Assignment 1 (CA1: 40%)

The objective of the assignment is to help you gain a better understanding of machine learning tasks of regression and classification.

Guidelines

1. You are to work on the problem set individually.

2. In this assignment, you will solve typical machine learning tasks and write a report that describes your solution to the tasks.

3. Submit your Python code and the report in a compressed package (zip file). Alternatively, write a Jupyter notebook including your code and comments.

4. Students are required to submit their assignment using the assignment link under the Assignment folder. Please remember to include your student name and student admission number on the first page of your assignment report.

5. The normal SP's academic policies on Copyright and Plagiarism applies. Please note that you are to cite all sources. You may refer to the citation guide available at: http://eliser.lib.sp.edu.sg/elsr_website/Html/citation.pdf

Submission Details

Deadline: **June 5, 2022, 23:59**
Submit through: **POLITEmall**

Late Submission

50% of the marks will be deducted for assignments that are received within ONE (1) calendar day after the submission deadline. No marks will be given thereafter.
Exceptions to this policy will be given to students with valid LOA on medical or compassionate grounds. Students in such cases will need to inform the lecturer as soon as reasonably possible. Students are not to assume on their own that their deadline has been extended.

PART A: CLASSIFICATION (50 marks)

This part of the assignment is to be completed individually.

Background

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

Complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.

Dataset

You are to use the dataset.
https://www.kaggle.com/c/titanic/data

Tasks

1. Write the code to solve the prediction task. Normally you would be using scikit-learn, but if you'd prefer to work with your own implementation of learning algorithms, or some other toolkit, that is fine.
2. Use the Kaggle test set and submit to Kaggle site to obtain the test score. Screen shot the test score and submit with the Jupyter notebook used and the submission.csv file.
   Use the URL: https://www.kaggle.com/c/titanic/submit
3. Write a report detailing your implementation, your experiments and analysis in the Jupyter notebook (along with your python code and comments). In particular, we'd like to know:

   - How is your prediction task defined? And what is the meaning of the output variable?
   - How do you represent your data as features?
   - Did you process the features in any way?
   - Did you bring in any additional sources of data?
   - How did you select which learning algorithms to use?
   - Did you try to tune the hyperparameters of the learning algorithm, and in that case how?
   - How do you evaluate the quality of your system?
   - How well does your system compare to a stupid baseline?
   - Can you say anything about the errors that the system makes? For a classification task, you may consider a confusion matrix.

- Is it possible to say something about which features the model considers important? (Whether this is possible depends on the type of classifier you are using)

4. Create a set of slides with the highlights of your Jupyter notebook report. Explain the entire machine learning process you went through, data exploration, data cleaning, feature engineering, and model building and evaluation. Write your conclusions.

Submission requirements

1. Submit a zip file containing all the project files (Jupyter notebook), all data sets used, Kaggle test score screenshot and the slides (PPTX or pdf).
2. Submit online via the Assignment link.

Evaluation criteria:

Application of suitable algorithms	20%
Suitable evaluation of algorithms	20%
Background research	20%
Presentation/Demo	20%
Quality of report (Jupyter)	20%

PART B: REGRESSION (40 marks)

This part of the assignment is to be completed individually.

Background

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.

Dataset

You are to use the dataset.
https://www.kaggle.com/harlfoxem/housesalesprediction

Tasks

1. Write the code to solve the prediction task. Normally you would be using scikit-learn, but if you'd prefer to work with your own implementation of learning algorithms, or some other toolkit, that is fine.
2. Write a report detailing your implementation, your experiments and analysis in the Jupyter notebook (along with your python code and comments). In particular, we'd like to know:

    - How is your prediction task defined? And what is the meaning of the output variable?
    - How do you represent your data as features?
    - Did you process the features in any way?
    - Did you bring in any additional sources of data?
    - How did you select which learning algorithms to use?
    - Did you try to tune the hyperparameters of the learning algorithm, and in that case how?
    - How do you evaluate the quality of your system?
    - How well does your system compare to a stupid baseline?
    - Can you say anything about the errors that the system makes?
    - Is it possible to something about which features the model considers important? say

3. Create a set of slides with the highlights of your Jupyter notebook report. Explain the entire machine learning process you went through, data exploration, data cleaning, feature engineering, and model building and evaluation. Write your conclusions.

Submission requirements

1. Submit a zip file containing all the project files (Jupyter notebook), all data sets used, and the slides (PPTX or pdf).
2. Submit online via the Assignment link.

Evaluation criteria:

Application of suitable algorithms	25%
Suitable evaluation of algorithms	25%
Background research	25%
Presentation/Demo	12%
Quality of report (Jupyter)	13%

PART C: TECHNICAL PAPER (10 marks)

This part of the assignment is to be completed individually. This is a challenge task for students who wish to attempt it for higher marks.

Write a technical paper on any **ONE** of the following topics.

- Classification
- Regression

Find a suitable real-world dataset from Kaggle or other public repositories with real world data such as data.gov.sg.

The paper should have the following component:

1. Abstract
2. Introduction
3. Related Works
4. Dataset/Methodology/Experiment
5. Discussion
6. Conclusions
7. References

Submit the paper in Word or PDF format (page limit of 10 pages)

— End of Assignment —