Please complete the assigned problems to the best of your abilities. Ensure that the work you do is entirely your own, external resources are only used as permitted by the instructor, and all allowed sources are given proper credit for non-original content.

1 Recitation Exercises

These exercises are to be found in: **Introduction to Statistical Learning, 7th Printing (Online Edition)** by *Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani.*

1.1 Chapter 3

Exercises: 1,3,4

1.2 Chapter 4

Exercises: 4,6,7,9

2 Practicum Problems

These problems will primarily reference the *lecture materials and the examples given in class* using **R** and **CRAN**. It is suggested that a *RStudio* session be used for the programmatic components.

2.1 Problem 1

Load the *Boston* sample dataset into **R** using a dataframe (it is part of the **MASS** package). Use **lm** to fit a regression between *medv* and *lstat* - plot the resulting fit and show a plot of fitted values vs. residuals. Is there a possible non-linear relationship between the predictor and response? Use the **predict** function to calculate values response values for *lstat* of 5, 10, and 15 - obtain confidence intervals as well as prediction intervals for the results - are they the same? Why or why not? Modify the regression to include $lstat^2$ (as well *lstat* itself) and compare the R^2 between the linear and non-linear fit - use **ggplot2** and *stat_smooth* to plot the relationship.

2.2 Problem 2

Load the *abalone* sample dataset from the UCI Machine Learning Repository (**abalone.data**) into **R** using a dataframe. Remove all observations in the Infant category, keeping the Male/Female classes. Using the **caret** package, use *createDataPartition* to perform an 80/20 test-train split (80% training and 20% testing). Fit a logistic regression using all feature variables via **glm**, and observe which predictors are relevant. Do the confidence intervals for the predictors

contain 0 within the range? How does this relate to the null hypothesis? Use the *confusionMatrix* function in **caret** to observe testing results (use a 50% cutoff to tag Male/Female) - how does the accuracy compare to a random classifier ROC curve? Use the **corrplot** package to plot correlations between the predictors. How does this help explain the classifier performance?

2.3 Problem 3

Load the *mushroom* sample dataset from the UCI Machine Learning Repository (**agaricus-lepiota.data**) into **R** using a dataframe (**Note**: There are missing values with a *?* character, you will have to explain your handling of these). Create a Naive Bayes classifier using the **e1071** package, using the *sample* function to split the data between 80% for training and 20% for testing. With the target class of interest being *edible* mushrooms, calculate the accuracy of the classifier both in-training and in-test. Use the **table** function to create a confusion matrix of predicted vs. actual classes - how many false positives did the model produce?