University of Illinois Spring 2018

CS 446: Machine Learning
Homework 11

Due on Tuesday, April 24, 2018, 11:59 a.m. Central Time

1. **[8 points]** Generative Adversarial Network (GAN)

   (a) What is the cost function for classical GANs? Use $D_w(x)$ as the discriminator and $G_\theta(z)$ as the generator, where the generator transforms $z \sim Z$ to $x \in X$.

   Your answer:

   $$\min_\theta \max_w V(D_w, G_\theta) = \mathbb{E}_{x\sim p_{data}}[\log D_w(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - G_\theta(D_w(x)))]$$

   (b) Assume arbitrary capacity for both discriminator and generator. In this case we refer to the discriminator using $D(x)$, and denote the distribution on the data domain induced by the generator via $p_G(x)$. State an equivalent problem to the one asked for in part (a), by using $p_G(x)$ and the ground truth data distribution $p_{data}(x)$.

   Your answer:

   $$\max_D V(D, G) = \int_x p_{data}(x)\log(D(x)) + p_g(x)\log(1 - D(x))dx$$

(c) Assuming arbitrary capacity, derive the optimal discriminator $D^*(x)$ in terms of $p_{data}(x)$ and $p_G(x)$.
You may need the Euler-Lagrange equation:

$$\frac{\partial L(x, D, \dot{D})}{\partial D} - \frac{d}{dx}\frac{\partial L(x, D, \dot{D})}{\partial \dot{D}} = 0$$

where $\dot{D} = \partial D / \partial x$.

Your answer: Because D(x) can be any function, we can maximize it point-wise using the Euler-Lagrange equation where:

$$\begin{aligned} L(x, D, \dot{D}) &= p_{data}(x) \log D(x) + p_g(x) \log(1 - D(x)) \\ \frac{\partial L(x, D, \dot{D})}{\partial D} &= \frac{p_{data}(x)}{D(x)} - \frac{p_g(x)}{1 - D(x)} \\ \frac{\partial L(x, D, \dot{D})}{\partial \dot{D}} &= 0 \end{aligned}$$

Substituting into the Euler-Lagrange equation yields:

$$\begin{aligned} \frac{\partial L(x, D, \dot{D})}{\partial D} + \frac{d}{dx}\frac{\partial L(x, D, \dot{D})}{\partial \dot{D}} &= 0 \\ \frac{p_{data}(x)}{D(x)} - \frac{p_g(x)}{1 - D(x)} &= 0 \\ \frac{p_{data}(x)}{D(x)} &= \frac{p_g(x)}{1 - D(x)} \\ p_{data}(x) - p_{data}(x) D(x) &= p_g(x) D(x) \\ p_g(x) D(x) + p_{data}(x) D(x) &= p_{data}(x) \\ D^*(x) &= \frac{p_{data}(x)}{p_g(x) + p_{data}(x)} \end{aligned}$$

(d) Assume arbitrary capacity and an optimal discriminator $D^*(x)$, show that the optimal generator, $G^*(x)$, generates the distribution $p_G^* = p_{data}$, where $p_{data}(x)$ is the data distribution
You may need the Jensen-Shannon divergence:

$$\text{JSD}(p_{\text{data}}, p_G) = \frac{1}{2} D_{KL}(p_{\text{data}}, M) + \frac{1}{2} D_{KL}(p_G, M) \quad \text{with} \quad M = \frac{1}{2}(p_{\text{data}} + p_G)$$

Your answer: For $p_g(x) = p_{data}(x), D^*(x) = \frac{1}{2}$, at this point:

$$\begin{aligned}
C(G) = V(D,G) &= \mathbb{E}_{\mathbf{x}\sim p_{data}}\left[\log \frac{p_{data}(\mathbf{x})}{p_{data}(\mathbf{x}) + p_g(\mathbf{x})}\right] + \mathbb{E}_{\mathbf{x}\sim p_g}\left[\log \frac{p_{data}(\mathbf{x})}{p_{data}(\mathbf{x}) + p_g(\mathbf{x})}\right] \\
&= \mathbb{E}_{\mathbf{x}\sim p_{data}}\left[\log \frac{1}{2}\right] + \mathbb{E}_{\mathbf{x}\sim p_g}\left[\log \frac{1}{2}\right] \\
&= -\log 4
\end{aligned}$$

This is a candidate value for the global minimum for C(G).

Solving for $C(G) = V(D^*, G)$ yields:

$$\begin{aligned}
C(G) &= \int_x p_{data}(x)\log\left(\frac{p_{data}(x)}{p_g(x) + p_{data}(x)}\right) + p_g(x)\log\left(\frac{p_g(x)}{p_g(x) + p_{data}(x)}\right) \mathrm{d}x \\
&= \int_x (\log 2 - \log 2)p_{data}(x) + p_{data}(x)\log\left(\frac{p_{data}(x)}{p_G(x) + p_{data}(x)}\right) \\
&\quad + (\log 2 - \log 2)p_G(x) + p_G(x)\log\left(\frac{p_G(x)}{p_G(x) + p_{data}(x)}\right) \mathrm{d}x \\
C(G) &= -\log 2 \int_x p_g(x) + p_{data}(x)\,\mathrm{d}x \\
&\quad + \int_x p_{data}(x)\left(\log 2 + \log\left(\frac{p_{data}(x)}{p_g(x) + p_{data}(x)}\right)\right) \mathrm{d}x \\
&\quad + \int_x p_g(x)\left(\log 2 + \log\left(\frac{p_g(x)}{p_g(x) + p_{data}(x)}\right)\right) \mathrm{d}x
\end{aligned}$$

Since the integral of a pdf over its support is always 1 the first term reduces to $-\log 4$, additionally using log properties we can reformulate the second and third terms into KL divergences.

$$\begin{aligned}
C(G) = &-\log 4 \\
&+ D_{KL}\left(p_{data}(x)||\frac{p_{data}(x) + p_g(x)}{2}\right) \\
&+ D_{KL}\left(p_g(x)||\frac{p_{data}(x) + p_g(x)}{2}\right)
\end{aligned}$$

These divergences actually form the Jensen-Shannon divergence:

$$\begin{aligned}
C(G) = &-\log 4 \\
&+ D_{KL}\left(p_{data}(x)||\frac{p_{data}(x) + p_g(x)}{2}\right) \\
&+ D_{KL}\left(p_g(x)||\frac{p_{data}(x) + p_g(x)}{2}\right)
\end{aligned}$$

$$C(G) = -\log 4 + 2 \cdot JSD\left(p_{data}(x)|p_g(x)\right)$$

Since JSD between two distributions is always non-negative and zero only when the arguments are equal, the global minimum value of $C(G)$ is $-\log 4$. The global minimum occurs iff is $p_g(x) = p_{data}(x)$ meaning $p^*_G(x) = p_{data}(x)$.

(e) More recently, researchers have proposed to use the Wasserstein distance instead of divergences to train the models since the KL divergence often fails to give meaningful information for training. Consider three distributions, $\mathbb{P}_1 \sim U[0,1]$, $\mathbb{P}_2 \sim U[0.5,1.5]$, and $\mathbb{P}_3 \sim U[1,2]$. Calculate $D_{KL}(\mathbb{P}_1, \mathbb{P}_2)$, $D_{KL}(\mathbb{P}_1, \mathbb{P}_3)$, $\mathbb{W}_1(\mathbb{P}_1, \mathbb{P}_2)$, and $\mathbb{W}_1(\mathbb{P}_1, \mathbb{P}_3)$, where $\mathbb{W}_1$ is the Wasserstein-1 distance between distributions.

Your answer:

$$\begin{aligned}
D_{KL}(U[0,1]||U[0.5,1.5]) &= -\int_x U[0,1] \log \frac{U[0.5,1.5]}{U[0,1]} dx \\
&= \sum \begin{cases} -\int_0^{0.5} 1 \cdot \log 0 \quad dx \\ -\int_{0.5}^{1} 1 \cdot \log 1 \quad dx \\ -\int_1^{1.5} 0 \cdot \log \infty \quad dx \end{cases} \\
&= \text{undef} \\
D_{KL}(U[0,1]||U[1,2]) &= -\int_x U[0,1] \log \frac{U[1,2]}{U[0,1]} dx \\
&= \sum \begin{cases} -\int_0^{1} 1 \cdot \log 0 \quad dx \\ -\int_1^{2} 0 \cdot \log \infty \quad dx \end{cases} \\
&= \text{undef} \\
\mathbb{W}_1(U[0,1]||U[0.5,1.5]) &= \int_0^{0.5} x dx + \int_{0.5}^{1} 0.5 dx + \int_1^{0.5} 1.5 - x dx \\
&= \frac{x^2}{2}\bigg|_0^{0.5} + \frac{x}{2}\bigg|_{0.5}^{1} + (1.5x - \frac{x^2}{2})\bigg|_1^{1.5} \\
&= \frac{1}{8} + \left[\frac{1}{2} - \frac{1}{4}\right] + \left[\frac{9}{8} - \frac{8}{8}\right] \\
&= \frac{1}{2} \\
\mathbb{W}_1(U[0,1]||U[1,2]) &= \int_0^{1} x dx + \int_1^{2} 2 - x dx \\
&= \frac{x^2}{2}\bigg|_0^{1} + (2x - \frac{x^2}{2})\bigg|_1^{2} \\
&= \frac{1}{2} + \left[(4-2) - (2 - \frac{1}{2})\right] \\
&= 1
\end{aligned}$$