MATH/STAT 4450/8456 Homework 4

Due date: March 28, 11:59 pm

1. (10 points) Textbook exercises.

   (a) (Chapter 9.7) 3, 7

2. (5 points) Redo contest 1. In this homework, you can NOT re-classify the dataset. Use the R code from Team Undergrad 1 ("undergrad1.R") to create some new variables that provide regional summary. You can create more variables if they will give a better result. Only use random forest with the following parameters to train the model: `mtry=4, ntree=1000, nodesize=3, strata=train$class`.

   (a) Use 5-fold cross validation to choose another parameter `sampsize`, which is a vector of length 8, for example, `c(5,6,5,4,5,6,4,7)`. They are the number of observations that will be randomly drawn from each class.

   (b) Read the help document of the `combine` function in the package `randomForest`. Use this function to combine a few random forest models, these models can have slightly different parameters.

   (c) Now set a random seed before each random forest model you combined. For example,

   ```
   set.seed(1)
   rf1 = randomForest(...)
   set.seed(2)
   rf2 = randomForest(...)
   set.seed(3)
   rf3 = randomForest(...)
   set.seed(4)
   rf4 = randomForest(...)
   set.seed(5)
   rf5 = randomForest(...)
   rf = combine(rf1,rf2,rf3,rf4,rf5)
   ```

   (d) Use the combined models built in parts (b) and (c) to predict on the test set. The true classes ("result.csv") have been provided. Compare the accuracy rate with your Kaggle submission.

3. [Additional problem only for graduate students. Undergraduate students will receive 2 extra credits if solving the additional problems correctly.] (5 points)

   Make a set of "boosted" predictions on the response variable `Sales` for the `Carseats` dataset in the `ISLR` package, by first fitting the random forest, then correcting the residuals using a support vector machine with the shrinkage parameter $\lambda = 0.1$, then correcting the residuals using a KNN model with shrinkage $\lambda = 0.1$. In each step, use 5-fold cross validation to tune the parameters: mtry (m = 3, 5, 10) in the random forest, gamma (γ= 0.01, 0.1, 1, 10) in the support vector machine, and k (k = 1, 5, 10, 20) in KNN. At the end, compare the three sets of residuals by some plots.