Variational Autoencoders (VAEs)

Allen C.

September 2023

1 Autoencoder

Illustration of an autoencoder:

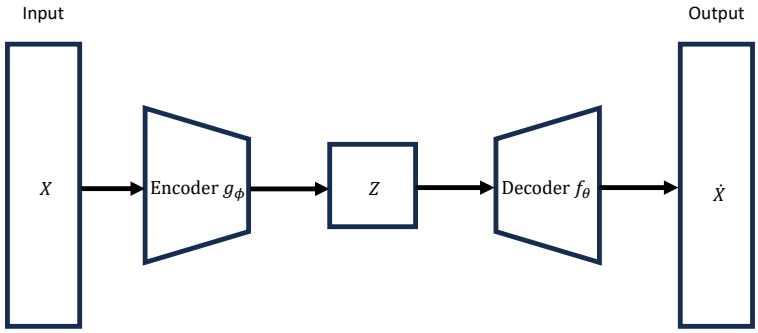

Figure 1: Autoencoder

As an applied mathematician, we understand the autoencoder model as follows:

- X is a vector.
- g_ϕ can be a matrix and nonlinear functions (ReLU, tanh).
- z is a reduced vector.
- f_θ can be a matrix and nonlinear functions (ReLU, tanh) which play the inverse role of g_ϕ.
- $\dot{X}$ is a vector.

The loss function of the autoencoder can be defined as

$$L(\theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} [X^i - f_\theta(g_\phi(X^i))]^2 \tag{1}$$

Here, the loss function compares X and $\dot{X}$ with the L^2 norm.

2 Variational Autoencoder

Illustration of a Variational Autoencoder:

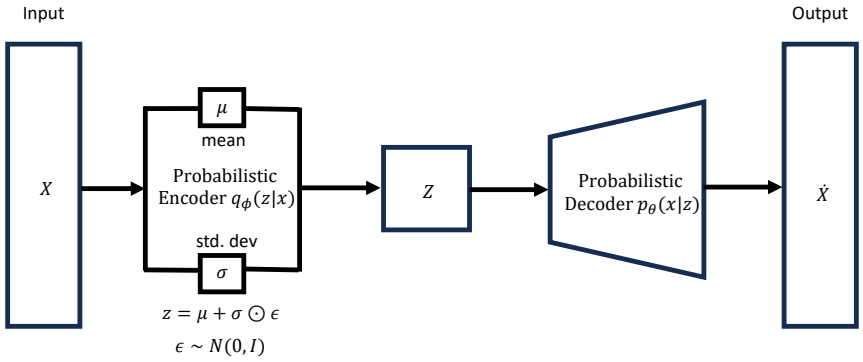

Figure 2: Variational Autoencoder

The loss function for the Variational Autoencoder model is given as follows, and it will be derived later:

$$L(\theta, \phi) = -E_{z \sim q_\phi(z|x)} \log(p_\theta(x|z)) + D_{KL}(q_\phi(z|x)||p_\theta(z)) \tag{2}$$

3 Probability

- $P(x)$: defines the probability of random variable x.
- $P(x|y)$: defines the probability of random variable x provided y has happened. Also called conditional probability.
- Bayes's Theorem: $P(y|x) = \frac{P(x|y)P(y)}{P(x)}$

4 Kullback-Leibler Divergence

KL divergence is a measure of how one probability distribution is different from the other.

$$D_{KL}(P\|Q) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right) \tag{3}$$

Properties of KL divergence:

- $KL(P\|Q) \geq 0$ or $KL(Q\|P) \geq 0$.
- $KL(P\|Q) \neq KL(Q\|P)$

Lemma 1. *Given a multivariate normal density function as follows:*

$$N(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^\intercal \Sigma^{-1}(x-\mu)\right) \tag{4}$$

where k is the dimension of x and Σ. With $p(x) = N(x; \mu_1, \Sigma_1)$ and $q(x) = N(x; \mu_2, \Sigma_2)$, where μ_1 and μ_2 are the means and Σ_1 and Σ_2 are the covariance matrices, the KL divergence for $p(x)$ and $q(x)$ can be written as follows:

$$D_{KL}(p(x)\|q(x)) = \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} - k + tr(\Sigma_2^{-1}\Sigma_1) + (\mu_1-\mu_2)^\intercal\Sigma_2^{-1}(\mu_1-\mu_2)\right] \tag{5}$$

The above equation will be applied to rewrite the loss function of VAE.

Proof. With the definition of KL divergence, we have

$$KL(p(x)\|q(x)) = \sum_x p(x)\log\left(\frac{p(x)}{q(x)}\right)$$

The probability density functions $p(x)$ and $q(x)$ are given by

$$p(x) = \frac{1}{\sqrt{(2\pi)^k|\Sigma_1|}}\exp\left(-\frac{1}{2}(x-\mu_1)^\intercal\Sigma_1^{-1}(x-\mu_1)\right)$$

$$q(x) = \frac{1}{\sqrt{(2\pi)^k|\Sigma_2|}}\exp\left(-\frac{1}{2}(x-\mu_2)^\intercal\Sigma_2^{-1}(x-\mu_2)\right)$$

Then we have

$$\log p(x) = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(x-\mu_1)^\intercal\Sigma_1^{-1}(x-\mu_1)$$

$$\log q(x) = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_2| - \frac{1}{2}(x-\mu_2)^\intercal\Sigma_2^{-1}(x-\mu_2)$$

Then KL divergence can be rewritten as

$$\begin{aligned}
&KL(p(x)||q(x)) \\
&= \sum_x p(x)\big(\log p(x) - \log q(x)\big) \\
&= \sum_x p(x)\Big[-\frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(x-\mu_1)^\intercal\Sigma_1^{-1}(x-\mu_1) \\
&\qquad\qquad + \frac{k}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma_2| + \frac{1}{2}(x-\mu_2)^\intercal\Sigma_2^{-1}(x-\mu_2)\Big] \\
&= \sum_x p(x)\Big[\frac{1}{2}\log\Big(\frac{|\Sigma_2|}{|\Sigma_1|}\Big) + \frac{1}{2}(x-\mu_2)^\intercal\Sigma_2^{-1}(x-\mu_2) - \frac{1}{2}(x-\mu_1)^\intercal\Sigma_1^{-1}(x-\mu_1)\Big]
\end{aligned} \tag{6}$$

Consider the above right hand side part by part:

$$\sum_x p(x)\frac{1}{2}(x-\mu_1)^\intercal\Sigma_1^{-1}(x-\mu_1) = E_{p(x)}\Big[\frac{1}{2}(x-\mu_1)^\intercal\Sigma_1^{-1}(x-\mu_1)\Big]$$

To prove the lemma, we need to use the following trace and expectation identity:

$$\begin{aligned}
E(X^\intercal AX) =& E(tr(X^\intercal AX)) \\
=& E(tr(AXX^\intercal)) \\
=& tr(E(AXX^\intercal))
\end{aligned}$$

Then we have

$$\begin{aligned}
&\frac{1}{2}E_{p(x)}\Big[(x-\mu_1)^\intercal\Sigma_1^{-1}(x-\mu_1)\Big]\\
=&E_{p(x)}\Big[tr\Big(\frac{1}{2}(x-\mu_1)^\intercal\Sigma_1^{-1}(x-\mu_1)\Big)\Big]\\
=&E_{p(x)}\Big[tr\Big(\frac{1}{2}(x-\mu_1)(x-\mu_1)^\intercal\Sigma_1^{-1}\Big)\Big]\\
=&tr\Big[E_{p(x)}\Big(\frac{1}{2}(x-\mu_1)(x-\mu_1)^\intercal\Sigma_1^{-1}\Big)\Big]\\
=&tr\Big[E_{p(x)}\Big((x-\mu_1)(x-\mu_1)^\intercal\Big)\frac{1}{2}\Sigma_1^{-1}\Big]\\
=&tr\Big[\Sigma_1\frac{1}{2}\Sigma_1^{-1}\Big]\\
=&\frac{k}{2}
\end{aligned}$$

Then for the second term of the right hand side of Equation (6) can be rewritten as:

$$\begin{aligned}
&\sum_x p(x)\Big[\frac{1}{2}(x-\mu_2)^\intercal\Sigma_2^{-1}(x-\mu_2)\Big]\\
=&\sum_x p(x)\Big[\frac{1}{2}\big[(x-\mu_1)+(\mu_1-\mu_2)\big]^\intercal\Sigma_2^{-1}\big[(x-\mu_1)+(\mu_1-\mu_2)\big]\Big]\\
=&\sum_x p(x)\Big[\frac{1}{2}(x-\mu_1)^\intercal\Sigma_2^{-1}(x-\mu_1)+(x-\mu_1)^\intercal\Sigma_2^{-1}(\mu_1-\mu_2)+\frac{1}{2}(\mu_1-\mu_2)^\intercal\Sigma_2^{-1}(\mu_1-\mu_2)\Big]\\
=&E_{p(x)}\Big[\frac{1}{2}(x-\mu_1)^\intercal\Sigma_2^{-1}(x-\mu_1)+(x-\mu_1)^\intercal\Sigma_2^{-1}(\mu_1-\mu_2)+\frac{1}{2}(\mu_1-\mu_2)^\intercal\Sigma_2^{-1}(\mu_1-\mu_2)\Big]
\end{aligned}$$

Since we have the following identity:

$$\begin{aligned}
&E_{p(x)}\Big[(x-\mu_1)^\intercal\Sigma_2^{-1}(\mu_1-\mu_2)\Big]\\
=&\Big(E_{p(x)}x-\mu_1\Big)^\intercal\Sigma_2^{-1}(\mu_1-\mu_2)\\
=&\Big(\mu_1-\mu_1\Big)^\intercal\Sigma_2^{-1}(\mu_1-\mu_2)\\
=&0
\end{aligned}$$

Then we have

$$\begin{aligned}
&\sum_x p(x)\Big[\frac{1}{2}(x-\mu_2)^\intercal\Sigma_2^{-1}(x-\mu_2)\Big]\\
=&tr(\frac{\Sigma_2^{-1}\Sigma_1}{2})+(\mu_1-\mu_2)^\intercal\Sigma_2^{-1}(\mu_1-\mu_2)
\end{aligned}$$

To conclude, we prove the lemma:

$$D_{K|L}(p(x)||q(x))=\frac{1}{2}\Big[\log\frac{|\Sigma_2|}{|\Sigma_1|}-k+tr(\Sigma_2^{-1}\Sigma_1)+(\mu_1-\mu_2)^\intercal\Sigma_2^{-1}(\mu_1-\mu_2)\Big]$$

□

In the following, we will derive the loss function of VAE:

$$L(\theta,\phi)=-E_{z\sim q_\phi(z|x)}\log(p_\theta(x|z))+D_{K|L}(q_\phi(z|x)||p_\theta(z))$$

We have Baye's theorem:

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} \tag{7}$$

Here $p(x)$ can be written as:

$$p(x) = \int_z p(x,z)dz \tag{8}$$

which is difficult to integrate if z is in high dimension. To obtain $p(z|x)$, the alternative way is to approximate $p(z|x)$ by another distribution $q(z|x)$ which is defined in a way that has tractable solution, such as Gaussian distribution. As discussed above, we use KL divergence to measure the 'distance' between $p(z|x)$ and $q(z|x)$:

$$\begin{aligned} & KL(q_\phi(z|x)||p_\theta(z|x)) \\ =& \sum_z q_\phi(z|x) \log\Big(\frac{q_\phi(z|x)}{p_\theta(z|x)}\Big) \\ =& E_{z\sim q_\phi(z|x)}\Big[\log\Big(\frac{q_\phi(z|x)}{p_\theta(z|x)}\Big)\Big] \\ =& E_{z\sim q_\phi(z|x)}\Big[\log q_\phi(z|x) - \log p_\theta(z|x)\Big] \end{aligned} \tag{9}$$

Substituting Equation (7) into Equation (9) gives:

$$\begin{aligned} & KL(q_\phi(z|x)||p_\theta(z|x)) \\ =& E_{z\sim q_\phi(z|x)}\Big[\log q_\phi(z|x) - \log \frac{p_\theta(x|z)p_\theta(z)}{p_\theta(x)}\Big] \\ =& E_{z\sim q_\phi(z|x)}\Big[\log q_\phi(z|x) - \log p_\theta(x|z) - \log p_\theta(z) + \log p_\theta(x)\Big] \end{aligned} \tag{10}$$

Then we have

$$KL(q_\phi(z|x)||p_\theta(z|x)) - \log p_\theta(x) = E_{z\sim q_\phi(z|x)}\Big[\log q_\phi(z|x) - \log p_\theta(x|z) - \log p_\theta(z)\Big] \tag{11}$$

Multiplying -1 at both sides of Equation (11) gives:

$$\begin{aligned} & \log p_\theta(x) - KL(q_\phi(z|x)||p_\theta(z|x)) \\ =& E_{z\sim q_\phi(z|x)}\Big[\log p_\theta(x|z)\Big] - E_{z\sim q_\phi(z|x)}\Big[\log q_\phi(z|x) - \log p_\theta(z)\Big] \\ =& E_{z\sim q_\phi(z|x)}\Big[\log p_\theta(x|z)\Big] - D_{KL}\big[q_\phi(z|x)||p_\theta(z)\big] \end{aligned} \tag{12}$$

If we optimize the right hand side of Equation (12), it implies to optimize $\log p_\theta(x)$ and minimize $KL(q_\phi(z|x)||p_\theta(z|x))$. Then the loss function can be defined as:

$$L(\theta,\phi) = -E_{z\sim q_\phi(z|x)} \log(p_\theta(x|z)) + D_{K|L}(q_\phi(z|x)||p_\theta(z)) \tag{13}$$

where $q_\phi(z|x)$ is a neural network - encoder; $p_\theta(x|z)$ is a neural network - decoder (it could be a normal distribution). $p_\theta(z)$ is the latent variable distribution. The simple choice is $N(0,1)$ (it can also be any distribution: flower shaped distribution). Then D_{KL} can be simplified as:

$$\begin{aligned} & D_{KL}(N(\mu_\phi(x), \Sigma_\phi(x))||N(0,1)) \\ =& \frac{1}{2}\Big[tr(\Sigma_\phi(x)) + \mu_\phi(x)^\intercal \mu_\phi(x) - k - \log|\Sigma_\phi(x)|\Big] \end{aligned}$$

where k is the dimension of the Gaussian. $tr(\Sigma_\phi(x))$ is the trace function, which is the sum of the diagonal matrix of $\Sigma_\phi(x)$. Furthermore the determinant of $\Sigma_\phi(x)$ is the product of a diagonal matrix. So the above equation can be written as

$$\begin{aligned}
&D_{KL}(N(\mu_\phi(x), \Sigma_\phi(x))||N(0,1)) \\
=&\frac{1}{2}\Big(\sum_k \Sigma_\phi(x) + \sum_k \mu_\phi(x)^2 + \sum_k 1 - \log\big(\prod_k \Sigma_\phi(x)\big)\Big) \\
=&\frac{1}{2}\sum_k \Big(\Sigma_\phi(x) + \mu_\phi(x)^2 - 1 - \log \Sigma_\phi(x)\Big)
\end{aligned}$$

In practice, we use $\exp(\Sigma_\phi(x))$ instead of $\Sigma_\phi(x)$. Then the KL divergence can be written as:

$$\begin{aligned}
&D_{KL}(N(\mu_\phi(x), \Sigma_\phi(x))||N(0,1)) \\
=&\frac{1}{2}\sum_k \Big(\exp(\Sigma_\phi(x)) + \mu_\phi(x)^2 - 1 - \Sigma_\phi(x)\Big)
\end{aligned}$$

Correspondingly the loss function is written as

$$L(\theta, \phi) = -E_{z\sim q_\phi(z|x)} \log(p_\theta(x|z)) + \frac{1}{2}\sum_k \Big(\exp(\Sigma_\phi(x)) + \mu_\phi(x)^2 - 1 - \Sigma_\phi(x)\Big) \tag{14}$$

5 Optimization of the loss function

Find $\theta^*, \phi^* = \text{argmin}_{\theta,\phi} L(\theta, \phi)$. In variational Baysian method, this loss function is known as the variaional lower bound or evidence lower bound (ELBD). We know that

$$L(\theta, \phi) \leq \log p_\theta(x) \tag{15}$$

Recall the loss function

$$L(\theta, \phi) = -E_{z\sim q_\phi(z|x)} \log(p_\theta(x|z)) + \frac{1}{2}\sum_k \Big(\exp(\Sigma_\phi(x)) + \mu_\phi(x)^2 - 1 - \Sigma_\phi(x)\Big) \tag{16}$$

$$\begin{aligned}
\theta^* &= \theta - \eta \nabla_\theta L(\theta, \phi) \\
\phi^* &= \phi - \eta \nabla_\phi L(\theta, \phi)
\end{aligned}$$

where

$$\begin{aligned}
&\nabla_\theta L(\theta, \phi) \\
=&\nabla_\theta \Big[- E_{z\sim q_\phi(z|x)} \log(p_\theta(x|z))\Big] + \nabla_\theta \Big[\frac{1}{2}\sum_k \Big(\exp(\Sigma_\phi(x)) + \mu_\phi(x)^2 - 1 - \Sigma_\phi(x)\Big)\Big] \\
=&\frac{1}{L}\sum_{l=1}^{L} \nabla_\theta \log p_\theta(x|z^l) \textit{(Monte Carlo estimation)} + 0
\end{aligned}$$

where $z^l \sim q_\phi(z|x)$. For the derivative with respect to ϕ, $\nabla_\phi L(\theta, \phi)$ can be written as

$$\nabla_\phi L(\theta, \phi) = \nabla_\phi \Big[- E_{z\sim q_\phi(z|x)} \log(p_\theta(x|z))\Big] + \nabla_\phi \Big[\frac{1}{2}\sum_k \Big(\exp(\Sigma_\phi(x)) + \mu_\phi(x)^2 - 1 - \Sigma_\phi(x)\Big)\Big]$$

For the second term of the right hand side, it is easy to calculate. However for the first term of the right hand side, we face some challenge, since the derivative for the first term is difficult to take to the sample points.

$$\nabla_\phi E_{z\sim q_\phi(z|x)} f(z) \neq E_{z\sim q_\phi(z|x)} \nabla_\phi f(z)$$

To overcome this, the expectation is rewritten by linear transformation:

$$E_{z\sim q_\phi(z|x)} f(z) = E_{p(\epsilon)} f(g_\phi(\epsilon, x))$$

where $z = g_\phi(\epsilon, x)$, with $\epsilon \sim N(0, 1)$ and

$$g_\phi(\epsilon, x) = \mu_\phi(x) + \epsilon \odot \Sigma_\phi^{1/2}(x) = z \sim N(\mu(x), \Sigma(x))$$

This transformation makes sure that the distribution is independent of ϕ:

$$\begin{aligned}
&\nabla_\phi L(\theta, \phi) \\
=&\nabla_\phi \Big[- E_{z^l\sim p(\epsilon)} \log(p_\theta(x|z^l)) \Big] + \nabla_\phi \Big[\frac{1}{2} \sum_k \Big(\exp(\Sigma_\phi(x)) + \mu_\phi(x)^2 - 1 - \Sigma_\phi(x) \Big) \Big] \\
=&- \frac{1}{S} \sum_{l=1}^{S} \nabla_\phi \log p_\theta(x|z^l) + \nabla_\phi \Big[\frac{1}{2} \sum_k \Big(\exp(\Sigma_\phi(x)) + \mu_\phi(x)^2 - 1 - \Sigma_\phi(x) \Big) \Big]
\end{aligned}$$

If the loss function exclusively incorporates the log entropy term, VAEs will generate only one value (vector) for each input. Conversely, if the loss function solely incorporates the KL divergence term, VAEs will attempt to approximate $q_\phi(z|x)$ to be as close to the normal distribution $N(0, 1)$ as possible. This can potentially result in image blurriness and hinder effective learning from the training data.