CPSC 440/540 Machine Learning (January-April, 2022)
Variational Autoencoders Assignment

1 "MLE" Derivation

Recall the Kullback-Leibler-divergence,

$$\mathcal{KL}(p \,||\, q) = E_{x\sim p}[\log(p(x)) - \log(q(x))]$$

as well as the definition of the ELBO function,

$$\text{ELBO}(\theta, \phi) = E_{z\sim q_\phi(z \mid x)}[\log(p(x, z)) - \log(q_\phi(z \mid x))]$$

1.1 Evidence Lower Bound

Show that the ELBO function can be written as

$$\text{ELBO}(\theta, \phi) = E_{z\sim q_\phi(z \mid x)}[\log(p(x \mid z))] - \mathcal{KL}(q_\phi(z \mid x) \,||\, p(z))$$

1.2 Log-Evidence

Starting from the KL-divergence between $q_\phi(z \mid x)$ and $p(z \mid x)$, derive the following formula for the log-evidence:

$$\log(p(x)) = \text{ELBO}(\theta, \phi) + \mathcal{KL}(q_\phi(z \mid x) \,||\, p(z \mid x))$$

Hint: use Bayes rule on the $p(z \mid x)$ term, along with the form of the ELBO *function you derived in the previous part*

1.3 Loss Function

Looking at the formula from the previous part, we still have the intractable $p(x)$ term lying around in the KL-divergence term. However, we can safely ignore the $\mathcal{KL}(q_\phi(z \mid x) \,||\, p(z \mid x))$ term. Recall from the lecture that ELBO is supposed to be a lower bound for the log evidence, ELBO $\leq \log(p(x))$. This allows us to try to maximize the evidence by instead maximizing ELBO.

Your task is to prove that ELBO is indeed a lower bound for $\log(p(x))$. Observing the formula for the log-evidence you derived in 1.2, notice that if we can show the KL-divergence term is nonnegative, then we will have proven ELBO $\leq \log(p(x))$. Show that the KL-divergence between any two distributions is always nonnegative, $\mathcal{KL}(p \,||\, q) \geq 0$.

Hint: Start by showing $\mathcal{KL}(p \,||\, q) = E_{x\sim p}\left[-\log\left(\frac{q(x)}{p(x)}\right)\right]$ *and apply Jensen's inequality. (2.12 in the link)*

2 Short Answer

1. Why do VAEs tend to perform better at generating new samples compared to traditional autoencoders?
2. Write a function in Julia or Python which uses the reparametrization trick for sampling z from $q_\phi(z \mid x)$ and submit the code.
3. What is the main advantage of β-VAEs as opposed to VAEs?
4. Suppose we have a perfect optimization algorithm which can find a unique maximum of the ELBO function. If we maximize ELBO in this manner, have we necessarily found a maximum of $\log(p(x))$? Why or why not?