EMATM0044 Introduction to AI
Coursework Part 1
Due date: Friday 7th May 13:00

Question 1 (35 pts)

Download the dataset `coursework1.csv` from Blackboard. This dataset consists of measurements of nitrous oxide (NOX) around a gas power station, together with data from 9 other sensors around the power station. Your task is to build a model to predict the amount of nitrous oxide emitted, given values on the other sensors.

Your 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 (not including the dummy baseline) and compare their performance on the dataset. 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 18 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 to building a system to predict the amount of nitrous oxide emitted. 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 (35 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) Picking a suitable type of algorithm (classification/regression/clustering) and justifying this choice. The lectures and worksheet from week 13 will be helpful here.

(2 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.

(3 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. The lectures and worksheet from week 13 will be helpful here.

(10 pts) 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.

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: Report the results of your algorithms in predicting the levels of nitrous oxide. Use tables or plots as appropriate to illustrate your results.

Question 2: (35 pts)

Download the dataset `coursework2.csv` from Blackboard. This dataset details the demand for hire bicycles, dependent on certain other factors. The bicycle hire company is interested in whether the demand for bicycles is going to be high, medium, or low on a given day. The number of bicycles used each day is given in `'count'`.

We will be using the attributes `'season'`, `'workingday'`, and `'weathersit'` to try to predict whether the demand is high, medium, or low. The attribute `'season'` contains an integer between 1 and 4 inclusive denoting the season. The attribute 'working day' contains a 1 if the day is Monday - Friday, and 0 if the day is Saturday or Sunday. The attribute `'weathersit'` contains an integer between 1 and 3 inclusive ranging from sunny to stormy.

You will be writing a decision tree to predict demand using these three attributes. The scikit-learn decision tree algorithm does not deal well with categorical data.

Submit your answers in the form of a short report (2 pages), using plots and images where appropriate, together with commented code.

5 pts Plot a histogram of the counts and use this information to divide the counts into three categories 'high', 'medium' and 'low'. Create a new feature in the dataset called `'usage'` which contains values `'high'`, `'medium'`, and `'low'` depending on the value in `'count'`.

4 pts Write a function to calculate the entropy of a subset of the target values.

6 pts Write a function to calculate the information gain of an attribute. Which attribute has the highest information gain over the whole dataset?

10 pts Write a function to build a decision tree. You may want to use a library to build trees such as `treelib`.

5 pts Display the decision tree (using, for example, the `show` function in `treelib`). Do you think some techniques should be applied to improve the structure of the tree? What do you suggest?

5 pts Create a new attribute called `'tempbins'` by binning the `'temp'` attribute into 'low' if `'temp'` < 12, 'medium' if $12 \leq$ `'temp'` < 24, and 'high' if `'temp'` ≥ 24. Does this attribute alter the structure of the decision tree?

Q2 mark scheme

(5 pts) Plot a histogram of the counts and use this information to divide the counts into three categories 'high', 'medium' and 'low'. Create a new feature in the dataset called `'usage'` which contains values `'high'`, `'medium'`, and `'low'` depending on the value in `'count'`.

- 2 pts histogram
- 3 pts for dividing counts and adding new feature to dataset.

(4 pts) Write a function to calculate the entropy of a subset of the target values.

- 4 pts for the function

(6 pts) Write a function to calculate the information gain of an attribute. Which attribute has the highest information gain over the whole dataset?

- 5 pts for the function
- 1 pt for the correct answer.

(10 pts) Write a function to build a decision tree. You may want to use a library to build trees such as `treelib`.

- For full marks in this question you should write a general function that could be applied to other datasets.
- If you write a function that builds the decision tree in a concrete fashion that could not be applied to other datasets, partial credit will be awarded.
- If you just run the sklearn decision tree function, no marks will be awarded.

Material from week 18 will be useful here.

(5 pts) Display the decision tree. Do you think some techniques should be applied to improve the structure of the tree? What do you suggest?

- 2 pts for displaying the decision tree.
- 3 pts for discussion of whether or not to apply techniques to improve the structure of the tree and which if so.

(5 pts) Create a new attribute called `'tempbins'` by binning the `'temp'` attribute into 'low' if `'temp'` < 12, 'medium' if $12 \leq$ `'temp'` < 24, and 'high' if `'temp'` ≥ 24. Does this attribute alter the structure of the decision tree?

- 2 pts for creating the new attribute
- 2 pts for building a new decision tree including that attribute
- 1 pt for correct answer to whether the attribute alters the structure of the tree.

Recommended structure of the short report

The short report should be around 2 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.

- Histogram of the counts, and statement of how the values are divided into high, medium, and low.
- Image of your decision tree and discussion of whether or not to apply techniques to improve the structure of the tree and which if so.
- Discussion of whether the new attribute alters the structure of the decision tree.

Question 3: 30pts

Write a datasheet for the dataset `http://saifmohammad.com/WebPages/wikiartemotions.html`

Questions for the datasheet are available in Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. `https://arxiv.org/abs/1803.09010`

You should cover 5 questions from each section, and explain if a question is not applicable.

As well as the dataset and accompanying paper, you should look at the Terms of Use and the Ethics and Data Statement which are linked at `http://saifmohammad.com/WebPages/lexicons.html`

Question 3 Mark Scheme

- Section 3.1: Motivation. This section has fewer questions: 3 pts
- Section 3.2: Composition. Answer 5 questions from this section: 5 pts
- Section 3.3: Collection Process. Answer 5 questions from this section: 5 pts
- Section 3.4: Preprocessing/cleaning/labelling. This section has fewer questions: 2 pts
- Section 3.5: Uses. Answer 5 questions from this section: 5 pts
- Section 3.6: Distribution. Answer 5 questions from this section: 5 pts
- Section 3.7: Mintenance. Answer 5 questions from this section: 5 pts

In order to get full marks for this section, students should answer the questions correctly. Also, if students choose to answer 5 questions that are not applicable to the dataset, by saying 'not applicable', this will not be viewed as an adequate answer. If there are fewer than 5 questions in a section and some are not applicable, then it is acceptable to answer 'not applicable'

The worksheet from week 16 will be helpful here. Example datasheets can also be seen in the appendix to the paper.