IE 510 project Proposal

Team number: Muxia Yi, Tianqi Wu

Overview of the Proposal

In this project, we aim to study an application of Kaggle competition. Dealing with real world problem is pretty challenging since the dataset from Kaggle competition is taken from real-world including number of samples with various features. Also, we have to focus more on the efficiency because it is important for practical problems. The project will focus on finding a most suitable algorithm to provide predictions by applying GD method, HB method and Nesterov's method.

Background and Motivation

This project will implement and compare different algorithms to solve real world problems provided by Kaggle. The topic is of interest to the technical community since it would validate the theoretical results derived from the algorithms. We are particularly interested in since it would give us a perfect chance to apply what we have learned from lectures and enhance our understanding. Potential problems worth further investigating includes comparing more algorithms (CD, SGD).

Proposed Project

After carefully examining several datasets, the dataset decided to be study is "Breast Cancer Wisconsin Data Set"[1]. Logistic model would be used to fit the data and optimization of the algorithms is then carries out. Finally, this project will compare different behaviors of the three optimization algorithms and conclude the best one. We believe the project can be completed by the end of the semester since some groups have already used the logistic regression model to predict the dataset and we can focus more on the optimization part. The algorithms to be studied were thoroughly discussed in class and homework.

Timeline

4/7: Clean the data
4/14: Fit the data using logistic regression
4/22: Implement the CD method
4/29: Implement HB method
5/6: Implement Nesterov's method.
5/13: Compare the results and finish report

---

[1] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/data

Expected Results
Compare the three algorithms considering the complexity, convergence speed and computation time and conclude the best one.

Initial Progress
After making the proposal of the tentative topic, our group has examined several datasets from Kaggle and we have decided the suitable dataset to be studied.