Homework 4
Variational Autoencoders

November 14, 2020

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,

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

Answer:

$$
\begin{aligned}
\log p_\theta(\mathbf{x}) &= \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log p_\theta(\mathbf{x}) d\mathbf{z} \\
&= \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{p_\theta(\mathbf{z} \mid \mathbf{x})} d\mathbf{z} \\
&= \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log \left[\frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})} \frac{q_\phi(\mathbf{z} \mid \mathbf{x})}{p_\theta(\mathbf{z} \mid \mathbf{x})} \right] d\mathbf{z} \\
&= \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})} d\mathbf{z} + \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log \frac{q_\phi(\mathbf{z} \mid \mathbf{x})}{p_\theta(\mathbf{z} \mid \mathbf{x})} d\mathbf{z} \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})} \right] + \mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}|\mathbf{x}) \right] \\
&= \mathrm{ELBO}(\theta, \phi; \mathbf{x}) + \mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}|\mathbf{x}) \right]
\end{aligned}
$$

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

   Answer:

$$
\begin{aligned}
-\mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}|\mathbf{x}) \right] &= \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log \frac{p_\theta(\mathbf{z} \mid \mathbf{x})}{q_\phi(\mathbf{z} \mid \mathbf{x})} d\mathbf{z} \\
&\leq \int q_\phi(\mathbf{z} \mid \mathbf{x}) (\frac{p_\theta(\mathbf{z} \mid \mathbf{x})}{q_\phi(\mathbf{z} \mid \mathbf{x})} - 1) d\mathbf{z} \\
&= \int p_\theta(\mathbf{z} \mid \mathbf{x}) d\mathbf{z} - \int q_\phi(\mathbf{z} \mid \mathbf{x}) d\mathbf{z} \\
&= 0
\end{aligned}
$$

   so we have $\mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}|\mathbf{x}) \right] \geq 0$, and therefore $\log p_\theta(\mathbf{x}) = \mathrm{ELBO}(\theta, \phi; \mathbf{x}) + \mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}|\mathbf{x}) \right] \geq \mathrm{ELBO}(\theta, \phi; \mathbf{x})$.

   $\log p_\theta(\mathbf{x}) = \mathrm{ELBO}(\theta, \phi; \mathbf{x})$ if and only if $\mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}|\mathbf{x}) \right] = 0$, i.e. $p_\theta(\mathbf{z} \mid \mathbf{x}) = q_\phi(\mathbf{z} \mid \mathbf{x})$.

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

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

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

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

Answer:

$$
\begin{aligned}
\mathrm{ELBO}(\theta, \phi; \mathbf{x}) &\equiv \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})}\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}, \mathbf{z})\right] - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log q_\phi(\mathbf{z} \mid \mathbf{x})\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{z})\right] - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log q_\phi(\mathbf{z} \mid \mathbf{x})\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] + \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log \frac{p_\theta(\mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})} d\mathbf{z} \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] - \mathrm{KL}\left[q_\phi(\mathbf{z} \mid \mathbf{x}) \| p_\theta(\mathbf{z})\right]
\end{aligned}
$$

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 $-\mathrm{KL}[...]$ term? (*Any interpretation that is consistent with what was presented in the lectures or labs is fine.)*

   Answer: The $\mathbb{E}[\log p_\theta(\mathbf{x} \mid \mathbf{z})]$ term functions as the reconstruction error. The $-\mathrm{KL}[...]$ term functions as a regularization term.

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,

$$
\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] \approx \log p_\theta(\mathbf{x} \mid \tilde{\mathbf{z}}), \tag{7}
$$

   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:

$$
p_\theta(\mathbf{x} \mid \tilde{\mathbf{z}}) = \prod_{d=1}^{D} \mathrm{Bernoulli}(x_d\,; \hat{x}_d). \tag{8}
$$

   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 $1 \ldots D$.

Answer: Given that $x_d \in \{0, 1\}$,

$$\begin{aligned}
-\log p_\theta(\mathbf{x} \mid \tilde{\mathbf{z}}) &= -\log \prod_{d=1}^{D} [x_d \hat{x}_d + (1 - x_d)(1 - \hat{x}_d)] \\
&= -\sum_{d=1}^{D} \log [x_d \hat{x}_d + (1 - x_d)(1 - \hat{x}_d)] \\
&= -\left[\sum_{x_d=1} \log \hat{x}_d + \sum_{x_d=0} \log(1 - \hat{x}_d) \right] \\
&= \sum_{d=1}^{D} -[x_d \log \hat{x}_d + (1 - x_d) \log(1 - \hat{x}_d)]
\end{aligned}$$

which is the BCE summed over dimensions 1 ... 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:

$$p_\theta(\mathbf{x} \mid \tilde{\mathbf{z}}) = \prod_{d=1}^{D} \mathcal{N}(x_d \,; \hat{x}_d, \sigma^2), \tag{9}$$

   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 $1 \ldots D$.

Answer:

$$
\begin{aligned}
-\log p_\theta(\mathbf{x} \mid \tilde{\mathbf{z}}) &= -\log \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x_d - \hat{x_d})^2}{2\sigma^2}) \\
&= -\sum_{d=1}^{D} \log \left[\frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x_d - \hat{x_d})^2}{2\sigma^2}) \right] \\
&= -\sum_{d=1}^{D} \left[-\frac{(x_d - \hat{x_d})^2}{2\sigma^2} - \log \sqrt{2\pi}\sigma \right] \\
&= \frac{1}{2\sigma^2} \sum_{d=1}^{D} (x_d - \hat{x_d})^2 + D \log \sqrt{2\pi}\sigma \\
&= \frac{1}{2\sigma^2} \sum_{d=1}^{D} MSE_d + D \log \sqrt{2\pi}\sigma
\end{aligned}
$$

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.

   Answer: Because sampling operation that cannot be differentiated through. If we sample z from $\mathbf{z} \sim \mathcal{N}(\cdot\,; \mu, \sigma^{\mathbf{2}} = \texttt{Encoder}_\phi(\mathbf{x}))$, it's hard to compute $\frac{\partial \mathbf{z}}{\partial \mu}$ for backprobagation; but if we sample z from $\mathbf{z}_{\text{reparam}} = \mu + \sigma \odot \epsilon$, it's easy to have $\frac{\partial \mathbf{z}}{\partial \mu} = 1$.

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?

   Answer: If $q_\phi(\mathbf{z} \mid \mathbf{x}^{(1)}) \approx q_\phi(\mathbf{z} \mid \mathbf{x}^{(2)})$ for different images $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$, then it's hard to reconstruct the images from the latent variable with the same trained decoder.

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.

   Answer:

Pretraining-Finetuning. First pretrain all images without labels on a self-supervised task (e.g. predicting rotation angle, denoising), in order to initialize a good parameter region. And then finetune the parameters to train the classifier on labeled data.

Semi-supervised VAE. Use an encoder with two heads: the first head infers latent z and the second head infers the label y. The reconstruction error and the regularization term are different for the labeled and the unlabeled respectively.

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.

   Answer: Regularization term would be problematic. An estimator could be reinforce estimator.