Final Project Specification

Introduction to Data Science (34100393-0)
Fall 2019
Tsinghua University

1 Important Dates (due 23:59)

- **12/03/2019 (Tue. 13th week)** - Proposal Due
- **12/23/2019 (Mon. 16th week)** - Presentation Day
- **01/07/2020 (Tue. 18th week)** - Report & Code Due

2 Introduction and Requirements

A major component of this course is a final project that allows you to investigate some applied data science problem in more detail. While we are very open regarding the topic and type of analysis that you do in the project, we require the following:

(i) The focus of the project should be on analyzing a dataset to ask some insightful questions about the data itself. While you can use algorithms learned from this course to help you answer these questions, the focus of the final project should not be solely on the algorithms themselves, but should **be grounded in some practical questions you want to understand from the data itself**. In other words, you should uncover some nontrivial insights from the data.

(ii) Owing to the above, the project must analyze a **real data set**. You cannot generate a purely synthetic data set for the project and simply call the APIs in Scikit-learn. To be clear, though, you can still collect your data from some computational process ("real" does not mean that it has to be generated by physical, non-computational systems). If you have any doubts about this point and your topic, talk to the instructor/TAs.

3 Proposal

You should submit a proposal at least covering the aspects below, but these contents should be covered in two A4 pages:

- Problem Background
- Evaluation Metrics
- Dataset Specification
- Your Preliminary Technical Solution

4 Presentation

At least one member in each group is required to present your work in class within 5 minutes. Your presentation should at least cover the following contents:

- Problem Background
- Final Technical Solution

  - Data Preprocessing
  - Exploratory Data Analysis
  - Feature Engineering
  - Algorithm Design
  - Performance Evaluation
  - Insights Analysis

5 Report

At the end of the semester, you should submit a report of the project. Apart from expanding the aforementioned points in your presentation, we encourage you to include other parts such as ablation studies of your algorithms and important cues of the particular dataset which inspire you to design powerful features.

Honor Code: please explain which parts of your project are borrowed from third parties.

6 Grading Policy

The final score of the projects will be calculated as

$$\mathbf{Proposal}(5\text{pnts}) + \mathbf{Presentation}(10\text{pnts}) + \mathbf{Code\&Report}(15\text{pnts}) = 30\text{pnts}.$$

7 Hints

Before you start to manipulate your dataset, we suggest you to think about these tips again:

- Make sure to preprocess your data before training your models
- Choose appropriate feature engineering methods
- Choose a simple baseline, e.g. logistic regression
- Consider regularization terms, e.g. L1 and L2
- Consider nonlinear feature transforms
- Compare performance of different learning models
- Pick up hyper-parameters through cross-validation
- Compare different evaluation metrics, e.g. AUC and confusion matrix.