CMPT459 Fall 2018
Data Mining
Martin Ester
TA: Sahand Khakabimamaghani

Programming Assignment 3

Total marks: 100
Due date: November 16, 2018

Data

Our dataset is the Titanic dataset, which we have already used for Programming Assignment 2. Use the training dataset Titanic.Train.csv and test dataset Titanic.Test.csv posted on CourSys under datasets.

Tasks

In this assignment, you will gain practical experience with further classification methods. Solve the tasks using R and answer the questions.

1. Use package randomForest to learn a random forest from the training data with the number of trees set to 100. Apply the random forest model to predict the class labels of the test data. What is the accuracy of the model? How does the accuracy of the random forest model compare to that of the best decision tree model from Programming Assignment 2? Using the pROC package, plot the ROC curve of your random forest model. What is the AUC?

2. Use functions **importance()** and **varImpPlot()** to analyze the importance of the different attributes across the whole forest. Report the top three most important attributes in decreasing order of importance and explain their relevance for the classification task.

3. Learn a logistic regression model from the training data using the **caret** package and the **glmnet** method of the function **train()**. What are the most significant three attributes of your model?

4. Apply the logistic regression model to predict the class labels of the test data. Plot the confusion matrix. What is the accuracy of the model? Plot the ROC curve of your logistic regression model. What is the AUC of your logistic regression model?

5. Using the function **tune.svm()** from package **e1071** and the training dataset, obtain the best parameters for a linear SVM and for a radial kernel SVM. What are the best parameters for the linear and for the radial kernel?

6. Apply the tuned linear SVM model and the tuned radial kernel SVM model to predict the class labels of the test data. What is the test accuracy of the two models? Plot the ROC curve of both SVM models. What is the AUC of the two models?

Submission

Submit your R Markdown code in pa3.Rmd following the format of programming assignment 2.