DEEP LEARNING PHD COURSE: ASSIGNMENT 3

Jalil Taghia

Department of Information Technology, Uppsala University (jalil.taghia@it.uu.se)

Due: June 2, 2019 at 23:59

VARIATIONAL AUTOENCODERS

For this part, the goal is to implement the vanilla VAE model and carry out experiments on the MNIST dataset. You are free to use high level or low-level toolboxes (packages) for this task.

Reading. Auto-Encoding Variational Bayes, Kingma and Welling[1].

MODEL

Consider a set of n i.i.d. observed variables $\{\mathbf{x}^{(i)}\}_{i=1}^{n}$ where $\mathbf{x}^{(i)}$ is a d-dimensional vector. The generative process assumes that data are generated by a random process involving a latent variable $\mathbf{z} \in \mathbb{R}^k$ on the continuous space, admitting the joint distribution

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} \,|\, \mathbf{z}) p(\mathbf{z}), \tag{1}$$

where $p(\mathbf{x} \,|\, \mathbf{z})$ is the likelihood, known as the generative network or decoder, and $p(\mathbf{z})$ is the prior. The goal is to find an approximation to the true intractable posterior $p(\mathbf{z} \mid \mathbf{x})$ using the VAE. Let $q(\mathbf{z} \,|\, \mathbf{x})$ be the variational posterior which approximates the true posterior $p(\mathbf{z} \,|\, \mathbf{x})$. Within the VAE framework, $q(\mathbf{z} \,|\, \mathbf{x})$ is known as the recognition network or the decoder.

The first step is to assign a prior distribution over the latent variable $\mathbf{z}$. Following standard formulation of the VAE, the prior is set and fixed to the centered isotropic multivariate Gaussian,

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k), \tag{2}$$

where $\boldsymbol{I}_k$ is a $(k \times k)$-dimensional identity matrix.

Next we need to assign parametric family of distributions for the generative and the recognition networks. The recognition network is assumed to take on a Gaussian distribution with a diagonal covariance matrix,

$$q_{\boldsymbol{\phi}}(\mathbf{z} \,|\, \mathbf{x}^{(i)}) = \mathcal{N}\left(\boldsymbol{\mu}^{(i)}, \mathrm{diag}(\boldsymbol{\sigma}^{2^{(i)}})\right), \tag{3}$$

where the mean $\boldsymbol{\mu}^{(i)}$ and the standard deviation $\boldsymbol{\sigma}^{(i)}$ of the approximate posterior are outputs of the encoding neural network, $\mathfrak{M}_{\boldsymbol{\phi}} : \mathbf{x} \rightarrow \{\boldsymbol{\mu}, \log\boldsymbol{\sigma}^2\}$, defined as:

$$\boldsymbol{\mu} = \boldsymbol{W}_{\boldsymbol{\mu}}\mathbf{h} + \boldsymbol{b}_{\boldsymbol{\mu}}, \quad \log\boldsymbol{\sigma}^2 = \boldsymbol{W}_{\boldsymbol{\sigma}}\mathbf{h} + \boldsymbol{b}_{\boldsymbol{\sigma}}, \quad \mathbf{h} = f_{\tanh}(\boldsymbol{W}_q\mathbf{x} + \boldsymbol{b}_q), \tag{4}$$

where the set $\boldsymbol{\phi}$ includes all the neural network parameters related to the recognition network, $\boldsymbol{\phi} = \{\boldsymbol{W}_{\boldsymbol{\mu}}, \boldsymbol{W}_{\boldsymbol{\sigma}}, \boldsymbol{W}_q, \boldsymbol{b}_{\boldsymbol{\mu}}, \boldsymbol{b}_{\boldsymbol{\sigma}}, \boldsymbol{b}_q\}$.

Exercise 1: Define a parametric family of distributions for the generative network (the decoder) that is suitable for the MNIST dataset. Motivate your choice.

[1] https://arxiv.org/abs/1312.6114

INFERENCE AND LEARNING

The exact marginal likelihood $\log p(\mathbf{x}^{(i)})$ for the ith observed variable is intractable. Instead we need to define a lower bound on the marginal likelihood, commonly known as the evidence lower bound (ELBO). We derive our lower bound based on the Kullback Leibler divergence. That leads to

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)} \,|\, \mathbf{z})] - \Delta_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z} \,|\, \mathbf{x}^{(i)}) || p(\mathbf{z})). \tag{5}$$

We want to optimize the lower bound $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)})$ w.r.t. both $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$.

Exercise 2: Explain why directly working with (5) is problematic.

We can derive a stochastic estimator which approximates (5) using the reparametrization trick and Monte Carlo samples for the evaluation of the expectations:

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = -\Delta_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z} \,|\, \mathbf{x}^{(i)}) || p(\mathbf{z})) + \frac{1}{r} \sum_{l=1}^{r} \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)} \,|\, \mathbf{z}^{(i,l)})$$
$$\text{where} \quad \mathbf{z}^{(i,l)} = g_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}^{(i,l)}, \mathbf{x}^{(i)}), \quad \text{and} \quad \boldsymbol{\epsilon}^{(l)} \sim p(\boldsymbol{\epsilon}). \tag{6}$$

Exercise 3:

- Compute $\Delta_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z} \,|\, \mathbf{x}^{(i)}) || p_{\boldsymbol{\theta}}(\mathbf{z}))$ in (6) for our choice of prior and posteriors in (2) and (3), respectively.
- Write down the exact functional form of the mapping function $g_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}^{(i,l)}, \mathbf{x}^{(i)})$ given our parametric choice of the variational posterior in (3);
- Indicate $p(\boldsymbol{\epsilon})$.

Exercise 4: Implementation of the VAE model and experiments.

- Implement the VAE using estimator $\tilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)})$ as defined in (6).
- Train your VAE model on the MNIST training set. You may define your stopping criteria based on the ELBO estimates on a validation set. But of course, you are free to use other techniques.
  (1) Compute the reconstruction error (the second term in (6)) and the regularization term (the first term in (6)) on the test set.
  (2) Generate random samples from the generative network, and visualize them. Explain briefly the process of generating random samples from the generative network.
  (3) Encode and decode a few randomly selected examples from your test set. Visualize your results.
- Repeat (1) to (3) for two different dimensionality of the latent variables: $k = 3$, and $k = 20$. Discuss your observations.