An Empirical Study of Approximation Inference Algorithms on Bayesian Logistic Regression

Wei Dai
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
wdai@cs.cmu.edu

Abhimanu Kumar
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
abhimank@cs.cmu.edu

Jinliang Wei
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
jinlianw@cs.cmu.edu

1 Project Idea

Latent variable models have gained significant popularity due to their ability to model data that potentially arise due to latent causes. These models are also helpful in case of missing data as these missing entries can be modelled via hidden variables in graphical models. The parameter estimation for this class of models are in general intractable, and numerous approximation algorithms are widely employed. These approaches fall broadly into two categories: 1) variational inference, and 2) sampling-based. While it is generally believed that sampling produces better approximation than variational inference, albeit at a higher computational cost, to the best of our knowledge there is no comprehensive empirical study that compares these approximation inference schemes.

We plan to carry out an empirical comparative study of these approximation algorithms on Bayesian logistic regression, a well-studied minimal model with only one latent variable: the regression coefficients. The approximation algorithms we will use are:

1. Variational inference [1]:
    (a) A (near) close-form variational bound [2]
    (b) Laplace variational inference and delta method variational inference [3]
2. Sampling-based: MCMC using Gibbs sampling

We would apply the above approximation algorithms to the following estimation problems for a comparative study:

1. Maximum a posteriori (MAP) from variational and sampling algorithms
2. Maximum likelihood estimation (MLE)
3. Laplace approximation

2 Software, Datasets and Midterm Milstone

We plan to write all the code in matlab or C++ so that we have language agnostic comparison results. We will use 5 UCI classification datasets: 1) Farms Ads dataset (4,143 instances, 54,877 text features), 2) Amazon Commerce reviews dataset (1,500 instances, 10,000 real features) 2) p53 Mutants dataset (16,772 instances, 5,409 real attributes) 4) Human Activity Recognition using Smartphones dataset (10,299 instances, 561 real features) 5) URL Reputation dataset (2,396,130 instances, 3,231,961 real features).

We plan to finish MCMC sampling for all the three estimation by midterm.

See reference for papers to read.

References

[1] Michael Braun and Jon McAuliffe. Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335, 2010.

[2] Tommi S. Jaakkola and Michael I. Jordan. A variational approach to bayesian logistic regression models and their extensions, 1996.

[3] C. Wang and D. M. Blei. Variational Inference in Nonconjugate Models. *ArXiv e-prints*, September 2012.