Math 5364 Homework 5

1. Create a function called `splitdata` that splits data into training and test sets.

   - The inputs should be a dataframe `data` and a number `trainfrac` between 0 and 1, representing the fraction of data that should be put in the training set.
   - The function should return a list with components `traindata` and `testdata`, which are the training and testing sets.
   - For example, the following code should split `iris` into 70% training and 30% test data.

     ```
     splitlist=splitdata(iris,.7)
     traindata=splitlist$traindata
     testdata=splitlist$testdata
     ```

   - Note that the following code will **not** work.

     ```
     traindata=splitdata(iris,.7)$traindata
     testdata=splitdata(iris,.7)$testdata
     ```

     The problem with this code is the random splitting will occur twice, so the training and test sets will not match.

2. Download the file `wdbc.data` from the Breast Cancer Wisconsin (Diagnostic) data set on the UCI Machine Learning Repository. Give a general description of the data, and determine what columns 1, 2, 6, 16, and 26 of this data represent.

3. (a) Now that we know what column 1 is, we know that we don't want any algorithm using this column to make predictions, so remove it from the data.
   (b) Use `splitdata` to split the data into 70% training and 30% test data.
   (c) Find `colSums` and `dim` of the original data and of the training and test data to verify that the splitting was done correctly.

4. (a) Use `rpart` to fit a tree called `tree1` to this data, plot it, and calculate its training and test error rates.
   (b) Use `ctree` to fit a tree called `tree2` to this data, plot it, and calculate its training and test error rates.
   (c) Intuitively, does there appear to be a statistically significant difference between the accuracies of `tree1` and `tree2`?
   (d) Test whether the difference in accuracies is statistically significant.

5. Estimate the accuracy of `tree1` using the following types of cross-validation.

   (a) 10-fold cross-validation
   (b) 20-fold cross-validation
   (c) Leave-one-out cross-validation
   (d) Delete-d cross-validation with $d = 20$ and $m = 100$.
   (e) The bootstrap with $b = 100$.