University of Illinois Spring 2018

CS 446: Machine Learning
Homework

Due on Tuesday, April 10, 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:

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

   (b) [1 point] Show that the KL divergence is non-negative. You can use Jensen's inequality.

   Your answer:

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

   Where the inequality follows from Jensen given the fact that $\ln(\cdot)$ is concave and, therefore, $-\ln(\cdot)$ is convex. And the one before last equality follows because $p(x)$ is a probability density function so it must sum up to 1.

2. **[4 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 (b) to give a lower bound for $\log p_\theta(x)$.

Your answer:

$$\begin{aligned}\log p_\theta(x) &= \int_z q_\phi(z|x) \log \frac{p_\theta(x,z)}{q_\phi(z|x)} dz + \int_z q_\phi(z|x) \log \frac{q_\phi(z|x)}{p_\theta(z|x)} dz \\ &= \int_z q_\phi(z|x) \log \frac{p_\theta(x,z)}{q_\phi(z|x)} dz + D_{KL}\left(q_\phi(z|x)||p_\theta(z|x)\right) \\ &\geq \int_z q_\phi(z|x) \log \frac{p_\theta(x,z)}{q_\phi(z|x)} dz\end{aligned}$$

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

Your answer: ELBO: Empirical Lower Bound

(c) [1 point] When will the bound be tight?

Your answer: When $q(z|x)$ approximates $p(z|x)$ (almost) everywhere, which is the same as having KL-divergence to be zero.

(d) [1 point] Write down the objective function for maximizing the lower bound formally.

Your answer:

$$\max_{\phi,\theta} \int_z q_\phi(z|x) \log \frac{p_\theta(x,z)}{q_\phi(z|x)} dz \approx \max_{\phi,\theta} \left(-D_{KL}(q_\phi||p) + \frac{1}{N} \sum_{i=1}^{N} \ln p_\theta(x|z^i) \right)$$

Where N is the number of samples to be drawn to approximate the expected distribution of $\log p(x|z)$ and z^i the corresponding latent variable for the drawn sample.

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: Lets remember the definition of KL divergence:

$$D_{KL}(X||Y) = \mathbb{E}_X\left[\log \frac{X}{Y}\right] = \mathbb{E}_X\left[\log X - \log Y\right]$$

In this particular case $X = q(z|x) \sim \mathcal{N}(\mu_z, \sigma_z^2)$ and $Y = p(z) \sim \mathcal{N}(0,1)$.
So lets first look at $\log \mathcal{N}(\mu, \sigma^2)$ in general:

$$\begin{aligned}
\log \mathcal{N}(\mu, \sigma^2) &= \log\left[(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\sigma^{-2}(x-\mu)^2\right)\right] \\
&= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sigma^{-2}(x-\mu)^2 \\
&= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\sigma^{-2}(x-\mu)^2
\end{aligned}$$

So

$$\begin{aligned}
&D_{KL}(\mathcal{N}(\mu_z, \sigma_z^2)||\mathcal{N}(0,1)) \\
&= \mathbb{E}_{\mathcal{N}(\mu_z,\sigma_z^2)}\left[\log \mathcal{N}(\mu_z, \sigma_z^2) - \log \mathcal{N}(0,1)\right] \\
&= \mathbb{E}_{\mathcal{N}(\mu_z,\sigma_z^2)}\Bigg[\left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma_z^2 - \frac{1}{2}\sigma_z^{-2}(x-\mu_z)^2\right) - \\
&\qquad\qquad \left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log 1 - \frac{1}{2}1^{-2}(x-0)^2\right)\Bigg] \\
&= \mathbb{E}_{\mathcal{N}(\mu_z,\sigma_z^2)}\left[\left(-\frac{1}{2}\log\sigma_z^2 - \frac{1}{2}\sigma_z^{-2}(x-\mu_z)^2\right) + \frac{1}{2}(x-0)^2\right] \\
&= \mathbb{E}_{\mathcal{N}(\mu_z,\sigma_z^2)}\left[-\frac{1}{2}\log\sigma_z^2 - \frac{1}{2}\sigma_z^{-2}(x-\mu_z)^2 + \frac{1}{2}(x-0)^2\right] \\
&= -\frac{1}{2}\log\sigma_z^2 - \frac{1}{2}\sigma_z^{-2}\mathbb{E}_{\mathcal{N}(\mu_z,\sigma_z^2)}\left[(x-\mu_z)^2\right] + \mathbb{E}_{\mathcal{N}(\mu_z,\sigma_z^2)}\left[\frac{1}{2}(x-0)^2\right] \\
&= \frac{1}{2}\log\sigma_z^2 - \frac{1}{2}\sigma_z^{-2}\sigma_z^2 + \frac{1}{2}\left(\sigma_z^2 + \mu_z^2\right) \\
&= -\frac{1}{2}\log\sigma_z^2 + -\frac{1}{2} + \frac{1}{2}\left(\sigma_z^2 + \mu_z^2\right) \\
&= \frac{1}{2}\left[-\log\sigma_z^2 - 1 + \sigma_z^2 + \mu_z^2\right]
\end{aligned}$$

The same result can be obtained by plugging the given values into the known, general, form of KL given in https://en.wikipedia.org/wiki/Kullback-Leibler_divergence#Multivariate_normal_distributions

4. **[1 points]** In VAEs, the encoder computes the mean μ_z and 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., model $\log\sigma_z^2$ instead of σ_z^2 when implementing it using neural nets?

Your answer: Because a neural network is not guaranteed to output positive values so if we model σ^2 we cannot really calculate σ. By modeling $\log \sigma^2$ we can correctly handle negative outputs from the network.

5. **[2 points]** Reparameterization trick

   (a) [1 point] 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: Because it allows us to do backwards propagation since the sampling has been moved outside of the propagation "path" in the neural network. A good picture illustrating this is both given in lecture and found in https://stats.stackexchange.com/a/205336/175920

   (b) [1 point] In the reparameterization trick, what is the sampled latent representation z given the mean μ_z and the variance σ_z^2 of $q_\phi(z|x)$?

   Your answer: $z = \mu_z + \sigma_z \cdot \epsilon$ where $\epsilon \sim \mathcal{N}(0, 1)$