**Final Project**

**Important Dates:**

Project proposal due by April 26.  
Final project paper due May 2.

**Overview**

The final project will give you the opportunity to use the techniques covered in the course to:

* prepare and analyze a collection of data that interests you
* draw conclusions based on your analysis
* present your results

**Requirements**

Your final project must include the application of three or more data-mining techniques to a dataset that you choose. In addition, you must write a clear and compelling presentation of the results that you obtain, both from the data mining and any other analysis that you perform.

**Choosing a Dataset and a Problem to Solve**

You should begin by selecting a dataset to analyze. ***Each student in the class must choose a different dataset.*** Possible sources of datasets include:

* the [datasets](http://www.google.com/publicdata/directory) associated with Google's Public Data Explorer
* [FedStats](http://www.fedstats.gov/), which includes a large number of datasets compiled by federal agencies
* NASA's [National Space Science Data Center](http://nssdc.gsfc.nasa.gov/)
* datasets available from the [U.S. Census Bureau](http://www.census.gov/)
* [UNdata](http://www.census.gov/), a collection of databases compiled by the United Nations
* [The UCI datasets](http://archive.ics.uci.edu/ml/index.php)
* [The WEKA datasets](http://www.cs.waikato.ac.nz/ml/weka/index_datasets.html)
* [The Journal of Statistics Education data archive](http://www.amstat.org/publications/jse/jse_data_archive.htm)
* a dataset that you find using a search engine. Try a search using the keywords "dataset" and whatever subject you are interested in (e.g., "population dataset").
* a dataset of your own creation

You are welcome to choose any dataset that interests you and that has enough data to enable a meaningful analysis.

In making your choice, you should be sure consider what problem or problems you would be able to solve by employing data mining on the dataset. In other words, you should ask yourself: How could I use data mining to answer one or more questions about this dataset?

The problems/questions will typically be of two basic types:

1. classification/estimation: Is it possible to determine, predict, or estimate some attribute in the dataset, based on the values of other attributes? Medical diagnosis and credit-card application ratings that we have covered in lecture are examples of this type of problem. If you choose a problem of this type, you will employ some type of classification learning and/or numeric estimation.
2. finding associations: Are there non-obvious associations or relationships between attributes in the dataset? Market-basket analysis (e.g., finding products that customers tend to purchase together) is one example of this type of problem. If you choose a problem of this type, you will employ some type of association learning.

**Important notes:**

You will use the Weka data-mining software to perform the actual data mining. Don't worry if you are uncertain at this point about the particular algorithms that you will use. All that matters for now is that you have a sense of the general data-mining approach that you will employ.

It will almost certainly be necessary to transform the dataset in some way before performing data mining. You should consider what transformations will be needed on the datasets that you are considering.

**Writing a Project Proposal**

Before beginning work on your project, you will need to submit a brief proposal outlining what you intend to do. This will allow us to make sure that you are on the right track, and to give you some initial guidance. Your proposal should include the following:

* a description of the dataset that you will be analyzing, including information about where it can be obtained (e.g., a URL). Include in your description a list of the key attributes that are present in the dataset.
* a description of the problem that you are intending to solve -- i.e., the question or questions that you hope to answer using data mining.
* the type(s) of data mining that you intend to perform (classification learning, numeric estimation, or association learning), and an explanation of your choice of approach(es). We highly recommend that you chose either classification learning or numeric estimation as one of your approaches.
* a description of any transformations that you will need to perform on the dataset before you perform data mining. If no transformations are needed, you should briefly explain why the current format of the dataset is amenable to the type of data mining that you will perform.

Your project proposal should be submitted to me by email.

**Splitting Your Dataset**

Your project will probably include an application of data mining to learn a model that predicts the value of some output or class variable. In order to validate the model or models that you produce, you should separate your dataset into two files. One file -- containing N percent of the examples (where N is typically somewhere between 70 and 90) -- should be used for training, and the remaining 100 - N percent should be used for testing. The testing examples should not be touched until you have developed a model using the training examples and are ready to test it. Note that you should save your dataset as an ARFF file before splitting it.

**Writing the Report**

You should submit a written report on your final project that incorporates all of the items mentioned in the Requirements section above. Your report should include at least the following sections:

* Introduction: a one- or two-paragraph overview of your project that summarizes the key points found in the rest of the report, including the problem or problems that you attempted to solve and a high-level (not too specific) description of the results that you obtained. Think of this section as a brief preview of what the rest of the report will contain.
* Dataset description: information about of the dataset that you analyzed, including its key attributes and details about where it was obtained (e.g., a URL). A table that includes a brief description of each attribute is often helpful.
* Data preparation: a description of any steps that you took to prepare your data for analysis and mining, using steps such as the ones described in the lecture notes on this topic. Discuss the steps at a high level that would make sense to someone who is not familiar with Weka. For example, rather than saying "I applied the unsupervised/attribute/Discretize filter to the following attributes:...", it would be better to say "I used Weka to perform equal-height discretization of the following attributes:...".
* Data analysis: a description of the analysis that you performed -- including the data-mining algorithm or algorithms that you employed. You should include a brief description of each data-mining algorithm -- enough so that someone who is not already familiar with it can understand what it does. You can often find some information about a given algorithm by clicking on its name in Weka and then clicking the More button in the window that pops up. You can also try using a search engine to find out more information about the algorithm. Include references to any sources that you use.
* Results: a summary of the results of your analysis. The exact form of this section will depend on the type of analysis that you performed, but make sure that the results are presented in a clear and compelling way. This section should also include the model(s) that you produced using data mining. In the case of a classification/estimation model, you should specify how well the model performed on the test examples. If you include a confusion matrix, please turn it into a nicely formatted table. Don't just copy the text version of the matrix from the Weka window into your report. You should also include a brief discussion of each model. Does the model make intuitive sense? Why or why not? How well does the model generalize? In discussing your results, beware of making overly confident claims. It is better to be realistic and cautious. For example, instead of saying "The results clearly show that attribute A is determined by the value of attribute B", it would be better to say something like "The results suggest that attribute B may have an impact on the value of attribute A."
* Conclusions: a one-paragraph summary of the report, reminding the reader of the key points that you want him or her to remember.
* Appendix (optional): You may want to include other information, e.g., the text version of large models produced by Weka.

Your report does not need to be overly long. Just make sure that it is a clear and complete presentation of the steps that you took and the conclusions you reached.

**Submitting Your Work**

You should email me the following files:

* your final report as a Word document or PDF file; please include your username in the name of the file
* a file containing the dataset that you analyzed (in its final form, after any transformations that you applied)