INFS4203/7203 Project

Semester 2, 2021

Due dates:

16:00 on 17th September 2021 for project proposal (Phase 1, 15%)

16:00 on 29th October 2021 for project report (Phase 2, 20%)

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

Overview

The assignment is designed to test the ability to apply data mining techniques to solve real-world problems. This is an individual assignment. The completion of the assignment should be based on your own design.

In this assignment, you will be asked to individually complete a project proposal and implement your proposal to have data mining models which could be applied to test data. You need to choose **either**

- a data-oriented project, or
- a competition-oriented project.

To complete the project, you need to submit a proposal **in the 1st phase** describing clearly and thoroughly the data pre-processing, model training and evaluation techniques you plan to apply, and based on the above proposal **in the 2nd phase** a project implementation and a report on the final test result.

Track 1: Data-oriented project

In this project, you will be provided with a dataset named *Ecoli.csv*. Except for the first row, each row in the data file corresponds to one data point. There are 1500 data points in this dataset, formed by micro-array expression data and functional data of 1500 genes of E. coli, a bacterium that is commonly found in the lower intestine of warm-blooded organisms. The first 103 columns are numerical features describing their expression level. The following 13 columns (from Column 104 to Column 116) are nominal features describing the gene functional. If the gene has a special functional, it will be denoted 1; otherwise, 0. NaN denotes that the feature value is missing at the position. The final column “Target” (Column 117) is the label for the gene indicating whether the gene has the function “Cell communication”. In this column, the positive class is denoted as 1, and the negative class is 0.

Based on the provided labelled data, the overall objective is to design a classifier with good generalization to differentiate whether a given gene has the function “Cell communication”. Note that the test data (without ground truth) for measuring the generalization ability will be released in Week 9.

Phase 1: project proposal (15 marks)

In the first phase of the project, a proposal is needed to be submitted by 16:00, 17th September 2021. The proposal takes 15 marks in total. 12 marks could be earned by describing clearly and comprehensively the following aspects and how to make use of them to achieve the best generalization performance

1. (**3 marks**) Based on your analysis of the dataset, discuss whether the following pre-processing techniques should be considered: outlier detection, normalization, imputation, etc. Describe how to determine appropriate techniques by cross validation, and how to apply them to the current data.
2. (**5 marks**) Based on the above pre-processed data, describe the procedure of applying the four classification techniques learned in lectures (decision tree, random forest, k-nearest neighbor and naïve bayes) to the data, including necessary model selection and hypermeter tuning by cross validation. You also need to consider an ensemble of the classification results from different classifiers at the end of learning.
3. (**3 marks**) Describe the process of evaluating the model given the current dataset using cross validation. Based on your analysis of the data, answer explicitly which metric is the most appropriate for measuring the classification performance of the current dataset.
4. (**1 mark**) Give your timeline for the implementation of your project in the 2nd phase. The timeline should include a justified, comprehensive and feasible list of milestones.

The **final 3 marks** will be given to the presentation of the proposal. We expect the proposal to have good structure, which helps comprehension. The presentation should be neat and professional, with bibliography, which is correctly formatted following the examples in the provided template and appropriately referenced. Marks will be deduced if there are formatting, spelling, grammar, bibliography, referencing or punctuation errors which impact the understanding of the proposal.

Format

The proposal should follow the style of ***Proposal_Template.doc***. The submission should be **within four pages, including all references**.

Submission

The proposal should be submitted

- in PDF or Doc (Docx) format, other formats are not acceptable, and
- through the “Proposal submission” Turnitin link provided at Blackboard -> Assessment -> Project -> Proposal submission before the deadline.

Only your submitted version will be marked. A penalty will be applied to the late submission (see ECP).

Phase 2: Project report (20 marks)

In this phase, you will need to implement the ideas in your proposal and use them to classify the test data which will be provided in Week 9. Details on format, marking standard and submission will be released in Week 9 as well.

Track 2: Competition-oriented project

In this project, you will need to complete a data mining-related online competition and achieve satisfactory test performance. The competition should **end no later than Oct. 1st, 2021** and be related to the learning object of this course. After you have successfully targeted a competition **with a minimum of ten competitors**, please Express of Interest **(EOI) by this link** (or https://forms.gle/BMJcCAXNkivSfg4B7). There are **limited spots** for this project of up to 20 students, determined by the time you submit your EOI and whether the project fits the learning objective of this course.

Phase 1: project proposal (15 marks)

In the first phase of the project, you need to submit a proposal. The proposal takes 15 marks in total. 12 marks could be earned by describing clearly and comprehensively the following aspects and how to make use of them to achieve the best generalization performance

1. (**2 marks**) Describe the task of the competition and the basic statistics of the provided dataset.
2. (**3 marks**) Based on your analysis of the dataset, discuss whether the following pre-processing techniques should be considered: outlier detection, normalization, imputation, etc. Describe how to determine appropriate techniques by cross validation, and how to apply them to the current data.
3. (**5 marks**) Based on the above pre-processed data, describe the procedure of applying the four classification techniques learned in lectures (decision tree, random forest, k-nearest neighbor and naïve bayes) or beyond (SVM, logistic regression, neural networks, boosting etc.) to the data, including necessary model selection and hypermeter tuning by cross validation. You also need to consider an ensemble of the classification results from different classifiers at the end of learning.
4. (**1 mark**) Describe the process of evaluating the model given the current dataset using cross validation.
5. (**1 mark**) Give your timeline for the implementation of your project in the second phase. The timeline should include a justified, comprehensive and feasible list of milestones.

The **final 3 marks** will be given to the presentation of the proposal. We expect the proposal to have good structure, which helps comprehension. The presentation should be neat and professional, with bibliography, which is correctly formatted following the examples in the provided template and appropriately referenced. Marks will be deduced if there are formatting, spelling, grammar, bibliography, referencing or punctuation errors which impact the understanding of the proposal.

Format

The proposal should follow the style of ***Proposal_Template.doc***. The submission should be **within six pages, including all references**.

Submission

The proposal should be submitted

- in PDF or Doc (Docx) format, other formats are not acceptable, and
- through the "Proposal submission" Turnitin link provided at Blackboard -> Assessment -> Project -> Proposal submission before the deadline.

Only your submitted version will be marked. A penalty will be applied to the late submission (see ECP).

Phase 2: Project report (20 marks)

In this phase, you will need to implement the ideas in your proposal and use the implemented models to achieve a good position in the competition's public leading board.

Format and submission details will be released in Week 9.

Marking standard

You need to submit the evidence of your achievements in the public leading board by the end of the project deadline to earn your marks. Your username in the public leading board must be your student username (sxxxxxxx, each x represents a digit).

If your targeted competition ends before the project deadline, you could show by cross validation that you have achieved comparable performance to a particular competitor on the public leading board before the project deadline. Your project could then be assessed by the competitor's corresponding rank percentage on the public leading board.

1. For three entry-level Kaggle competitions

a. Titanic - Machine Learning from Disaster

https://www.kaggle.com/c/titanic

You need to get an accuracy through the Public Leader Board by the project deadline.

Earned marks = min (20 – (1 - PublicLeaderBoardAccuracy) *200, 0)

That is, you earn 20 marks when having Public Leader Board Accuracy 1.0.

b. House Prices - Advanced Regression Techniques

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Note that this project is not related to the learning objective of the course. It is not eligible as a competition-based project.

c. Digit Recognizer

https://www.kaggle.com/c/digit-recognizer

You need to get an accuracy through the Public Leader Board by the project deadline

Earned marks = min (20 – (1.0-PublicLeaderBoardAccuracy) *200, 0)

That is, you earn 20 marks when having Public Leader Board Accuracy 1.0.

2. For other prized competitions

You have to earn a public Leader Board top ranking index (your rank divided by the total number of competitors) by the project deadline

Earned marks = min (20 – max (public_LB_top_ranking_index – 0.5, 0)*40, 0)

That is, you earn 20 marks when having Public Leader Board top ranking to be within 50% of all competitors.

---End---