Assessment Autumn 21/22

Due: last commit on or before August 21st, 2022

To submit, please email your repository URL to `ian.mcloughlin@atu.ie`.

These are the instructions for the assessment of Machine Learning and Statistics in Autumn 2021/2022. The assessment is worth 100% of the marks for the module. Please read the *Using git for assessments* [1] document on the Moodle page which applies here. As always, you must also follow the code of student conduct and the policy on plagiarism [2].

Instructions

The purpose of this assessment is to ensure that you have achieved the learning outcomes of the module while also providing you with sample work to show prospective employers. The overall assessment is split into the three components as detailed below. The percentages beside each heading give the weighting of each of the three components. You may assume that each bullet point has an equal weighting within its component. Note, however, that the examiners' overall impression of your submission may override the individual weightings where deemed appropriate.

GitHub Repository (20%)

Create a GitHub repository containing two Jupyter notebooks – these are described further down. The repository should contain the following:

- A clear and informative `README.md` explaining why the repository exists, what is in it, and how to run the notebooks.
- A `requirements.txt` file that enables someone to quickly run your notebooks with minimal configuration. You should also include any other required files such as data files and image files.

Scikit-Learn Jupyter Notebook (40%)

Include a Jupyter notebook called `scikit-learn.ipynb` that contains the following.

- An overview the `scikit-learn` Python library [3].
- An example of using `scikit-learn` to perform regression. You may use any open dataset you wish for this purpose, including any built-in to the library. Appropriate plots and explanations in MarkDown should be included.

- An example of using `scikit-learn` to perform classification. You may use any open dataset you wish for this purpose, including any built-in to the library. Appropriate plots and explanations in MarkDown should be included.
- An appropriate Dockerfile for the repository.

ANOVA Notebook (40%)

Include a Jupyter notebook called `anova.ipynb` that contains the following.

- A comparison of performing ANOVA using the `scipy.stats` Python library [4] versus the `statsmodels` Python library [5].
- An example hypothesis test using ANOVA with `scipy.stats`. You should find an appropriate open data set online, ensure the assumptions underlying ANOVA are met, and then perform and display the results of your ANOVA.
- The same analysis done using `statsmodels`.
- Appropriate plots and other visualisations to enhance your notebook for viewers.

More information about marking

In completing each component of the assessment, you should consider the following four overall aspects of academic work. It is important that your submission provides direct evidence of each aspect. For instance, your commit history should demonstrate that you were consistent in your work. Likewise, your submission should have references in it to demonstrate that you considered the literature and the work of others.

Research

Evidence of research performed on topic; submission based on referenced literature, particularly academic literature; evidence of understanding the documentation for any software or libraries used.

Development

Environment can be set up as described; code works without tweaking and as described; code is efficient, clean, and clear; evidence of consideration of standards and conventions appropriate to code of this kind.

Consistency

Evidence of planning and project management; pragmatic attitude to work as evidenced by well-considered commit history; commits are of a reasonable size; consideration of how commit history will be perceived by others.

Documentation

Clear documentation of how to create an environment in which any code will run, how to prepare the code for running, how to run the code including setting any options or flags, and what to expect upon running the code. Concise descriptions of code in comments and README.

References

[1] Ian McLoughlin, "Using git for assessments," https://github.com/ianmcloughlin/using-git-for-assessments/.

[2] ATU, "Quality Assurance Framework," https://www.gmit.ie/general/quality-assurance-framework.

[3] "scikit-learn: machine learning in python — scikit-learn 0.24.2 documentation," 2021. [Online]. Available: https://scikit-learn.org/stable/index.html

[4] "Statistical functions (scipy.stats)," 2022. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/stats.html

[5] "statsmodels," 2022. [Online]. Available: https://www.statsmodels.org/stable/index.html