Fundamentals of Data Mining and Information Management

Course Project Instruction

1. Team structure

You can choose to work either individually or collaboratively as a team (with up to four members). Your instructor will post the deadline for submitting your team member information on the course wall. Let the instructor know if you want to work in a team but can't find teammates; the instructor will place you on a team if possible. Toward the end of the semester the instructor will provide an assessment form for team members to evaluate other teammates' contribution to the development of the final project deliverables.

2. Choose a data mining problem and dataset

For this project, you must choose your own dataset. It can be one found from an online source, one of your own, or one of the ones from the UCI repository (http://archive.ics.uci.edu/ml/).

Some rules/tips about choosing datasets:

a. Do not choose datasets that we have already analyzed in class.
b. It should not be a small or made-up dataset. For this semester, "small" is defined as fewer than 500 examples in the dataset.
c. Choose a dataset 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 at least comparable to one homework assignment.

The problem may use at least two different types of data mining algorithms that we have studied this semester such as classification, clustering, and association rules, in an investigation of the analytics 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, such as different types of classifiers, and some experiments with tuning parameters of the algorithms. Alternatively, if your problem is suitable, you may use multiple algorithms (clustering and classification, for example). 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.

4. Project proposal

24 hours before your Week 5 Live Session, you should submit a brief project proposal (less than one page) describing the dataset you plan to use and, tentatively, what type(s) of data preparation and algorithms (which need not be final as you could add new methodologies you learn later in the semester) you would like to investigate. Please provide brief justifications regarding your choices of the data mining algorithms for the problem set. Post your project proposal to the corresponding discussion forum in Blackboard. If multiple teams choose the same dataset, the first team posting the proposal will have the pick, and other teams will need to choose a different dataset. Therefore post your proposal early. You are encouraged to read other's project proposals to learn from each other. Your proposal might be commented on but not graded.

5. Project presentation

Each team should prepare a PowerPoint file with up to 10 slides to present the data mining problem, the analytical approach, and major findings in your project. Your presentation should be posted to the **LMS 24 hours prior to your Week 10 Live Session** so that your instructor can comment on it. You will have a few more days to incorporate the comments into your final project report, **due 24 hours prior to your Week 11 Live Session.**

6. Final project report

To complete this project, write a final report that conforms to general research paper format. See Pang, Lee, and Vaithyanathan (2002) as an example. Your report should be no more than eight pages, with one-inch margin on all sides, and at least 12-point Arial or Times New Roman. Remember that your project report serves as the tour guide for your readers to be able to repeat your journey and discover the same patterns that you did. It is very important to cite and paraphrase relevant work appropriately.

Submit your final project report to the LMS **24 hours prior to your Week 10 Live Session**. This is the final report that will be graded.

You will also submit a peer evaluation form at this time.

References:

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of EMNLP 2002*, 79–86.