University of Illinois Spring 2018

CS 446: Machine Learning
Homework 11

Due on Tuesday, April 17, 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.

   Solution:

   $$\max_\theta \min_w -\sum_x \log D_w(x) - \sum_z \log(1 - D_w(G_\theta(z)))$$

   (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 $p_G(x)$.

   Solution:
   (1 point)

   $$\max_{p_G} \min_D -\int_x p_{\text{data}}(x) \log D(x) dx - \int_x p_G(x) \log(1 - D(x)) dx \tag{1}$$

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

   Solution:
   (2 points) Use the Euler-Lagrange formalism which says that the stationary point of $S(D) = \int_x L(x, D, \dot{D}) dx$ can be obtained from

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

   to demonstrate that a stationary point of GANs (use Eq. 1) is obtained for $p_D = p_G$. Note that $\dot{D} = \partial D / \partial x$.

   $$\frac{\partial L(x, D, \dot{D})}{\partial D} = -\frac{p_{\text{data}}}{D} + \frac{p_G}{1 - D} = 0$$

   Consequently:

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

(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:

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

Solution:
(2 points)

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

Therefore:

$$\begin{aligned} & -\int_x p_{\text{data}}(x) \log D^*(x) + p_G(x) \log(1 - D^*(x)) dx \\ = & -\int_x p_{\text{data}}(x) \log \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_G(x)} + p_G(x) \log \frac{p_G(x)}{p_{\text{data}}(x) + p_G(x)} dx \\ = & -2\,\text{JSD}(p_{\text{data}}, p_G) + \log(4) \end{aligned}$$

Hence:

$$p_{\text{data}} = p_G$$

(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.

Solution:
(2 points) Since the distributions don't overlap ,

$$D_{KL}(\mathbb{P}_1, \mathbb{P}_2) = \infty$$

$$D_{KL}(\mathbb{P}_1, \mathbb{P}_3) = \infty$$

For the Wasserstein distance, an optimal transport map between the 1-d Uniform distributions is simply a horizontal shift of appropriate size.

$$\mathbb{W}_1(\mathbb{P}_1, \mathbb{P}_2) = 0.5$$

$$\mathbb{W}_1(\mathbb{P}_1, \mathbb{P}_3) = 1$$