Please sign on the right of the Stanford logo to signify you will abide by Stanford's honor code. Please write your name at the top of page 2. For this exam you agree not to use any electronic device (phone, laptop or calculator) and only your two pages of notes.

Signature:
Please write your name on the second page and do not write any answers on this first page (it will be archived but not graded). Please make your answers brief and do not write on the reverse of the pages (you may use the reverse of the pages as rough drafts, but it will not graded).
1. (15 pts)
   (a) Figure 1 shows the output from a data analysis performed on a matrix of ordinal variables. What is this figure called, what R function would generate it and what does it show? (explain its three main elements ie the two trees and the central rectangle)

   (b) Interpret briefly what you see for this particular data.

2. (10 pts)
   (a) What two R commands would provide Figure 2 on the first page.

   (b) What is the main problem with doing the analysis this way?

   (c) How would you fix it?

3. (20pts) Write a R function that takes as input a data frame \( \text{adf} \) with 1200 rows and 21 columns, the last column being a nominal variable giving the group label. A tuning parameter \( k \) whose default value is 10 is added as a second argument to the function. The function should compute the nearest neighbors from the training set for the 200 test points, use these nearest neighbors to predict the group label and compare the prediction with the truth. The function should:

   - Split randomly the data into two (same number of columns):
     \[
     \text{train} \text{ with 1000 rows} \\
     \text{test} \text{ with 200 rows.}
     \]
   - Compute the \( k \) nearest neighbors (from the training set) of the 200 points in the test data, use these to infer the label of the 200 points, call this vector of size 200 \( \hat{y} \).
   - Return \( \hat{y} \) and the misclassification rate in a list object \( \text{result} \) containing two components, one named "\( \hat{y} \)" and the other "misclass".
4. (22 pts) Choose TRUE or FALSE on the following (circle the correct answer)

(a) Ward's method is used in classification. TRUE or FALSE.

(b) K-means is well adapted to clusters of unequal densities. TRUE or FALSE.

(c) The gap statistic can be used to test separation of eigenvalues in PCA. TRUE or FALSE.

(d) Random forests use complete trees on subsets of the variables as weak learners. TRUE or FALSE.

(e) A dataframe with both continuous and nominal variables can be analyzed using MDS. TRUE or FALSE.

(f) An itemset A contained in an itemset B has to have its support at least as large as that of B. TRUE or FALSE.

(g) Data imputation is an iterative process for reconstructing missing values. TRUE or FALSE.

(h) An estimator with a low breakdown point is robust. TRUE or FALSE.

(i) In R a dataframe is a matrix. TRUE or FALSE.

(j) The interquartile range is a robust estimate of variability. TRUE or FALSE.

(k) Three binary variables can be summarized by a VCD plot. TRUE or FALSE.

5. (18 points)

(a) Give the k-means algorithm in few lines of pseudocode.
(b) What are the two important steps in the E-M algorithm?

(c) Describe one similarity and one difference between these algorithms.

6. (15 pts) You are given a dataset that needs to be analyzed. Show a binary decision tree that would help you decide what method to use on the data. Make sure the tree has at least six methods at its leaves.