CSCI 6350 Semester Project

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

Timeline:

Step 1 Problem Selection and Data Selection – beginning of class, Tuesday March 29th

Each student will look for a Data Mining project of interest and prepare a 8 minute presentation to be given in class on Thursday March 24th. Finding a data mining problem and the data to be used are the first step of your semester project. **Submit the presentation to the Dropbox.**

Here are some places where you can look for the project:

- KDD cups, **KDD Cup** is the annual Data Mining and Knowledge Discovery competition organized by ACM Special Interest Group on Knowledge Discovery and Data Mining, the leading professional organization of data miners. **http://www.kdd.org/kdd-cup**
  - KDD-Cup 2016: Whose papers are accepted the most: towards measuring the impact of research institutions
  - KDD-Cup 2015 – Predicting dropouts in MOOC
  - KDD-Cup 2014 - Predict funding requests that deserve an A+
  - KDD-Cup 2013 (multiple tracks)
  - KDD-Cup 2012 – User Modeling based on Microblog Data and Search Click Data, http://kdd.org/kdd2012/kddcup.shtml
  - KDD-Cup 2011 – Recommending Music Items based on Yahoo!Music Dataset. http://www.kdd.org/kdd2011/kddcup.shtml
  - Info for Year 1997-2010 can be found at: http://www.kdd.org/kddcup/index.php
  - KDD-Cup 2010 - Student performance evaluation
  - KDD-Cup 2009 - Customer relationship prediction
  - KDD-Cup 2008 - Breast cancer
  - KDD-Cup 2007 - Consumer recommendations
  - KDD-Cup 2006 - Pulmonary embolisms detection from image data
  - KDD-Cup 2005 - Internet user search query categorization
  - KDD-Cup 2004 - Particle physics; plus protein homology prediction
  - KDD-Cup 2003 - Network mining and usage log analysis
  - KDD-Cup 2002 - BioMed document; plus gene role classification
  - KDD-Cup 2001- Molecular bioactivity; plus protein locale prediction
  - KDD-Cup 2000 - Online retailer website clickstream analysis
  - KDD-Cup 1999 - Computer network intrusion detection
  - KDD-Cup 1998 - Direct marketing for profit optimization
  - KDD-Cup 1997 - Direct marketing for lift curve optimization

- Kaggle Datasets **https://www.kaggle.com/datasets**

- UCI Machine Learning repository https://archive.ics.uci.edu/ml/index.php

- **KDnuggets https://www.kdnuggets.com, look under the "Datasets" tab**

- **Other sites -- Data Quest, Data.gov, DrivenData, Quandl, etc.**

In your presentation, explain:

- Describe the domain chosen for your data mining task
- What is the main goal for your data mining task?
- Why is this project of interest and of importance to you?
- What data has been used, or will be used? Give details of the data, include past usage and results obtained from this data.

Step 2: Project Progress Report. – Thursday, April 14th

In class project progress presentation 10 min talk. At this presentation, you should describe:

- Data finalized
- Data cleaning and pre-processing steps performed
- The experiments designed, Program code developed, or Experiments performed

Submit Progress report presentation to Dropbox.

Step 3: Project Final Report dur Tuesday May 3rd, Project presentation May 3rd and 5th
Submit project report, data and code to Dropbox.

Project final report due (at least 5 pages)
Final report should include:

- Introduction:
  Project description, including the motivation and goal of the project;

- Background:
  Brief literature survey in the chosen domain, data mining research performed in this area, brief description of the data mining methods your project uses, or based on which your method is developed;

- Methodology:
  Describe your work, e.g., how you designed your new method, designed your experiments, what are the questions you plan to answer with your experiments, etc….

- Results and Discussion:
  Present the results, i.e., discuss what data has been used, any data pre-processing steps performed, experimental results. If possible, compare the results obtained from your work to those documented from previous usage of the data by other data mining researchers.

- Conclusion:
  What have you been able to achieve in this project?
  What conclusions were you able to draw from the experiments?

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

Possible project ideas:

1. Recommendation
   a. Incorporate classification or clustering techniques to improve the recommendation quality

   b. Apply feature selection and incorporate demographic information in recommendation
   c. Video recommendation
   d. …
2. Classification/Regression
   a. improve existing classification system implementation on scikit learn, e.g., bayes classifier with mixture typed attributes (numeric, nominal, categorical, binary, etc)
   b. an easy to use tool for DM researcher to study cross-validation based hyper-parameters for neural network for classification
   c. implement your own text classification system with Naïve Bayes
   d. compare text classification results for two existing text classification systems, e.g., Naïve Bayes and K nearest neighbor, compare the performance of the system
   e. comparing the bagging, boosting ensemble methods with basic classification systems.
   f. Study the characteristics of data, i.e., imbalanced data, on the results of classification
   g. …
3. clustering
   a. perform a comparative study of clustering with various dissimilarity/distance measures on a selected database. Use the same clustering control structure.
   b. Use clustering results as pre-processing step, and perform classification with the clusters identified as class label. Apply to web page classification, or email classification etc.
   c. Explore fuzzy clustering, i.e., a data may belong to multiple clusters with different probability
   d. …
4. Association rule
   a. Parallel discovery of association rules
   b. Association rule discovery in sequence data, or temporal data. Identify co-occurring pattens in sequences.