Final Project – MATH 574M: Statistical Data Mining and Machine Learning

You only need to complete ONE of the following four options. The “audit” are not required to do the project. Email me ANYTIME with your questions.

1. Literature Review
   Identify a machine learning topic you are interested. Read at least 2 relevant papers on the topic and write a literature review. Include the following elements in the report.
   
   • clearly define the statistical problem and specify the goal of this analysis
   • summarize main ideas and results in each paper using the same mathematical notations defined by you.
   • compare these methods regarding their formulations, model assumptions, computational cost (if relevant), statistical properties of the estimators, and etc.
   • show some numerical examples or codes (which can be extracted from the papers)
   • address the questions such as: What are their strengths and limitations? Is there room for further improvement? Which one do you recommend in practice?

   Suggestion: The students are welcome to talk with me on identifying the topic and papers. The choice depends on your interest, background, and expertise. You can start with the “suggested reading list” available on the website.

2. Real Data Analysis
   Choose one data set and apply at least two machine learning methods to the data. The size of the dataset should not be too small or simple, so that the learning task is non-trivial. You need to make comparisons of different methods in terms of their model assumptions, methods, tuning procedures, and algorithms. Write a report which should include the following elements
   
   • describe the data set and the background of this study.
   • introduce the key idea of each method, its performances, advantages, and limits.
   • summarize the results given by different methods and major discovery.
   • address the questions like: Do all the methods give consistent results? Does one method perform significantly better than others? Why?
3. **Creative Work**
Develop a new method or idea that may improve existing methods or provide a new solution to a challenging machine learning problem. The method is not necessarily entirely novel, as long as it contains some new thoughts. The focus can be on theoretical, algorithmic, exploratory, or graphical. The report should

- introduce and define the statistical learning problem
- review the literature briefly and your motivation
- describe the main idea, model assumption, fitting and tuning, and algorithms
- discuss statistical properties (if possible)
- implement the method on some simulated or real examples

4. **Related to Thesis Research**
This option gives you an opportunity to apply the techniques learned in this class to your on-going or future thesis research in your field. In the report, you should

- describe the dataset and the background of this study.
- clearly define the goal of this research, major motivations and main challenges.
- review existing methods (if there are any) and state their advantages, and limits.
- introduce the machine learning methods you plan to implement and justify
- summarize the results given by your and any discovery about this data.
- address the questions like: Do the methods look promising? Is there room for further improvement? Remaining issues?

**Format and Requirements:**

- The length of your report should not exceed 10 pages, and the letter size is $\geq 10$ fonts.
- You should provide complete and accurate references for all of your citations.
- Submit your report, along with your data set, code, or reviewed papers, on D2L.
- The code, references, and reviewed paper are NOT counted for the page limit.
- The report is due by **May 12nd, 2020 (Midnight)**.