INFS4203/7203 Project Phase II (20 marks)

Semester 2, 2022

Due date:

16:00 on 28th October 2022 (Brisbane Time) (Phase II, 20%)

All assignments should be submitted to UQ Blackboard. If any assignment is failed to be submitted appropriately before due, a penalty will be applied according to ECP. Please take the responsibility to ensure your submission is successful before due time. Email submission will not be accepted.

Overview

In Phase II, you will implement your proposal submitted in Phase I, with necessary **adjustment** according to the empirical performance and the feedback from the proposal. This is an individual assignment. The completion of the assignment should be based on your own design and feedback from the proposal.

Track 1: Data-oriented project

In Phase II, you will be provided with the test data named *Ecoli_test.csv*. The first row describes features' names. Except the first row, each row in the data file corresponds to one data point. There are 917 test data points in this file, and each column represents the same feature as the training data *Ecoli.csv*. Note that the test data only has 106 columns, without labels, i.e., without the final column "Target (Column 107)" in *Ecoli.csv*. Labels for the test data will not be released and will be used by the teaching team for marking only.

In this phase, you will need to implement the ideas in your proposal and classify the test data. In the marking phase, "F1" of the test data will be used for making. When calculating F1, "1" is counted as positive label, and "0" as negative label. You need to submit

- A *result report* on
  - Test result: the prediction on test data (in integer type) and
  - Evaluation result: the evaluated accuracy and F1 on the training data using cross-validation (in float type).
- A *readme* file with clear and thorough *description of your coding environment* (operation system, programming language and its version, additional packages installed etc.) and instructions on how to run the codes such that your reported test and evaluation results can be reproduced. If you need any additional justification or references for your implemented methods, please also include them in the readme file. The readme file can be in any text format, such as *.docx*, *.pdf*, or *.txt*.
- Your *implemented codes* which include all your training procedures and a main function in a main file (for example, *main.py*) to pre-processing and train on the training data, prediction on the test

data, and generate the submitted *result report sxxxxxxx.csv file*. Codes for the pre-processing and training procedure are required to be submitted. But the hyperparameter tuning or model selection procedure are NOT required to be submitted. You are recommended to include the best hyperparameters, models, pre-processing methods etc. directly in your submitted codes.

Additional requirements

- Please include the provided training and test files into your submitted *.zip* file for reproducing your results in the marking phase. The generated result report *sxxxxxxx.csv* file should be in the root directory.
- Any programming language can be used. However, if you use Python, please submit *.py* files instead of any other formatted files. If you use the Jupyter Notebook or Colab, please submit *.py* file instead of *.ipynb* file.
- Please submit your best prediction according to your cross-validation results. Multiple test results submitted will not be marked

Format

- The result report should be named as *sxxxxxxx.csv (sxxxxxxx is your student username)* with the same **Submission Title** when submitting through the "Report Submission" Turnitin link provided. For example, if your student username is *s1234567*, then the result report should be named as *s1234567.csv* and submitted with the same Submission Title.
- The result report should be composed of 918 rows. For the first 917 rows, the ith row gives the prediction of the ith test instance, either 1 or 0 (in integer type). The last row (row 918) gives the accuracy (first column, rounded to the nearest **3rd decimal place**) and F1 (second column, rounded to the nearest **3rd decimal place**) evaluated by yourself through cross-validation on the training data, both in float type. You could refer to *result_report_example.csv,* which provides an example (NOT groundtruth) of the result report.

Note that result report submitted in other forms or names will not be accepted or marked.

- Together with the result report, you need to submit a *readme* file and all your *codes*.
- The *readme* file and your *codes* should be compressed into **one** zip file named *sxxxxxxx.zip (sxxxxxxx is your student username)* with the same **Submission Title** when submitting through the "Readme and code submission" Turnitin link provided.

Note that code and readme file submitted in other forms or names will not be accepted or marked.

We recommend you follow the Google Style Guides (https://google.github.io/styleguide/) for the programming style. Following such style is not mandatory for this assignment but using it may benefit your future career as a data scientist!

Submission

Only your submitted version will be marked. All required files need to be submitted before due. Otherwise, penalty will be applied according to ECP, i.e.,

10% of the maximum possible mark for the assessment item will be deducted per calendar day (or part thereof), up to a maximum of seven (7) days. After seven days, no marks will be awarded for the item. A day is considered to be a 24-hour block from the assessment item due time. Negative marks will not be awarded.

- Result report should be submitted through the "Report submission" Turnitin link provided on Blackboard -> Assessment -> Project Phase II -> Report submission before the deadline, with the Submission Title *sxxxxxxx.csv*.
- Compressed file of readme and codes should be submitted through the "Readme and code submission" Turnitin link provided on Blackboard -> Assessment -> Project Phase II -> Readme and code submission before the deadline, with the Submission Title *sxxxxxxx.zip*.

Marking standard

Submissions satisfying the following four conditions will be accepted and marked

1. The classifiers used to do classification can be reproduced by the submitted readme file and codes.
2. The classifier is generated by using only techniques delivered in INFS4203/7203 lectures.
3. The test and evaluation results can be reproduced by the submitted readme file and codes.
4. The test and evaluation results are generated by applying the learned classifiers to the data.

When the above four conditions are satisfied, the result report will be marked according to the F1 result on the test data in the following way (rounded to the nearest 1st decimal place)

- For F1 less than or equal to 0.15: Mark = max[(F1-0.05) *80, 0]
- For F1 greater than 0.15 but less than 0.65: Mark = (F1-0.15) * 10 + 8
- For F1 greater than or equal to 0.65 but less than 0.85: Mark = (F1-0.65) * 35 + 13
- For F1 greater than or equal to 0.85: Mark = 20
- Please see the example below

F1	Mark
0.05	0
0.15	8
0.25	9
0.35	10
0.45	11
0.55	12
0.65	13
0.75	16.5
0.85	20

Training time or prediction time will not be counted into marking.

Track 2: Competition-oriented project

In this phase, you need to submit:

- A *result report* of the Public Leader Board results, including a screenshot and an URL of the Public Leader Board.
- A *readme* file with clear and thorough *description of your coding environment* (operation system, hardware requirement, programming language and its version, additional packages installed etc.) and instructions on how to run the code such that your final submission to Kaggle can be reproduced
- Your *implemented codes* including training and test codes which have a main function to generate the final submission to Kaggle.

Your submission will be marked according to the marking standard specified in "Project Specification" released in Week 2.

Format

- The result report should be named as *sxxxxxxx.pdf or sxxxxxxx.doc/docx (sxxxxxxx is your student username)*. For example, if your student username is s1234567, then the result report should be named as *s1234567.pdf/doc/docx*.

Note that result report submitted in other forms or names will not be accepted or marked.

- Together with the report, you need to submit all your *code* and a *readme* file. The *readme* file and your *code* should be compressed into **one** zip file named *sxxxxxxx.zip (sxxxxxxx is your student username)*.

Note that code and readme file submitted in other forms or names will not be accepted or marked.

Submission

Only your submitted version will be marked. All required files need to be submitted before due. Otherwise, penalty will be applied according to ECP.

- Result report should be submitted through the "Report submission" Turnitin link provided at Blackboard -> Assessment -> Project Phase II -> Report submission before the deadline with the Submission Title *sxxxxxxx.pdf or sxxxxxxx.doc/docx*.
- Compressed readme file and code should be submitted through the "Readme and code submission" Turnitin link provided at Blackboard -> Assessment -> Project Phase II -> Readme and code submission before the deadline with the Submission Title *sxxxxxxx.zip*. Note that the zip file should be smaller than *100MB*. **If your file is larger than *100MB*, please contact Zijian Wang (zijian.wang@uq.edu.au) before due time by email in case there is any penalty applied to later submission.**

End of Specification for Phase II