CITS5508 Machine Learning
Semester 1, 2019
Lab Sheet 2

Assessed, worth 5%. Due: 11:59pm, Friday 22$^{\text{nd}}$ March 2019

1 Outline

In this lab sheet, you will learn to develop Python code for a small classification task. You will also learn the *Markdown* syntax to put comments and explanation in your Jupyter Notebook file to improve the presentation of your work.

2 Submission

Name your Jupyter Notebook file as **lab02.ipynb** and submit it to LMS before the due date and time shown above. You can submit your file multiple times. Only the latest version would be marked.

3 Dataset

We will use the **Forest type mapping dataset** supplied on the UCI Machine Learning website:

http://archive.ics.uci.edu/ml/datasets/Forest+type+mapping#

However, do NOT download and use the training and test sets from this website, as there are some undesirable blank characters in the class label column of these files. The correct `training.csv` and `testing.csv` files are available for download on the *Schedule and Material* page of the unit on LMS. Make sure that you do not rename the files to something else. You should save both files to the same directory with your `lab02.ipynb` file.

The training set (`training.csv`) contains 198 instances of multivariate remote sensing data of some forest areas in Japan. There are 4 different forest types labelled in the first column (the column heading is *class*), as described in the link above. The test set (file `testing.csv`) contains 325 test instances. This file has the same format as `training.csv`.

4 Tasks

Your tasks for this lab sheet are:

1. Read in the contents of both csv files. Inspect what the columns are by displaying the first few lines of the file. Use appropriate functions to display (visualize) the different features (or attributes / columns). Display some plots for visualizing the data. Describe what you see.
2. To simplify the classification task, write Python code to remove all the columns whose names begin with `pred_minus_obs`. You should have only 9 features (`b1`, `b2`, $\cdots$, `b9`) left for both the training and test sets. The class labels can be extracted from the first column of both datasets.
3. Write Python code to count the number of instances for each class label. Do you have an imbalanced training set?

4. Perform appropriate data normalization before performing classification. You can use `MinMaxScaler`, `StandardardScaler`, or any suitable normalization function in the `sklearn.preprocessing` package. You can also write your own normalization code if you prefer. Either way, ensure that you normalize the training data and the test data consistently.
5. Use the *stochastic gradient descent* classifier to perform one-versus-all binary classification on the 4 class labels. Show the confusion matrix on the test set. You should try experimenting with some hyperparameters to see if you can improve the performance of the classification.
6. Repeat the above step using the *logistic regression* classifier for multi-class classification with the *Softmax* function.
7. What is your conclusion? Which classifier gave better performance?

5 Presentation

A few tips on the presentation of your `ipynb` files:

- Present your `ipynb` file as a portfolio, with *Markdown* cells inserted appropriately to explain your code. See the following links if you are not familiar with *Markdown*:

  `https://www.markdownguide.org/cheat-sheet/` (basic)
  `https://jupyter-notebook.readthedocs.io/en/stable/examples/Notebook/Typesetting%20Equations.html` (more advanced)

- Dividing the portfolio into suitable sections and subsections (with section and subsection numbers and meaningful headings) would make your portfolio easier to follow.
- Avoid having too many small *Markdown* cells that have only one short sentence. In addition to *Markdown* cells, some short comments can be put alongside the Python code.
- Use meaningful variable names.
- When printing out your results, provide some textual description so that the output is meaningful.

6 Penalty on late submissions

See the *Unit Outline* about the penalty on late submissions.