Prof. Sergio A. Alvarez
St. Mary's Hall, room S255
Computer Science Department
Boston College
Chestnut Hill, MA 02467 USA

http://www.cs.bc.edu/~alvarez/
alvarez@cs.bc.edu
voice: (617) 552-4333
fax: (617) 552-6790

CSCI 3346, Data Mining
Fall 2018
Preliminary proposal for term project

Scope

As described in class, an important component of the course is a project to be developed over much of the semester. Working in groups of three or four persons is encouraged. The scope of the project is flexible, subject to the requirement that data mining concepts and / or techniques must play an important role in it. Options range from coding-intensive projects that aim to compare different implementations of a particular data mining approach (e.g., training predictive neural networks), to theoretical projects that explore fundamental performance limits (e.g., minimum attainable prediction error for a particular data model), to applied data mining projects that seek good results on a given problem (e.g., diagnosis of leukemia, recommendation of music), to projects that explore social and ethical aspects of data mining (e.g., privacy-preserving data mining techniques).

Choosing a topic

Start from the ideas that you mentioned in PS1, and refer to my feedback (if you haven't gotten feedback, please remind me, and I'll be happy to comment). If you are uncertain about a topic area in mind, consider starting by browsing publicly available dataset collections such as the Machine Learning Repository at the University of California - Irvine. See the links on the CSCI 3346 web page. You may find a dataset or two that seem appropriate, or at least a particular application area or "flavor" may suggest itself. Feel free to run ideas by me if you'd like. Once you have an idea of what you'd like to do, proceed with the preparation of the preliminary proposal as described below.

Directions follow on the next page.

Directions for preliminary proposal

Please submit only one proposal per group. This should be a typed document, preferably a PDF file, that includes the following information about your proposed term project.

1. Names of group members (3–4 persons).

2. Tentative title.

3. Main aims of the project. State whether the main task will involve classification, regression, or clustering (or other, but explain).

4. Specific source(s) of data for the project. The dataset is a crucial component that constrains what tasks you would be able to address.

5. Data mining techniques that you are considering using. The choice should relate to the specific dataset and target task.

6. Thoughts about how you might evaluate how well the aims were accomplished. These should be as specific as possible. For example, suppose that your project aims to develop and code a data mining technique that recognizes handwritten characters from images. A concrete way of assessing performance here would be the percentage of input characters that are labeled incorrectly by your program. If competing methods in published papers label 10% of a standard dataset of characters incorrectly, you might use that value as a benchmark to be matched or improved upon. If the aim is instead to code a known data mining technique efficiently, the relevant performance measure could relate to running time on a standard platform. These are just examples. Be creative, and specific.

7. Anticipated challenges (e.g., "many values are known to be missing from this dataset, so dealing with missing values will be an important pre-processing step", or "the number of attributes is very high, and attribute selection or feature extraction will be needed", or "the theoretical framework is well understood, but it will be necessary to estimate the probabilities precisely").

8. A list of references that are relevant to the proposed project. These could be books, papers, or web pages. Provide full publication information for each (authors, title, venue, volume and issue number, date, pages, publisher). Include a few words about *what* you hope to extract from each reference (e.g., "this book should be helpful in understanding the XYZ implementation of the ZigZag algorithm", or "this paper describes factors that are important in interpreting customer satisfaction surveys".)