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:

   (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:

(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:

(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:

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

Your answer:

(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: