ISyE 7406 — Data Mining & Statistical Learning

Yajun Mei (ymei@isye.gatech.edu)

Team Project

For the project, you are encouraged to work in a team of $2-5$ students, but it will be fine if you plan to work alone. **If you decide to work in a team, you will need to submit only one report per group.** You are encouraged to choose a project related to your own research interests, and please feel free to briefly discuss your project with the instructor via piazza if you want.

The Project Proposal (3 points) is due at Canvas (One submission per team). Also see the following page of this pdf file for possible datasets of your project.

Please make sure that all teammates sign up to the same group at Canvas (through "People" and "7406 Project Group"), and you need to sign up at Canvas even if you are working alone. Please do not create your own group or change the group names, so that it is easier for TAs and instructor to manage group grading.

The purpose of the proposal is to get you started, and also allows the TAs and other students to provide feedback to your project. It shall **be** $1 \sim 2$ **pages**. You will need to provide the following information:

1. Your name(s)
2. Project description
3. How and where you obtained the data. For the data set, you can just direct to a website where we can find them.
4. Scientific Research questions you may want to address and corresponding data mining & statistical learning methods

One submission per team, please! I.e., please designate one team member to submit ALL materials on the behalf of your team (since otherwise the Canvas system might be confused and treat your team's multiple submissions as different teams). Everyone on your team will receive full credits on the proposal as long as your team's proposal provides all these information. For the data set, you can just direct us to a website where we can find them, or provide some high-level statement to make sure that you will have access to the dataset.

Peer review comments: when you are assigned to grade proposals, please provide comments to the proposal, (e.g., on problem formulation whether the project sounds interesting, on dataset whether the dataset can help answer the questions, on the proposed methods, etc.). It is crucial to provide construct feedback comments. Note that all teams will receive full credits (3 points) on the project proposal as long as the team provide all these information.

Possible Topics of Your Project

The objective of a class project is to help you gain experience with research, and to relate what you learn to real life problems which may require you learn new techniques (or develop new methods by yourself). You are expected to present the project findings during the class and submit a summary report at the end of the semester. Below are the two types of possible projects (you only need to choose one of them).

1. **Solving a real life data mining problem.** A typical report includes problem formulation, data analysis, proposed solutions, and interpretation of results. The data set can be from your own research or the public domain, see the information below. As an example, you can choose to participate a data mining competition such as the Knowledge Discovery and Data Mining (KDD) cup, see the link below for the past KDD Cup <http://www.kdd.org/kdd-cup>, or the KDD CUP 2017, <http://www.kdd.org/kdd2017/>. Another example is "2017 Data Challenge" sponsored by the Government Statistics Section of the American Statistician Associations (ASA) that analyzes the Consumer Expenditure Survey (CE) data on the Bureau of Labor Statistics website, see <http://magazine.amstat.org/blog/2017/01/01/data-challenge-on-tap-for-jsm2017> for the announcement and <https://www.bls.gov/cex/pumd.htm> for the datasets.
2. **Numerical study of data mining methods using well-known data sets in the literature.** Note that when dealing with well-known data sets, your approach needs to be substantially different from the literature, i.e., you should do more than repeating the analysis there. Some examples are

   - Compare performance of competitive data mining techniques;
   - Ask different questions or investigate new ideas of data mining methods;
   - Identify optimal parameters of specific data mining techniques;

Note that the crucial aspect of your project is to **analyze some data sets and justify your conclusions,** not using some specific statistical models or methods we discussed in class.

Datasets: You can collect the data by yourself, use the data set from your own research or the public domain. One way to find online datasets is to use the search engineer such as google. The followings are some examples of online datasets (you can use google or other search engineer to find more):

1. http://kdd.ics.uci.edu/ or http://archive.ics.uci.edu/ml/
   One example is the KDD cup 1999 data at http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
   More KDD cup data can be found at http://www.kdd.org/kdd-cup
2. http://www.quandl.com/ (financial and economic time-series datasets)
3. Data sets from some government websites such as <http://www.cdc.gov/surveillancepractice/data.html> or <http://www.ngdc.noaa.gov/stp/satellite/goes/dataaccess.html>.
4. http://lib.stat.cmu.edu/DASL/
5. http://www.kdnuggets.com/datasets/index.html (links to more data repositories.)
6. http://www.dmoz.org/Computers/Artificial_Intelligence/Machine_Learning/Datasets/

To inspire your projects, some concrete examples can be as follows:

- analyze some data sets in some competitions, see the links < http://www.kaggle.com/competitions>
- find the traffic or crash pattern near Georgia Tech or your appartment/home by using data from <http://www.dot.ga.gov/DS/Data>

- predict Allergy season by using Atlanta Pollen count data from <http://www.atlantaallergy.com/PollenCount.aspx> .
- derive the relationship between sleep and selected health risk behaviors, see the paper <http://www.cdc.gov/nchs/data/hestat/sleep04-06/sleep04-06.pdf>

To further motivate your projects and encourage you to write up a solid project report, try to think that you want to publish your project report as a paper. There are two possible kinds of data mining or statistical learning papers (you only need to choose one).

- **Application Papers:** apply standard methods to analyze some datasets, thereby answering some important questions in real-world applications such as bioinformatics, economic, finance, banking, healthcare, online advertisements, manufacturing, music, natural disasters, social networks, (bio)surveillance, warehouse, logistics, etc.
- **Methodology Papers:** develop new methodologies and demonstrate their advantages as compared to the standard methods when analyzing some data sets, say, in the context of temporal data mining, spatial data mining, spatio-temporal, streaming data mining, web or graphic mining, etc.