Variational Auto-encoder Bayes

Zhong Wenjie
zvengin@nii.ac.jp

April 2, 2018

1 General Introduction

1.1 Problem Definition

This paper is going to tackle problem about inference in a directed graph model shown in figure 1. This problem can be formalized as following, let X denote variable, continous or discrete, and let Z denote continous latent variable. So model shown in figure 1 can be expresses as some probability distribution, priori probability $P_\theta(z)$, marginal probability $P_\theta(x)$, likelihood probability $P_\theta(x|z)$, posterior probability $P_\theta(z|x)$. Here likelihood probability is often regarded as a probabilistic encoder since it gives probability distribution of X when Z is given, and similarly, posterior probability is regarded as decoder. Therefore, model in figure 1 actually is auto-encoder and inference in this model refers to estimating posterior probability $P_\theta(z|x)$. So what this paper did is about how to estimate posterior probability and likelihood probability.

1.2 Methods

given dataset $X = \{x_i\}_{i=1}^N$ and a model, usually it is possible to get likelihood probability $P_\theta(x|z)$ and priori probability $P_\theta(z)$ parametralized by θ. But posterior probability $P_\theta(z|x)$ is intractable, since $P_\theta(x) = \int P_\theta(x|z)P_\theta(z)dz$ is intractable. Therefore, in this paper a recognition model $Q_\phi(z|x)$ is adopted to approximate posterior probability $P_\theta(z|x)$. once parameter ϕ is obtained, recognition model will be known since we all already know its format. In this paper, given dataset X, parameter ϕ and θ can be jointly estimated by optimizing lower bound of marginal probability of X. During the process of optimizing lower bound of $P_\theta(x)$, Monte Carlo gradient estimator is not used directly since its large variance, instead variable $Z \sim Q_\phi(z|x)$ is reparametralized as $Z = g_\phi(\epsilon, x)$ and $\epsilon \sim P(\epsilon)$. After that, we optimize newly obtained lower bound by using Monte Carlo methods.

1.3 Details

Marginal probability $P_\theta(x)$ can be written as following:

$$\begin{aligned} logP_\theta(x) &= \sum_z Q_\phi(z|x)logP_\theta(x) && (1)\\ &= \sum_z Q_\phi(z|x)log(\frac{P_\theta(x)*P_\theta(z|x)}{P_\theta(z|x)}) && (2)\\ &= \sum_z Q_\phi(z|x)log(\frac{Q_\phi(z|x)*P_\theta(z,x)}{P_\theta(z|x)*Q_\phi(z|x)}) && (3)\\ &= D_{KL}(Q_\phi(z|x)||P_\theta(z|x)) + E_{Q_\phi(z|x)}(-logQ_\phi(z|x) + logP_\theta(x,z)) && (4) \end{aligned}$$

Here lower bound of $P_\theta(x)$ is $L(\phi,\theta;X) = E_{Q_\phi(z|x)}(-logQ_\phi(z|x) + logP_\theta(x,z))$. given X, $logP_\theta(x)$ is a fixed value, we want to use $Q_\phi(z|x)$ to approximate $P_\theta(z|x)$, that is, we have to minimize D_{KL}, which is equvalent to maximize lower bound. If we directly use Monte Carlo gradient methods to construct estimator, it will cause large variance. Therefore, in this paper, latent variable $Z \sim Q_\phi(z|x)$ is reparameterized by a differentiable transforms $z = g_\phi(\epsilon, x)$ and $\epsilon \sim P(\epsilon)$. we can now form Monte Carlo estimates of expectation of some function w.r.t $Q_\phi(z|x)$

$$E_{Q_\phi(z|x)}(f(x)) = E_{p(\epsilon)}(f(g(\epsilon,x))) \approx \frac{1}{L}\sum_{l=1} f(g(\epsilon^{(l)},x)) \text{ where } \epsilon^{(l)} \sim P(\epsilon) \tag{5}$$

we apply this to the variational lower bound, yielding statistic gradient variational bayes.

$$L(\phi,\theta;X) \approx \hat{L}(\phi,\theta;X) = \frac{1}{L}\sum_{l=1}\{logP_\theta(x,z^{(l)}) - logQ_\phi(z^{(l)}|x)\} \text{ where } z^{(l)} = g_\phi(\epsilon^{(l)},x) \text{ and } \epsilon^{(l)} \sim P(\epsilon) \tag{6}$$

Given multiple data points from a dataset with N datapoints, we can construct an estimator of marginal likelihood lower bound of the full dataset, based on minibatches.

$$L(\phi,\theta;X) \approx \hat{L}^M(\phi,\theta;X^M) = \frac{N}{M}\sum_{i=1}^{M}\hat{L}(\phi,\theta;x^{(i)}) \tag{7}$$

where mini batch $X^M = \{x_i\}_{i=1}^M$ is randomly drawn M samples from X dataset with N datapoints. Now we can use calculate gradient of lower bound, and use SGD and Adapm methods to update parameters.