The goal of the course project is to implement and investigate the behavior of a statistical technique that interests you and to use it to analyze some non-trivial datasets (at least 10,000 data points or more). You can choose to tackle a problem in machine learning, data mining, or statistical computing of your own choosing. Most likely you will be able to identify a problem that is inherent to your own research area. However, remember that identifying a suitable problem may take you some thought. Note that it might be hard work to think about what you are really interested in, but it is crucial for the project remain open ended and in your hands. It is up to you to pick a problem you would like to solve, and then discover what it would really take to tackle it successfully. Along the way you might learn useful things, like:

- How to take a poorly defined task and turn it into a concrete problem to solve (and then a concrete implementation and test).
- What is involved in solving the specific statistical problem you are interested in, how hard it is to actually succeed.
- How partial progress can be sensibly made on difficult problems (or perhaps how not).
- How to read the research literature to find the ideas and techniques that might be relevant to your needs.

Experience in these areas will help you in whatever direction you follow in your professional career. In the end, however, the project is just a chance to learn more about some particular statistical computing techniques that might be very useful to you in your research. Some ideas to get your creative juices flowing are as follows:

1. Tackling a specific optimization problem that arises in an application of interest.
2. Experimental investigation of a technique (or techniques) in an abstract setting.
3. Theoretical investigation of a technique or problem.

Your project will be evaluated along the following metrics:

- Amount of effort you put into the project as compared to the other students in the class.
Creativity in setting up your experiments, and
Analysis of the results.

Deliverables

Write a initial project of no more than 3 pages outlining the problem you intend to solve, the various tools you are going to employ, and different hypothesis you are going to test. Also make three separate lists: a) things that you will try out at the very least b) things that you will try out if you have more time and c) things that you would love to try out if everything else works great (bonus goodies). Put some thought into coming up with this list. You will be evaluated on what you promised vs what you delivered. Over promising and under delivering is not desirable and neither is under promising and over delivering.

Your final report is due on or before the 29th of April 2010. It should be no more than 5 pages, and give details of the dataset used in your experiments, the analysis you performed, your inference of the results, and if deemed necessary some special implementation details. Only PDF files will be accepted. Mail your report and source code to vishy@stat.purdue.edu.

Marking

Your final project report should clearly articulate the rationale behind your project, in addition to simply reporting what you did and the results that were achieved. Below is a rough idea of the things that I will be looking for when grading the project reports. Note that this scheme is not intended to be a binding contract for exactly how I will mark the final project! It is only meant to give you guidance on the elements that I will typically be looking for. The specific marking scheme will be adapted to the different types of projects that are submitted.

Background

Problem: As clearly as possible, describe the general problem you are working on, and explain the specific refinements and special cases that you addressed.

Literature Survey: Briefly survey the existing work that has been done on your problem, as well as the existing work that has been done on the approach(es) you are considering. You should cite some relevant references. Your survey need not be exhaustive, but you should try to cover the most important prior work if you can.
Methods

• **Approach:** As clearly as possible, describe the statistical technique(s) that you applied to the problem, and clearly specify the final system(s) that were implemented.

• **Rationale:** Explain why you implemented the systems that you did. Specifically, explain the simplifications you made along the way, and why you made these particular choices (and not others). (For example, you might detail other approaches that could be applied to the problem and explain why these were not pursued.)

Plan

The most important (and most difficult) part of any research project is figuring out how to evaluate the results. You should state concrete goals for your project. In particular, you should identify at least three concrete hypotheses that you think would be interesting to test with your implemented system, and outline a specific plan for experiments that you would need to conduct in order to test each of these hypotheses. Testing these hypotheses should be stated as the concrete goals for your project.

**Note:** You should not phrase your project goals as “my proposed approach will work great at solving problem X”. Rather, you should phrase your goals as tests of specific hypotheses whose outcome would be interesting no matter what happened (so that your project would succeed however the tests turned out). For example, rather than simply claiming my technique will work great, it is better to formulate and test a hypothesis that would teach all of us something useful about the technique and problem, like: technique blah is robust to certain kinds of perturbations in the problem (or perhaps not), or technique blah scales up feasibly in the size of problem X (or not), or technique blah is inherently limited in the ultimate performance capabilities it can achieve on problem X because of blah blah about the problem (or not), etc. This requires a little more thought than perhaps you are used to in planning a project.

• **Hypotheses:** Clearly state the main questions that you investigated in this project. These should be identified before you run the experiments! Ideally, these should be interesting questions whose answer is not obvious beforehand, but the answer would be interesting no matter how the experiments turned out.

• **Experimental Design:** Once you have settled on some good questions, it is important to figure out how to answer them. For each main question, describe a series of experiments that is designed to answer the question. Explain the difficulties faced in designing these tests and explain why your experiments will overcome these difficulties to yield a definitive answer.
Experiments

• **Results:** As clearly as possible, describe the results you obtained with your tests. Use plots, graphs, and tables, if necessary. Make sure it is easy to understand what happened.

• **Critical Evaluation:** In the end it is important to critically evaluate the results of the experiments. Did the experimental results answer the questions? If so, then what were the answers? If not, then why did the results fail to yield definitive answers? Is it then possible to formulate a new experimental strategy?

Note that it is important to demonstrate critical thought in this part of your assessment. You will be graded more on the strength of your reasoning rather than how the tests actually turned out. For example, it is perfectly acceptable if you did not achieve definitive answers to your questions, so long as you can recognize this, explain why it happened, and suggest additional tests that might yield more definitive results.

Conclusion

• **Lessons Learned:** What specific things did you learn from doing this project? Did you learn anything about the problem itself, the approaches you tried, or the experiments you conducted? If there are any good ideas you came up with in the end but did not have time to pursue, this would be a good place to mention them.