Yiling Chen

Dec.1, 2020

Hw#6 write-up

**Question 1**

The purpose for this problem is to apply and analysis on random forest, bagging, boosting, and kNN method on pima dataset and evaluate them on test set to see the error rate. The pima dataset consists of 532 observations on 9variables but for the benefit of this problem, we are eliminating the last column of the variable. The first 7 variables are diagnostic measurements, the 8th variable is the result of the observant whether she has diabetes or not.

Graphical user interface, text, application, chat or text message

Description automatically generated

My first step is to set working directory, load pima dataset, make sure all the packages are available, check to see there are no missing data, and take a brief look at our dataset.

Graphical user interface, text, application

Description automatically generated

Next, I split the pima dataset into 80% training and 20% testing and set classdigit into 0,1s for future usage of calculating test error.

Graphical user interface, text, application

Description automatically generated

I first apply random forest method and set the parameters using training data, 1000 trees, and classdigit as predictor. Then I calculated the error rate using test set and the result is the 24.29%

Graphical user interface, text, application

Description automatically generated

Next, I applied bagging set m = 3, which means to have three randomly choose variables to build the forest. Then I again use test set to check my error rate. Which is 22.42%, it is performing better than random forest.

Graphical user interface, text, application, email

Description automatically generated

I applied boosting where I tried shrinkage both for .1 and .6 to see which one would perform better. After getting the result of the shrinkage, it seems like when shrinkage is at .1 the misclassification error base on our testing set is lower than when shrinkage is .6. However, it is similar to random forest and bagging with slightly different error rate.

Graphical user interface, text, application

Description automatically generated

Now, I want to compare my previous results with the more simplistic method: kNN. By using kNN and set the k = 20, I am predicting the classdigit using only glucose, which is the highest ranking on the variable importance. The result I got back is not optimal. The error rate for the test set turns out the be 77%, which is much worse than random forest method. This make sense because kNN only use one variable as the predictor whereas in random forest, there are multiple variables to predict the result.

Some advantage of bagging relating to this dataset is that it improved the accuracy of random forest. The error rate for random forest is 24.29%, which is better than the kNN error rate. However, the bagging error rate is 22.42%. On the other hand, a disadvantage is the training period is a long time especially if I increase the number of trees.

In conclusion, after applying bagging, boosting, random forest and kNN methods to predict whether a person is diabetic or not, bagging’s prediction is most accurate comparing with other methods. However, the error rate of bagging, boosting, and random forest are not very different than each other and are all around 25%. kNN on the other hand, has a very high error rate. Indeed, random forest will give use a more accurate prediction however, the complexity and amount of time for it to run is the opportunity cost.

**Question 2**

For this problem, I will need to use the pima dataset to predict whether the observant is diabetic or not base on 7 variables. I will use boosting, random forests and CART model to make the prediction, then look at the performance and finally, explore the partial dependence plots for variables that are listed as high ranking on variable importance table.

The pima dataset consists of 532 observants and 7variables of measurements where the 8th variable shows whether the person is diabetic or not.

Graphical user interface, text, application, chat or text message

Description automatically generatedGraphical user interface, text, application

Description automatically generated

What I did above is to set working directory, load pima dataset and some libraries needed for this question. Next, I take a closer look at the pima dataset and deleted the last column of pima because it is similar with the 8th variable which is a numeric value of prediction.

Then, I split the pima dataset into test and training so later I can use the training to predict and test to look at the accuracy of my prediction.

Graphical user interface, text, application

Description automatically generated

Next, I grew a single tree by first set the parameter of the tree using the function rpart(), and I want to use classdigit as the predictor, trainning data as the data, etc. After growing the full tree, I pruned it base on the min\_cp which is at 3.

Finally, I compute the test error for CART method and the result is 23.36%.

Diagram

Description automatically generated

I then use the random forest method to compare the difference with the CART method. For random forest, I again use classdigit as the predictor, data as the training set, and I used 1000 trees to generate my forest. Next, I use the test set to compute my prediction error rate, which is also 23.36%.

Graphical user interface, text, application

Description automatically generated

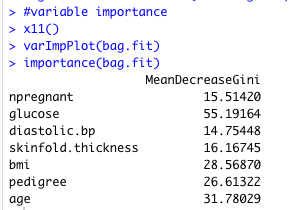
Finally, I applied boosting method where I tried shrinkage both for .1 and .6 to see which one would perform better. After getting the result of the shrinkage, it seems like when shrinkage is at .1 the misclassification error base on our testing set is lower than when shrinkage is .6.

Graphical user interface, text, application, email

Description automatically generated

In order to explore the partial dependence plots, I need to first identify the top three importance variables base on the variable importance plot and importance function. Which is shown below. I notice that the level of glucose, age is the most important variable for bagging fit and glucose, pedigree is most importance for random forest fit.

A picture containing chart

Description automatically generated

The x-axis of each of the partial dependence graph shows each of the important variables base on variable importance function. The y-axis shows how impactful age/pedigree/glucose is in term of classdigit probability.

Text

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generated

In conclusion, I grew and pruned a single tree(CART) and the result of error rate is the same as I use random forest when number of trees is 1000, both test error rate is 23.36%. Then, I used boosting to calculate the error rate when shrinkage is .1 and .6, the result of shrinkage is .1 is better than .6. Finally, I have computed the partial plot base on the top three variables that are important to predict classdigit, which are age, pedigree, and glucose.

**Question 3**

For this question, we are fitting a series of random-forest classifiers to the SPAM dataset, explore the sensitivity to m as well as computing both OOB error and test error and plot the graph. The SPAM dataset contains 4601 observations and 58variables where V58 is the predictor. This dataset classifies emails as spam or non-spam with 57varaibles indicating the frequency of certain words and characters in the email.

Graphical user interface, text, application

Description automatically generated

My first step is to load the data, check to see if there are any missing values, then split the data into test and training where train contains 80% and test contains 20% of the total SPAM database.

Graphical user interface, text

Description automatically generatedText

Description automatically generated

Next, I compute four bagging random forest models and changing only the sensitivity to 3,54,10,30.

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application

Description automatically generated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| mtry | 3 | 10 | 30 | 54 |
| OOB | 5% | 4.65% | 5.19% | 5.33% |
| Test error rate | 5.2% | 5.4% | 5.4% | 5.75% |

As we can see, as m increase the OOB and test error rate starts to decrease, however, at somewhere when m is between 10~30 the out of bagging error rate start to increase again, then the test error rate as well. Therefore, the most optimal model (where error rate is lowest) for bagging method is when mtry is between 10~30.