ASI assessed exercise 2018/2019

7th May 2019

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 yourselves - you may not use existing implementations (specifically, for this exercise you are not allowed to use any off-the-shelf implementation of logistic regression) - and should submit all of your code.

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 either Matlab or Python. 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 through Moodle at: https://cloud-platform.eurecom.fr/moodle/

If you intend to submit collections of files for the code, please only use .zip or .tar formats

The deadline is Wednesday 29th May 2019 at 4:00pm.

Exercises

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

1. (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]

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

3. Bayesian Linear Regression

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

    b) (text) Discuss how can you select the (hyper-)parameters for the Gaussian prior [5]

    c) (code) Write code that calculates the N-th order polynomial transformation of the input data. For simplicity, do not consider polynomials of more than one variable (such as x^2y), but raise each input variable to the power of N individually. Consider N=1, 2, 3, and 6. [5]

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

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

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

    g) (text) Discuss the performance, compare it against a classifier that outputs random class labels. [5]

4. Logistic Regression

    a) (code) The goal is to implement a logistic regression classifier that optimizes for the *Maximum a Posteriori* (MAP) estimate; assume a Gaussian prior on the parameters. As a first step, write a function that calculates the gradient of the joint likelihood. [10]

    b) (code) Write a simple gradient descend algorithm that uses the gradients calculated by the function of previous question to converge to the MAP estimate. [10]

    c) (text) Comment on the convexity of the problem; do you need multiple restarts in order to obtain a solution sufficiently close to the global optimum? [5]

    d) (text) Report the confusion matrix and classification accuracy on the test data. Discuss logistic regression performance with respect to the performance of Bayesian linear regression [5]

    e) (text) Laplace approximation is an efficient way to obtain an approximate posterior for logistic regression. Describe the steps of this approach. What is the

form of approximation obtained? [5]

5. Bonus question

(code) Implement the Laplace approximation and compare the predictive mean and variances with the ones obtained by linear regression.

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

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