STAT 479: Mini-Project

You will be expected to apply the techniques introduced in the class to a real-world data problem. **The focus is on understanding and interpretation rather than just making something work**. The mini-project can be done **by yourself or in a group of two**. Both R and Matlab are allowed. You can consult with the instructor or the TA for a problem (and data set) or define your own problem (and data set). One advantage of kernel methods is that you can use when data is non-Euclidean (that is, something other than $\mathbb{R}^d$; as an example, I have provided link to network data). If you are interested in such data sets, feel free to talk to me about it and I could point out to kernels that work for your data.
A place to look for data sets is:

- http://archive.ics.uci.edu/ml/datasets.html
- http://snap.stanford.edu/data/ for network based data.

You can browse the data sets and find some that excites you and ask the instructor or TA for problems based on the data set. You have to fix your mini-project (dataset and problem) by March 30. You should send an email confirming your problem and dataset to the instructor and the TA. On April 20, I will be asking about your progress in-class. By this point, you should feel confident that you can complete the mini-project.

The results of the project should be reported in the following two ways:

1. A 10 minute presentation during the last week of the course - you can use 5-6 slides (ignoring title/names etc). Roughly, your first slide should introduce the problem considered, the second slide should describe the dataset used, the next few slides should discuss your approach for solving the problem (report the methods tried and reasons for trying and issues you faced while implementing, etc), and the last few slides should report results and your interpretation of the results. Plan your presentation for 7-8 minutes. The last few minutes are reserved for questions.
2. A 2-3 page report of due May 4th (last day of class). It should have three sections: Problem statement, Method description (along with dataset used) and Results and Interpretation. The results could be reported as numbers, tables, or graphs or other ways.

Your presentation and report will be graded out of 30 points (15 each), which will be used in the final grade calculation. Things that will be considered are: how organized and clearly written is the report ? Were the analyses chosen and carried out properly? Were your conclusions about the dataset sensible and clearly justified by numerical or graphical evidence ?