EMATM0044 Introduction to AI
Coursework Part 1
Due date: 7th August 2023 13:00

Question 1 (40 pts)

For students NOT on the programme Data Science with Financial Technology

Download the dataset `coursework_resit_other.csv` from Blackboard. This dataset contains data on the chemical properties of some red wines. Your task is to predict the quality of the wine, as given in the final column labelled *quality*.

For students ON the programme Data Science with Financial Technology

Download the dataset `coursework_resit_fintech.csv` from Blackboard. This dataset has information on features of cars. Your task is to predict the price of the car, as given in the first column labelled *Price*.

All students:

You should consider the following aspects:

- The kind of algorithm to use (e.g.: classification/regression/clustering)
- The metric to use to measure the performance of the model
- What sort of baseline to compare the model to (sklearn has a module `sklearn.dummy` which may be useful in generating a baseline)
- How to choose the hyperparameters of your model
- How to test the performance of your model

Concretely, you should use two algorithms from scikit-learn and compare their performance on the dataset. You should also compare the performance of your chosen models against a baseline–i.e. a simple model that more complex models should be able to beat. sklearn has a module `sklearn.dummy` which may be useful in generating a baseline. You should use techniques to assess the ability of the models to generalise to unseen data and to ensure that your assessment of the models' performance is robust.

Material from worksheets 13, 14, and 16 will be helpful here.

Your answer to this question should take the form of a short report (maximum 4 pages), together with commented code, detailing the approach you will take. Make sure you address all the bullet points above, and explain your decisions. For example: 'I chose to use a X algorithm because Y'. 'Because of Z, I used metric M'. You should use plots and figures as appropriate to illustrate your decisions.

The code will not be marked for elegance, but it should run correctly. If you are using jupyter, a good tip is to make sure you have restarted the kernel and made sure that the code can run from scratch before submitting.

Q1 mark scheme (40 pts)

At least 2 algorithms should be tested. If only 1 is tested then the maximum points for the question is 20. You can obtain full marks using 2 algorithms plus the baseline.

(5pts) Overall presentation of the report, including use of appropriate sections, plots, diagrams, or tables to make your point. Do not include code snippets in the report. Instead, describe in words or equations what you are implementing. Format equations correctly.

(3pts) Picking a suitable type of algorithm (classification/regression/clustering) and justifying this choice. The lectures and worksheet from week 13 will be helpful here.

(3 pts) An appropriate choice of performance metric (e.g.: accuracy/precision/mean squared error etc) and justification. The lectures and worksheet from week 13 will be helpful here.

(4 pts) Discussion of the kind of baseline to compare against. (sklearn has a module `sklearn.dummy` which may be useful in generating a baseline)

(10 pts) Use of an appropriate method to select the hyperparameters of the chosen algorithms. The explanation of which hyperparameters are selected should be backed up with e.g. tables and plots to show which hyperparameter values were chosen and why. Please choose at least one model that uses hyperparameters so that you can show your knowledge in this area. If you choose one model without hyperparameters then please explain in a couple of sentences what the benefits of choosing a model without hyperparameters are. The lectures and worksheet from week 13 will be helpful here.

Breakdown

- 2 pts: Show that you understand what hyperparameters are and how they can be selected
- 3 pts: Look at the effects of different hyperparameter choices on the performance of your models
- 4 pts: Present the effects of the different hyperparameter choices on the performance of your models using tables, plots, or other presentation.
- 1 pts: State what choices you make and why

(10 pts) Training and testing the performance of the models in a way that shows whether the models are able to generalise to unseen data and that ensures that the performance of the models is robust. The lectures and worksheet from week 13 will be helpful here.

- 4 pts: Train models and select hyperparameters in a way that gives robust performance
- 3 pts: Test the performance of your models and compare their performance
- 3 pts: Make sure your models are tested in a way that shows whether they are able to generalise to unseen data

(5 pts) Analyse your results. What conclusions can be drawn?

Recommended structure of the short report

The short report should be no more than 4 pages. Shorter is fine. You should use LaTeX, MS Word, or a similar text editor to prepare the report and submit it as a pdf document.

- Introduction: State what the problem is. State what kind of algorithm needs to be used (classification/regression/clustering) and explain why that kind of algorithm needs to be used.
- Methods: State which specific algorithms you will use. State which performance metric you will use and why. Describe the baseline that you will measure your algorithms against. Describe how you will choose the hyperparameters of the algorithms. Explain which hyperparameters you have selected for each model using tables or plots to illustrate your decision.
- Results and Analysis: Report the results of your models. Use tables or plots as appropriate to illustrate your results.

Question 2: 10pts

The VGGFace2 dataset is a collection of images of faces. The website for the dataset is at `https://www.robots.ox.ac.uk/~vgg/data/vgg_face2/` and it is described in the paper Cao et al. [2018]. NB: You do not need to read the whole paper - the most important sections are sections I to IV. Provide answers to the datasheet questions in the template provided on Blackboard. **Please ensure you only answer the questions in the template.** Datasheets are described in the paper Gebru et al. [2021] available at `https://arxiv.org/abs/1803.09010`

Question 2 Mark Scheme

- Section 3.2: Composition. 5 pts
- Section 3.3: Collection Process. 3 pts
- Section 3.5: Uses. 2 pts

A template containing just the relevant questions is available on Blackboard.

The worksheet from week 17 will be helpful here. Example datasheets can also be seen in the appendix to the paper.

References

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. URL `https://arxiv.org/abs/1803.09010`.