ASI assessed exercise

29th April 2022

Introduction and Instructions

In this work you will analyze the Santander Customer Transaction Prediction dataset, which is available to download from:

https://www.kaggle.com/c/santander-customer-transaction-prediction/data

Listed below are various exercises to undertake. Note that in each case you should implement the algorithms yourself - you may not use existing implementations (specifically, for this exercise you are not allowed to use any off-the-shelf implementation of Bayesian linear regression and Bayesian logistic regression) - and should submit all of your code. You should have most of this code ready from the exercises that you have done during the labs.

Note that you are not allowed to work in groups for this assessed exercise - each student is required to submit her/his own work having worked on the exercise individually

Submission

You are free to use any programming language of your choosing, but it is your responsibility to ensure that we can run your code. We recommend you use the code you have developed during the lab sessions. Please submit either:

- Your code (including instructions for running - there should be one script that answers all the questions) and a .pdf report documenting your answers to the exercises.

- Or (preferably) a single iPython notebook that we can run. If you take this route, please *also* submit a .pdf output of the script (print the html to pdf). Your notebook should include any text descriptions required in the answers. (iPythons markdown cells allow you to add text)

Please submit your work by following the assignment link on the ASI course page on Moodle: https://moodle.eurecom.fr .

If you intend to submit collections of files for the code, please ONLY use .zip or .tar formats (PLEASE AVOID .rar and other unpopular formats)

The deadline is Wednesday 25th May at 4:00pm.

Exercises

Note (code) and (text) before each task indicate whether the corresponding part involves coding or writing.

A. (code) Download and import the Santander dataset. The labels of the test data are not publicly available, so create your own test set by randomly choosing half of the instances in the original training set. [3]

B. (text) Comment on the distribution of class labels and the dimensionality of the input and how these may affect the analysis. [7]

1. Bayesian Linear Regression

   A. (code) Implement Bayesian linear regression (you should already have an implementation from the lab sessions) [10]

   B. (text) Describe any pre-processing that you suggest for this data [5]

   C. (code) Treat class labels as continuous and apply regression to the training data. Also, calculate and report the posterior variance of the weights [10]

   D. (text) Suggest a way to discretize predictions and display the confusion matrix on the test data and report accuracy [5]

2. Logistic Regression

   A. (code) The goal is to implement a Bayesian logistic regression classifier; assume a Gaussian prior on the parameters. As a first step, implement a Markov chain Monte Carlo inference algorithm to infer parameters (you should already have an implementation of the Metropolis-Hastings algorithm from the lab sessions). [10]

   B. (code) Implement the variational approximation we studied in the course to obtain an approximation to the posterior over model parameters (you should already have an implementation of the from the lab sessions). [10]

   C. (code) Based on samples from the posterior over model parameters, write a function that computes the predictive distribution, and write the necessary functions to evaluate classification metrics such as the log-likelihood on test data and error rate. [10]

   D. (text) Comment on the tuning of the Metropolis-Hastings algorithm, and how to guarantee that samples are representative of samples of the posterior over model parameters. [5]

   E. (text) Comment on the tuning of the variational inference algorithm, and discuss the behavior of the optimization with respect to the choice of the optimizer/step-size. [5]

   F. (text) Report the error metrics implemented in point 2.B. above and the confusion matrix on the test data. Discuss logistic regression performance with respect to the performance of Bayesian linear regression. [5]

   G. (text) Compare the uncertainties on predictions obtained by the Metropolis-Hastings algorithm and variational inference. First, compare the log-likelihood on test data as a global metric to assess which inference method yields better uncertainty quantification. Second, pick a few test points for which the mean of the predictive distribution for Metropolis-Hastings is (a) around 0.5 (b) giving a correct prediction (c) giving a wrong prediction, and visualize/discuss what the predictive distribution looks like. Discuss the difference between the Metropolis-Hastings algorithm and variational inference. [15]

3. Bonus questions

   A. (code/text) Implement the Laplace approximation and compare the predictive mean and variances with the ones obtained by the other approximations.

   B. (code/text) Suggest and implement ways to improve performance.

Numbers at the end of each section are the number of marks available.

Be concise - a complete solution should be around 10/15 pages (including figures) and no more than 20/25 pages.