SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

ENG335
Machine Learning

Group-based Assignment

July 2021 Presentation

GROUP-BASED ASSIGNMENT

This mini-project assignment is worth 15% of the final mark for ENG335 Machine Learning. The total mark assigned to this assignment is 100 marks.

This is a group-based assignment. You should form a group of **minimum** 2 and **maximum** 4 members from your seminar group. Each group is required to upload a single report to Canvas Turnitin via your respective seminar group. Please elect a group leader. The responsibility of the group leader is to upload the report on behalf of the group. In your 1-page cover sheet, please include all project partners' names and student PI numbers.

Note to Students:

You are to submit the GBA assignment i.e. using Canvas in the form of a single MS Word file. It should be saved as ENG335_GBA01_yournames.doc Submission in any other manner like a hardcopy or any other means will not be accepted. You are to ensure that the file to be submitted does not exceed 20MB in file size.

Additional Instructions for Submission:
Please follow the submission instructions stated below:

1. Please submit all Program Code / Answers in the form of a Juypter Notebook file (i.e. .ipynb File) for all the programming questions via the additional submission link found under Assignments in your respective ENG335 T group course site.

2. All Answers for each question should be indicated clearly using the Comments section / markups in the Notebook so that the marker can see clearly which code is for which Question. (e.g. # Answer for Q1a).

The submission deadline for this assignment is announced on L01 course site**. Late submissions will carry mark penalty.**

Questions: (Total 100 marks)

Question 1

(a) Read about the term "Explainable AI". In your own words explain this term and appraise the need for it.

(5 marks)

(b) Read about Waze carpool from https://cloud.google.com/blog/products/ai-machine-learning/how-waze-predicts-carpools-using-google-cloud-ai-platform.

(i) Formulate the problem statement that the AI platform is trying to solve and how it is being used for the carpool.

(4 marks)

(ii) From the information provided in the blog, give the number of parameters and number of records used for training the model. If you think the values are not stated in the article, then state that the values are not available.

(4 marks)

(iii) What is the maximum latency provided by the Google AI platform for this application. Provide a numerical value. If its not available then state that the latency information is not available.

(3 marks)

(iv) In the high level schema architecture, there is offline and online processing. Explain what you infer from this schema architecture. Specify the algorithms and framework being used.

(2 marks)

(v) Is the learning supervised or unsupervised? Explain your answer.

(2 marks)

(c) Construct a problem statement relevant to Singapore that can be solved using AI platform and cloud infrastructure offered by AWS or Google or other vendors. You are required to provide the following details:

(i) State the problem/scenario. You can construct your own problem or discuss any AI solution relevant to Singapore.

(4 marks)

(ii) List any **FOUR (4)** parameters that will be present in your dataset used for training.

(4 marks)

(iii) Is the learning supervised or unsupervised? If supervised, then provide the target variable.

(2 marks)

Question 2

Download the real estate dataset from the Kaggle link https://www.kaggle.com/quantbruce/real-estate-price-prediction

(a) Perform exploratory data analysis and identify the parameters.

(7 marks)

(b) Design a linear regressor to predict the price of the house using only **TWO (2)** parameters. Specify the linear regression equation obtained from learning the dataset. Explain what you infer from observing the linear regression equation.

Note: You are required to select the best **TWO (2)** parameters and justify your selection.

(10 marks)

(c) Assess the performance of the linear regressor by getting the relevant performance metrics. You need to provide any **THREE (3)** metrics and explain the importance of these metrics.

(3 marks)

Question 3

Download the Iris dataset from the scikit-learn package.

(a) Perform exploratory data analysis and understand the dataset. Select any **TWO (2)** classes from the dataset. Implement a suitable algorithm from what you have learned in the class for predicting the target in the Iris dataset.

(14 marks)

(b) Implement a Naïve Bayes classifier for the Iris dataset.

(5 marks)

(c) Compare the performance metrics of the algorithm in Question 3(a) and Naïve Bayes classifier. Does the scaling of the parameters have any impact on the performance? (Justify your answer)

(6 marks)

Question 4

Use the breast cancer dataset available in sklearn package. You are required to show the steps in loading the data set, perform exploratory data analysis, identify the algorithm (from what has been covered in the seminars) suitable for detecting breast cancer. Present appropriate performance metrics. You can use the following Python code to load the dataset.

(25 marks)

-----END OF GBA ASSIGNMENT-----