Midterm Project - Gould Rush

Stats 101C Lecture 3

Andy Shen, Ethan Allavarpu, Varan Nimar

Fall 2020

Parts I changed are bolded

The purpose of this analysis is to identify certain genes that play a role in cancer. We apply statistical learning techniques to **a dataset of genes and a large number of mutation-related, genomic, phenotype, and epigenetic features; with the goal of identifying oncogenes (OGs), tumor suppressor genes (TSGs), and neutral genes (NGs)**, ultimately aiding future **research into cancer prevention, diagnosis, and treatment**.

Upon plotting each predictor against its respective gene class, we noticed that there existed outliers for many predictor variables in the dataset. Most plots saw clusters of points in certain locations without much variability, but there was always at least one stray point in many of the plots that stood out and did not fit the general trend of the plot. We decided to remove the top 50 observations containing the greatest number of outliers, as well as extreme points that clearly stood out as unusual when examining the scatterplots for each predictor. Because there were no unknown (`NA`) values in the dataset, we did not remove any observations on the basis of missing values.**Transformations???**

Once the data were cleaned and transformed, we selected our predictor variables of interest by first determining the significance of each predictor. We used an Analysis of Variance (ANOVA) approach to determine if there existed a significant difference amongst the predictors. While a vast majority of the predictors were statistically significant, we still had a large number of predictors since the dataset had over 90. To further refine our predictors to the most important ones, we visualized the correlation amongst our subset of predictors to see which variables exhibited high correlation. After refining our predictors to those that were both highly significant and largely uncorrelated were we able to begin fitting our models.

We ended up utilizing the Linear Discriminant Analysis technique to predict the type of gene based on the other observations and predictors. We prefer this method due to its relatively low flexibility compared to its quadratic counterpart, as well as its reasonable, but not exorbitant, test prediction rate when testing out our model. We used a weighted test prediction rate by placing extra emphasis on correctly identifying oncogenes and tumor suppressor genes, genes that play the largest role in detecting cancer. We subsequently placed less weight on the neutral genes, since their relevance in cancer research was not high.

Evaluation Metrics. We also can look at True/False Positve/Negative rates for each class (ROC?). If we implement a way to try and maximize point value (NGs worth 1, others worth 20) other than selecting a more/less strict threshold we can discuss that here too.

Go into more detail about LDA Centering/scaling Parameters Cross-validation test/train split Why LDA over the others?

Appendix

Statement of Contributions