Chenxi Yang & Yi Tang

Professor Zhang

11 November 2018

Report: Statistical Modelling and Analysis Results for the HIV

Abstract

Like Assignment One, we choose 5 drugs (3TC, ABC, AZT, D4T and DDI) in the drug class NRTIs. Differently, in Assignment Two, except for linear regression and penalized linear regressions (Ridge, LASSO, and Elastic Net) that we used in the last assignment, we apply more models including logistic regression, logistic regression version penalized regressions, Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), classification tree and regression tree to calculate the value of misclassification rate, which is the comparative indice in this analysis.

Criterion for best fitting model

As we learnt from the lecture, the misclassification rate is calculated by

$$\frac{bad\ prediction}{total\ prediction}$$

which indicates the lower misclassification rate indicates better prediction result.

Data Cleansing

After loading data with the help of HelperFunction.R, we observe that each drug has different sample sizes, which may result in errors in making matrices. Therefore, we keep common variables and isolates and eliminate different terms. In the end, each drug has 228 variables and 1246 observations used in the following test. The order of the 5 drugs is 3TC, ABC, AZT, D4T and DDI.

Working Process

1) Setting drug cutoffs for 5 drugs

According to the research paper and online data given, for the five drugs (3TC, ABC, AZT, D4T and DDI), we set cutoff with 3, 2, 3, 1.5 and 1.5 respectively, which means

$$\begin{cases} IC_{50} \leq cutoff \rightarrow Susceptible \\ IC_{50} > cutoff \rightarrow Resistance \end{cases}$$

2) Logistic regression model and logistic glmnet model : setting proportion cutoff

We use proportion cutoffs in the logistic regression model and logistic glmnet model.

For each drug, we firstly choose the optimal proportion cutoff by calculating misclassification rates on the training data in the range between its maximum and minimum IC_{50} values among them and get predicted values. Then, the proportion cutoff that cut the training data the best is the one we choose in this fold. Each fold of each drug has the different proportion cutoff. We hold proportion cutoffs and predicted values for each fold. The average proportion cutoff of all folds is the final one we use to cut for that drug.

3) Logistic glmnet model and glmnet model : choosing α value

Choosing α value is a way to optimize the model. We apply this method in both glmnet model and logistic glmnet model. We provide two functions: one is choosing α by users themselves (we use 0.3 in this case); the other one is using a function to choose α, which will take much more time to get the outcome. When we observe the difference between misclassification rates of two glmnet model outcomes, it is small enough to ignore, which means whether we choose optimal α or not it will not significantly affect the final result. Therefore, in our analysis, we only apply α-choosing function in glmnet model to calculate the final rate. For logistic glmnet model, since cv.glmnet runs very slow, we set an example seed to find the best one in logistic glmnet model in order to reduce the time taken. If you want to find the optimal α, you can use the function for choosing α provided.

4) KNN and LDA

K-Nearest Neighbors and Linear Discriminant Analysis are both classification models. KNN is a time-

consuming model for finding optimal k while LDA is a fast classification model that is easy to implement. However, KNN is beneficial to provide notable results without assumptions about the shape of the decision boundary while LDA is restrict to Gaussian assumptions.

In our KNN model, we construct a k-selection function in KNN for choosing best k all odds between 1 and 19.

5) Regression tree and classification tree

The regression and classification trees are machine-learning methods to building the prediction models from specific datasets. The data is split into multiple blocks recursively and the prediction model is fit on each of such partition of the prediction model. Both classification and regression decision trees have dependent variables and predictor variables. The primary difference between classification and regression decision trees is that, the classification decision trees are built with unordered values with dependent variables. The regression decision trees take ordered values with continuous values. (Pulipaka 2016)

5) Linear model and glmnet model: using log-value

In linear model and glmnet model, we use log-value to predict first because there exist large difference in the origin data, in order to avoid the situation that some of values are identified as outliers. Then, we make predicted value exponentiate back, which will get better misclassification rate.

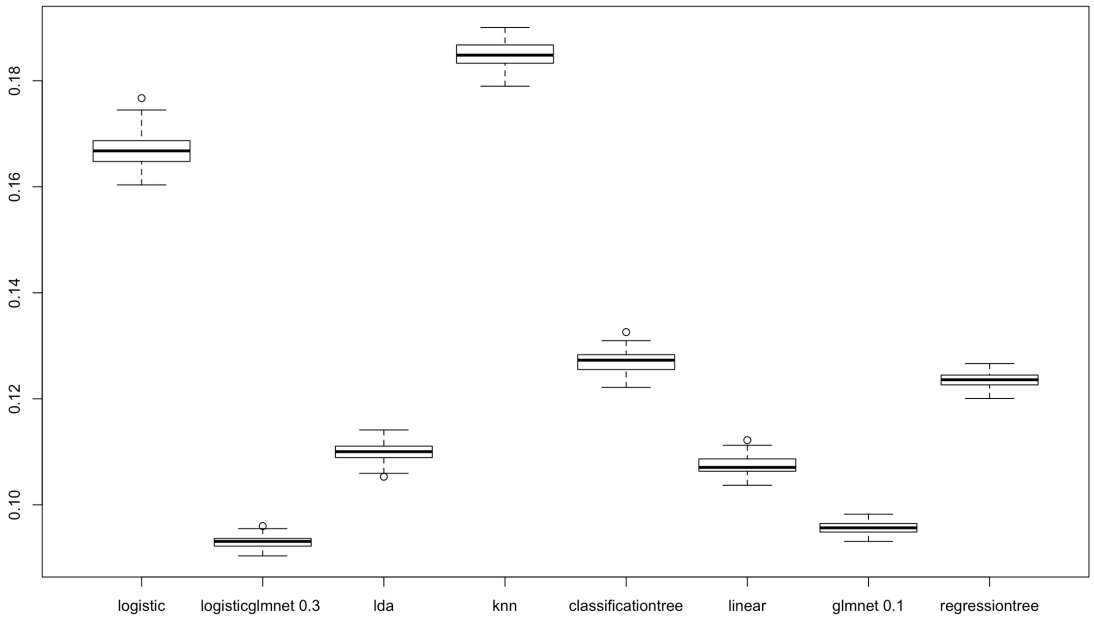
6) Run 100 times for comparison

We set 100 seeds for running the final result and construct 7*100 matrix in order to find the best performing model (the least misclassification rate).

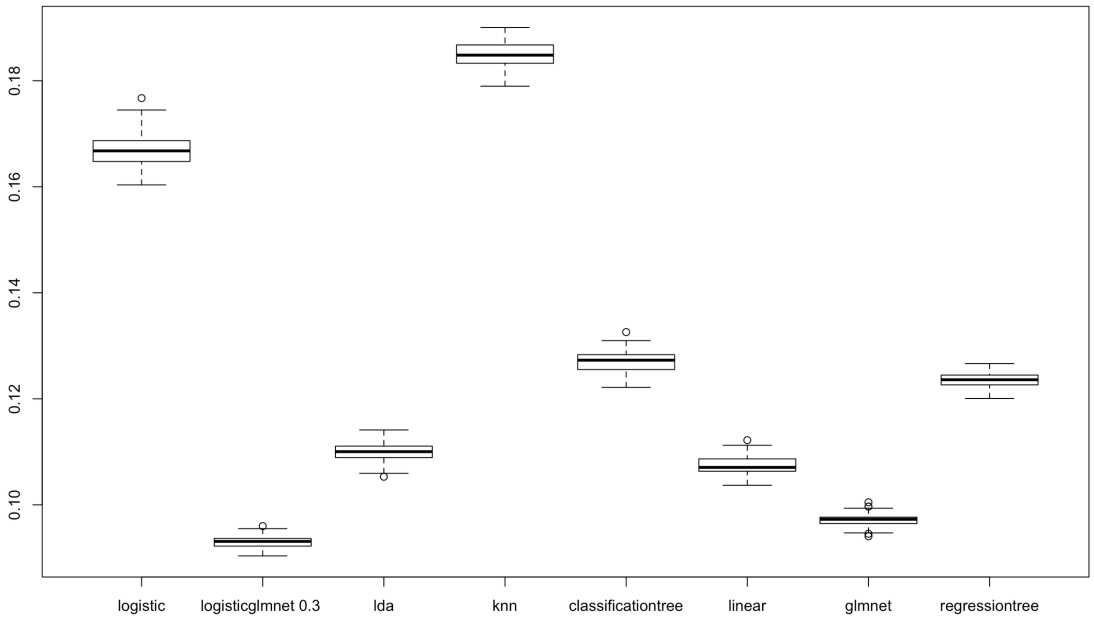
Conclusion

After applying 7 types of models ($\alpha = 0.3$ performs best in logistic glmnet model and $\alpha = 0.1$ performs

best in glmnet model), we construct boxplots below, which shows the result that logistic glmnet model with $\alpha = 0.3$ is the best prediction model.



Also, when we pick optimal α for glmnet function, we get boxplots below:



To sum up, regarding to the glmnet model, whenever α that is chosed is optimal one or not, its misclassification rate is larger than that of the logistic glmnet model. Logistic glmnet model is the best performed model among 7 models we analyze. The reason for that might be it has optimal proportion cutoff and avoid the risk of overfitting.

In addition, we notice that except for logistic glmnet model, all classification models (logistic model, LDA, KNN, and classification tree) perform worse than models with continuous variables (linear model, glmnet model and regression tree). Our guess is that classification models are consisted of binary data, which may have large collinearity between 0 and 1.

The last point is that as we conclude in Assignment One, the MSE of ElasticNet model with 70% proportion is smaller than that of linear model while in Assignment Two, the misclassification rate of ElasticNet model is also smaller than that of linear model. Based on this phenomenon, we may suppose that lower MSE comes with lower misclassification rate.

Work Cited

Pulipaka, GP. An Essential Guide to Classification and Regression Trees in R Language, 2016, medium.com/@gp_pulipaka/an-essential-guide-to-classification-and-regression-trees-in-r-language-4ced657d176b. Accessed 19 Nov 2018.