Due: Saturday, February 23, by 11:59pm

Note: For handwritten questions, please scan (or photo-scan) and incorporate into your answer.

1. We will use the `bank` data again from the `gclus` library. We will again not use the "Bottom" variable, and this time we will perform hierarchical clustering. The `Status` variable should, of course, be removed as well since we are doing unsupervised learning.

   (a) What is an appropriate distance measure to use, and why?

   (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?

2. Find `lots.Rdata` on github. 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(461)` and run kmeans with k=20. Report the adjusted Rand index (function available in `mclust` library.

   (c) Use `set.seed(41)` and run kmeans with k=20. Report the adjusted Rand index (function available in `mclust` library.

   (d) Use `set.seed(461)` and run kmeans with k=20 and nstart=1000. Report the adjusted Rand index (function available in `mclust` library.

   (e) Use `set.seed(41)` and run kmeans with k=20 and nstart=1000. Report the adjusted Rand index (function available in `mclust` library.

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

3. Pull the mickey mouse simulation code from lab and regenerate the associated data. Load the `mclust` library and run `Mclust` on the data under all default settings. Provide a scatterplot with groups discovered by mclust given different colours. Is the result more sensible than k-means results that were seen in lab? Why or why not? It may help to reference the chosen model's constraints on the covariance matrix.

4. Find `asim.Rdata` on github. 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 the `lm` function. 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.

5. Handwritten question: Below is a (condensed) pair-wise distance matrix for 4 observations created in R.

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

   (a) Manually perform hierarchical clustering using complete linkage on this distance matrix.

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

   (c) How many groups does the dendrogram suggest?