outline

Probability Density Estimation

- **Problem**: $\max_\theta \mathbb{E}_{x \sim p_{\text{data}}} \Big[\log p_\theta(x) \Big]$.
- Generative Adversarial Networks (GAN): Define p_θ implicitly by a mapping $G : z \to x$.
- Variational Auto Encoder (VAE): Define p_θ explicitly, such as Gaussian, Laplace; Amortize the posterior $p_\theta(z|x)$ by an inference network $q_\phi(z|x)$ with additional parameter ϕ.

Generative Adversarial Networks

- The objective of GAN is inspired by logistic regression:

Logistic Regression

$$\max_{\theta} L(\theta) := \sum_i y_i \log p_\theta(y_i \mid x_i) + (1 - y_i) \log\left(1 - p_\theta(y_i \mid x_i)\right)$$
$$= \mathbb{E}_{i:\ y_i=1}\Big[\log p_\theta(y_i \mid x_i)\Big] + \mathbb{E}_{i:\ y_i=0}\Big[\log\left(1 - p_\theta(y_i \mid x_i)\right)\Big]$$

GAN

$$\min_G \max_D L(D, G) := \mathbb{E}_{x\sim p_{\text{data}}}\Big[\log D(x)\Big] + \mathbb{E}_{x\sim p_G}\Big[\log(1 - D(x))\Big]$$
$$= \mathbb{E}_{x\sim p_{\text{data}}}\Big[\log D(x)\Big] + \mathbb{E}_{z\sim p_z}\Big[\log\Big(1 - D\big(G(z)\big)\Big)\Big]$$

Ideal Training Process

GAN Objective

$$\min_G \max_D L(D, G) := \mathbb{E}_{x \sim p_{\text{data}}}\Big[\log D(x)\Big] + \mathbb{E}_{x \sim p_G}\Big[\log(1 - D(x))\Big]$$
$$= \mathbb{E}_{x \sim p_{\text{data}}}\Big[\log D(x)\Big] + \mathbb{E}_{z \sim p_z}\Big[\log\Big(1 - D\big(G(z)\big)\Big)\Big]$$

- Update D: Solve inner maximization to optimum: $D^*(G^t)$.
- Update G: $G^{t+1} \leftarrow G^t - \gamma \nabla_G L(D^*(G^t), G^t)$.

Ideal Training Process

Proposition

$$D^*(G)(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_G(x)}$$

Proof.

$L(D, \cdot) = \int_x f(D(x))dx$ with $f(q) = a \log q + b \log(1 - q)$, where $a, b \in (0, 1]$ are constants. Note that

$$\max_{q \in [0,1]} a \log q + b \log(1 - q) \Leftrightarrow q^* = \frac{a}{a + b}.$$

□

Ideal Training Process

GAN Objective

$$\min_G \max_D L(D, G) := \mathbb{E}_{x\sim p_{\text{data}}}\Big[\log D(x)\Big] + \mathbb{E}_{x\sim p_G}\Big[\log(1 - D(x))\Big]$$

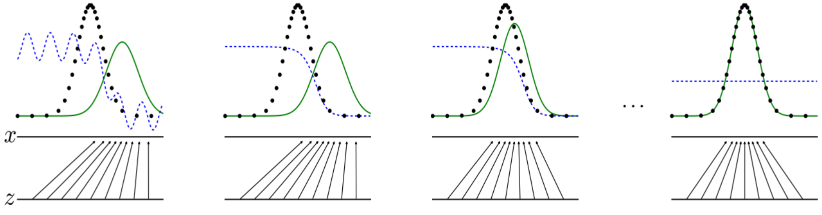
Given $D^*(G)$, we have

$$\begin{aligned} L(D^*(G), G) &= \mathbb{E}_{x\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_G(x)}\right] + \mathbb{E}_{x\sim p_G}\left[\log \frac{p_G(x)}{p_{\text{data}}(x) + p_G(x)}\right] \\ &= -\log 4 + \mathbb{E}_{x\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(x)}{\frac{p_{\text{data}}(x)+p_G(x)}{2}}\right] + \mathbb{E}_{x\sim p_G}\left[\log \frac{p_G(x)}{\frac{p_{\text{data}}(x)+p_G(x)}{2}}\right] \\ &= -\log 4 + D_{\text{KL}}(p_{\text{data}} \| \frac{p_{\text{data}} + p_G}{2}) + D_{\text{KL}}(p_G \| \frac{p_{\text{data}} + p_G}{2}) \\ &= -\log 4 + D_{\text{JSD}}(p_{\text{data}} \| p_G) \end{aligned}$$

Ideal Training Process

GAN Objective

$$\min_G \max_D L(D, G) := \mathbb{E}_{x \sim p_{\text{data}}} \Big[\log D(x) \Big] + \mathbb{E}_{x \sim p_G} \Big[\log(1 - D(x)) \Big]$$
$$= \mathbb{E}_{x \sim p_{\text{data}}} \Big[\log D(x) \Big] + \mathbb{E}_{z \sim p_z} \Big[\log \Big(1 - D(G(z)) \Big) \Big]$$



(a) (b) (c) (d)

Actual Training Process

for k steps **do**

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Sample minibatch of m examples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.
- Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \right].$$

end for

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right).$$

Actual Training Process

- **Issue**: Early in training, when G is poor, $D(G(z)) = 0$ for almost all z, which means

$$\log(1 - D(G(z))) \equiv 0 \quad \text{saturates.}$$

- So, instead of

$$\min_G \mathbb{E}_{z \sim p_z} \Big[\log \Big(1 - D\big(G(z)\big) \Big) \Big] \tag{1}$$

$$\longrightarrow \max_G \mathbb{E}_{z \sim p_z} \Big[\log D\big(G(z)\big) \Big] \tag{2}$$

Actual Training Process

- **An alternative of G update** (sum of (1) and (2)):

$$\max_G \mathbb{E}_{z\sim p_z}\left[\log \frac{D\big(G(z)\big)}{1-D\big(G(z)\big)}\right] \tag{3}$$

- When $D \to D^*(G)$, $D^*(G) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x)+p_G(x)}$.

$$\mathbb{E}_{x\sim p_G}\left[\log \frac{D(x)}{1-D(x)}\right] \simeq \mathbb{E}_{x\sim p_G}\left[\log \frac{D^*(x)}{1-D^*(x)}\right]$$
$$= \mathbb{E}_{x\sim p_G}\left[\log \frac{p_{\text{data}}(x)}{p_G(x)}\right] = -D_{\text{KL}}(p_G \| p_{\text{data}})$$

- **Interpretation**:
  - D-step makes a good approximation of the *density ratio* $\frac{p_{\text{data}}(x)}{p_G(x)}$.
  - G-step minimizes $D_{\text{KL}}(p_{\text{data}} \| p_G)$.

Variational Auto Encoder

VAE Objective

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)}\Big[\log p_\theta(x, z) - \log q_\phi(z|x)\Big]$$

$$\max_{\theta,\phi} L(\theta, \phi) := -D_{\mathrm{KL}}\Big(q_\phi(z|x) \| p(z)\Big) + \mathbb{E}_{q_\phi(z|x)}\Big[\log p_\theta(x|z)\Big]$$

- An example of p_θ and q_ϕ:
  - $p_\theta(x|z) := \mathcal{N}(x; f(z), \sigma^2 I)$, where $\theta = f$.
  - $q_\phi(z|x) := \mathcal{N}(z; \mu(x), \Sigma(x))$, where $\phi = (\mu, \Sigma)$.

Variational Auto Encoder