Varying Variational Autoencoders

Nathan Kong
University of Toronto
nathan.kong@mail.utoronto.ca

Jingyao (Jason) Li
University of Toronto
jasons_email@toronto.edu

Abstract

In variational inference, the approximate posterior distribution that is chosen is very important. It needs to be computationally tractable, yet flexible enough to approximate the true posterior. In this paper, we discuss an application of variational inference in dimensionality reduction. We experiment with the variational autoencoder (VAE), which was developed by Kingma and Welling (2013), by comparing two different variational inference methods. The first method is the vanilla VAE and the second method improves variational inference by introducing normalizing flows, developed by Rezende and Mohamed (2015), which increases the complexity of an initial simple distribution, so that more complex true posteriors can be potentially approximated.

1 Introduction

Nowadays, with increasingly large amounts of data, making posterior inferences is intractable since the evidence in Bayes' Rule consists of a computationally intractable integral. Stochastic variational inference was developed that makes this inference more tractable, by turning the inference problem into an optimization problem.

2 Formal Description

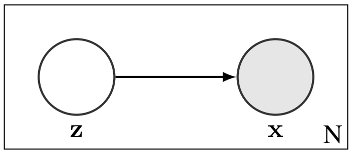

Figure 1: Probabilistic graphical model with latent variables, $\mathbf{z}$, and observed variables, $\mathbf{x}$.

From a probabilistic graphical model perspective, we have a latent space, governed by $\mathbf{z}$, and an observed space, which are our data, $\mathbf{x}$. The observed variables depend on the latent variables. Figure 1 illustrates this model. Using this framework, the joint density is: $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})\, p(\mathbf{z})$. $p(\mathbf{z})$ is the prior over the latent variables and $p(\mathbf{x} \mid \mathbf{z})$ is the likelihood of the data given the latent variables.

2.1 Variational Lower Bound

In order to obtain the marginal likelihood $p(\mathbf{x})$, we must integrate over $\mathbf{z}$, which is intractable. So, a posterior distribution, $q(\mathbf{z} \mid \mathbf{x})$, is introduced allowing us to obtain a lower bound on the marginal likelihood:

$$
\begin{aligned}
\log p(\mathbf{x}) &= \log \int_{\mathbf{z}} p(\mathbf{x} \,|\, \mathbf{z})\, p(\mathbf{z})\, \mathrm{d}\mathbf{z} && (1)\\
&= \log \int_{\mathbf{z}} \frac{q(\mathbf{z} \,|\, \mathbf{x})}{q(\mathbf{z} \,|\, \mathbf{x})} p(\mathbf{x} \,|\, \mathbf{z})\, p(\mathbf{z})\, \mathrm{d}\mathbf{z} && (2)\\
&\geq \mathbb{E}_q \left[\log p(\mathbf{x}, \mathbf{z})\right] - \mathbb{E}_q \left[\log q(\mathbf{z} \,|\, \mathbf{x})\right] && (3)\\
&= \log p(\mathbf{x}) - \mathbb{D}_{\mathrm{KL}} \left[q(\mathbf{z} \,|\, \mathbf{x}) \,||\, p(\mathbf{z} \,|\, \mathbf{x})\right] && (4)\\
&= -\mathbb{D}_{\mathrm{KL}} \left[q(\mathbf{z} \,|\, \mathbf{x}) \,||\, p(\mathbf{z})\right] + \mathbb{E}_{q(\mathbf{z} \,|\, \mathbf{x})} \left[\log p(\mathbf{x} \,|\, \mathbf{z})\right] && (5)
\end{aligned}
$$

Equation 3 follows from Equation 2 by applying Jensen's inequality. Equation 5 is known as the variational lower bound, which we want to maximize. Note that from Equation 4, maximizing the lower bound minimizes the Kullback-Leibler (KL) divergence between the approximate posterior and the true posterior and maximizes the marginal likelihood since the KL divergence is always positive. We want $q(\mathbf{z} \,|\, \mathbf{x})$ to be computationally tractable, but flexible enough to be able to match $p(\mathbf{z} \,|\, \mathbf{x})$ such that the KL divergence is close to 0.

2.2 Vanilla VAE

In an autoencoder, there are two main stages: the encoder stage and the decoder stage, which are both neural networks. The input to the encoder, or recognition, stage is the high dimension feature vector that we want to reduce to a space with lower dimensionality. In a VAE, the output of the encoder stage are parameters to the approximate posterior, $q(\mathbf{z} \,|\, \mathbf{x})$. If $q(\mathbf{z} \,|\, \mathbf{x})$ is a Gaussian, the parameters would be the mean, $\boldsymbol{\mu}$, and variance, $\boldsymbol{\sigma}^2$. In this case, the covariance matrix is a diagonal matrix.

We want to minimize the negative ELBO (negative of Equation 5) to train the VAE. It is: $\mathbb{D}_{\mathrm{KL}} \left[q(\mathbf{z} \,|\, \mathbf{x}) \,||\, p(\mathbf{z})\right] - \mathbb{E}_{q(\mathbf{z} \,|\, \mathbf{x})} \left[\log p(\mathbf{x} \,|\, \mathbf{z})\right]$ The first term measures how different the approximate posterior is from $p(\mathbf{z})$, so that a smaller value indicates a better approximation. The second term is the reconstruction error.

2.3 VAE with Normalizing Flow

3 Related Work

4 Comparison

4.1 Baseline VAE

For the vanilla VAE, we want to maximize the variational lower bound or minimize the negative variational lower bound.

4.2 VAE with Normalizing Flow

In Rezende and Mohamed (2015), variational inference is improved by introducing normalizing flows. Recall that in variational inference, we want $q(\mathbf{z} \,|\, \mathbf{x})$ to be flexible enough to approximate the true posterior, $p(\mathbf{z} \,|\, \mathbf{x})$, since the true posterior can be extremely complicated.

5 Limitations

6 Conclusions

References

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, pages 1530–1538, 2015.