COMM054 Coursework Assignment

Submission deadline – 4:00pm 28th November 2024.

There are 100 marks in total available, contributing a total of 50% of the final module mark. The remaining 50% of the final module mark is contributed by an exam.

You must not copy and paste code or written sections of your answers from other sources, or from other students. This includes explanations of models and metrics, and discussions of results. **The Turnitin score should not be taken as an indication that a certain amount of copying is acceptable** – even copying a few sentences or lines of code from another source could still count as academic misconduct.

Coursework Tasks – Regression and Classification

You should produce a Jupyter Python notebook where you analyse the following two datasets. You can use any of the Python libraries covered in the module for these tasks, and you do not need to implement the regression and classification models yourself, you can use the scikit-learn library.

Dataset 1 - Regression

The data for this task is provided on SurreyLearn in the file `syield.csv`

The file is a CSV format file with two columns, one labelled as `DOY`, which contains the day of the year recorded as an integer, and another labelled `Yield` which contains a measure of the yield of a sorghum crop. The aim is to use the `DOY` feature to predict the `Yield` column.

The coursework task is to apply and evaluate linear and polynomial regression methods on this data. You do not need to understand the interpretation of the variables provided in the data to complete the coursework.

Task 1.1

- Train a linear regression and polynomial regression models of degree ranging from 2 to 5, to predict yield (`Yield` column) based on the day of the year (`DOY` column).

Task 1.2

- Use appropriate methods to evaluate the performance of the regression models when applied to data not used in training, and discuss the results.
- Explain any metrics that are used. [Around 200 words]

Task 1.3

- Discuss the performance of the regression models, and whether you think each of them is overfitting or underfitting the data, explaining why you think this may be the case. [Around 500 words]

Task 1.4

- Produce plots to illustrate any metrics you have calculated, and to show the fit of the models to the data.

Dataset 2 – Classification

The data for this task is provided on SurreyLearn in the file `ATLAS.csv`

The file is a `CSV` file, with columns for the features `V1` to `V29` and a single target feature `Signal`.

This dataset was originally produced as part of an open machine learning challenge to develop methods for detecting Higgs bosons from data generated by the ATLAS detector. You can read more about ATLAS here: https://atlas.cern.

Each event belongs to one of two classes, either background or signal, and has several associated measurements that can be used to predict the type of an event. Each row in the data file corresponds to an individual event, with columns for each of the measurements. For each event, several measurements from the ATLAS detector are provided, and these can be used to predict the class of the event. You are not expected, and do not need to understand the meaning of the different measurements for this coursework task.

Task 2.1

- Train a logistic regression model to predict the class of events from the data provided.

Task 2.2

- Use appropriate methods to evaluate the performance of the classification when applied to data not used in training, and discuss the results.
- Explain any metrics that are used. [Around 200 words]

Task 2.3

- Discuss the performance of the classification model, in terms of the metrics calculated, explaining which may be the most important in this context. [Around 300 words]

Task 2.4

- Produce plots to illustrate any metrics you have calculated.

Coursework Task Marks Breakdown:

- Application of linear regression:
  - Python code to perform linear regression.

    [10 marks]
- Application of polynomial regression:
  - Python code to perform polynomial regression.

    [10 marks]
- Application of logistic regression:
  - Python code to perform logistic regression.

    [10 marks]
- Performance comparison of the models used in the tasks:
  - Use of appropriate methods for validating the performance of the methods applied for regression and classification.

    [20 marks]
  - Selection and explanation of appropriate metrics to measure performance for regression and classification.

    [20 marks]
- Discussion of the performance of the models, and discussion of overfitting and underfitting in the regression models used.

  [20 marks]
- Use of visualisation to compare metrics between the models and illustrate the fit of the models to the data.

  [10 marks]

Assignment Submission Requirements

- You should submit one Jupyter notebook containing your answers to the task **and** a PDF file version of your Jupyter notebook. Do not submit these as a single archive (zip/tar) file, but as two separate files.
- Please include your student ID number in the name of the file containing your work.