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.

Grading: The course project will be peer graded, and will have a weight of 25% = 3% (proposal) + 10% (presentation slides) + 12% (report), in your final course grade. Be aware of the following requirements and deadlines:

1. **The Project Proposal** (3 points) is due at Canvas during week #9 (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 manager 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:

   (a) Your name(s)

   (b) Project description

   (c) How and where you obtained the data. For the data set, you can just direct to a website where we can find them.

   (d) Scientific Research questions you may want to address and corresponding data mining & statistical learning methods

   Peer review comments: 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.).

   Note that all teams will receive full credits (3 points) on the project proposal as long as the team provide all these information.

2. The **final presentation file** (10 points) of your team project is due at Canvas during week #13 (either pptx or pdf version will be fine). One submission per team.

   (a) There are no specific guidelines on the presentations, and the commonsense applies, e.g., write all team members' names somewhere on the first slide, highlight your problem, data set, main ideas/methods, and conclusions.

   (b) There are no official guidelines how many slides your group might include. To give you a rough estimate, ideally you or your team should prepare for the slides so that each teammate member can present about 3 minutes with a standard deviation of 0.5 minutes, e.g., a team of 2 students will prepare for slides for about 6 ± 1 minutes presentation, and a team of 3 students should prepare for slides for 9 ± 1.5-minutes presentation, etc. Hopefully this gives you rough guidelines how to prepare for your slides.

   (c) (Optional, not Required, no credits): The team is encouraged to submit a recorded video/voice oral presentation (i.e., each student member gives a 3-minute oral presentation). We understand that it might be difficult to generate such video/voice presentation, and thus **this is optional, not required.**

Peer review comments: please write constructive comments to the presentation slides, e.g., whether it is easy to understand the presentation, whether the presentation is interesting, whether the methodology or main conclusions are reasonable, etc.

The TAs will assign a grade based on their own reading,

- 10 points (=100%,A+) if the presentation file is clear, the project sounds interesting, and the conclusion sounds reasonable, etc.
- 9 points (=90%, A) if there are some minor concerns on the presentation file
- 8 points (=80%, B) if there are some major concerns on the presentation file
- 7 points (=70%, C) if the presentation file contains some critical technique errors or has poor presentation
- 6 points (=60%, D) if the presentation file is not understandable or sloppy
- 0 points if no submission.

3. The **final written report** (12 points) of your team project is due at Canvas during week #14 (One submission per team). Either word or pdf file is fine. See the page #6 of the pdf file on some suggestions on the writeup of your report.

   (a) In your writeups, we expect clear explanations of models chosen, hypotheses tested, and findings analogous to what you would produce for a consulting project.

   (b) **Mandatory subsection in the final written report: the lessons you learned** (you can use any names for this subsection). For the purpose of this class, at the end of conclusion section of your final report, please **add a subsection for lessons you learned** from this project or this course. You can also write any suggestions to the instructor. The instructor/TAs will read this subsection, so that we can improve our teaching in the future.

   Peer Review Comments: please feel free to provide comments on the team's selecting and adhering to a logical and readable format for the report; on the appropriate use of whatever data mining technique the team uses; on the appropriateness in the conclusions of the report; and on the readability and understandability of the report when technical material is needed.

   The TAs will assign one of the following grades based on their own independent reading:

   - 12 points (=100%,A+) if you think this is an outstanding or excellent project, e.g., one that deserves possible publication
   - 10.8 points (=90%, A) if you have some minor concerns on the project or report (e.g., on either presentation or technique aspects)
   - 9.6 points (=80%, B) if you have some major concerns on the project or report
   - 8.4 points (=70%, C) if you think the project contains some critical technique errors or the report has poor presentation
   - 7.2 points (=60%, D) if you think the project or report is not understandable
   - 0 points if no submission
   - The Instructor/TA keeps the right to deduct 2 points if we find out that the team miss the mandatory subsection on the lessons learned.

   As always, if you or your team has a concern about peer grading, please feel free to let the instructor/TA know asap at piazza: we will double check to make a final decision, although please do understand that ultimately the grade on the final written project will be subjective.

4. **Peer evaluation form:** if there are two or more students on a team, each teammate should also independently submit the completed **peer evaluation** form at Canvas (with the same due date of the final written report). If you conduct the project by yourself without team, this is optional and not required.

   This peer evaluation is to discourage free ride, and allows the Instructor to adjust an individual student's score based on the teammates' peer evaluations if needed. In general, the ideal is for all team members to receive the same grade on the final project. However, individual deductions from the team's final project grade will be assessed for failing to contribute a fair and significant share to the team's project, as determined by the teammates' peer evaluation and the instructor.

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.

ISyE 7406 — Data Mining & Statistical Learning

Yajun Mei (ymei@isye.gatech.edu)

The final written report shall **not be longer than** 25 **pages**, and the main body of the report is generally $5 \sim 12$ pages. Only very relevant plots and tables shall be included in the body of the report, and the rest should go to Appendix. When writing up your summary report, it is useful to ask yourself the following questions: What is the work? Why is it important? What background is needed? How will the work be presented?

Here is a suggested format for your summary report.

1. **Title Page**: Project Title, author(s) (your name, the last three digits of your student ID, and email address), the submission date, course name/number;
2. **Abstract:** informative summary of the whole report (100-300 words).
3. **Introduction** includes problem description and motivation, data mining challenge(s), problem solving strategies, accomplished learning from the applications and outline of the report.
4. **Problem Statement or Data Sources**: cite the data sources, and provide a simple presentation of data to help readers understand the problem or challenge(s).
5. **Proposed Methodology**: explain (and justify) your proposed data mining strategies.
6. **Analysis and Results**: present *key findings* when executing the proposed data mining methods. For the benefit of readability, detailed results should be placed in the Appendix. Reference of computer softwares to implement your proposed data mining methods (even it is a web page) should be given.
7. **Conclusions:** Draw conclusions from your data mining practice. Unfinished or possible future work could be included (with proper explanation or justification).
   *__A Mandatory Subsection of "Lessons we have learned":__ *at the end of conclusion section, please add a subsection for lessons you or your team learned from this project or this course. Please feel free to write any comments/suggestions/remarks, or share your experiences of data mining.*
8. **Appendix:** This section only includes needed documents to support the presentation in the report. Feel free to divide it into several subsections if necessary. Do NOT dump all computer outputs unorganized here.
9. Bibliography and Credits.

Parts 3-6 constitute the main body of the paper for your primary audience. Usually, as with fictional boss in this example, your audience is intelligent but unschooled in Data Mining or Statistics. So these parts should have as little technical material as you can possibly get away with.

It is appropriate, and even recommended, to refer the reader to the appendix in part 8 if you need to provide a more technical explanation for something. Part 8 is your secondary audience - me - and should follow closely enough the "story" of parts $4 - 6$ that it is easy for me to see what technical material backs up with results and discussion.

It is not necessary to number these parts 1-9 or name them as-above-mentioned. Please feel free to merge some parts or provide more informative section names if it seems natural to do so.

A good on-line resource for writing reports is http://www.ccp.rpi.edu/. This site has links to writing centers at universities around the country, many of which in turn have pages that describe how to put together different types of reports.