**Module 5 Exercises**

1. (30 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)

> data(AppliedPredictiveModeling)

> # use ?hepatic to see more details

The matrices bio and chem contain the biological assay and chemical fin- gerprint predictors for the 281 compounds, while the vector injury contains the liver damage classification for each compound.

1. Given the classification imbalance in hepatic injury status, describe how you would create a training and testing set.
2. Which classification statistic would you choose to optimize for this exercise and why?
3. Split the data into a training and a testing set, pre-process the data, and build 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 contains the most information about hepatic toxicity?
4. For the optimal models for both the biological and chemical predictors, what are the top five important predictors?
5. 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 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?
6. 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.
7. (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.

1. Like the hepatic injury data, these data suffer from extreme imbalance. Given this imbalance, should the data be split into training and test sets?
2. Which classification statistic would you choose to optimize for this exercise and why?
3. Of the models presented in this chapter, which performs best on these data? Which oil type does the model most accurately predict? Least accurately predict?
4. (30 points) The “churn” data set was developed to predict telecom customer churn based on information about their account. The data files state that the data are “artificial based on claims similar to real world.” The data consist of 19 predictors related to the customer account, such as the number of customer service calls, the area code, and the number of minutes. The outcome is whether the customer churned.

The data are contained in the C50 package and can be loaded using:

> library(C50)

> data(churn)

> ## Two objects are loaded: churnTrain and churnTest > str(churnTrain)

> table(churnTrain$Class)

1. Explore the data by visualizing the relationship between the predictors and the outcome. Are there important features of the predictor data themselves, such as between-predictor correlations or degenerate distributions? Can functions of more than one predictor be used to model the data more effectively?
2. Fit some basic models to the training set and tune them via resampling. What criteria should be used to evaluate the effectiveness of the models?
3. Use lift charts to compare models. If you wanted to identify 80% of the churning customers, how many other customers would also be identified?