Data Mining Project Guidelines

This document provides some guidelines for writing your project proposal and then your final paper. Note that the project is a significant portion of your grade, so you are expected to devote a reasonable amount of time to it and to the write-up. It is difficult to quantify precisely the total amount of time you should spend on the project and write-up—but 5 hours might be far too few.

Project Proposal

Your project proposal must be typed and should be approximately 1 page long, single spaced. The purpose of the proposal is to make sure that you are on the right track and to give me enough information so that I can give you useful feedback. Proposal can be either in Hebrew or English (while the later option is preferred).

In your proposal you should cover the following items:

- Preliminary title and list of students working on the project (1 or 2 students per project without prior permission from me; 3 or more is *possible* for very ambitious projects – requires prior permission).
- Abstract: Similar to the abstract that will ultimately appear in your paper. It should be one paragraph long, for now perhaps only 5-10 lines. It should provide a high level summary of your project and outline your main goals.
- Brief description of what you plan to do.
    - What data sets do you plan to use? If you must do significant work to get the data or convert it into the proper format, then describe the process and approximate effort required.
    - What learning tools do you plan to use (e.g., scikit, statsmodels, weka, other) and what methods (e.g., clustering, classification, regression, what methods in each, etc.)?
    - How do you formulate the problem as a data mining problem (e.g., is it classification, association rule mining, etc.)? What *exactly* are you trying to predict (for prediction tasks) and how will you evaluate your results. How will you know if your results are good? What can you compare them to? It is critical that your problem is **well-defined**.

Types of Projects

There are two main types of projects. You can decide to do a research project, where you look at a research issue. This could be original research, but could also be something straightforward—such as an empirical evaluation of data mining methods or strategies for improving performance (e.g., a study about strategies for removing missing values). However, many of you will wind up examining real-world data sets. This is an application-based project. Ideally you should try to do something a bit interesting, like studying a data set that has not been thoroughly evaluated, or using a different approach. You should make sure that your analysis is not trivial. For example, running a data set through *sklearn* and spending an hour on the analysis and then doing a quick write-up would be considered trivial. You should study the dataset, determine the issues, address any preprocessing issues, try multiple modeling techniques, and perhaps take some creative steps to try to improve the predictive performance.

Full Project Writeup

The actual write-up of your project paper should be *roughly* 3-5 pages, single spaced. I suspect that not everyone will wind up doing a presentation in the class of their project, but that will be determined later in the semester. I do expect that a few students will be able to present their results during the last class, so if you want to volunteer—and help your grade—talk to me as we approach the end of classes. In any case, an in-person presentation is required after submission (Can be to me only). The paper need not be organized exactly as described below, but this should be taken as a reasonable template.

- Abstract: summarizes the paper and the goals of the work (required)
- Introduction: Introduces the project and what you are trying to do. May include some background.
- Background: Depending on the project, you may want a separate background section, depending on how much background you want to include. For example, it may provide domain information for the domain that you are studying.
- Experiments: Describes the experiments and the experimental setup. Will describe the data sets, the evaluation metrics, the data mining tools used, and any other details related to the experiments.
- Results: Includes the experiment results (which are typically not included in the experiments section). A discussion of the results may be included, or they could be included in a separate discussion section, which follows the results. For now, we will assume no separate discussion section.
- Conclusion: Provide your conclusion. For example, comment on the quality of your results. You may also want to include some material on future work, whether or not you intend to do such work. These suggestions could be followed up by someone else.