Take-home Assessment

1 Description

This final practical is based on the material covered in lectures and previous practicals. You should write a practical report that should include a description and evaluation of the work done of not more than 5000 words excluding tables, graphs and images. The final practical will contribute 80% of the final mark. The deadline for uploading completed reports as a PDF to Moodle is **Tuesday, 19th January, 2021, 4pm**. You should be spending no more than 40 hours on the practical and your report.

Additionally, you will need to submit your code as Jupyter notebook(s) to the Moodle webpage. Assessors may run your code, but you will **not** be assessed on the quality of code writing, **nor** will you be assessed on the basis of where your system's results rank amongst others. The assessment will be based on the report itself and on clarity of description of the work done, evaluation performed, and insights gained.

2 Dataset

The dataset is: Stock market prediction using a diverse set of variables

You should download the dataset from the UCI ML Repository and should read: This paper

Note that the description of the data is in the appendix of this paper.

3 Your task

Your task is to build a machine learning pipeline to predict the daily movement of a stockmarket up or down conditioned on the values of variables in the dataset over the previous N (trading) days. You can calculate this from the `Close` columm. You can choose the value of N based on your reading of the accompanying paper and your own experimentation. You can base your approach on insights gained from thinking about the authors' assumptions and approach in the paper above, but you do not need to replicate their work, results, or approach to evaluation. We suggest you focus on one stockmarket file to begin with, but you can test the extent to which your approach will generalise on the others if you wish.

Your implementation and report should include the following steps:

- *Data exploration*: explore the different features in the dataset, gain insights from the data, and report your findings. Note that there are missing values in the dataset, and that there are features that may correlate closely.
- *Machine learning algorithms*: apply some algorithms that you learned about in the course and practicals. Try to find out which algorithms, architectures, training regimes, hyperparameter settings, etc., work better, and report your findings.

- *Evaluation*: consider different methods and measures for evaluating the algorithms. Report your results and compare your findings for the better-performing ML algorithms.
- *Visualisation and dimensionality reduction*: look into dimensionality reduction. For instance, you may consider using PCA or t-SNE on a selected set of features, plotting a scatter plot of the components, dropping some features, etc.

4 Computational Resources

You can ask sys-admin to set you up with a departmental GPU virtual machine if you want: see here and for some guidelines here.