Instruction for the Final Project

Xin Liu

Due Date: TO BE DETERMINED

Overall Picture

In this final project, you will be using the materials and models we have discussed during the whole semester, to analyze a real life data that you may have to find yourself. The goal of the project is to test whether the students are able to employ the knowledge in the course, to deal with real life problem. You will analyze a real data during the project, and submit your report for marking. Usually, the project will be due immediately on due date that will be announced very soon.

The project will be **individual-oriented**, that is, each student has to complete and submit his or her own report. The data for your project will be selected by yourself. However, you are encouraged to group-discuss your work with others. Detailed planning of the project will be given in the following. The marking system will be attached in the appendix.

Instructions for Data

You are responsible for finding the suitable dataset you will use for your own project. Usually, the more complicated your data set is, a higher score you will potentially get. For example, one will earn more score if he or she plays with a data set of more than 10,00 sample points and more than 10 predicted variables, compared with someone else who only uses a dataset with 100 sample points and 3 predicted variables. Or, a project will be scored higher if it contains more data types (such as continuous variables, characters, factors, etc.).

Usually, you can find data from textbooks or other courses. However, as a bonus, you are encouraged to find real data from some popular online data repository such as **kaggle** and **KEEL**. Also, your project deserves a higher mark if you are using the data related to the following application areas:

- Image
- Transportation
- Environment-related data
- Health care
- Finance and economics

Analysis of the data

You will analyze your data using what we have learnt in this course. Basically, the more complicated your data set is, the more models you may be able to employ, as well as the visualization. As a mimimum of the analyze, you will choose some (or all) of the variables conduct the descriptive statistics, hypothesis test, confidence interval estimation and linear regression (if applicable), presenting results with visualization techniques. The following is a **MUST** for the project

- Descriptive statistics + plot
- Linear regression (if applicable) with regression diagnosis or logistic regression
- Visualization

Writing

- Well organized and clearly written
- If you choose one of the bonus methods, you may have to introduce the method as part of your report, instead of applications only.
- Write within the page limit of 8. Contents beyond 8 pages will not be read and scored.

Bonus Methods

- Polynomial regression and splines
- Support vector machine and support vector regression
- Decision tree and random Forest
- Variable Selection Method combo
- Neural Network and Multiple Perceptron Layer
- Principal compnonent analysis and independent component analysis
- EM algorithm and Bayesion Modeling
- Other deep learning method