University of Illinois Spring 2018

CS 446: Machine Learning
Homework

Due on Tuesday, April 17, 2018, 11:59 a.m. Central Time

1. **[2 points]** KL Divergence

   (a) [1 point] What is the expression of the KL divergence $D_{KL}(q(x)||p(x))$ given two continuous distributions $p(x)$ and $q(x)$ defined on the domain of $\mathbb{R}^1$?

   Your answer:

   $$D_{KL}(q(x)||p(x)) = -\int_{\mathbb{R}^1} q(x)\log\left\{\frac{p(x)}{q(x)}\right\}dx$$

   (b) [1 point] Show that the KL divergence is non-negative. You can use Jensen's inequality here without proving it.

   Your answer: According to Jensen's inequality:

   $$f(E(x)) \leq E(f(x))$$

   $$D_{KL}(p(x)||q(x)) = -\int p(x)\log\left\{\frac{q(x)}{p(x)}\right\}dx \geq -\log\int p(x)\frac{q(x)}{p(x)}dx = -\log\int q(x)dx = 0$$

   Similarly we can show for $D_{KL}(q(x)||p(x))$

2. **[3 points]** In the class, we derive the following equality:

   $$\log p_\theta(x) = \int_z q_\phi(z|x)\log\frac{p_\theta(x,z)}{q_\phi(z|x)}dz + \int_z q_\phi(z|x)\log\frac{q_\phi(z|x)}{p_\theta(z|x)}dz$$

   Instead of maximizing the log likelihood $\log p_\theta(x)$ w.r.t. θ, we find a lower bound for $\log p_\theta(x)$ and maximize the lower bound.

   (a) [1 point] Use the above equation and your result in 1(b) to give a lower bound for $\log p_\theta(x)$.

   Your answer:

   $$\log p_\theta(x) = \int_z q_\phi(z|x)\log\frac{p_\theta(x,z)}{q_\phi(z|x)}dz + \int_z q_\phi(z|x)\log\frac{q_\phi(z|x)}{p_\theta(z|x)}dz$$

   $$\log p_\theta(x) = \int_z q_\phi(z|x)\log\frac{p_\theta(x,z)}{q_\phi(z|x)}dz + D_{KL}(q_\phi(z|x)||p_\theta(x|z))$$

   But we have already shown that $D_{KL}(q(x)||p(x)) \geq 0$

   $$\log p_\theta(x) \geq \int_z q_\phi(z|x)\log\frac{p_\theta(x,z)}{q_\phi(z|x)}dz$$

   $$\log p_\theta(x) \geq \mathcal{L}(p_\theta, q_\phi)$$

(b) [1 point] What do people usually call the bound?

Your answer: ELBO : Empirical Lower Bound

(c) [1 point] In what condition will the bound be tight?

Your answer: The Bound is tight when the KL Divergence is 0 ($D_{KL}(q_\phi(z|x)||p_\theta(x|z)) = 0$) i.e. we have $q_\phi \sim p_\theta$ and then we maximize the ELBO.

3. **[2 points]** Given $z \in \mathbb{R}^1$, $p(z) \sim \mathcal{N}(0, 1)$ and $q(z|x) \sim \mathcal{N}(\mu_z, \sigma_z^2)$, write $D_{KL}(q(z|x)||p(z))$ in terms of σ_z and μ_z.

Your answer: In general

$$D_{KL}(p||q) = -\int p(x)\log q(x)dx + \int p(x)\log p(x)dx$$

$$D_{KL}(p||q) = \frac{1}{2}\log(2\pi\sigma_2^2) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}(1 + \log 2\pi\sigma_1^2)$$

$$D_{KL}(p||q) = \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

Therefore,
we have

$$D_{KL}(q(z|x)||p(z)) = \log\frac{1}{\sigma_z} + \frac{\sigma_z^2 + (\mu_z)^2}{2} - \frac{1}{2}$$

4. **[1 points]** In VAEs, the encoder computes the mean μ_z and the variance σ_z^2 of $q_\phi(z|x)$ assuming $q_\phi(z|x)$ is Gaussian. Explain why we usually model σ_z^2 in log space, i.e., modeling $\log\sigma_z^2$ instead of σ_z^2 when implementing it using neural nets?

Your answer: We model the σ_z^2 in log space so that our neural network cannot have restriction to work in positive space of σ_z^2. Working in log space will enable neural network to generate negative outputs as well and thereby we can search for σ_z^2 on entire space. This creates the numerical stability.

5. **[1 points]** Why do we need the reparameterization trick when training VAEs instead of directly sampling from the latent distribution $\mathcal{N}(\mu_z, \sigma_z^2)$?

Your answer: We cannot backpropagate with direct sampling as it is stochastic in nature. Hence, we use reparameterization trick of $\mu_\phi(x)+\sigma_\phi(x).\epsilon$ to remove stochasticity which helps us to use backpropagation in our deep net.