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:

   $$D_{KL}(q(x)||p(x)) = \int_{-\infty}^{\infty} q(x) \log \frac{q(x)}{p(x)} dx$$

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

   Your answer:

   $$D_{KL}(q(x)||p(x)) = -\int_{-\infty}^{\infty} q(x) \log \frac{p(x)}{q(x)} dx$$

   Since $-\log(\cdot)$ is convex, we can use Jensen's inequality to provide a lower bound

   ie. $E[f(g(x))] \geq f(E[g(x)])$

   $$\begin{aligned} -\int_{-\infty}^{\infty} q(x) \log \frac{p(x)}{q(x)} dx &\geq -\log \int_{-\infty}^{\infty} q(x) \frac{p(x)}{q(x)} dx \\ &\geq -\log \int_{-\infty}^{\infty} p(x) dx \\ &\geq -\log(1) \\ &\geq 0 \end{aligned}$$

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:

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

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

from 1b

$$\log p_\theta(x) - \int_z q_\phi(z|x) \log \frac{p_\theta(x,z)}{q_\phi(z|x)} dz \geq 0$$

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

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

Your answer: Evidence Lower Bound

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

Your answer: The bound will be tight when the KL divergence is low.

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:

$$\begin{aligned}
\int_z q(z|x) \log \frac{q(z|x)}{p(z)} dz &= \int_z q(z|x)(\log q(z|x) - \log p(z)) dz \\
&= \int_z \frac{1}{\sqrt{2\pi}\sigma_q} e^{-\frac{1}{2}(\frac{z-\mu_q}{\sigma_q})^2} \times \\
&\quad (-\frac{\log(2\pi)}{2} - \log(\sigma_q) - \frac{1}{2}(\frac{z-\mu_q}{\sigma_q})^2 + \frac{\log(2\pi)}{2} + \frac{1}{2}z^2) dz \\
&= \int_z \frac{1}{\sqrt{2\pi}\sigma_q} e^{-\frac{1}{2}(\frac{z-\mu_q}{\sigma_q})^2} (-\log \sigma_q + \frac{1}{2}[z^2 - (\frac{z-\mu_q}{\sigma_q})^2) dz \\
&= E_q[-\log \sigma_q + \frac{1}{2}(Z^2 - (\frac{Z-\mu_q}{\sigma_q})^2)] \\
&= -\log \sigma_q + \frac{1}{2}E_q[Z^2] - \frac{1}{2\sigma_q^2}E_q[(Z-\mu_q)^2] \\
&= -\log \sigma_q + \frac{\sigma_q^2 + \mu_q^2}{2} - \frac{1}{2}
\end{aligned}$$

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 usually model σ_z^2 in log space because it lowers the computational cost of training and also improves the overall 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 need the re-parameterization trick, because the loss is not differentiable when randomly sampling directly from the latent distribution. Instead we sample with $z = \mu + \sigma\epsilon$ where ϵ is random noise in the distribution $\mathcal{N}(0, 1)$ to move the randomness outside of this step, making the function differentiable.
Backprop cannot flow through a random node; ie. z is a random selection in the latent distribution. Introducing a new parameter ϵ allows us to re-parameterize z which allows back-propagation to flow to μ and σ.