Due: Friday, March 1, by 11:59pm (hard deadline, solutions posted shortly after)

Note: For handwritten questions, please photo/scan and incorporate into the output (html/pdf) of the Rmarkdown file. See this link for a rundown on how.

1. On Canvas in this assignment area, you will find a data set called "car93". It is a cleaned up version of the "Cars93" data in the "MASS" library.

   (a) Perform a principal components analysis on the numeric variables within the car93 data set. Provide a summary of the fitted model and a biplot. Ensure you scale the data.

   (b) Interpret the loadings of the first principal component.

   (c) Interpret the loadings of the second principal component.

   (d) How many principal components should be kept...

      i. according to the Kaiser criterion?

      ii. if we wish to retain at least 90% of the variance in the data?

      iii. according to the scree plot?

   (e) Keep the components suggested by the Kaiser criterion and...

      i. perform LDA (from MASS library, with built-in leave-one-out cross-validation) with our response being "Small" or "Not Small" for the "Type" of car and the predictors being the components retained. What is the cross-validated logloss of this model?

      ii. perform LDA (with built-in leave-one-out cross-validation) using all categories from the original "Type" variable as the response. What is the logloss of this model?

   (f) Do the results from the above classification runs approximately match the discussion surrounding the interpretation of the first and/or second principal components? Explain.

2. We will use the `banknote` data from the `mclust` library. Remove the `Length` and `Status` variables and perform hierarchical clustering. The `Status` variable is our likely target for benchmarking clustering methods.

   (a) Explore the data a bit. What is an appropriate distance measure to use, and why? Note: no need to adjust your answer if later questions make you second-guess your initial rationale.

   (b) Use the distance measure from above and apply hierarchical clustering with all three linkage types discussed in class. Provide the dendrograms for each.

   (c) Which linkage method would you choose, or do they all provide a similar outcome?

   (d) Give the classification table that results from cutting your chosen dendrogram at an appropriate level, and the misclassification rate, both with reference to the true `Status` variable.

   (e) Apply k-means using $K = 2$ and `set.seed(632)` prior to the analysis (for consistency) on the scaled data. Provide a classification table and the misclassification rate.

(f) Apply k-means using $K = 2$ and `set.seed(632)` prior to the analysis (for consistency) on the raw data. Provide a classification table and the misclassification rate. Give rationale as to why this performs better than the scaled data.

(g) Overall, what does the (generally) strong performance of unsupervised methods signify for this data set?

3. Find `lots.Rdata` on Canvas. There are two objects: `clusts` are the true groups and `datmat` is the data. This is a bivariate simulation with 20 groups under appropriate assumptions for k-means.

   (a) Provide a scatterplot with the observations coloured according to their real groups.

   (b) Use `set.seed(1026)` and run kmeans with k=20. Report the adjusted Rand index (function available in `mclust` library) between the clustering results and the true groups.

   (c) Use `set.seed(6201)` and run kmeans with k=20. Report the adjusted Rand index (function available in `mclust` library) between the clustering results and the true groups.

   (d) Use `set.seed(1026)` and run kmeans with k=20 and nstart=1000. Report the adjusted Rand index (function available in `mclust` library) between the clustering results and the true groups.

   (e) Use `set.seed(6201)` and run kmeans with k=20 and nstart=1000. Report the adjusted Rand index (function available in `mclust` library) between the clustering results and the true groups.

   (f) What, if anything, do you find interesting among all the above results?

4. Find `bsim.Rdata` on Canvas. This is data I simulated with one Y response variable and 9 predictors. For the supervised aspect, you are only permitted to fit linear models via A SINGLE `lm` function, and you may not have more than 25 coefficients estimated in the model. Using unsupervised methods on the predictors in tandem with linear modelling, find a model with an R^2 and adjusted R^2 both greater than 0.99 when predicting for the entire data set.

5. Here are some distances between 4 observations. Submission for this question should be handwritten.

```
     1    2    3
2 2.98
3 4.78 1.91
4 6.16 3.26 1.46
```

   (a) Perform (agglomerative) hierarchical clustering using complete linkage for the above distance matrix *by hand*.

   (b) Sketch a dendrogram for the process from part a).

   (c) How many groups does the dendrogram suggest?