STA314H: Final Project

Start: Sep 26th, 11:59pm Due: Dec 8th, 11:59pm

1 Project description

This is a group project and students are encouraged to form groups of size 1 to 4 on Quercus. The final delivery of the project should be one written report and one final prediction submission from each group (see details in Section 2). Your final score will not depend on group size. Note that it is your responsibility to make sure that you have the correct teammates on Quercus, by Nov 8th, 11:59pm.

After forming the groups, each group needs to choose ONE dataset from the three provided (see Section 4). For the chosen dataset, you should familiarize yourselves with the context and meaning of each measurement. The first thing you and your teammates need to decide is:

- formalize the goal(s) of your study, and think about the motivation.

You are expected to conduct statistial analysis which include, but are not limited to,

- exploratory analysis of the data,
- prediction of the specified target, (you may also predict other targets if you want)
- feature selection,
- statistical inference.

For either data set, prediction of the specified target is the only requirement. However, you may choose to conduct whatever other statistical analyses that you think meaningful and aligned well with the goal(s) of your study. You may refer to Section 3 for the evaluation criterion.

In terms of statistical methods you might use, there are no restrictions, and you are allowed to employ any statistical models or methods, including those not covered in this class. The goal of this project is to give you concrete applications to apply various machine learning algorithms you have learned in this course, as well as those we were unable to cover. You can also use this opportunity to explore more sophisticated algorithms that you and your teammates find interesting and wish to learn. There are no strict rules for this project, so feel free to explore, experiment, and enjoy the learning and application process!

2 Submission

There are two required submissions per team:

1. *Prediction of your chosen data set.* This is submitted via Kaggle (see the link of each data set in Section 4) and you can find detailed instruction of the submission therein. The competition is closed by Dec 6th, at 11:59pm, and you will not be able to update your prediction after that.
2. *Final report.* Your final report should be submitted by Dec 8th, at 11:59pm via Quercus. Note that we DO NOT accept any late report. A score of 0 will be assigned if we do not receive your report on time.

   In terms of format, your report should be in PDF format and must not exceed 8 pages on A4 paper, with a font size of 11, single line spacing, and margins set as follows: top = 2 cm, left = 2 cm, right = 2 cm, bottom = 2.5 cm.

Content of the final report. There are no specific rules regarding the content of your report. However, it should be a complete report; for instance, it needs to address the following aspects:

- Clearly state the problems you studied and explain your motivation.
- Clearly describe the statistical analysis you conducted and explain your results.
- For prediction performance, your report must include your team name on Kaggle, the final prediction accuracy of your model, and your final ranking.
- Statistical analysis can be done in either R or Python. However, all the code used in your data analysis must be included at the end of the report (code does not count toward the 8-page limit).

3 Evaluation and grading policy

The evaluation of the final report will depend on the statistical problems you have addressed. For instance, if you focus solely on prediction, your report will be evaluated based only on the predictive performance of your final model (e.g., your ranking in Kaggle) and the efforts you made to improve it. If you also consider, for example, selecting a subset of features and interpreting your model, your final report will be evaluated based on both prediction and the additional aspects you include.

The final report should be well-written and coherent, containing all necessary elements of a scientific report. For example, a top-ranked predictive model accompanied by a poorly written report will not result in a high grade. Conversely, even if your final model does not rank among the top, a coherent statement of your thought process and an exposition of the efforts you made to improve it can still contribute to achieving a high grade.

Since this is a group project, please make sure you and your chosen teammates are correctly assigned in the same group on Quercus, by Nov 8th, 11:59pm. All group members are expected to contribute equally in the project and the same grade will be assigned to all group members.

4 Datasets

There are three data sets provided below. Each team needs to choose and work on ONE data set.

- *Alzheimer's Disease dataset.* You can access the data set and join this Kaggle competition via this link. You can also find detailed information of this data set therein. Here are a few background on this dataset.

  Alzheimer's disease is a brain disorder that gets worse over time. It's characterized by changes in the brain that lead to deposits of certain proteins. Alzheimer's disease causes the brain to shrink and brain cells to eventually die. Alzheimer's disease is the most common cause of dementia a gradual decline in memory, thinking, behavior and social skills. These changes affect a person's ability to function.

  It is therefore important to detect Alzheimer's disease in an early stage and explore factors associated with Alzheimer's. This dataset contains extensive health information for 2,149 patients, each uniquely identified with IDs ranging from 1 to 2149. The dataset includes demographic details, lifestyle factors, medical history, clinical measurements, cognitive and functional assessments, symptoms, and a diagnosis of Alzheimer's Disease.

- *Dataset of three Cancer Diseases.* You can access the data set and join this Kaggle competition via this link.

  The Cancer Genome Atlas (TCGA), a collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), has generated comprehensive, multi-dimensional maps of the key genomic changes in 33 types of cancer. The TCGA dataset, describing tumor tissue and matched normal tissues from more than 11,000 patients, is publically available and has been used widely by the research community. The data have contributed to more than a thousand studies of cancer by independent researchers and to the TCGA research network publications.

In this data set, we focus on three cancer types: the glioblastoma multiforme (GBM), the lung squamous cell carcinoma (LUSC) and ovarian cancer (OV). The features contain gene expression data using the Affymetrix HT Human GenomeU133a microarray platform by the Broad Institute of MIT and Harvard University cancer genomic characterization center. Data are in log space. Genes are mapped onto the human genome coordinates using UCSC xena HUGO probeMap. Both data consist of 12,043 identifiers. We have 886 samples in total. To be specific, the GBM dataset has 376 samples, LUSC dataset has 90 samples and OV dataset has 420 samples.

- *Detection of YouTube Spam Comments.* You can access the data set and join this Kaggle competition via this link.

  YouTube is one of the most widely used video-sharing platforms globally, featuring millions of uploaded videos and billions of daily views. However, its popularity has also drawn in spammers, who exploit the platform by posting irrelevant or promotional comments that negatively impact the user experience.

  To tackle this issue, we aim to develop a classifier that is capable of automatically detecting and flagging spam comments on YouTube videos. The training data set contains 1369 labelled YouTube comments.

  This data set is related with natural language processing, to which you need to find a way of constructing your features from the input text. There are many representations out there to use, such as, the Bag-of-words representation, word embedding, etc. This blog has a very basic introduction but also provides useful references for more detailed explanations.