Homework 4
Variational Autoencoders

November 4, 2020

Submission Instructions You must typeset the answers to the theory questions using LaTeX or Microsoft Word and compile them into a single PDF file. Add your NetID to every filename you submit. Create a ZIP file containing both the PDF file and the completed `hw_4.ipynb` notebook. Name it `lastname-firstname-netid-hw4.zip`. Submit the ZIP file on NYU Classes. The due date is November 13, 2020, 11:55 PM (NYC time).

Show your work in the problems below!!

1 Variational autoencoders. (60 points).

This problem will help deepen your understanding of *variational autoencoders*.

Our goal is to learn a *latent variable model* of the form:

$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x} \mid \mathbf{z}) p_\theta(\mathbf{z}), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^D$, and $\mathbf{z} \in \mathbb{R}^z$ is a latent variable. For this model, $p_\theta(\mathbf{x})$ is of the form,

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x} \mid \mathbf{z}) p_\theta(\mathbf{z}) d\mathbf{z}, \tag{2}$$

which is difficult to approximate in high dimensions using samples from $p_\theta(\mathbf{z})$, implying that directly maximizing $\log p_\theta(\mathbf{x})$ is difficult.

1. ELBO. Instead, we will *maximize a lower bound* to $\log p_\theta(\mathbf{x})$,

$$\log p_\theta(\mathbf{x}) \geq \mathrm{ELBO}(\theta, \phi; \mathbf{x}), \tag{3}$$

where

$$\mathrm{ELBO}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})} \right], \tag{4}$$

and $q_\phi(\mathbf{z} \mid \mathbf{x})$ is a neural network. "ELBO" stands for "evidence lower-bound".

1. **Your task:** Show that,

$$\log p_\theta(\mathbf{x}) = \text{ELBO}(\theta, \phi; \mathbf{x}) + \text{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}|\mathbf{x})\right], \tag{5}$$

starting the derivation from:

$$\begin{aligned}\log p_\theta(\mathbf{x}) &= \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log p_\theta(\mathbf{x}) d\mathbf{z} \\ &= ...\end{aligned}$$

2. **Your task:** Explain why Equation 5 implies that $\log p_\theta(\mathbf{x}) \geq \text{ELBO}(\theta, \phi; \mathbf{x})$, and explain under what condition $\log p_\theta(\mathbf{x}) = \text{ELBO}(\theta, \phi; \mathbf{x})$.

2. ELBO surgery. You may have noticed that $\text{ELBO}(\theta, \phi; \mathbf{x})$ above (Equation 4) does not look like the VAE objective from the lecture or the lab.

1. **Your task:** Show that,

$$\text{ELBO}(\theta, \phi; \mathbf{x}) \equiv \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] - \text{KL}\left[q_\phi(\mathbf{z} \mid \mathbf{x}) \| p_\theta(\mathbf{z})\right], \tag{6}$$

which is the objective that we used to train VAEs in the lab.

2. **Your task:** What is the role of the $\mathbb{E}[\log p_\theta(\mathbf{x} \mid \mathbf{z})]$ term? What is the role of the $-\text{KL}[...]$ term? (*Any interpretation that is consistent with what was presented in the lectures or labs is fine.)*

3. ***Bonus:*** **Your task:** We'll look at one more form of the ELBO that gives some more insight into q_ϕ. Show that,

$$\text{ELBO}(\theta, \phi; \mathbf{x}) \equiv \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}, \mathbf{z})\right] + \mathbb{H}\left[q_\phi(\mathbf{z} \mid \mathbf{x})\right], \tag{7}$$

where $\mathbb{H}[\cdot]$ is the entropy.

Intuitively, the left-hand term encourages q_ϕ to assign most of its probability mass to regions of high joint density, while maximizing the entropy (right-hand term) discourages q_ϕ from concentrating its mass on a small number of regions.

3. Reconstruction loss. In the preceding problems, you derived the lower bound (ELBO, Equation 6) that we maximize to train a VAE. In practice, we equivalently *minimize* a loss that is equal to the negative ELBO.

In the VAE lab, the loss used `nn.BCELoss`, which is not explicitly written in Equation 6, so there's still more to understand to connect Equation 6 with the code.

1. First, assume that we approximate the reconstruction term in the ELBO with a single sample; that is,

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] \approx \log p_\theta(\mathbf{x} \mid \tilde{\mathbf{z}}), \tag{8}$$

   where $\tilde{\mathbf{z}} \sim q_\phi(\mathbf{z} \mid \mathbf{x})$. Second, assume that $\mathbf{x}$ has binary-valued elements, i.e. $\mathbf{x} = (x_1, \ldots, x_D)$ with each $x_d \in \{0, 1\}$.

   Third, denoting the decoder neural network's output as $(\hat{x}_1, \ldots, \hat{x}_D) = f_\theta(\tilde{\mathbf{z}})$, with each $\hat{x}_d \in [0, 1]$, assume that:

$$p_\theta(\mathbf{x} \mid \tilde{\mathbf{z}}) = \prod_{d=1}^{D} \text{Bernoulli}(x_d\,; \hat{x}_d). \tag{9}$$

   Your task: Show that under these assumptions, $-\log p_\theta(\mathbf{x} \mid \tilde{\mathbf{z}})$ equals the binary cross-entropy loss summed over dimensions $1 \ldots D$.

2. We'll now only assume that $\mathbf{x}$ has real-valued elements, $\mathbf{x} = (x_1, \ldots, x_D)$ with each $x_d \in \mathbb{R}$.

   Second, we now assume that the decoder neural network outputs Gaussian mean parameters, rather than Bernoulli parameters:

$$p_\theta(\mathbf{x} \mid \tilde{\mathbf{z}}) = \prod_{d=1}^{D} \mathcal{N}(x_d\,; \hat{x}_d, \sigma^2), \tag{10}$$

   where the decoder neural network's output is $(\hat{x}_1, \ldots, \hat{x}_D) = f_\theta(\tilde{\mathbf{z}})$, and σ^2 is fixed.

   Your task: Show that under these assumptions, $-\log p_\theta(\mathbf{x} \mid \tilde{\mathbf{z}})$ equals the MSE loss (up to a constant) summed over dimensions $1 \ldots D$.

Hopefully you now understand the VAE objective better.[1]

[1] If you are interested to see how to derive the closed-form KL term that we use in implementations, see Appendix B of the VAE paper (`https://arxiv.org/pdf/1312.6114.pdf`).

4. Short answer. *The following questions are intended to help you review and articulate concepts from the lectures and labs; we will grade these generously and accept answers as long as they are consistent with what was presented in the lectures and labs. 1-2 sentences should suffice.*

1. **Reparameterization.** Give one reason why VAEs use reparameterization.

2. **Overlapping latents.** Suppose that $q_\phi(\mathbf{z} \mid \mathbf{x}^{(1)}) \approx q_\phi(\mathbf{z} \mid \mathbf{x}^{(2)})$, where $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ are two distinct images. Why can this 'overlapping' be problematic?

3. **Missing labels.** Suppose you built a dataset of images, with the labels stored in ten different files. You accidentally delete 9 of the label files (and there's no way to restore the files, no extra copies, etc). Name two methods which leverage all of the images, and the labels when they are available, for training a classifier on your dataset.

4. ***Bonus:*** **Discrete latent variables.** Which term in the VAE objective becomes problematic when we want to use a VAE with a discrete, categorical latent variable? Name an estimator that allows for a VAE with a discrete, categorical latent variable.

2 Programming assignment. (40 points).

In the programming assignment, you will implement

1. a convolutional VAE
2. a conditional VAE: $q_\phi(\mathbf{z}|\mathbf{x}, c)$, $p_\theta(\mathbf{x}|\mathbf{z}, c)$, where $c \in \{0, 1, \ldots, 9\}$ is a class label.

Please complete `hw4_vae.ipynb`.