University of Illinois Fall 2023

ECE 544: Pattern Recognition
Problem Set 5

Due: Thursday, November 9, 2023, 11:59 pm

1. **[Variational Auto-Encoders (VAEs)]**

   (a) We want to maximize the log-likelihood $\log p_\theta(x)$ of a model $p_\theta(x)$ which is parameterized by θ. To this end we introduce a joint distribution $p_\theta(x, z)$ and an approximate posterior $q(z|x)$ and reformulate the log-likelihood via

   $$\log p_\theta(x) = \log \sum_z q(z|x) \frac{p_\theta(x, z)}{q(z|x)}.$$

   Use Jensen's inequality to obtain a bound on the log likelihood and divide the bound into two parts, one of which is the Kullback-Leibler (KL) divergence

   $$\mathrm{KL}(q(z|x), p(z)).$$

   (b) State at least two properties of the KL divergence.

   (c) Let

   $$q(z|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(z - \mu_q)^2\right).$$

   What is the value for the KL-divergence $\mathrm{KL}(q(z|x), q(z|x))$ and why?

   (d) Further, let

   $$p(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(z - \mu_p)^2\right).$$

   Note the difference of the means for $p(z)$ and $q(z|x)$ while their standard deviation is identical. What is the value for the KL-divergence $\mathrm{KL}(q(z|x), p(z))$ in terms of μ_p, μ_q and σ?

   (e) Now, let $q(z|x)$ and $p(z)$ be arbitrary probability distributions. We want to find that $q(z|x)$ which maximizes

   $$\sum_z q(z|x) \log p_\theta(x|z) - \mathrm{KL}(q(z|x), p(z))$$

   subject to $\sum_z q(z|x) = 1$. Ignore the non-negativity constraints. State the Lagrangian and compute its stationary point, *i.e.*, solve for $q(z|x)$ which depends on $p_\theta(x|z)$ and $p(z)$. Make sure to get rid of the Lagrange multiplier.

   (f) Which of the following terms should $q(z|x)$ be equal to: (1) $p(z)$; (2) $p_\theta(x|z)$; (3) $p_\theta(z|x)$; (4) $p_\theta(x, z)$.

   (g) Provide the code for implementing the 'reparameterize' function in `A8_VAE.py`. Report the final loss and provide the final sampled output of the network generated by the code.

2. **[Variational Autoencoders (VAEs)]**

Suppose we are given a dataset with N data points $\{x_i\}_{i=1}^N$ where each of the data points is a D-dimensional vector. We use VAEs to learn the distribution of the data. Let z denote the unobserved latent variable. We refer to the approximated posterior $q_\phi(z|x)$ as the encoder and to the conditional distribution $p_\theta(x|z)$ as the decoder. Use the above notations to answer the following questions.

(a) The empirical lower bound (ELBO) of $p_\theta(x_i)$ is

$$\mathcal{L}(\theta, \phi, x_i) = -D_{KL}(q_\phi(z|x_i)||p(z)) + E_{q_\phi(z|x_i)}\Big[\log p_\theta(x_i|z)\Big]$$

Write down the minimization program that VAEs solve in terms of $p_\theta(\cdot)$ and $q_\phi(\cdot)$ given the dataset $\mathbf{x} = \{x_i\}_{i=1}^N$.

(b) Write down the reconstruction error in the loss function of VAEs in your previous answer, again given the dataset $\mathbf{x} = \{x_i\}_{i=1}^N$.

(c) Write down the formula to compute the reconstruction error empirically by drawing M samples from the distribution $q_\phi(z|x_i)$. Denote these M samples as $z_{i,m}$, where $m = 1, 2, ..., M$.

(d) Let $f(z_{i,m}) \in \mathbf{R}^D$ be the reconstructed sample with respect to $z_{i,m}$, which is the output of the decoder. What is the empirical reconstruction error if we assume $p_\theta(x_i|z)$ to be a Gaussian distribution $\mathcal{N}(f(z), \sigma^2\mathbf{I})$, where σ is a constant and $\mathbf{I}$ is the D-by-D identity matrix (simplify as much as possible)?

(e) Now consider all the data points x_i to be binary, i.e., $\forall i$, $x_i \in \{0, 1\}^D$. If we want to have the empirical reconstruction error to be the cross entropy loss, what should we assume $p_\theta(x_i|z)$ to be? What is the name of the distribution? Let the output of the decoder be $g = f(z) \in [0, 1]^D$, where the values are all between 0 and 1. If you need, use $x_i^{(d)}$ to denote the d-th element in the vector x_i.

3. **[Generative Adversarial Nets (GANs) and Duality]**

Consider the following program for a dataset $\mathcal{D} = \{(x)\}$ of points:

$$\max_{\theta} \min_{w} -\sum_{x \in \mathcal{D}} \log p_w(y = 1|x) - \sum_{z \in \mathcal{Z}} \log(1 - p_w(y = 1|G_\theta(z))) + \frac{C}{2}\|w\|_2^2. \tag{1}$$

Hereby θ denotes the parameters of the generator $G_\theta(z)$, which transforms 'perturbations' $z \in \mathcal{Z}$ into artificial data, w refers to the parameters of the discriminator model $p_w(y|x)$, $y \in \{0, 1\}$ denotes artificial or real data, and $C \geq 0$ is a fixed hyper-parameter.

(a) Without restrictions on the generator model G_θ and the discriminator model p_w, what are challenges in solving the program given in Eq. (1)?

(b) We now restrict the discriminator as follows:

$$p_w(y = 1|x) = \frac{1}{1 + \exp w^\top x}.$$

Using this discriminator, write down the resulting cost function for the program given in Eq. (1).

(c) When is the function $\frac{C}{2}\|a\|_2^2 - a^\top b$ convex in a? Why?

(d) When is the function $\log(1 + \exp a^\top b)$ convex in a? Why?

(e) Assume we restrict ourselves to the domain (if any) where $\frac{C}{2}\|a\|_2^2 - a^\top b$ and $\log(1 + \exp a^\top b)$ are convex in a, what can we conclude about convexity of the function

$$\sum_{x \in \mathcal{D}} \log(1 + \exp w^\top x) + \sum_{z \in \mathcal{Z}} \log(1 + \exp(w^\top G_\theta(z))) - \sum_{z \in \mathcal{Z}} w^\top G_\theta(z) + \frac{C}{2}\|w\|_2^2$$

in w and why?

(f) Let us introduce variables $\xi_x = w^\top x$ and $\xi_z = w^\top G_\theta(z)$ and let us consider the following program:

$$\min_{w} \quad \sum_{x \in \mathcal{D}} \log(1 + \exp \xi_x) + \sum_{z \in \mathcal{Z}} \log(1 + \exp \xi_z) - \sum_{z \in \mathcal{Z}} w^\top G_\theta(z) + \frac{C}{2}\|w\|_2^2 \tag{2}$$

$$\text{s.t.} \quad \begin{cases} \xi_x = w^\top x & \forall x \in \mathcal{D} \quad \text{(C1)} \\ \xi_z = w^\top G_\theta(z) & \forall z \in \mathcal{Z} \quad \text{(C2)} \end{cases}.$$

What is the Lagrangian for this program? Use the Lagrange multipliers λ_x and λ_z for the constraints (C1) and (C2) respectively.

(g) What is the value of

$$\min_{w} \frac{C}{2}\|w\|_2^2 - w^\top b$$

in terms of b and C?

(h) What is the value of

$$\min_{\xi} \lambda\xi + \log(1 + \exp \xi)$$

in terms of λ? What is the valid domain for λ?

(i) Combine your results from the previous two sub-problems to derive the dual function of the program given in Eq. (2). Also state the dual program and clearly differentiate it from the dual function. State how this dual program can help to address a challenge in GAN training.

4. **[Generative Adversarial Network (GAN)]**

   (a) What is the key difference between VAE and GAN?

   (b) What is the cost function for classical GANs? Use $D_w(x)$ as the discriminator and $G_\theta(x)$ as the generator.

   (c) 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 $p_G(x)$.

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

   (e) 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:

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

   (f) What is the optimal discriminator $D^*(x)$, assuming arbitrary capacity and optimal generator?