Assignment 4 Machine Learning

Haoran Huang (20729702)

Nov 14th, 2017

This assignment was completed by my own efforts and I did not collaborate with any other person for ideas or answers.

This is the first time I have submitted this assignment or essay (either partially or entirely) for academic evaluation

* + - * 1. Using the starter script provided, use randomForest() to build a model to classify tree Type using all available predictors from the subset of the trees data. Create 5000 trees, and set the parameter *m* for subsetting predictor variables at each split to 2. [3 marks]

fit <- randomForest(Type ~ ., data = trees.sub.rm, ntree = 5000, mtry = 2)

* + - * 1. What is the accuracy of this model in predicting the training data (i.e. "in the bag" accuracy)? What is the accuracy for "out-of-the-bag" OOB predictions? [2 marks]

The accuracy tested on the original train data is 100%.

The “out of bag” prediction accuracy is 70.82%.

* + - * 1. Which tree types are generally well predicted, and which are less well predicted? [2 marks]

Call:

randomForest(formula = Type ~ ., data = trees.sub.rm, ntree = 5000, mtry = 2)

Type of random forest: classification

Number of trees: 5000

No. of variables tried at each split: 2

OOB estimate of error rate: 29.18%

Confusion matrix:

1 2 3 5 6 7 class.error

1 269 107 0 0 0 0 0.2845745

2 90 374 8 0 3 0 0.2126316

3 0 17 46 0 1 0 0.2812500

5 0 15 0 2 0 0 0.8823529

6 1 8 12 0 6 0 0.7777778

7 26 2 0 0 0 7 0.8000000

According to the confusion matrix of classes, we can see from the class error rate that type 1, 2, 3 are generally well predicted with the lowest error rates (28.4%, 21.3%, 28.1%). And type 5, 6, 7, are less well predicted with relatively high error rates (88.2%, 77.8%, 80.0%).

* + - * 1. Fit the model for a second time and note what the OOB error is -- what do you notice, and why does this happen? [1 mark]

OOB error is 28.28%, which is slightly different from 28.18% of the previous model.

This happens because, randomness is introduced during the process of bagging. The sample are randomly selected based on the “mtry” which determined how many predictors are available.

* + - * 1. This time running randomForest() from within the train() function of the caret package, perform repeated 10-fold cross-validation on a randomForest model to classify Tree type using all predictors. Test different values of the parameter *m* to tune the model by finding the parameter value that produces highest accuracy (lowest error rate). [4 marks]

Random Forest

994 samples

11 predictor

6 classes: '1', '2', '3', '5', '6', '7'

No pre-processing

Resampling: Cross-Validated (10 fold, repeated 2 times)

Summary of sample sizes: 896, 896, 893, 894, 893, 895, ...

Resampling results across tuning parameters:

mtry Accuracy Kappa

3 0.7088497 0.5118064

4 0.7174068 0.5275223

5 0.7184121 0.5291764

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was mtry = 5.

“mtry” equals to 5 produces the highest accuracy rate of 71.8%

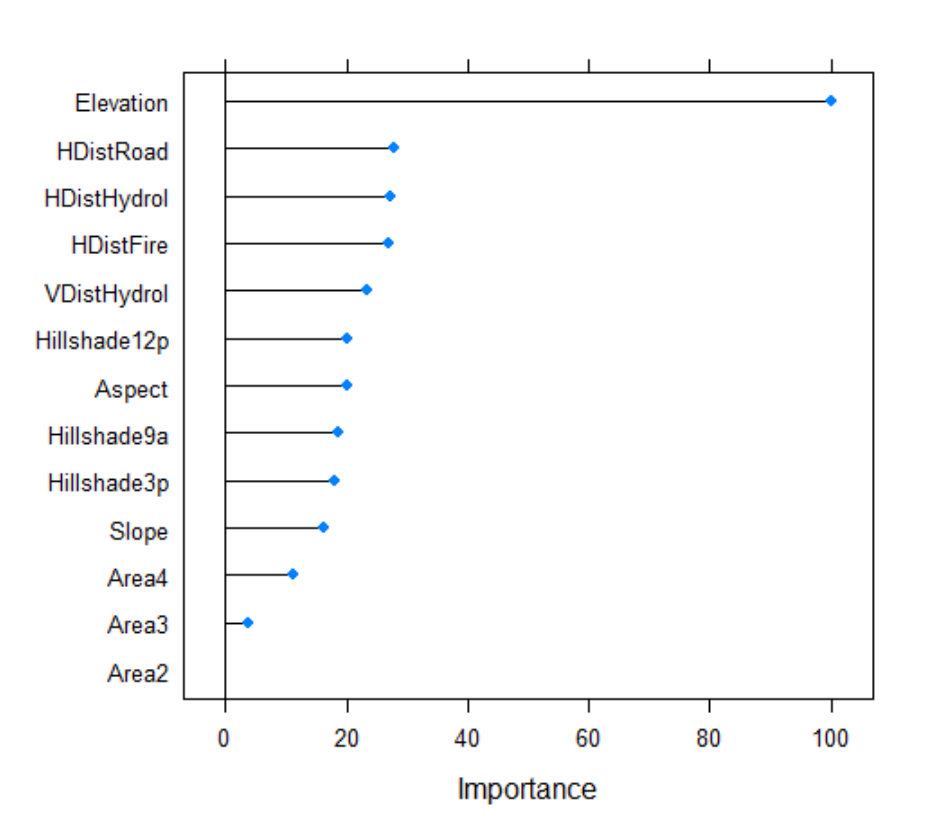
* + - * 1. How does this value of *m* differ from what the textbook would predict? [1 mark]

Number of variables randomly sampled as candidates at each split. Note that the default values are different for classification (sqrt(p) where p is number of variables in x) and regression (p/3)

“Typically we choose m ≈ √ p—that is, the number of predictors considered at each split is approximately equal to the square root of the total number of predictors”(James CH8).

We are performing classification using 11 predictors, where sqrt(11) = 3.31 is the default value of mtry. Our m equals to 5,which is greater than the textbook would predict.

* + - * 1. Create a plot showing variable importance from the cross-validated fit, as a percentage relative to the most important variable. Which variable(s) emerge as most important? Interpret the importance of the first two predictor variables. [4 marks]



As is shown in the table, Elevation, HDistRoad, HDistHydrol, HDistFire are the most important variables, without which, there will be a large increase on error rate. Elevation is the most important variable. Relatively to the Elevation, the second most important variable is about 35% in terms of importance.

The elevation determines the climate which have great influence on the type of tree.

The horizontal distance to road can also decide the tree type, because the trees near the road are planted by human. (It can be designed to be rows of same tree).

* + - * 1. Make a table showing the highest accuracy achieved for each method, and fill in the value for randomForest. [2 marks]

|  |  |
| --- | --- |
| randomForest | SVM |
| 71.8% | 63.6% |

* + - * 1. Run the svm() function in the e1071 library from within train() to tune the optimal SVM model to classify Type, using a radial kernel and all available predictors from the subset of the trees data. Perform a grid search to tune the cost (C) and gamma (G; referred to as sigma in train()) parameters over the ranges C=10^(-1:5) and G=2^seq(-15,3,2). Report the cross-validated accuracy from the best, and worst, combination of parameters you tried. [5 marks]

The best combination was sigma = 8 and C = 10 with the accuracy of 63.6%. (Accuracy was used to select the optimal model using the largest value.)

The worst combination was sigma = 8 and C = 0.1 with the accuracy of 47.8%.

* + - * 1. Record the highest accuracy value in the table you created in 3. [1 mark]

As the table shows.

* + - * 1. Compare the accuracy you achieved using the random forest and SVM methods, and comment on which method works best to classify tree type. Making reference to any relevant academic literature sources you can find, offer possible reasons for why one method may perform better than another for this type of problem. [4 marks]

The random forest produced a higher accuracy of 71.8%, while the accuracy of SVM was 63.6%. The difference is rather significant.

One reason is that SVM model applies to classification between two groups better.

The SVM is based on binary classification. When doing multi-class classification, it has two method one-versus-one and one-versus-all. One-versus-one constructs C(2,K) SVMs, and the final classification is performed by assigning the test observation to the class to which it was most frequently assigned in all SVMs.

The one-versus-all approach is an alternative procedure for applying SVMs one-versus-all

in the case of K> 2 classes. We fit K SVMs, each time comparing one of the K classes to the remaining K − 1 classes. Let β0k,β1k,...,βpk denote the parameters that result from fitting an SVM comparing the kth class (coded as +1) to the others (coded as −1). Let x∗ denote a test observation. We assign the observation to the class for which β0k+β1kx1∗ +β2kx∗ 2 +...+ βpkxp∗ is largest, as this amounts to a high level of confidence that the test observation belongs to the kth class rather than to any of the other classes (James CH9).

As a result, SVM may have a less satisfying result for multi-class classification (tree types).

It has been examined that, in the cancer diagnosis using gene expression microarray, random forest is outperformed by SVM.

Statnikov, A., & Aliferis, C. (2007). Are random forests better than support vector machines for microarray-based cancer classification? *AMIA ... Annual Symposium Proceedings. AMIA Symposium,* 686-90.

Norinder says that the machine learning techniques (random forest, SVM, and neural network) all do well in prediction, and it is hard to interpret which method is better for certain cases. Therefore we should carefully compare the accuracy of each method and even combine these method.

Eklund, M., Norinder, U., Boyer, S., & Carlsson, L. (2014). Choosing feature selection and learning algorithms in QSAR. *Journal of Chemical Information and Modeling,54*(3), 837-43.