INFS4203/7203 Project

Semester 2, 2022

Due dates:

16:00 on 16th September 2022 for project proposal (Phase 1, 15%)

16:00 on 28th October 2022 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 the 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 Phase 1** describing clearly and thoroughly the data pre-processing, model training and evaluation techniques you plan to apply. Based on the above proposal, **in Phase 2**, a project implementation and a report on the final test result will be submitted.

Track 1: Data-oriented project

This assignment is designed to mimic the real-world situation when some imperfect data has been collected and data mining techniques should be applied. In a real-world application, many considerations will be around deciding how and which data mining techniques to be applied to the given data to benefit the future prediction in testing phase.

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 3 columns (from Column 104 to Column 106) 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 107) 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) will be released in Week 9**.

Phase 1: project proposal (15 marks)

In the first phase of the project, a proposal should be submitted by 16:00, 16th September 2022. The proposal should tell your overall plan for the project, including the learning process and the timeline for Phase 2. You do not need to submit any codes nor report any training/validation/test results in the Phase 1 proposal. Abstract is not required in the proposal either.

The proposal takes 15 marks in total. 12 marks could be earned by describing clearly and comprehensively the following four aspects. These aspects provide guides on what you have to discuss (as a whole) in the proposal instead of serving as independent questions for you to address. Note that you should ONLY use techniques delivered in INFS4203/7203. Techniques beyond those delivered in INFS4203/7203 are NOT allowed.

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 hyperparameter tuning by cross-validation. You also need to consider an ensemble of the classification results from different classifiers at the end of learning. Discussing and giving the reasonable ranges to perform hyperparameters search are expected.
3. (**3 marks**) Describe the process of evaluating the model given the current dataset using cross-validation. Based on your analysis of the data distribution, 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 Phase 2. The timeline should include a justified, comprehensive and feasible list of milestones. The timeline is a succinct plan showing that the implementation and testing can be submitted on time before the Phase 2 due on 28 Oct.

The **final 3 marks** will be given to the presentation of the proposal. We expect the proposal to have good structure, that 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.

Hints:

1. You do not need to explain the mechanical of how each technique works or how to calculate each metrics unless needed. In the proposal, please focus more on practical aspects such as the criteria to determine which technique(s) to apply.
2. The pre-processing technique should be decided on whether they contribute to the classification. More specifically, the pre-processing and classification techniques are conjugated: one follows closely after the other. To achieve a good performance, you should use the cross-validation technique to select the *best combination* of pre-processing techniques and classification methods. Some of the combinations of algorithms and pre-processing techniques may achieve better performance than other combinations.
3. Consider using both mean and standard deviation to decide whether a result is better than another when using cross-validation.
4. Please bear in mind that ensemble of ensemble (such as combining the output of random forest and k-NN by majority voting) may also achieve a good result.
5. Note that in the testing phase, you usually apply the same pre-processing technique in the training phase.

Format

The proposal should follow the style of ***Proposal_Template.doc***. The submission should be **within four pages, including all references and illustrations (if needed)**. References should be properly provided once necessary, even if you use contents from lecture slides. Non-peer reviewed web sources could be used and should also be properly cited.

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.

You are allowed to submit the proposal multiple times before the due date. Only the last submitted version will be marked. A penalty will be applied to the late submission (see ECP Section 5 for details).

Do not submit codes in Phase 1, even if you have analysed data by programming.

Phase 2: Project report (20 marks)

In this phase, you will implement the ideas in your proposal and use them to classify the test data which will be provided in Week 9. Only techniques delivered in lectures are allowed to be used. Details on format, marking standard and submission will be released in Week 9.

We do not limit the type of programming languages in Phase 2. You could pick any language you want.

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, 2022** and be related to the learning objective 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 10 students, determined by the time you submit EOI and whether the project fits the learning objective of this course.

The following entry-level Kaggle competitions are NOT eligible for the project

1. Titanic - Machine Learning from Disaster: https://www.kaggle.com/c/titanic
2. House Prices - Advanced Regression Techniques: https://www.kaggle.com/c/house-prices-advanced-regression-techniques
3. Digit Recognizer: https://www.kaggle.com/c/digit-recognizer
4. Optiver Realized Volatility Prediction: https://www.kaggle.com/competitions/optiver-realized-volatility-prediction/overview/description

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 the process based on them to achieve the best generalization performance. In this track, you are allowed to use techniques based on your own exploration of the data mining subject beyond the techniques delivered in INFS4203/7203.

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 and illustrations (if needed)**. References should be properly provided once necessary, even if you use contents from lecture slides. Non-peer reviewed web sources could be used and should also be properly cited.

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.

You are allowed to submit multiple times before the due date. Only the last submitted version will be marked. A penalty will be applied to the late submission (see ECP Section 5 for details).

You do not need to submit codes in Phase 1, even if you analyse data by programming.

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.

We do not limit the type of programming languages in Phase 2. You could pick any language you are familiar with.

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.

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 = max (20 – max (public_LB_top_ranking_index – 0.4, 0)*30, 0)

That is, you earn 20 marks when having Public Leader Board top ranking to be within top 40% of all competitors.

---End---