B565: Homework 1

1. The "mlbench" R package contains a number of datasets used for Machine Learning benchmarking. Install this package with

   ```
   > install.packages("mlbench")
   ```

   When you want to access one of these datasets in your R source file, you need the command

   ```
   > library("mlbench")
   ```

   If you wanted to load, for instance, the "Ionosphere" datset, you would also need the command

   ```
   > data("ionosphere")
   ```

   as we have done before. You can see that the 35th attribute of this dataset gives a classification of "good" or "bad," thus dividing the dataset into two classes. For each of these classes compute the empirical covariance matrix, and use these to compute the Mahalanobis distance from each point to the two class means. A simple-minded classifier might associate each point with the class having the smaller Mahalanobis distance. Classify each point according to this rule, and tabulate the "confusion matrix." That is, create a 2x2 matrix where the i, j entry is the number of examples from class i (good or bad), which were classified as type j (good or bad).

2. Just as we have created Unif(0,1) random numbers in R, we can also compute Normal(0,1) numbers — that is, random numbers from a normal distribution with mean 0 and variance 1. R uses the command "rnorm" to do this. Sometimes these are called "standard normal" numbers. Consider dimensions $d = 2, 3, \ldots, 50$. For each dimension create 1000 random vector pairs, each with coordinates that are standard normal, and compute the cosine distance between each pair.

   (a) Create a graph that shows the mean and standard deviation of this cosine distance, as a function of the dimension, d.

   (b) Summarize in words what your graph says and explain this behavior.

3. Create a random sample $x_1, \ldots, x_n$ for $n = 1000$ from the Exponential(1) distribution using the R command "rexp" (analogous to runif and rnorm).

   (a) Create a plot of the empirical cumulative distribution function for this sample as follows. For each x_i in your sample, plot a point (x_i, y_i) where y_i is the empirical estimate of the probability of being less than or equal to x_i. That is, y_i is the proportion of your samples that were $\leq x$. Show the graph with a smooth curve, rather than an a sequence of points.

   (b) Create a new data set by transforming each point by F where F is the cdf you computed in the previous part. That is, your new data are $F(x_1), \ldots, F(x_n)$. What is the distribution of these new samples?

   (c) Suppose you do the same experiment using random numbers from a different distribution. Argue that the result of the *transformed* points will still have the same distribution as in the previous part.

4. The file "time_series.csv" on the Canvas site contains 200 time series generated from one of four different models. Each time series (each row of the .csv file) has length 200. The models are unknown to you and are chosen randomly for each time series.

   (a) Visualize these different time series in a way that allow you distinguish the 4 models. Submit a plot that illustrates the fundamental ways in which the 4 categories differ. Characterize this difference in words.

   (b) Derive two features that effectively separate the time series into four categories when shown in a scatterplot. There may by many ways to do this. Submit your scatterplot as well as the code that generates the features.

5. Suppose a pair of critics rate a collection of $N = 1000$ movies giving a numerical score for each movie. A reviewer cannot give two movies the same score. One critic is generally more positive than the other, so these scores are reported as *percentiles*. That is, each movie gets a score in $[0, 1]$ from each reviewer which is the overall fraction she judges no better than the current movie. Denote the reviewers' percentile scores as x_i, y_i, $i = 1, \ldots, N$.

   (a) Sketch a reasonable scatterplot for the the pair (x_i, y_i)

   (b) Would it be reasonable to model the two reviewers scores as independent random variables? Why or why not?

   (c) Is it possible that one reviewer always gives a higher percentile score than the other reviewer? Either give an example showing this is possible or argue that it is impossible.

   (d) Is it possible that difference in percentile scores, $x_i - y_i$, is, on average, greater than 0? Either give an example showing this is possible or argue that it is impossible.

   (e) Is is possible that $x_i > y_i$ for all but 1 movie? Either give an example showing this is possible or argue that it is impossible.

6. This problem is on distance functions.

   (a) Suppose you are given an $n \times n$ matrix of positive numbers, D. Is it always possible to find a collection of n points (x_i, y_i), $i = 1, \ldots, n$ so that the Euclidean distance between (x_i, y_i) and (x_j, y_j) is D_{ij}? Why or why not?

   (b) Suppose we create n unobserved points (x_i, y_i), and create the $n \times n$ distance matrix, D, where D_{ij} is the Euclidean distance between (x_i, y_i) and (x_j, y_j). Devise a numerical algorithm to find the two dimensional points $\{(x_i, y_i)\}$. You do not need to implement your algorithm, but rather describe how it would work.

7. In your book, problem 3.10.