CSCI 6350 Term Project
(counts 25% of your total grade)

You can select your term project from all data mining related topics, including but not limited to: association rule discovery, classification, clustering, causal analysis, trend analysis, data summarization, data cleaning, temporal data analysis, web data analysis, …

April 20th, 25th — In class project presentation (20 -- 25 min per talk)
The order of the talk will be determined alphabetically based on last name.

May 2nd — Project final report due (at least 5 pages)
Final report should include :

- Description of your project, including the motivation of the project
- Brief literature survey in the chosen topic
- Describe your work, i.e., the algorithm developed, the software implemented, special experiments designed, and carried out, etc….
- Present the results, i.e., discuss what data has been used, any data preprocessing performed, experimental setup, experimental results. If possible, compare the results obtained from your work to those documented from existing data mining systems.
- Conclusion

You are encouraged to come and discuss with me your choice of term project topic.

Pointers to possible projects:

1. classification
    a. implement your own text classification system
    b. collect text data of your choice, obtain two existing text classification systems, e.g., Naïve Bayes and K nearest neighbor, compare the performance of the system
    c. applying and comparing bagging, boosting ensemble methods with basic classification systems.
2. clustering
    a. see attached old project 3 description
    b. clustering temporal data using Markov chain model based clustering
    c. perform a comparative study of clustering with various dissimilarity/distance measures on a selected database. The same clustering control structure will be used.
3. Association rule
    a. Parallel discovery of association rules
    b. Association rule discovery in temporal data

Old project 3

This is a three part assignment:

- part one, you are required to re-implement the K-means clustering algorithm.
  Test your program with the two data sets provided.
  Copy the two data sets into your directory by:
  % cp ~cli/CSCI6350/iris.dat .
  % cp ~cli/CSCI6350/ch8ox.dat .
  Both data sets contain objects described by interval-based features only

- part two, compare the clustering results obtained from your K-means algorithm to the those obtained by using AUTOCLASS clustering system.
  Copy the AUTOCLASS system into your directory by:
  % cp ~cli/CSCI6350/autoclass.tar.gz .

  unzip and extract the files(and directories) by:
  % gunzip autoclass.tar.gz
  % tar -xvf autoclass.tar

- part three, select one dataset from University of California, Irvine data repository, and run AUTOCLASS on the data. The data you choose should contain data objects described by both numeric and nominal valued features.
  % ftp ftp.ics.uci.edu
  Then, log in with "anonymous" as user name and your email address as password
  % cd pub/machine-learning-databases
  read the file "SUMMARY-TABLE" which gives the description of all data in that directory.

Notes:

1. Prepare data for clustering:

The two data sets provided, as well as most of the data sets in the Irvine data repository are labeled, which means that for each object in the data, there is a class label associated with it. The objective of clustering is to discover the structure of data by finding the underlying classes in data, i.e., it partitions the data into homogeneous groups, each of which corresponds to a class, and study the description of the classes.
Therefore, we do not want to have the class labels to be used as one of the features of the data to be clustered. Before applying the clustering algorithms to the data, remove the class label of the objects.

2. Analyze the clustering results

For K-means algorithm, (1) compare the object to cluster memberships to the true class labels of the data.
(2) show the description of individual clusters in terms of the definition of the cluster centroids.
Show the final mean squared error result.

For AUTOCLASS, (1) compare the object to cluster memberships to the true class labels of the data.
(2) show the description of individual clusters in terms of the models, and the parameters of the models in individual clusters.

Things to turn in :

1. the hardcopy of your K-means implementation
2. the experimental results using K-means on iris and ch8ox data sets
3. the experimental results that compares K-means and AUTOCLASS on the two data sets
4. the analysis of clustering results using AUTOCLASS on data selected from data repository.