Data Mining

Rationale

Datasets consist of observations sampled from a population. They can be as large as terabytes with many variables and many records. The population may consist of subpopulations with each subpopulation having different sets of dependencies among the variables. Data Mining has tools and techniques to identify the structure that enable making valid predictions.

Course Description:

Data mining is the name given to a variety of new analytical and statistical techniques that are already widely used in business, and are starting to spread into social science research. Other closely-related terms are machine learning 'pattern recognition' and predictive analytics. Data mining methods can be applied to visual and to textual data, but the focus of this class is on the application of data mining to symbolic or numerical data. In this area, data mining offers interesting alternatives to conventional statistical modeling methods such as regression and its offshoots.

Each student will undertake a data mining analysis project as a final paper, typically analyzing a dataset chosen by the student.

Topic List

The topic list may include but is not limited to:

- Exploratory Data Analysis
- Association Rules
- Distance and Similarity Measures
- Clustering
  - K-means
  - Hierarchical Clustering: Agglomerative and Divisive

    - Subspace Clustering
    - Linear Manifold Clustering
    - Graph Theoretic Clustering
    - Spectral Clustering
    - Mixture Models
    - Biclustering
    - Density-based Clustering

- Prediction and Classification with K-Nearest Neighbors
- Discriminant Analysis
- Classification and Regression Trees
- Random Forests
- Logistic Regression
- Validation Techniques
    - Training and Test Sets
    - Permutation Tests
    - Bootstrap Resampling

Learning Goals

- Understand the mathematical and statistics foundations of the methodology and algorithms of data mining techniques
- Become proficient with data mining software such as WEKA and R
- Given a dataset, be able to discover patterns and relationships in the data that may be used for descriptive modeling or to make valid predictions

Assessment

Assessment of understanding the mathematical and statistic foundations will be done through a midterm (40%) and homeworks (20%). Assessment of proficiency in using data mining software and discovery of patterns and relationships in a data set will be done by a project (40%).