Final Project

Final projects can be done individually or in groups of up to three students. Each group or individual must submit a project proposal for approval by the submission deadline below. Additional information and resources relevant to project will be posted on this page.

PROJECT TYPES

The final project for the class may involve one or a combination of the following.

- **Data Analysis**: The application of the knowledge discovery process to one or more real-world data sets. The tasks must include preprocessing and preparation of the data, data explorations (using statistical approaches to provide an overview of data characteristics), data visualization, and the application of two or more machine learning techniques on the data (e.g., classification, estimation, clustering, association rule discovery, etc.). At least one of the machine learning techniques used must involve building and evaluating a predictive model. Unless otherwise approved (as part of the project proposal), the project must involve the use of Python scripts to perform various data analysis or mining tasks (including available modules or libraries such as NumPy, Scipy, Mathplotlib, Pandas, scikit-learn, and others. In addition to Python tools, you may also use other third-party tools (preferably open-source) to assist with tasks such as preprocessing, data storage and management, and visualization. The deliverables for the project must include a detailed data analysis report, including relevant findings an conclusions about the data, as well as documented code used as part of the project.

- **Application Development**: The development and evaluation of an original application using machine learning and data mining techniques. The goal of this type of project is not to perform a full analysis of a given data set, but rather to perform useful tasks in a given application domain. The application must be tested and evaluated using a specific data set. The application must also involve the use of one or more of the modeling techniques relevant to the course topics. Your application may also include a significant extension of an existing application discussed in class materials or other sources (in this case, the application must be extended to include additional or more sophisticated types of modeling and analysis). The deliverable for the project must include the fully documented code, distribution files, including any third party sources, installation/deployment documents (including examples, screen shots of test runs, etc.), data used for the application, and a project report providing a description of the components of the application and the results of any evaluation. Many different types of applications are possible, but some examples of such applications include (but are not limited to):

  - Recommender Systems: applications that learn from user profiles to provide personalized recommendations for items in a given domain such as movies, books, products, documents, stocks, twitter feeds, etc.

  - Social Computing Applications: applications that analyze social network data, including social connections, social annotations (such as tags), microblog feeds (e.g., on Twitter, Facebook, etc.), blog posts or customer review, and other sources in order to aggregate and present users with useful information or predictions.

- Business Analytics: applications involving the use of machine learning and statistical analysis in order to derive business intelligence and assist in business decision making, including tasks such as customer segmentation, predicting customer behavior, market analysis, price prediction, inventory management, Web site analytics, etc.

- Document Filtering and Analysis: applications involving the use of machine learning and text mining techniques to identify or filter relevant documents, analyze the content of documents to discover interesting patterns (such as identifying topics or events in news stories, analyzing features of items based on customer reviews, spam filtering, etc.), recommending news items, tweets, etc., based on predictive models of users, etc