Bayesian Neural Networks

Ilya Zharikov, Roman Isachenko, Artem Bochkarev

Abstract

The recent advances in deep learning allow us to solve really hard real word problems with high accuracy. What these methods lack is interpretability of the model and uncertainty guarantees for the predictions. Bayesian framework allows to solve both this two issues, and with recent development in variational inference technique it is possible to take the best of two worlds and implement bayesian neural network. Our project is dedicated to testing different priors for neurons weights and exploring advantages and disadvantages of bayesian approach in deep learning.

Introduction

Problem Statement

Bayesian approach

Let suppose that $\mathbf{X}$ is the observed data which comes from the unknown distribution $p(\mathbf{X})$. This distribution defines our model and is called model evidence. We assume that our model also contains latent variables $\boldsymbol{\theta}$. The likelihood function is given by conditional probability distribution $p(\mathbf{X}|\boldsymbol{\theta})$. If we know the prior distribution $p(\boldsymbol{\theta})$ over hidden variables, we could use Bayes' theorem to derive the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{X})}, \tag{1}$$

where the evidence function $p(\mathbf{X})$ could be obtained by integration over hidden variables

$$p(\mathbf{X}) = \int p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

The posterior distribution shows the transformed initial prior knowledge of $\boldsymbol{\theta}$ given the observed data $\mathbf{X}$.

The main goal of bayesian inference to compute the posterior distribution. In some cases we could derive the posterior in the closed form formula. However, it is impossible for complex models where the latent variables lies in high-dimensional space. The problem is the denominator of (1), since it is obtained by integrating over all possible values of $\boldsymbol{\theta}$.

Suppose that we have the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$, then we could use the usual Maximum A Posteriori (MAP) approach to obtain point estimate for the parameters

$$\boldsymbol{\theta}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \left[\ln p(\boldsymbol{\theta}|\mathbf{X})\right] = \arg\max_{\boldsymbol{\theta}} \left[\ln p(\mathbf{X}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})\right].$$

In the case of flat prior distribution $p(\boldsymbol{\theta}) = \text{const}$, this estimate coincides with the Maximum Likelihood Estimation (MLE) approach

$$\boldsymbol{\theta}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \left[\ln p(\mathbf{X}|\boldsymbol{\theta})\right].$$

Variational inference

One of the widely used approach to get posterior distribution is sampling methods such as Metropolis-Hastings, Gibbs, NUTS algorithms. The general idea behind sampling methods is to obtain procedure for generating samples from the true posterior distribution. Most of these methods is based on Monte Carlo Markov Chains (MCMC) procedures.

References