Midterm Project - Gould Rush

Stats 101C Lecture 3

Andy Shen, Ethan Allavarpu, Varan Nimar

Fall 2020

Introduction and Setup

The purpose of this analysis is to identify certain genes that play a role in cancer. We apply statistical learning techniques to a data set 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.

After plotting each predictor against its respective gene class, we noticed that there existed outliers for many predictor variables in the data set. 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. We also remove all rows with more than 50 predictors recorded as 0 in case these metrics might correspond to genes that are irrelevant to the predictions of this analysis. This will hopefully reduce variability and allow predictions to be more accurate.

Because there were no unknown (`NA`) values in the data set, we did not remove any observations on the basis of missing values. While a vast majority of the predictors were statistically significant, we still had a large number of predictors since the data set 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.

Evaluation Metrics

In terms of evaluating our model, we selected a model using the Linear Discriminant Analysis (LDA) method as well as with Logistic Regression. We prefer this method due to its relatively low flexibility compared to its quadratic counterpart, as well as its reasonable, but not exorbitantly high, test prediction rate. For all techniques used, we ran our model numerous times with different seeds and submitted the model which had a weighted categorization accuracy that was closest to around 0.78, in order to avoid submitting an inaccurate or over-fitted model.

We calculated the performance of our model using the weighted categorization accuracy (WCA). This score places extra emphasis on correctly identifying oncogenes and tumor suppressor genes, genes that play the largest role in detecting cancer. Less weight is placed on the neutral genes, since their relevance in cancer research was not high.

The distribution of the response variable in the training set is shown in Table 1.

Table 1: Percentage of Gene Type in Training Set.

NG	OG	TSG
89.39%	5.29%	5.32%

After testing various thresholds, predictor combinations, and training/test data sets, LDA proved to be the most consistent when it came to the weighted test error rate. Other techniques, such as Quadratic Discriminant Analysis (QDA) and K-Nearest Neighbors (KNN), saw test error rates that fluctuated when the training and test data were changed. We use 5-fold cross-validation to validate each of our models. Moreover, the sporadically low test error rates seen in QDA and KNN indicate overfitting of the data, while the sporadically high test error rates indicate a poor model fit. Both the consistency of the the LDA technique and the inconsistency of the more flexible techniques led us to conclude that the relationship of the data is likely a linear one.

Candidate Models

The predictors for our first candidate model (model 13) were selected by examining a correlation heat map and fitting a logistic regression model with all predictors. Based on the predictors that were largely uncorrelated from the heat map, we examine all of the significant predictors from the logistic regression model and extract the top 29 predictors that were both significant in distinguishing between each of the three response classes and had low correlations with other predictors. For instance, many of the predictors beginning with `Broad_` were statistically significant, but were very closely correlated with other predictors of the same `Broad_` family. This led us to only select the most significant predictors from this family that had the least correlation with other predictors of the data set.

There were many predictors in this data set that were of the same family and closely correlated as well, so we repeated this methodology for all of the significant predictors of the same family. In this analysis, our significance level α was 0.05 divided by the number of predictors in the data set. We scale α because we want to decrease the likelihood of a false positive that could result from running the model multiple times across many different seeds.

To select a threshold, we calculated the weighted categorization accuracy using 5-fold cross-validation across arbitrarily set seeds and selected the model with the WCA that was neither the maxiumum nor minimum to account for possible over fitting. The thresholds with the most consistent performance across all seeds seems to occur at 0.01 or less.

As a result of the different predictors, our weighted categorization accuracy calculations during 5-fold cross-validation revealed that our candidate model performed most accurately and consistently with a threshold of approximately 0.005 to 0.01 as seen in Figure 1. The threshold is regarding the probabilities for classes 1 and 2. If either of these classes exceeds the threshold, we take the higher probability; if neither exceeds it, we choose class 0.

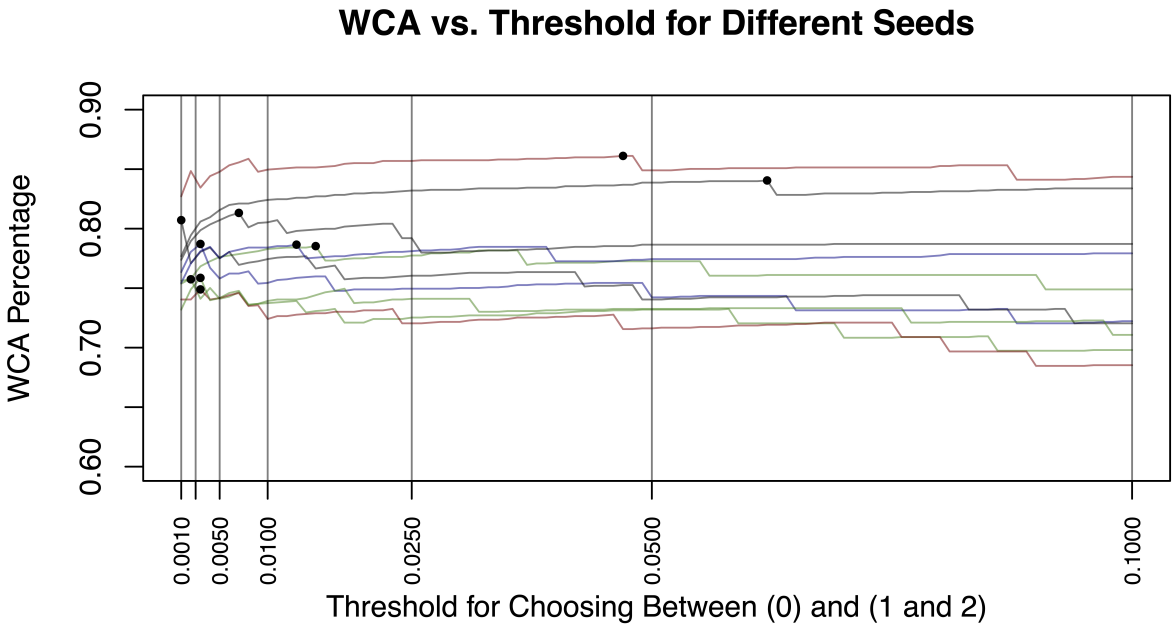
Our candidate model's WCA for the training data was 0.81, which resulted in a WCA of 0.852 on the public leaderboard.

Predictor selection for our second candidate model was done similarly to our first model: we utilized the same predictors in this model as the first candidate. The key difference between the two models was the chosen threshold of 0.01. As such, our second model was also less likely to choose genes of classes 1 and 2 unless the probabilities were greater than 1%, as opposed to 0.5% in candidate model 1.

Our second candidate model's WCA for the training data was 0.824, which resulted in a WCA of 0.811 on the public leaderboard.

We selected the 0.005 threshold as our primary candidate model because as shown in Figure 1, many of the best threshold were below 0.005. However, we still included a threshold of 0.01 as our secondary candidate model because there were a few seeds for which a higher threshold performed better and we want to account for that.

Figure 1 below shows how the WCA varies with increasing threshold for both candidate LDA models. This measurement greatly aided our selection of a candidate model.



The reasoning for selecting LDA as our two model selections is due to LDA's limited flexibility, which accounts for less variability and provides more consistent WCA scores across various subsets of training and test data. Nonetheless, we believe our LDA model's threshold provides enough flexibility to account for its strong performance: by lowering our threshold to a value no higher than 1%, we account for the uneven distribution and scoring of predicting different responses while simultaneously placing greater emphasis on the less likely responses (OGs and TSGs) because they are weighted more heavily. This allows us to predict the less-frequent responses at more appropriate intervals, ultimately accounting for the WCA.

SOC

Code Appendix