SEIS 763 Machine Learning
Class Project

At the outset, note that this project handout only provides broad requirements so as to permit each team with sufficient flexibility in the course project. As the project handout is not meant to be fully prescriptive, each team should meet me on an ongoing basis to discuss their project and ensure that their work is appropriate and deserving of an appropriate grade. Feel free to make appropriate decisions, but clearly document each decision and its rationale, and include these in the final project report.

Project Specifics:

- The class project will involve the application of Machine Learning techniques discussed in class to a dataset of your choice. There are lots of data sets available online. Pick something that you will enjoy working on. The only requirement is that the dataset needs to be of a moderately large size (at least 3,000 instances before class balancing).
- Consider your data carefully. Even if you downloaded it, you should look for information about it. *How was it collected? What are the data quality issues? Are there biases inherent in who collected the data or how it was collected? How might this impact the subsequent conclusions?* A good report will dive into biases that may exist or data quality issues.
- Formulate questions that you would like to answer about this data set. *What is the dependent variable? What are the predictors?*
- Implement your analysis using machine learning techniques. These should have some relation to what we have learned in the class! Are you doing a regression, classification or association or clustering task? Are there interesting visualizations to do? How will you evaluate the performance of your model, or choose between competing models?

Project Proposal (Assignment 7)

Your proposal should be approximately 1 page long, single spaced. The purpose of the proposal is to make sure that you are on the right track and to give me enough information so that I can give you useful feedback. In your proposal you should cover the following items:

- What data set is being used - where does the data come from, and what are some characteristics of it (size, missing values, continuous vs. categorical).
- How many instances? If you are doing classification, how many classes do you have, and is the dataset approximately balanced?
- Is there a reason you picked this data set?
- What is the question(s) of interest - be specific. You should avoid generic questions like "*I want to look for patterns in stock prices*". Devise a specific question you can answer with the data.
- What methods do you plan to use - understanding that this might change and that we have yet to cover many methods in class.

This is a proposal, and I expect that your question and your approaches will likely change as the class progresses.

Project Report

- You will be able to produce an excellent write-up only if you have something to write about, i.e. completed a moderately sized decent project with some exciting results.

- Write a report summarizing your data, your question of interest, and your findings. Reference other existing work which has analyzed your data or addressed similar topics. The report should include specification of the problem (based on application objectives), data collection, data preparation, machine learning and statistical techniques used, interpretation of results, and conclusions). The report should be 6 pages in length (without any code) in **ACM Conference style format.**

Evaluation: Overall project grade will be based on:

- **Experiments and Results**
    - This component will be evaluated based on the size of the dataset, whether extensive experiments were performed, convincing results obtained, etc.
    - You should complete the following:
        - Class balancing (if you are doing classification)
        - Use all features of the data
        - Use feature selection (backward elimination)
        - Dimensionality Reduction:
            - PCA, LDA, and Kernel PCA (if you are doing classification)
            - PCA and Kernel PCA (if you are doing regression)
        - You should perform k-fold cross validation for all the experiments.
    - I highly recommend making use cloud resources for large scale experiments.
- **Presentation**
    - You will be presenting to the class explaining the dataset, models built, and results obtained.
    - Presentation will be a maximum of 20 minutes
- **Final Project Report**

Deadlines:

- **Project Submission (Mon 4/28)**
- **Class Presentation (Tue 4/29)**

Expected Deliverables:

- Project Report
- Project presentation
- You need to turn in all the code that you will be using for the project (preferably as a jupyter notebook)
- **Very important:** I need to be able to reproduce the same results using your code. As a result, if there are specifics steps to follow then your Jupyter notebook should specify that as a markdown cell. In addition, optionally if you want you can also submit a README file explaining how your program needs to be run to verify your results.