Notes on Derivation of the Variational Autoencoder Evidence Lower Bound Loss

Alina Selega

March 2022

1 Introduction

In a variational autoencoder (VAE) (Figure 1), we want to learn a variational posterior distribution $q(z|x)$ that is as close as possible to the target distribution $p(z|x)$, for input x and a latent variable describing the input data, z. Let's recall our terminology for the main components of VAE:

- $p(x|z)$: decoder that reconstructs input x from a sample z from the variation posterior;
- $p(z|x)$: encoder that encodes input x into a latent representation governed by the variational posterior;
- $p(z) \sim N(0, 1)$: prior of the latent variable z assumed to be a standard Normal for analytical properties.

Our aim is to learn a well-regularized latent space such that each input sample from our data x is encoded by its variational distribution $q(z|x)$, samples from which can be decoded back into the original input. The latent variable z that aims to capture underlying information about our input data is typically of lower dimensionality compared to x, enabling dimensionality reduction.

2 Loss as an optimization problem

We typically learn the variational posterior $q(z|x) \sim N(g(x), h(x))$ with the help of neural networks approximating functions g, h. In order to learn these functions, we express the optimization problem over g, h as functions that minimize the Kullback-Leibler (KL) divergence between variational and target distributions. KL divergence measures the "distance" between two distributions and our goal is to bring $q(z|x)$ as close as possible to $p(z|x)$.

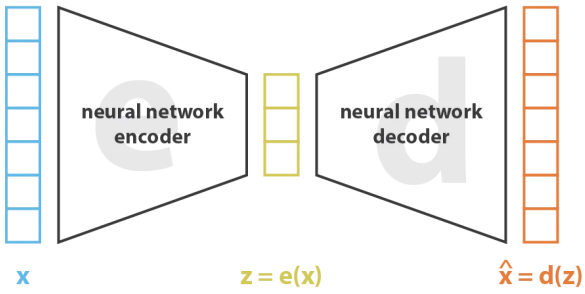

Figure 1: An illustration of a variational autoencoder. Source: Understanding Variational Autoencoders (VAEs), Rocca J, Towards Data Science article (2019): accessed in 03/2022.

Let's derive the evidence lower bound (ELBO, called so because it represents a lower bound on the evidence) loss from the KL divergence between the variational and target distribution:

$$\begin{aligned}
KL(q(z|x), p(z|x)) &= \mathbb{E}_{q(z|x)}\left[\log \frac{q(z|x)}{p(z|x)}\right] \\
&= \mathbb{E}_{q(z|x)}\left[\log q(z|x) - \log \frac{p(x|z)p(z)}{p(x)}\right] \\
&= \mathbb{E}_{q(z|x)}\left[\log q(z|x) - \log p(x|z) - \log p(z) + \log p(x)\right] \\
&= \mathbb{E}_{q(z|x)}\left[\log \frac{q(z|x)}{p(z)} - \log p(x|z) + \log p(x)\right] \\
&= KL(q(z|x), p(z)) - \mathbb{E}_{q(z|x)}\left[\log p(x|z)\right] + \mathbb{E}_{q(z|x)}\left[\log p(x)\right] \\
&= KL(q(z|x), p(z)) - \mathbb{E}_{q(z|x)}\left[\log p(x|z)\right] + \log p(x)
\end{aligned}$$

Recall that we want to minimise the KL divergence above w.r.t. g, h. As the distribution of the input data $p(x)$ is independent of g, h we can drop the last term from the optimization, which now becomes:

$$\min KL(q(z|x), p(z)) - \mathbb{E}_{q(z|x)}\left[\log p(x|z)\right]$$

When modelling a Gaussian likelihood of the data $p(x|z) \sim N(f(z), cI)$, we can drop the normalization constant and get:

$$\min KL(q(z|x), p(z)) - \mathbb{E}_{q(z|x)}\left[-\frac{(x - f(z))^2}{2c}\right]$$
$$= \min KL(q(z|x), p(z)) + \mathbb{E}_{q(z|x)}\left[\frac{(x - f(z))^2}{2c}\right]$$

The second term is the reconstruction error and, for a Gaussian likelihood, it is typically written as $C(x - \hat{x})^2$ for input x and its decoding $\hat{x}$, where $C = \frac{1}{2c}$ represents a balancing term between minimising the reconstruction error and the KL divergence between the variational posterior and prior.

Now let's derive the KL divergence term.

3 KL divergence between two Gaussians

This StackExchange answer[1] gives the general form for KL divergence between two Gaussians. Here I derive a more detailed working. Let $p_1(x) \sim N(\mu_1, \sigma_1^2)$ and $p_2(x) \sim N(\mu_2, \sigma_2^2)$.

$$\begin{aligned} KL(p_1(x), p_2(x)) &= \int p_1(x) \log \frac{p_1(x)}{p_2(x)} dx \\ &= \int p_1(x) \log p_1(x) dx - \int p_1(x) \log p_2(x) dx \end{aligned}$$

3.1 First term

Let us derive the expression for the first term above. I will denote the normalization constants for the distributions as Z_1 and Z_2, respectively.

$$\begin{aligned} \int p_1(x) \log p_1(x) dx &= \int p_1(x) \left(\log \frac{1}{Z_1} - \frac{(x - \mu_1)^2}{2\sigma_1^2} \right) dx \\ &= \log \frac{1}{Z_1} \int p_1(x) dx - \int p_1(x) \frac{(x - \mu_1)^2}{2\sigma_1^2} dx \\ &= -\log Z_1 - \frac{1}{2\sigma_1^2} \int p_1(x)(x^2 - 2x\mu_1 + \mu_1^2) dx \\ &= -\log Z_1 - \frac{1}{2\sigma_1^2} \left((\sigma_1^2 + \mu_1^2) - 2\mu_1^2 + \mu_1^2 \right) \end{aligned}$$

The last step is achieved by computing $\int p_1(x) x^2 dx = \mathbb{E}_{p_1(x)}[x^2] = \sigma_1^2 + \mu_1^2$ and similarly, computing expected values of $-2x\mu_1$ and μ_1^2 under $p_1(x)$.

Putting this together, we get

[1] https://stats.stackexchange.com/questions/7440/kl-divergence-between-two-univariate-gaussians

$$\int p_1(x) \log p_1(x) dx = -\log Z_1 - \frac{1}{2} = -\frac{1}{2}(\log 2\pi\sigma_1^2 + 1)$$

3.2 Second term

Let us derive the expression for the second term.

$$\begin{aligned}
-\int p_1(x) \log p_2(x) dx &= -\int p_1(x) \left(\log \frac{1}{Z_2} - \frac{(x-\mu_2)^2}{2\sigma_2^2} \right) dx \\
&= \log Z_2 + \frac{1}{2\sigma_2^2} \int p_1(x)(x^2 - 2x\mu_2 + \mu_2^2) dx \\
&= \log Z_2 + \frac{1}{2\sigma_2^2}((\sigma_1^2 + \mu_1^2) - 2\mu_2\mu_1 + \mu_2^2) \\
&= \frac{1}{2} \log 2\pi\sigma_2^2 + \frac{1}{2\sigma_2^2}(\sigma_1^2 + (\mu_1 - \mu_2)^2)
\end{aligned}$$

3.3 Final KL expression

We now add the two expressions together to derive the final form of a KL divergence term between any two Gaussians, $p_1(x)$ and $p_2(x)$.

$$\begin{aligned}
KL(p_1(x), p_2(x)) &= -\frac{1}{2}(\log 2\pi\sigma_1^2 + 1) + \frac{1}{2} \log 2\pi\sigma_2^2 + \frac{1}{2\sigma_2^2}(\sigma_1^2 + (\mu_1 - \mu_2)^2) \\
&= \log \frac{\sigma_2}{\sigma_1} + \frac{1}{2\sigma_2^2}(\sigma_1^2 + (\mu_1 - \mu_2)^2) - \frac{1}{2}
\end{aligned}$$

4 Final expression for the ELBO loss

In the ELBO loss, the KL divergence term we are interested is between $q(z|x) \sim N(g(x), h(x))$ and $p(z) \sim N(0, 1)$. Using the result from the previous section, we get:

$$\begin{aligned}
KL(q(z|x), p(z)) &= \log \frac{1}{\sqrt{h(x)}} + \frac{1}{2}\left(h(x) + g(x)^2\right) - \frac{1}{2} \\
&= \frac{1}{2}\left(h(x) + g(x)^2 - \log h(x) - 1\right)
\end{aligned}$$

Finally, the ELBO loss for a variational autoencoder with a Gaussian likelihood is then given by:

$$C(x - \hat{x})^2 + \frac{1}{2}\left(h(x) + g(x)^2 - \log h(x) - 1\right)$$

This is the loss minimized by a neural network that implements a variational autoencoder model.