CS66-Final Project Proposal
Luke Pietrantonio

Dataset and Goal

The proposed data set for this final project is the a Bike Sharing Dataset from Hadi Fanaee-T of The Laboratory of Artificial Intelligence and Decision Support. It contains 17389 data points, collected from Capital Bikeshare in Washington DC, between the years 2011 and 2012. There are 16 features associated with each data point, including but not limited to: date, season, weather, temperature, number of riders etc. Not all of the features will be utilized in the modeling of the data. The dataset starts on the January 1st 2011 and ends on December 31st 2012, each data point represents one hour of each day and the corresponding weather and usage information on that day and during that hour. The goal is to predict bike share usage based on weather patterns in the DC region.

Software and Models

Given the regression nature of this problem, it would be possible to build various models. Initially, a Sklearn linear regression will be used to predict the usage of bikes. Random forests, also from Sklearn, will then be used as a linear regression model. Finally, if time permits, a neural network with a linear regression final layer will be used to fit the data.

Motivation and Scientific Question

Personal motivation for the project stems from my love for biking and witnessing first hand the growth of Capital Bikeshare in my hometown, the DC-Metro area. The scientific question that I want to address is how data set sizes can improve the predictive qualities of models. This could be achieved not only by performing training with small subsections of the data, but also by data augmentation as to increase the number of datapoints. If time permits, data augmentation may look like utilizing more Capital Bikeshare usage information, in conjunction with weather datasets, to create another dataset, similar to that of the one used for initial training.

Results, Evaluation, and Interpretation

The expected result is that larger datasets will increase testing accuracies in the models and data augmentation will only continue to increase these accuracies. However, there will be a plateau where adding more training data points will no longer add benefit to the model and thus

the testing accuracy will not continue to increase. Evaluation of this hypothesis will be in the form of graphs of the number of training data points versus the test accuracies of the models. There will also be comparisons of the various models' performance at the various dataset sizes, as to explore the proficiency of the different models at the different dataset sizes.

References

D.F. Specht, "A general regression neural network", IEEE Transactions on Neural Networks, 1991

Andy Liaw and Matthew Wiener, "Classification and Regression by randomForest", R news, 2002