Math 5364 Homework 6

1. This problem will apply knn to the wdbc.data data set.
   
   (a) Standardize the data, and verify that the column means are equal to zero and the column standard deviations are equal to one afterwards.
   
   (b) Split the data into 70% training and 30% test data.
   
   (c) Calculate the test error rate for predicting breast cancer diagnosis using knn with $k = 3$, and find a 95% confidence interval for this error rate.
   
   (d) Compare the test error rates of knn with $k = 3$ and rpart, and determine if there is a statistically significant difference between them.

2. (a) Use knn.cv to estimate the error rate when $k = 3$.

   (b) Use knn.cv to find the value of $k = 1, 2, \ldots, 10$ that minimizes the error rate. Let the optimal value be $k_0$.

   (c) Estimate the error rate of knn with $k = k_0$ using 10-fold cross-validation.

   (d) Estimate the error rate of knn with $k = k_0$ using the bootstrap with $b = 100$.

3. Bonus: Write your own function for performing $k$-nearest neighbors classification and compare the results you obtain with knn. Here are some guidelines.

   (a) To keep things simple, you can assume there are only two class labels and $k$ is odd, so you don’t have to worry about ties. On the other hand, breaking ties isn’t too hard. It might be interesting to figure out how to do it (hint: use runif).

   (b) A good starting point would be to write a function called distancematrix, which accepts matrices $X_{train}$ and $X_{test}$. It returns a matrix $D$, such that $D_{ij}$ is the Euclidean distance between the $i$th row of $X_{train}$ and the $j$th row of $X_{test}$. Once you have the distance matrix $D$, finding nearest neighbors and so on should be pretty easy.