University of San Diego
SHILEY-MARCOS
SCHOOL OF ENGINEERING

Assignment 5.1

In this assignment, you will apply your knowledge and understanding of the topics presented in the module readings and materials.

Instructions:

1. Answer the assignment questions on the following pages.
2. Create a single document that combines your solutions to all question prompts.

Final Format for Submission:

Submit one assignment file using these guidelines:

- The file must be either a **Microsoft Word** or a **PDF** document file.
- **Do not** combine and submit files into a zipped compressed folder.
- Please use the following naming conventions for your **Word** or **PDF** assignment file:
  File Naming: **`LastName_FirstName_Assignment<Number>.pdf`**
  Example: **`Smith_James_Assignment5.1.pdf`**
- Answer all parts of a question in one place and answer questions in the order they appear in the assignment.
- For programming answers using **R**, it is recommended that the answers are written in R Markdown and 'knitted' to a Word/PDF file.
  - Do not print data frames in your submission, if you want to make a point about data you can use head(df) to print the first few rows.
  - Submit the code used to answer the questions in the assignment with your name on it, answers without code and appropriate results will not get full credit.
  - It is not a professional practice, but in case of difficulty, you can take screenshots of code and outputs and submit them in a Word/PDF file.
- Maximum number of pages should be 15. Any submissions that exceed 15 pages will **not be graded.**

Grading and Scoring:

- Use common sense to gauge the expectations of the answer to the number of points assigned to the question. See the Scoring Rubric in Blackboard for details.

Assignment 5.1 - Questions

Read the questions carefully and respond to all prompts.

1. (60 points) The hepatic injury data set was described in the introductory chapter and contains 281 unique compounds, each of which has been classified as causing no liver damage, mild damage, or severe damage. These compounds were analyzed with 184 biological screens (i.e., experiments) to assess each compound’s effect on a particular biologically relevant target in the body. The larger the value of each of these predictors, the higher the activity of the compound. In addition to biological screens, 192 chemical fingerprint predictors were determined for these compounds. Each of these predictors represent a substructure (i.e., an atom or combination of atoms within the compound) and are either counts of the number of substructures or an indicator of presence or absence of the particular substructure. The objective of this data set is to build a predictive model for hepatic injury so that other compounds can be screened for the likelihood of causing hepatic injury.

   Start R and use these commands to load the data:

   ```
   library(caret)
   library(AppliedPredictiveModeling)
   data(hepatic)
   # use ?hepatic to see more details
   ```

   Apply the following pre-processing steps:

   ```
   # Lump all compounds that cause injury into a "Yes" category:
   #
   any_damage = as.character( injury )
   any_damage[ any_damage=="Mild" ] = "Yes"
   any_damage[ any_damage=="Severe" ] = "Yes"
   any_damage[ any_damage=="None" ] = "No"

   # Convert our response to a factor (make the first factor correspond to the event of interest):
   any_damage = factor( any_damage, levels=c("Yes","No") )
   ```

   The dataframes bio and chem contain the biological assay and chemical fingerprint predictors for the 281 compounds, while the factor variable injury contains the liver damage classification for each compound.

a) Given the size of the dataset and the injury status distribution, describe if you would create a separate training and testing data set? (5 points)
b) Which classification statistic would you choose to optimize for this exercise and why? (5 points)
c) Perform appropriate pre-processing of data and build logistic regression, linear discriminant analysis, penalized logistic regression and nearest shrunken centroids models described in this chapter for the biological predictors and separately for the chemical fingerprint predictors. Which model has the best predictive ability for the biological predictors, and what is the optimal performance? Which model has the best predictive ability for the chemical predictors, and what is the optimal performance? Based on these results, which set of predictors (biological or chemical) contains the most information about hepatic toxicity? (20 points)
d) For the optimal models for both the biological and chemical predictors, what are the top five important predictors? (5 points)
e) Now combine the biological and chemical fingerprint predictors into one predictor set. Retrain the same set of predictive models you built from part (c). Which model yields the best predictive performance? Is the model performance better than either of the best models from part (c)? What are the top five important predictors for the optimal model? How do these compare with the optimal predictors from each individual predictor set? (20 points)
f) Which model (either model of individual biology or chemical fingerprints or the combined predictor model), if any, would you recommend using to predict compounds' hepatic toxicity? Explain. (5 points)

2. (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 a food sample. To create their model, they used 96 samples 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).

   We would like to use these data to build a model that predicts the type of oil-based on a sample's fatty acid percentages.

a) Like the hepatic injury data, these data suffer from imbalance. Given this imbalance, should the data be split into training and test sets? (5 points)

b) Which classification statistic would you choose to optimize for this exercise and why? (5 points)

c) Build linear discriminant analysis, penalized multinomial regression, and nearest shrunken centroids models to this data; predict the fitted models on the training data set and evaluate which model performs best on these data? Which oil type does the optimal model most accurately predict? Which oil type does the optimal model least accurately predict? (20 points)