CIS 431 - Machine learning Project

An important aspect of learning in this course is to use the knowledge gained from the class and put it together to solve a large problem as part of a project. The ideal project will require you to first identify a data set that is appropriate for machine learning. The ideal data set will have Big Data (defined by number of data points exceeding 10,000 to 100,000 and/or having a large number of features) and is freely available for analysis. Some sources of data include: UCI Machine Learning Repository.

Deliverables:

Title of project – Due Friday March 27

Short abstract – Due Tuesday March 31; Briefly list your planned approach, what data set you will draw upon, provide some idea of model results

Extended Abstract – Due Friday April 3; a few sentences on the problem you are solving; What machine learning models are you likely to use; What are your features; How will you test your model; what are the expected results

Detailed analysis plan – Due Friday April 10

Report 1 – preliminary analysis of your data set; some basic statistical analysis of different features, etc – Due Friday April 17

Report 2 – Extended analysis; First model from the data set; additional statistical analysis of data. Due Thursday April 30

Final Report – Detailed analysis of data set selected; Full explanation of approach followed; Model description; Quality of fit measured using different statistical approaches; Predictions using your model; Discussion about what worked and what did not; How could the model be better; And what are you concerned about with respect to the model (missing features, lack of data, etc). Due Tuesday May 19th.

Final Presentation – Finals week May 18th week

The grading rubric for final report:

Project description (10 points): Abstract, Background information, Problem description;

The key is make it clear to the reviewers as to what your project is about and what you are trying to achieve.

Approach (20 points): How did you go about building your model; What analysis did you conduct of the features and the data set in general; What different modeling approaches did you consider and why did you pick the final approach?

At the end of this section, you will have convinced the reader as to why you chose the approach you did and what you are trying to model.

Results (25): Performance of the model. Predictions, etc.

How did the model perform under different test measures (e.g. training error, etc). What are the major predictions?

Discussion (20): Tests of major predictions.

What do the results mean. Do the models work. What else needs to be done to ensure that the models are more robust. What could go wrong.

Overall (15): Individual effort, scale of the problem, organization of the report, quality of figures, write-up quality

Presentation (10): We will have project presentation during finals week (May 18th week).