IST 707 Data Mining
Course Project Instruction

1. Project Format

The objective of the project is to use the main skills taught in this class to solve a real data mining problem. Students can choose to work individually or pair up with another student.

2. Choose data mining problem and data set

For this project, you must choose your own dataset. It can be one that you created yourself or found from other resources, such as the Kaggle competitions and the UCI repository (http://archive.ics.uci.edu/ml/).

Some rules/tips about choosing data sets:

a. Do not choose the data sets that we have analyzed in class, such as the Titanic data, the zoo data, etc.
b. It should not be a small or made-up dataset. For this semester, "small" is defined as fewer than 50 examples in the dataset.
c. Choose a data set that does not require excessive data preprocessing.

3. Experiment design

Define a problem on the dataset and describe it in terms of its real-world organizational or business application. The complexity level of the problem should be comparable to HW4 or HW7 assignment.

The problem may use one or more of the types of data mining algorithms that we have studied this semester: Classification, Clustering and Association Rules, in an investigation of the solution to the problem.

This investigation must include some aspects of experimental comparison: depending on the problem, you may choose to experiment with different types of algorithms, e.g. different types of classifiers, and some experiments with tuning parameters of the algorithms. Alternatively, if your problem is suitable, you may use more than one of the algorithms (Clustering + Classification, e.g.). If there are a larger number of attributes, you can try some type of feature selection to reduce the number of attributes. You may use summary statistics and visualization techniques to help you explain your findings.

No need to use all of them. Some explanation is needed to justify your choice of algorithms.

4. Project idea proposal

- Submit a one-pager to describe the data mining problem, the data set, and your initial strategies for data analysis 48 hours prior to the Week 7 Live Session ("Naïve Bayes").
- You are encouraged to read each other's proposal to learn from each other.
- It's OK to choose a data set that another student also chooses to use, as long as your work is independent from each other.
- Your proposal will be commented but not graded. Comments will focus on whether the problem modeling is appropriate, whether the project complexity is appropriate (if not, suggestion for adjustment), and whether the initial data analysis strategy is reasonable.

5. Project progress report

Each student should submit a project progress report 48 hours prior to the Week 9 Live Session. You are expected to have finished all experiments and a draft project report. This round of feedback is expected to help you fix major problems in your research design and result analysis. Your report will be commented but not graded.

6. Final project report

To complete this project, write a final report that conforms to general academic paper format. Make sure to cite relevant work appropriately. Your report should be within 8 pages, 1 inch margin on all sides, and at least 12 point Arial or Times New Roman. Submit your final report 48 hours after the Week 10 Live Session.