Project Proposal: Comparison of Regression Algorithms

Mirandi Dallas Fuge mrdallasfuge@ucdavis.edu
Sean Deely spdeely@ucdavis.edu
Xiaozhu Zhang xizzhang@ucdavis.edu

Motivation

As a widely used technique in the field of traditional statistics and machine learning, regression analysis is utilized to study and model the relationship between predictors and dependent variables to generate better prediction or forecast. There are a number of algorithms for regression. Some algorithms may perform better than the others in a specific setting, and this emphasizes the importance of choosing the suitable regression technique based on the dataset.

Project Outline

In this project, we are going to work on comparing different algorithms in Python for regression. We will be examining the performance of linear regression, non-linear regression, and random forests (CART) by applying them on datasets with different features. Under linear regression, we will use multiple linear regression and lasso regression. Under non-linear regression, we will use Kernel regression and Spline regression.

We have found eight data sets that are all different in the scale and dimensions to compare the performance of these different regression algorithms. We have attached an appendix that includes each dataset. Specifically, we are interested to investigate the performance of the above regression technique when the number of samples is much larger than the predictors, the dataset contains attributes with low correlation with the dependent variable and when the sample size is small.

The evaluation metric we may use are R square and mean squared error. In addition, we will also look at the computation time of the regression algorithms on large scale dataset.

We will be using Python's 'sklearn' package to apply the regression technique. One of the references for the concepts behind the algorithms is Tibshirani's Introduction to Statistical Learning.

Appendix: Possible Data Sets

https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized
(2215x125) moderate high in feature
Summary: This dataset is on the community information(socioeconomic, ethnic) and the instances of crimes in the community. There are 125 (predictive) attributes and 18 possibles target attribute(we choose 1 of the 18 to predict). The detailed description of the attributes can be found in the link. This dataset contains many extra predictive columns.

https://www.kaggle.com/c/zillow-prize-1/data
3 million by 58 (high sample ie:n>>d)
Summary: This is a kaggle challenge which contains datasets on home value and associated features. This dataset might be an example of the type of data that has a lot more samples than features.

https://archive.ics.uci.edu/ml/datasets/Insurance+Company+Benchmark+%28COIL+2000%29
9,000 by 86
Summary: This data set used in the CoIL 2000 Challenge contains information on customers of an insurance company. The data consists of 86 variables and includes product usage data and socio-demographic data.

https://archive.ics.uci.edu/ml/datasets/KDD+Cup+1998+Data
191,779 by 481
Summary: This is the data set used for The Second International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-98.

https://archive.ics.uci.edu/ml/datasets/Solar+Flare
1389 by 10
Summary: Each class attribute counts the number of solar flares of a certain class that occur in a 24 hour period.

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Prognostic%29
198 by 34
Summary: Prognostic Wisconsin Breast Cancer Database

https://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring
5875 by 26
Summary: Oxford Parkinson's Disease Telemonitoring Dataset

https://archive.ics.uci.edu/ml/datasets/YearPredictionMSD
515,345 by 90
Summary: Prediction of the release year of a song from audio features. Songs are mostly western, commercial tracks ranging from 1922 to 2011, with a peak in the year 2000s

Appendix: Reference

Introduction to statistical Learning: Chapter 6,7,8
https://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf