Statistics 6861/4861 Assignment 6

This is the Final Project for this class.

Due: December 8

200 points

1. Modify the ticket.line function you wrote for assignment 3. This time there are still n people with a $5 bill, and n people with a $10 bill. But this time the ticket taker may start with some $5 bills in hand. The new function prototype is
   ticket.line <- function(n,change = 0,sim.length).
   Here n and sim.length have the same meaning as in assignment 3. The number of $5 bills the ticket taker initially has is given by the change parameter. Note that change defaults to 0. If change = 1, then the ticket taker initially has one $5 bill. Obviously change must be a non-negative integer.

2. For this problem your inputs are a training data set and a validation set. Both data sets have coordinates in n-space and the corresponding classification for each observation in the data set. In practice, one has a set of points for which the classification is known. We divide this set into two sets: one is a training data set and one is a validation set. We then use the training set to devise an algorithm to predict which classification should be assigned to new observations. We can then run the algorithm on the data we have in the validation set. Since we know the correct classification for these points, we can determine the error rate of our algorithm had these data been new observations.

   In this problem you will write code to test the k-nearest-neighbors (knn) regression algorithm. The algorithm works as follows: we are given a new observation and an odd number, k. We find the k nearest neighbors from the training set for the new observation. We then assign the new observation to whichever classification the majority of the k-nearest-neighbors has. For example, if k = 5, and three of the nearest neighbors have a classification of 2, then the new observation is assigned to class 2.

   Write two functions. The first has prototype
   k.nn <-function(k,v.data,t.data)

   Here k is a positive integer to denote the numbers of nearest neighbors to find. The v.data parameter is a matrix. Each row contains the coordinates of an observation. If v.data has 100 observations in 2 dimensional space, then v.data is a matrix of dimension 100x2. The t.data is formed similarly. We must have ncol(t.data) equal to ncol(v.data).
   The output is a nrow(v.data) by k matrix. Each row of the output contains the k nearest neighbors for the corresponding observation in v.data. Store the observation number (or row number) from t.data in the output matrix.

Example: Suppose the first observation of v.data has as its 3 nearest neighbors the 10^{th},23^{rd}, and 45^{th} observation from t.data. If k = 3, then the first row of the output matrix will be 10,23,45.

The second function has prototype
vote <- function(class.id,knn.out)
Here knn.out is the output of your k.nn function, and class.id is a vector containing the class id of the training data set. The output is a vector of length nrow(knn.out).
Example:
The first row of knn.out is the vector (10,23,45). The classification for the tenth and twenty-third training observation is 2, and the classification of the forty-fifth training observation is 1. Then the first element of the output vector is 2.

I have placed two files on Blackboard under Course Materials. These are data frames and can be read with the read.table function. Download them from Blackboard and read them into the R environment. Use your knn and vote functions to determine the classification to the validation data points given by the knn regression algorithm. Then compare them to the actual (known) classification recorded in the validation data frame. Then print the coordinates of the validation observations that were misclassified by the knn regression algorithm. Run this procedure for k =3, k = 9 and k = 23.

Save the console output from these runs to a data file. Print out the data file.

What to turn in:

(1) Email a file to me containing the new version of the ticket.line function, your knn function and your vote function. The first line of this file should contain
name = “Albert Einstein”. The file name is AlbertEinsteinAssign6.R for a student with the name Albert Einstein. This file should not contain any other data or runs of tests for your code.

(2) A print out of your console run of the knn regression algorithm. This print out should begin with your knn and vote function definitions. It should include the read.table command you used to read the files “train.set.1.dat” and “validate.set.1.dat”. After reading the data, do the knn regression algorithm for k =3, k = 9 and k = 23. Print to the console the coordinates of the validation points that were misclassified. Of course the print out should have your name at the top. This item is a hard copy.