Variational autoencoders, Kullback-Leibler divergence and Evidence Lower Bound

Quick note

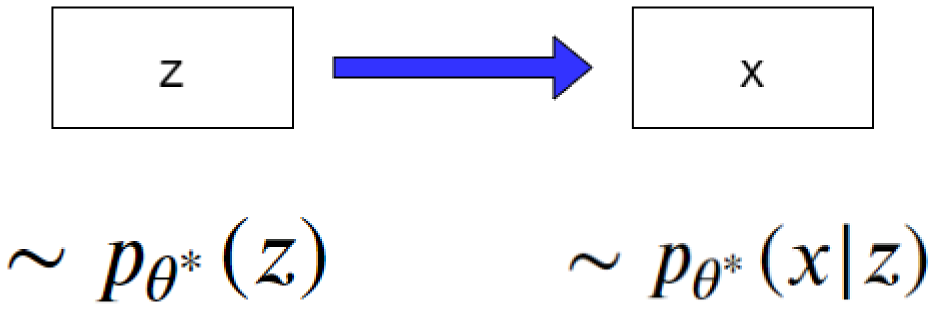
Abdelwahid Benslimane

wahid.benslimane@gmail.com

A variational autoencoder aims to learn how to generate the data it receives as input. To be a little more precise it will seek to learn the distribution of these data. I won't go further into the details of variational autoencoders because it is a supposedly acquired notion, but if you don't quite know yet what it is and what it is used for, it shouldn't be a hindrance to understand what follows. Just know that we are talking about generative AI.

The input variables x, whose true distribution is unknown, are reconstructed from latent variables z. The parameterization of the model (θ^*) and the latent variables are also unknown at the beginning.

The diagram below summarizes the situation.



A possible objective function will seak to maximize the log-likelihood on the definition set D of the variables x, i.e., to maximize the marginal likelihood of $x \sim p_\theta$ over the set of observations.:

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, E_{p_D}[log\, p_\theta(x)] = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \frac{1}{N} log\, p_\theta(x_i) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \frac{1}{N} log \int p_\theta(z) p_\theta(x_i|z) dz$$

but there are some challenges in getting there !

Among these chalenges:

$p_\theta(x) = \displaystyle\int p_\theta(z) p_\theta(x|z) dz$: we have no analytical expression either for this function or for its gradient.

An alternative approach would be to maximize the posterior expectation of the log-likelihood :

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \tfrac{1}{N} \sum_{i=1}^{N} E_{p_\theta(z|x_i)}[log\, p_\theta(x_i, z)].$$

We approach the posterior distribution $p_\theta(z|x)$ by a distribution $q_\phi(z|x)$ parameterized by ϕ. Provided that $q_\phi(z|x)$ is well constructed, this distribution gives access to values of z that are likely to be at the origin of an x.

It turns out that:

$$\hat{\phi} = \underset{\phi}{\operatorname{argmin}}\, KL(q_\phi(z|x)||p_\theta(z|x))$$

where KL is the Kullback-Leibler divergence, which is defined (in the discrete case) as:

$$KL(p||q) = \sum_i P(i) log \frac{P(i)}{Q(i)}.$$

This quantity measures how different distributions p and q are. The KL divergence is not symmetrical and should not be qualified as a distance as it is often seen.

With the evidence lower bound (usually just called ELBO) defined (in the discrete case) as:

$$\mathcal{L}(\theta, \phi, x) = E_{q_\phi(z|x)}[log\, p_\theta(x|z)] - KL(q_\phi(z|x)||p_\theta(z))$$

we can write:

$$\hat{\phi} = \underset{\phi}{\operatorname{argmax}}\, \mathcal{L}(\theta, \phi, x).$$