MSCI446 Project Information

The goal of the term project is to apply machine learning techniques to a real-world problem of your choosing. Your project may show how to predict something given a set of data that you will collect, or how to discover previously unknown structure or patterns in your data (or both!). The project will be done in groups of up to four students and will be worth 30 percent of the course grade. The deliverables and deadlines are as follows:

Proposal (5%) – electronic copy due by 2pm on Thursday, September 27

Please use the project proposal template (see course website). Marking scheme: 2 points for presentation, organization and clarity of writing, 4 points for novelty and feasibility (problem selection, chance of success), 4 points for technical content (understanding of what machine learning is and what it can do, literature survey).

Status Update (2%) – electronic copy due by 2pm on Tuesday, November 13

Please submit a document summarizing your progress. Please discuss your data set (how much data you have collected) and describe at least one result you have obtained. Was it surprising? Have you done or will you do any deeper analysis to understand this result? Also, please include an appendix with all the source code you used to pre-process and analyze the data. Remember to comment your code!

Project Presentation (3%) – In class, Tuesday, November 27 and Thursday, November 29

Each group will give a five-minute oral presentation in class. Please include the following information: Data, Hypotheses (what were your expectations before you started analyzing the data) and Results (choose 1-2 most interesting or surprising results to present).

Project Write-up (20%) – electronic copy due by 2pm on Tuesday, November 27

Marking scheme: 15 points for technical content, 5 points for presentation, organization and clarity. Here is a suggested outline:

Abstract: briefly summarize your project and the main findings

Introduction: Explain the purpose of your project and highlight the main hypotheses and results. Argue why your project is interesting and non-trivial.

Related work: Give an overview of prior scholarly work on your topic. Use Google scholar, IEEE Explore, ACM Digital Library, magazine articles, blogs written by reputable authors, etc. Has anyone worked with similar data before? How is your analysis different? Have you discovered any new patterns or insights in the data?

Data: Explain how you collected the data. Explain any data cleaning or data transformation that you had to do. Explain how you dealt with missing data, if applicable. Present some statistics about the data using tables and/or graphs.

Results: Every group must do at least one classification or prediction task. If you are doing classification, compare several algorithms and feature sets. If you are doing prediction, you will probably only use linear regression, but you can try different features. Additionally, every group must do one of association or clustering (preferably both, if you can). If you are doing association rule mining, try different support and confidence thresholds. For association, you can use all the features that you used for classification plus the class variable. For clustering, please remove the class variable and only use the remaining features. If you want to visualize your clusters using 3-d plots but have more than 3 feature variables, pick up to 3 features at a time, and experiment with different subsets of 3 features. Use one section per task/experiment, and clearly state the purpose of each experiment, your hypothesis and your findings. Explain every result and emphasize surprising and unexpected results.

Conclusions: Summarize the main findings. Discuss lessons learned (specific lessons from the data and general lessons about data mining) and discuss who can benefit from the results of your analysis and how.

Appendix: Include all your source code and make sure it is well commented.

Report Length: there is no page limit but please be as concise as possible.

Project Proposal Tips

You will be graded on technical content, novelty/originality/creativity, and writing quality. Please proofread your work, define all non-obvious terms and provide citations to any facts or numbers you mention. Stick to facts and avoid opinions.

Clearly explain the business problem or scientific hypothesis you want to solve using data mining.

Your problem must be complex enough to require a data mining solution. There must be a variable whose value is unknown and cannot be easily guessed. For example, it's not easy to guess what engineering program is right for a prospective student!

If you can solve your problem using basic SQL, it's probably not complex enough. For example, suppose you wanted to recommend a restaurant based on location, cuisine and price. If someone tells you they want an inexpensive (say under $20) Italian restaurant in Waterloo, you could simply return all Italian restaurants in your database whose average dinner price is <$20. You don't need a model to learn which restaurants in Waterloo serve inexpensive Italian food; you can create a database with all the information you need.

You may collect data through surveys, obtain data from a company or organization, or download data from the Web. However, if you want to use data available online, consider merging several datasets to come up with your own unique dataset, and make sure you can formulate an interesting business problem or scientific hypothesis. Avoid copying datasets and problems from online data mining competitions.

If you already have the data, or know what it looks like, describe it in as much detail as possible: number of rows, number of columns, name/data type/semantic meaning of each column, semantic meaning of each row (i.e., one row = one event, or one purchase transaction, or one response to a survey, or one loan application, etc.). You don't have to include graphs or illustrations in the project proposal (but you do in the final report).

If you plan to collect data through surveys, describe the variables you want to collect, and, for at least 1-2 variables, explain what questions you will ask to obtain these variables. You don't have to include the complete survey with your project proposal. Try not to make the questions too obvious! Also, explain who will fill out your survey: students, your facebook friends, random people at the university plaza?

Please be as clear as possible about your variables and why they make sense. For example: for the engineering quiz, we wanted to predict the best academic program (which is a discrete variable with 12 values corresponding to the 12 engineering programs offered at UW) using explanatory variables corresponding to a student's personality and interests. We believe that different engineering programs are suitable for different types of students and therefore students' personality and interests should be highly correlated with their program of choice. To obtain these explanatory variables, we designed 10 questions about where a student would like to work, what they would like to have invented, who they would like to work with, whether they would like to be a specialist or a multi-disciplinary generalist, etc. Of course, the survey also asked about the academic program (and whether the student is happy with their choice) - the training dataset must include the class/dependent variable, not just the independent/explanatory variables.

How many explanatory variables do you need for a good model? It depends on the application but I would say at least 10 or so.

You may not be able to collect all relevant explanatory variables, or your dataset may be missing some important information. This is OK as long as you identify what is/might be missing and argue that you have enough variables for a reasonable model.

For prior work, try Google Scholar or the IEEE Explore library to find at least 10 scientific publications related to your topic. Additionally, you may cite blogs (only if written by reputable authors) and technical reports if they contain relevant material.

Examples of Successful Data Mining Projects

- Analyzing the mental health of undergraduate Engineering students: http://www.educationaldatamining.org/EDM2013/papers/rn_paper_34.pdf
- Mining electric vehicle opinions: http://tommyjcarpenter.com/papers/2014/eenergy_sentiment.pdf
- Analyzing satisfaction with co-operative education: https://www.ijwil.org/files/APJCE_16_4_225_240.pdf
- Data mining of undergraduate course evaluations: https://www.mii.lt/informatics_in_education/pdf/infedu.2016.05.pdf
- Predicting peak electricity demand: http://www.engineering.uwaterloo.ca/~lgolab/jiang_6page.pdf
- Analyzing the impact of time-of-use electricity pricing: https://ece.uwaterloo.ca/%7Ecath/miller2017.pdf
- Graph mining to characterize competition for employment: http://www.engineering.uwaterloo.ca/~lgolab/toulis_nda17.pdf
- Impact of entrepreneurship on co-op job creation: https://www.ijwil.org/files/IJWIL_19_1_51_68.pdf
- Job description mining: http://educationaldatamining.org/files/conferences/EDM2018/papers/EDM2018_paper_42.pdf
- Gender differences in engineering applicants: http://educationaldatamining.org/files/conferences/EDM2018/papers/EDM2018_paper_43.pdf