**Reminder**: All homework solutions must be written up independently, even though you are allowed to discuss with other students. You need to save your homework assignment in a pdf/html format and upload it with the R code (.R or .rmd) into the Canvas before 11:59pm CT on the due day. No late homework assignment will be graded in any circumstance.

**Problem 1(30 points)**: This exercise involves the **Boston** housing data set.

1. To begin, load in the Boston data set. Since the Boston data set is part of the MASS library, you need to install the MASS package into R/Rstudio and then access the package as follows:

|  |
| --- |
| >install.packages('MASS') #if you do not install before  > library(MASS) |

Now the data set is contained in the object Boston using

|  |
| --- |
| >Boston |

Read about the data set using

|  |
| --- |
| >?Boston |

How many rows are in this Boston data set? How many columns? What do the rows and columns represent?

1. Make *some* pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.
2. Are any of the predictors associated with per capita crime rate? If so, explain the relationship.
3. Do any of the census tracts of Boston appear to have particularly high crime rates? Tax rates? Comment on the range of each predictor.
4. How many of the census tracts in this data set bound the Charles river? (
5. What is the median pupil-teacher ratio among the towns in this data set?

**Problem 2 (30 points)**: The soybean data can also be found at the UC Irvine Machine Learning Repository. Data were collected to predict disease in 683 soybeans. The 35 predictors are mostly categorical and include information on the environmental conditions (e.g., temperature, precipitation) and plant conditions (e.g., left spots, mold growth). The outcome labels consist of 19 distinct classes.

|  |
| --- |
| > library(mlbench)  >data(Soybean)  > ## See ?Soybean for details |

1. Investigate the frequency distributions for the categorical predictors. Are any of the distributions degenerate in the ways discussed earlier in this chapter?
2. Roughly 18 % of the data are missing. Are there particular predictors that are more likely to be missing? Is the pattern of missing data related to the classes?
3. Develop a strategy for handling missing data, either by eliminating predictors or imputation.

**Problem 3 (30 points)**: Brodnjak-Vonina et al. (2005) develop a methodology for food laboratories to determine the type of oil from a sample. In their procedure, they used a gas chromatograph (an instrument that separates chemicals in a sample) to measure seven different fatty acids in an oil. These measurements would then be used to predict the type of oil in food samples. To create their model, they used 96 samples2 of seven types of oils.

These data can be found in the **caret** package using data(oil). The oil types are contained in a factor variable called oilType. The types are pumpkin (coded as A), sunflower (B), peanut (C), olive (D), soybean (E), rapeseed (F) and corn (G). In R,

|  |
| --- |
| >library(caret)  > data(oil)  > str(oilType)  Factor w/ 7 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...  > table(oilType)  oilType  A B C D E F G  37 26 3 7 11 10 2  > |

1. Use the sample function in base R to create a completely random sample of 60 oils. How closely do the frequencies of the random sample match the original samples? Repeat this procedure several times of understand the variation in the sampling process.
2. Use the caret package function createDataPartition to create a stratified random sample. How does this compare to completely random samples?
3. With such a small samples size, what are the options for determining performance of the model? Should a test set be used?
4. One method for understanding the uncertainty of a test set is to use a confidence interval. To obtain a confidence interval for the overall accuracy, the based R function binom.test can be used. It requires the user to input the number of samples and the number correctly classified to calculate the interval. For example, suppose a test set sample of 20 oil samples was set aside and 76 were used for model training. For this test set size and a model that is about 80 % accurate (16 out of 20 correct), the confidence interval would be computed using

|  |
| --- |
| >binom.test(16, 20)  Exact binomial test  data: 16 and 20  number of successes = 16, number of trials = 20, p-value = 0.01182  alternative hypothesis: true probability of success is not equal to 0.5  95 percent confidence interval:  0.563386 0.942666  sample estimates:  probability of success  0.8 |

In this case, the width of the 95% confidence interval is 37.9 %. Try different samples sizes and accuracy rates to understand the trade-off between the uncertainty in the results, the model performance, and the test set size.

**Problem 4 (10 points)**: Briefly discuss what is the bias-variance tradeoff in statistics and predictive modeling.