Exercises 1
Introduction to CARET package based on Cars93 dataset

1 November 2016

> d = d <- read.table("Cars93.txt", header=T, sep="\t")

Exercise 1. Install and load caret package in your R environment.
Exercise 2. Use train() function in caret to build a linear model similar to m2 model you created before (m2 <- lm(Price Horsepower + Weight, data=d).

Exercise 3. Compare the two models using summary().
Exercise 4. Use the following command to select all numeric features in your dataset. The new features subset would be used to train predictive models of Price in the following exercises.

> dn = d[,sapply(d, is.numeric)]

Exercise 5. Remove Min.Price and Max.Price from the features set as they are highly correlated with Price.
Exercise 6. remove rows with NA values.
Exercise 7. partition the data into train(70%) and test (30%) sets using createDataPartition() method.
Exercise 8. On the training set, build a linear model of the Price, using all the numerical features.
Exercise 9. How much variance does the model explain?
Exercise 10. Predict the Price for the samples in the test set.
Exercise 11. Plot the predicted and observed value.
Exercise 12. Define a function which computes RMSE, i.e. given vectors $Y = (Y_1, ..., Y_n)'$ and $Yhat = (\hat{Y}_1, ..., \hat{Y}_n)'$, it computes rooted mean squared differences of coefficients

\[ RMSE = function(Y,Yhat) = \sqrt{\sum_{i=1}^{n} |Y_i - \hat{Y}_i|^2} \]

Exercise 13. Use your RMSE function to compute the prediction error (between predicted and observed Price).
Exercise 14. Use method=lmStepAIC to simplify the model by removing features of low importance based on AIC criteria. Compare the variance explained by the reduced model and the saturated model.
Exercise 15. Use the reduced model to predict the Price for the samples in the test set. Compare the prediction error between the reduced and saturated models.
Exercise 16. On the training set, build a linear svm (method=svmLinear) model of the Price, using Horsepower, Wheelbase and Luggage.room features. Apply the model on test data and compute the prediction error.

Exercise 17. Use cross validation as the resampling method in the training phase (trControl parameter in train function).

Exercise 18. Tune the model parameter using tuneGrid parameter. Plot the result.

Exercise 19. Try other models in caret such as Neural Networks (nnet), Random forests (rf), generalized Boosted Regression (gbm) and other methods.

Exercise 20. Use median cut on Price to add a new feature PriceCategory with two levels high and low to your train and test sets.

Exercise 21. Use featurePlot method to visualize the previous feature set for the PriceCategory classes in the training set.

Exercise 22. Use the previous setup to train a SVM classifier for predicting the PriceCategory.

Exercise 23. Predict the PriceCategory for the test data and evaluate the prediction accuracy using confusionMatrix method.