Machine Learning and Statistics (Autumn)

Autumn 2023/2024

These are the assessment instructions for Machine Learning and Statistics in Autumn 2023/24. They cover 100% of the marks for the module.

Students who have deferred or failed the assessment of this module during the semester usually complete this Autumn assessment.

The assessment consists of three parts: a repository, a set of tasks, and a small project.

The deadline for all elements is **Friday, 23 August 2024**.

Repository Component (20%)

The first component is to set up a GitHub repository. This repository should include all work you are submitting for this assessment. Make sure your work is in the `main` branch of the repository, which is usually the default.

After creating the repository, immediately submit your repository URL using the form on the module page. Your submission will be graded based on the last commit in GitHub on or before the deadline day.

Your repository should be well-organized. At a minimum, it should have a clear `README.md`, a relevant `.gitignore` file, and any requirements below. Avoid including unnecessary files or folders.

Try to use lower case for file and folders names, except for the `README.md`. Avoid using spaces or other unusual characters in file and folder names. However, it is okay to use underscores, hyphens, and full stops.

Your completed repository should be easy to present during technical job interviews. An interviewer should be able to understand your work and how to interact with it without your assistance. This will significantly impact your grade for this component, and all the other components.

Tasks (40%)

Create a Jupyter notebook called `abalone.ipynb` within your repository. Use this single notebook for all tasks and the project below. For each task, include explanations in Markdown cells alongside any code cells. Ensure all code cells include comments, and try to break your work into smaller code cells where possible.

1. Download the abolone.csv file and save to your repository. You can find the file here and more information about the dataset is available from the UC Irvine Machine Learning Repository.
2. Use the `read_csv()` function from the `pandas` library to load the data into your notebook.
3. Within the notebook, discuss the classification of each variable in the dataset based on common variable types and measurement scales in mathematics, statistics, and Python.
4. Select, demonstrate, and explain the most appropriate summary statistics to describe each variable.
5. Select, demonstrate, and explain the most suitable plot(s) for each variable.

Project (40%)

In the same notebook where you completed the above tasks, conduct an analysis of classification algorithms applied to the `abalone.csv` data set. Begin by explaining supervised learning and the concept of classification algorithms.

Describe at least one common classification algorithm and demonstrate it using the `scikit-learn` Python library. Throughout the notebook, use appropriate plots, mathematical notation, and diagrams to explain the relevant concepts.

Marking Scheme

Each component will be assessment based on the following four categories, each carrying equal weight. Remember, your repository is what will be evaluated. It should demonstrate evidence of the criteria outlined for each category.

In line with ATU policy, the examiners' overall impression of the submission may affect marks in each category. At any stage you may be asked to discuss the work to date in your repository.

Research

- Evidence of research on relevant topics.

- Appropriate referencing.
- Building upon the literature and documentation.
- Comparisons to similar work.

Development

- Clear, concise, and correct code.
- Appropriate tests.
- Knowledge of different approaches and algorithms.
- Clean architecture.

Documentation

- Clear explanations of concepts in notebooks.
- Concise comments in code and elsewhere.
- Appropriate README for repository.

Consistency

- Tens of commits, each representing a reasonable amount of work.
- Literature, documentation, and code evidencing work on the assessment.
- Evidence of reviewing and refactoring.

Advice

In open-style assessments like this one, students may find it challenging to navigate the freedom provided. Guided by the module's materials, you'll need to determine where and how to begin, what content is relevant for your submission, how much is appropriate, and how to personalize your work. This level of autonomy is intentional and meant to foster independent thinking and decision-making skills.

Companies value graduates who can take initiative, work autonomously, and make design decisions with minimal guidance. We assume you have a reasonable knowledge of programming and an ability to source your own information. You need a plan, you cannot just start coding straight away.

Remember you must adhere to ATU policies and regulations. You can view these on the Student Hub. Pay special attention to the Policy on Plagiarism and the Student Code. If you have any questions about what is permitted, reach out the lecturer by email.

Purpose

The purpose of the assessment is to ensure students can demonstrate the following.

1. Describe the stochastic nature of real-world measurements.
2. Select an appropriate mathematical model of a real-world problem.
3. Select an appropriate cost function for a given machine learning task.
4. Apply an optimization technique to the parameters of a model.

16 June 2024 #assessments