Term Projects:

Written Term Projects Reports are due by Monday, May 7. They may be submitted earlier. You should also plan to make a 15-20 minute project presentation to the class during the last two weeks of the semester starting the first or second week of April.

Please turn your written reports into Jiang Du or Edo Liberty. Make sure your name is on the cover sheet and you include an "executive summary" which outlines the problem you addressed, your approach, and a summary of your results/conclusions.

If you wish, you can work in teams of 2-4 people. But if you select to do a multi person project, you must accomplish proportionally more than a single person would. Each team may turn in one project report or individual reports. In either case, the team members should be clearly listed on the first page.

The following projects are examples. Some of them are very loosely defined which gives you the opportunity to be creative in driving the projects in directions that you find interesting. You can find many more online materials via Google searches. You may also completely define your own project. For example, you may wish to explore the use of data mining for an appropriate problem of your choice or you may wish to investigate a particular data mining algorithm or implementation.

Please send email to us (martin.schultz@yale.edu, mark.gerstein@yale.edu and jiang.du@yale.edu, edo.liberty@yale.edu) by Tuesday March 27 with a one paragraph description of your project. If it is a team project identify all the team members.

(1) Bioinformatics application. Try to use R or Matlab to reproduce the datamining results in one or more the following famous bioinformatics papers. How do the techniques proposed in the paper compare to other datamining techniques discussed this semester.

http://www.pnas.org/cgi/content/full/97/1/262

http://bioinfo.mbb.yale.edu/papers/spine-nar/index-all.html

Sandrine Dudoit, Jane Fridlyand & Terence P. Speed (June 2000) Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data [Abstract, PostScript, PDF]

(2) Explore fast algorithms for computing Support Vector Machines, cf. http://research.microsoft.com/~jplatt/smo-book.pdf. Explore ideas for handling training sets that are too large for memory.

A research group in Wisconsin has advocated using datamining approaches based on linear programming. Try to mine the Wisconsin Breast Cancer data set using SVMs and compare with the results obtained by the linear programming approach. See the following web page for details: http://www.cs.wisc.edu/~olvi/uwmp/cancer.html

(3) Explore fast algorithms for computing decision trees for large training sets, see http://www.cs.cornell.edu/johannes/papers/1998/vldb1998-rainforest.pdf http://www.almaden.ibm.com/u/ragrawal/pubs.html#classification

(papers on sliq, sprint)

(5) Explore techniques for handling problems with training sets which are missing data, see http://bioinformatics.oupjournals.org/cgi/reprint/17/6/520.pdf.

http://binf.gmu.edu/%7Ejweller/pages/BINF733_s2005_pdf/Kim_mvLocalLSQ_Bioinformatics2005.pdf

Try out these techniques on some training sets of interest by simulating the loss of data. Compare with how well these algorithms do with data sets that aren’t missing data.

(6) Bayesian Networks is a fashionable approach to many bioinformatics problems. Read the following papers and explore the use of winMine on some problems of interest.

http://research.microsoft.com/~dmax/winmine/tooldoc.htm

http://www.pnas.org/cgi/reprint/100/14/8348.pdf

Investigate the use of "Markov Blankets" (MB)for feature selection cf. http://citeseer.ist.psu.edu/aliferis03hiton.html What is the MB algorithm? How does it compare to other approaches? Create or acquire a MB code and run some benchmarks on prototypical problems to get quantitative comparisons. What do you conclude?

(7) Attribute (feature) selection is an approach for dealing with problems whose training sets have a large number of attributes. Read the following survey paper and explore the effectiveness of this approach for some interesting problems.

www.cs.iastate.edu/~honavar/Papers/bookfinal.ps

Explore the use of Genetic Algorithms for feature selection.

(8) Explore the use of parallel computing for compute intensive datamining components such as cross validation, clustering, Random Forest, random decision stumps, and Genetic Algorithm (for the feature selection problem) using parallel Matlab, parallel Octave, or parallel R.

(9) Investigate the use of global optimization, eg the genetic algorithm in datamining. Examples include feature selection and algorithm optimization.

http://www.geatbx.com/docu/index.html

(10) Explore the methods of "random projections" for classification problems. Implement a random projection method with your favorite machine learning algorithm in Matlab/Octave or R and benchmark on some typical problems.

http://cm.bell-labs.com/who/tkh/papers/df.pdf

http://cm.bell-labs.com/who/tkh/papers/rnn.pdf

(11) Explore data mining for one of the following applications:

(a) Sports

http://www.virtualgold.com/customers_sstories.html

http://citeseer.ist.psu.edu/cache/papers/cs/20768/http:zSzzSzwww.cs.bilkent.edu.trzSz~guvenirzSzcourseszSzCS558zSzSeminarPaperszSznba.pdf/bhandari97advanced.pdf

(b) Bioinformatics

Use R or Matlab for the following Patient outcome problems:

http://www.kdnuggets.com/dmcourse/data_mining_course/assignments/assignment-2.html

http://www.kdnuggets.com/dmcourse/data_mining_course/assignments/assignment-5.html

Use R or Matlab for the following Microarray analysis project

http://www.kdnuggets.com/dmcourse/data_mining_course/assignments/final-project.html

(c) Text Mining

http://intelligent-web.org/wsm/

(d) Using Data mining for SPAM filtering, cf.
http://seattlepi.nwsource.com/national/213232_microsofthiv23.html

(13) See http://www-users.cs.umn.edu/%7Ekumar/dmbook/projects.htm

for numerous ideas for projects and references.

(14) Compare and contrast the results of using the binary decision tree and random forest codes in R to the results presented in

http://www.pnas.org/cgi/content/full/98/12/6730 and

http://www.pnas.org/cgi/content/full/100/7/4168.

The training set for the first paper can be found in

http://www.sph.uth.tmc.edu/hgc/default.asp?id=2775

(15) The "GUIDE" package in Matlab is designed to support the construction of Matlab GUIs. Most of the data mining algorithms discussed in class have implementations in Matlab toolboxes such as Statistics, Optimization, and Bioinformatics. Design and implement, using GUIDE, a prototype "data mining GUI" in Matlab that integrates the appropriate Matlab codes similar in style and function to the Rattle package in R or Weka package in Java.