Course 2025 - chapter 1 - VAE

1. Generative models.

Goal of generative models.

Given some observed samples x from an unknown distribution of interest, the goal of a generative model is to learn the true data distribution $p(x)$ underlying the data or, at least, to generate data having the same distribtuion as the observed ones.

This is the objective of models such as Generative Adversarial Networks (GAN), autoregressive models, normalising flows, variational autoencoders (VAE), or diffusion denoising probabilistic models.

The use of latent variables.

Data can be thought as been represented or produced by some unobserved variables (called latent variables). Assume for instance, we observe image of handwritten digits (from 0 to 9). A generative model could benefit from any information about the number which is drawn, information which is not given by observation but latent. Any features (curvatures, presence of loops) may also help generating the number.

The fundamental assumption is that there exists a setting of latent variables from which the data can be generated. Formally, there should be a space $\mathcal{Z}$ of latent variables of dimension d, and a parametric function f_θ mapping $\mathcal{Z}$ into the data space $\mathcal{X}$ such that if Z is sampled from a distribution $p(z)$ than $X_\theta = f_\theta(Z)$ should be approximately sampled from the distribution $p(x)$ for some appropriate parameter values θ.

The function f is not observed and should not be constructed manually. It is inferred from data using some optimization procedure.

Optimization

Finding f_θ is not straightforward. In a likelihood-based approach, we shall relate the probability P_X of the data to the latent variable. This can either be done by integrating the the joint distribution of data and latent variables

$$p(x) = \int_{\mathcal{Z}} p(x, z) dz$$

or using a Bayes formula

$$p(x) = \frac{p(x, z)}{p(z|x)}.$$

But, the integration over latent variables is usually intractable and a ground truth distribution $p(z|x)$ is not available.

2. VAE.

Definition of the ELBO.

In the variational autoencoder approach, the posterior distribution $p(z|x)$ is approached by some parametric distribution $q_\phi(z|x)$.

Finding the appropriate parameters ϕ is related to the maximization of the evidence $\log p(x)$ with respect to ϕ. The parameter ϕ can be trivially introduced in the expression of the evidence as follows.

$$\begin{aligned}\log p(x) &= \int \log(p(x)) q_\phi(z|x) dz \\ &= \mathbb{E}_{q_\phi(z|x)}(\log p(x)).\end{aligned}$$

Using a Bayes' formula, it follows that

$$\begin{aligned}\log(p(x)) &= \mathbb{E}_{q_\phi(z|x)}\left(\log \frac{p(x,Z)}{p(Z|x)} \frac{q_\phi(Z|x)}{q_\phi(Z|x)}\right), \\ &= \mathbb{E}_{q_\phi(z|x)}\left(\log \frac{p(x,Z)}{q_\phi(Z|x)}\right) + \mathbb{E}_{q_\phi(z|x)}\left(\log \frac{q_\phi(Z|x)}{p(Z|x)}\right).\end{aligned}$$

This can also be written as

$$\log p(x) = \mathrm{ELBO}(x) + D_{KS}(q_\phi(Z|x)||p(Z|x)). \tag{1}$$

where ELBO stands for the Evidence Lower Bound and is defined as

$$\mathrm{ELBO}(x) = \mathbb{E}_{q_\phi(z|x)}\left(\log \frac{p(x,Z)}{q_\phi(Z|x)}\right), \tag{2}$$

and $D_{KS}(q||p)$ stands for th Kullbach-Leibler divergence between two distributions p and q and is defined as

$$D_{KS}(q||p) = \int \log \frac{q(y)}{p(y)} q(y) dy = \mathbb{E}_{q(y)}\left(\log \frac{q(y)}{p(y)}\right). \tag{3}$$

The KL-divergence measures the divergence between two distributions and is always non-negative. Hence,

$$\log p(x) \geq ELBO(x),$$

which explains why the ELBO is a lower bound of the evidence. Equation (1) reveals the interest of optimizing the ELBO for finding q_ϕ: since $\log p(x)$ is constant with respect to ϕ, increasing the value of the ELBO makes the KL-divergence between the approximate and the true posterior distributions decrease. Thus optimizing the ELBO can be used as a proxy to approximate the posterior distribution $p(z|x)$ by $q_\phi(z|x)$ without knowning the true posterior distribution.

Decomposition of the ELBO.

As mentioned in the introduction, a VAE involves a underlying mechanism to convert a latent variable z to a data x. We shall denote by $p_\theta(x|z)$ the parametric distribution describing this mechanism. This distribution can be introduced in the expression of the ELBO as follows.

$$\begin{aligned}\mathrm{ELBO}(x) &= \mathbb{E}_{q_\phi(z|x)}\left(\log \frac{p(x,Z)}{q_\phi(Z|x)}\right) \\ &= \mathbb{E}_{q_\phi(z|x)}\left(\log \frac{p_\theta(x|Z)p(Z)}{q_\phi(Z|x)}\right) \\ &= \mathbb{E}_{q_\phi(z|x)}\left(\log p_\theta(x|Z)\right) + \mathbb{E}_{q_\phi(z|x)}\left(\log \frac{p(Z)}{q_\phi(Z|x)}\right) \\ &= \mathbb{E}_{q_\phi(z|x)}\left(\log p_\theta(x|Z)\right) - D_{KL}(q_\phi(Z|x)\|p(Z)).\end{aligned}$$

The two terms of this expression of the ELBO has an interpretation. The first term, called the reconstruction term, ensures that the latent variables generate variables faithful to data. The second, called the regularization term, encourages the approximate posterior distribution to be close to a reference distribution, avoiding degenating into unwanted distributions. The reference distribution $p(z)$ is usually chosen as a standard multivariate Gaussian distribution $\mathcal{N}(0, \mathbf{I})$.

Optimization.

To find parameters ϕ and θ in a VAE, we shall maximize the expectation of the ELBO with respect to these parameters. According to the previous computations, this amounts to minimize

$$\mathbb{E}_{p(x)}\left(\mathbb{E}_{q_\phi(z|X)}\left(-\log p_\theta(X|Z)\right)\right) + \mathbb{E}_{p(x)}\left(D_{KL}(q_\phi(Z|X)\|p(Z))\right). \tag{4}$$

In a usual approach, the distribution $q_\phi(z|x)$ is assumed multivariate Gaussian $\mathcal{N}(\mu_\phi(x), \Sigma_\phi(x))$ with a mean $\mu_\phi(x)$ of size d and a diagonal covariance matrix $\Sigma_\phi(x)$ of size $d \times d$.

Due to this assumption, the regularization term can be explicitely written as

$$D_{KL}(q_\phi(Z|x)\|p(Z)) = \frac{1}{2}\left(|\mu_\phi(x)|^2 - d + \mathrm{tr}\left(\Sigma_\phi(x)\right) - \log\det\Sigma_\phi(x)\right)$$

Depending weither the values of x are binary or not, the distribution $p_\theta(x|z)$ is assumed Bernoulli $\mathcal{B}(f_\theta(z))$ or Gaussian $\mathcal{N}(f_\theta(z), \sigma^2\mathbf{I})$. As a result, the log of the distribution will be either the binary cross entropy or the mean square error, up to a constant. In the non binary case, we have

$$\mathbb{E}_{q_\phi(z|x)}\left(-\log p_\theta(x|Z)\right) = \mathbb{E}_{q_\phi(z|x)}\left(\frac{1}{\sigma^2}\left|f_\theta\left(z\right) - x\right|^2\right) + \mathrm{cst}.$$

To ease the computation of the gradient of such a term, we use another parametrization of Z. Let

$$Z = \mu_\phi(x) + \Sigma_\phi^{\frac{1}{2}}(x)\,\varepsilon,$$

with $\varepsilon \sim \mathcal{N}(0, \mathbf{I_d})$. Then the variable Z has the distribution $\mathcal{N}(\mu_\phi(x), \sigma_\phi^2(x)\mathbf{I})$. Hence, we have

$$\mathbb{E}_{q_\phi(z|x)}\left(-\log p_\theta(x|Z)\right) = \mathbb{E}_{\mathcal{N}(0,\mathbf{I})}\left(\frac{1}{2\sigma^2}\left|f_\theta\left(\mu_\phi(x)+\Sigma_\phi^{\frac{1}{2}}(x)\,\varepsilon\right)-x\right|^2\right)+\text{cst}.$$

Therefore, in the non binary case, we end up with minimizing

$$\begin{aligned}&\mathbb{E}_{p(x)}\left(\mathbb{E}_{\mathcal{N}(0,\mathbf{I})}\left(\left|f_\theta\left(\mu_\phi(X)+\Sigma_\phi^{\frac{1}{2}}(X)\,\varepsilon\right)-X\right|^2\right)\right)\\&+\beta\mathbb{E}_{p(x)}\left(\frac{1}{2}\left(|\mu_\phi(X)|^2-d+\text{tr}\left(\Sigma_\phi(X)\right)-\log\det\Sigma_\phi(X)\right)\right),\end{aligned}$$

where we have set $\beta = 2\sigma^2$.

Empirically, this criterion is approximated by

$$\sum_{i=1}^{n}\left|f_\theta\left(\mu_\phi(x_i)+\Sigma_\phi^{\frac{1}{2}}(x_i)\,\varepsilon_i\right)-x_i\right|^2+\frac{\beta}{2}\left(|\mu_\phi(x_i)|^2-d+\text{tr}\left(\Sigma_\phi(x_i)\right)-\log\det\Sigma_\phi(x_i)\right),$$

where x_i are some independent observations and ε_i i.i.d. $\mathcal{N}(0,\mathbf{I_d})$.

Link with an auto-encoder.

A VAE is analogous to a classical autoencoder (AE) in the sense that it includes two successive data transforms, a first one that encodes the data x into the latent variable z, and a second one that decodes the latent variable z to recover the data x. However, whereas the AE is built to estimate these transforms, the VAE aims at characterizing their probability distributions. Still, as for AE, layers of neural networks are used to estimate the probabality distributions in a VAE.
In the encoding part of the VAE, the distribution $q_\phi(z|x)$ can be represented by a succession of dense or convolution layers mapping the input data x into the couple of variables ($\mu_\phi(x)$, $\Sigma_\phi(x)$). In the decoding part of the VAE, the mapping f_θ is defined as a succession of layers mapping the latent variable z into x.

Data generation.

Once the VAE model has been learned, we can generate new data by removing the encoding part and encoding new realizations $\hat{\varepsilon}$ of ε

$$\hat{x} = f_\theta\left(\hat{\varepsilon}\right), \hat{\varepsilon}\sim\mathcal{N}(0,\mathbf{I_d}).$$