COMP 1433: Introduction to Data Analytics

Final Project Specifications

Important Notes (Read before you get started!).

1. The project will account for 45% of the final grade.
2. The project is to be completed in groups, i.e., one group only needs to submit one project report (and one attachment with R scripts if needed).
3. Paper size: A4; font: 12-pt Times New Roman; language: English; margin: 2.5cm (each of the four sides); line spacing: 1.0 (single line). [1]
4. Page limit: at most 8 pages (*no lower bound*).
5. Submission deadline: **23:59 May 31, 2020** (Sunday)
6. Late submissions will have a 30% penalty in the first 12 hours (from 23:59 May 31 to 11:59 AM June 1 (noon)). Submissions on or after 11:59 AM June 1, 2020 (noon) will not be accepted.
7. There three options to be introduced in below and each group can pick up any one of them. No matter what option you go for, a project report is needed to be submitted. For option 2 and 3, in addition to the report, please also submit an attached zipped file with the R scripts used to do data analysis and model implementation (so we can review the contributions from the group members working on programming).
8. The project report should be submitted via *Blackboard* (https://learn.polyu.edu.hk) and the entry is *Assessments/Project/Report*. The attachment (R scripts for option 2 and 3) should be zipped and submit to *Assessments/Project/Attachment.*
9. Plagiarism is not allowed. Any suspected cases will be reported to PolyU and processed according to the university's regulations.

Detailed Requirements.

- ***Option 1: Literary Review***

This option is for students who are not a big fan of programming yet like to learn new things for data analysis. In the literary review project, each group should first pick up a topic related with this course, e.g., *Comparison of Diverse Clustering Models*, *Development of Statistical Models*, *PageRank Model and its Expansions*, etc.

Then, read at least three research papers centring around the topic you select. The papers should have at least 100 citations on Google Scholar and were published in a year after 1980. Here are some suggested conferences and journals where you can find high quality papers:

- **Data Mining**: SIGKDD (https://www.kdd.org/), WWW (https://www.iw3c2.org/), TKDE (https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=69)
- **Information Retrieval**: SIGIR (https://sigir.org/), CIKM (https://dl.acm.org/conference/cikm), TOIS (https://dl.acm.org/journal/tois).
- **Machine Learning**: NIPS[2] (https://nips.cc/), ICML(https://icml.cc/), PAMI (https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=34).

[1] Slightly changing the fonts for titles, footnotes, figure and table captions as well as tuning paragraph spacings to allow better display are acceptable.

[2] Changed to NeurIPS recently.

- **Artificial Intelligence**: AAAI (https://www.aaai.org/), IJCAI (https://www.ijcai.org/), TIST (https://dl.acm.org/journal/tist)

Before start writing the report, you'll have a chance to email the title (topic) and the papers to me and I'll examine whether the papers and the titles are expected. Please do so as early as possible (and by **23:59 May 20, 2020**) to allow sufficient time to change the title or papers if not suitable, such as low quality papers or too challenging topic.

Afterwards, the team work together to write a report that summarize the papers and present a discussion of your thoughts. Your report is suggested to be organized in the following way (you may add other parts to make the story goes more smoothly):

- **Motivation**. The reason why you select the topic, e.g., why the topic is useful and who can be benefited from the task?
- **Background**. The history and development of the related area, e.g., what previous research did on the task. Such information can be found on related work or background study of the research paper.
- **Description.** Introduce the methods proposed by the papers in an easy-to-understand way (which means that more description (with figures) and less formulas). Summarize the contributions made by the papers you read. The goal here is to make sure that people who have not read the papers are able to capture the ideas quickly.
- **Data.** A brief description about what the data is used for experiment in the papers you read.
- **Discussions.** Your thoughts after reading the papers. You can share the inspirations gathered from the papers or the suggestions you propose to further improve the work.
- **Reference.** List of the papers you surveyed and cited.
- **Teamwork.** Describe the role of each team members. A simple example:
  - *Member A: search and find related papers.*
  - *Member B: read paper I.*
  - *Member C: read paper II.*
  - *Member D: read paper III.*
  - *Member E: write the report.*

- ***Option 2: Data Analysis for Titanic Survivals***

This option is for students who are a big fun of programming for data analysis. The topic is inspired by Kaggle Titanic competition (https://www.kaggle.com/c/titanic). Here is the start of the story --- *The sinking of the Titanic is one of the most well-known shipwrecks in history. On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone onboard, resulting in the death of 1,502 out of 2,224 passengers and crew. While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others. For example, in the 1997 movie (shown in* Figure 1*), we have seen that gentlemen gave the chance to survive to ladies and children (as Jack did to Rose at the touching end of the romantic story).*

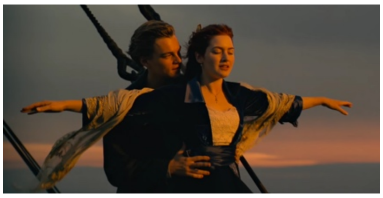

Figure 1 The famous scene in the 1997 Titanic movie.

In this project, you will be able to figure out what sorts of people were more likely to survive. You will be given the information of some passengers in Titanic (in the attached file *train.csv*), such as their genders, ages, socio-economic class, and whether they survived the shipwreck or not, and will have the chance to use the tools (e.g., statistics, regression, graphs) learned in COMP 1433 to analyze the key factors (features) resulting in survivals in the Titanic shipwreck. Afterwards, please learn how to implement Naïve Bayes model (from the reading material *nb.pdf*) and the use the features (genders for example) you consider important to produce a classification model that is able to predict whether other passengers (in *test.csv*) survived the shipwreck. At last, you can submit your results to Kaggle to get the results (Please go to the Kaggle page https://www.kaggle.com/c/titanic to see how to submit the predictions and refer to a baseline prediction example that assumes no one survives in *survive_none.csv* to understand the format for prediction submission). You will be allowed to have 10 submissions per day on Kaggle.

In the report to be submitted, the organization is suggested to be in this way:

- **Motivation**. The reason why we should understand the Titanic data (from the perspectives of history, social science, and so forth).
- **Description.** Describe the methods (or tools) you used to analyse the features and why they can be helpful. Also, give a brief introduction of Naïve Bayes models (in your own words).
- **Implementation.** How you implement the Naïve Bayes models (describing both variables and functions used in the R scripts).
- **Data.** Present the key statistics in the data you are working on (both train.csv and test.csv), such as the average value, range, distributions, etc.
- **Results and Observations.** Show your analysis results (in figures, tables, or numbers) and list the observations drawn from the results. Here please also provide the screenshot of your team profile and the scores you got from Kaggle.
- **Discussions.** Your thoughts and opinions after analysing the data (can be from different perspectives of history, social science, etc.).
- **Teamwork.** Describe the role of each team members. A simple example:
  - *Member A: Process the data and analyse features.*
  - *Member B: Read and learn Naïve Bayes.*
  - *Member C: Implement Naïve Bayes and predict the new data.*
  - *Member D: Write the report.*

In addition to the report, please also submit a zipped file with the R scripts used to analyse the data and implement the Naïve Bayes models. Please make sure that the codes are consistent with the implementation parts in the report and commented well to allow reviewers to capture the key idea.

- ***Option 3: DIY your own task***

This option is for those who want to challenge themselves and are not interested in either option 1 or 2. So, you can have the choice to DIY your own task, including problem definition, data collection, methodology design, results analysis, etc. For option 3, you are required to submit a report covering the following parts:

- **Motivation**. The reason why the task you pick up is important and who can be benefited from the task you tackle.
- **Background**. The history and development of the related area, e.g., what previous research did on the task. Present a short literary review with brief description, which is different from option 1 (where you need a detailed introduction).
- **Description.** First define task you are exploring. Then, describe the methods (or tools) you used to analyse the data and explain why they can be helpful.
- **Implementation.** How you implement the data analysis tools (describing both variables and functions used in the R scripts).
- **Data.** Describe the data you are using for analysis and the way you collect them. Present the key statistics in the data you are working on, such as the average value, value range, distributions, etc.
- **Results and Observations.** Show your analysis results (in figures, tables, or numbers) and list the observations drawn from the results.
- **Discussions.** Your thoughts and opinions after analysing the data (can be from different perspectives such as public health, business, etc.
- **Teamwork.** Describe the role of each team members. A simple example:
    - *Member A: Data Processing*
    - *Member B: Survey the Previous Work*
    - *Member C: Analysis Tool Implementation*
    - *Member D: Result Analysis*
    - *Member E: Write the project report.*

In addition to the report, please also submit a zipped file with the R scripts used to analyse the data. Please make sure that the codes are consistent with the implementation parts in the report and commented well to allow reviewers to capture the key idea.