Midterm Project - Perfectly Imperfect Models

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 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.** We also remove all rows with more than 50 predictors recorded as 0 in case these metrics might have been recorded as missing or unknown. This will hopefully reduce variability and allow predictions to be more accurate.

We simultaneously used an Analysis of Variance (ANOVA) approach to determine if there existed a significant difference among the predictors. By simultaneously using the ANOVA approach as well as visually examining the scatterplots, we were able to use multiple techniques to refine our predictors. This prevented us from relying too heavily on a single method and provided us more insight into which predictors would be best. Because there were no unknown (`NA`) values in the dataset, 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 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.

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. In general, we prefer these parametric techniques since they are generally less variable than non-parametric techniques. These two methods also produced sufficiently high scores with the test data, but the scores did not appear to be excessively high to the point where we believed we were over-fitting the data. For both LDA and Logistic Regression, 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 selected a threshold of 0.024 due to the distribution of the response variable in the training set shown in Table 1. 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. Less weight is placed on the neutral genes, since their relevance in cancer research was not high.

Table 1: Percentage of Gene Type in Training Set.

NG	OG	TSG
89.39%	5.29%	5.32%

Linear Discriminant Analysis

We utilized the LDA 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 exorbitantly high, test prediction rate.

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.

Mention Candidate LDA models and which ones were the best. Ask Ethan.

Logistic Regression

We also noticed that implementing a logistic regression model proved to have a relatively high weighted score, making it a viable technique to investigate. We utilize the same threshold to ensure consistency across all models.

In this method, we used the `nnet` package in R to fit a logistic regression model with the same set of predictors as we used in LDA. We also noticed that our weighted category score was consistently between 0.75-0.82, so after running the model multiple times with different seeds, we decided on a model that had a score of 0.77, which translated to a WCA of 0.74 on the test data. This model was fit using a 0.05 threshold, but we trained 76% of the training set instead of the standard 80%.

Talk about separate model using `broom` to organize `multinom` predictors by p-value?

Mention Candidate Logistic models and which ones were the best. Ask Ethan.

Measuring Error

Discuss graphics, parameter specifics, for both models

Appendix

Statement of Contributions

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