HOMEWORK 5
BIOSTATISTICS 825
DUE NOVEMBER 26TH, 2023

1. **(40 points)** The data we'll use for this problem originated in a Kaggle competition. The goal was to predict success or failure of a grant application based on information about the grant and the associated investigators. It is very important to read about the variables which you can do here. Think carefully about how to include them in your model. You *may* use listwise deletion for some missing values (but be careful with this).

   You will train a neural network using the data `unimelb_training.csv`, then you will predict for the `unimelb_test.csv` which does not include the outcome (I do have the outcomes to this data). Here's what you'll hand in:

   - Turn in (via R markdown or copy paste) the code you used to fit your neural network. This code is work 15 points.
   - An estimate of your misclassification rate on the test data (worth 10 points)
   - An estimate of the 95% CI for the misclassification rate (worth 5 points)
   - A csv file that will have two rows – 'Proposal.num', and 'Pred.Grant.Status' – for all observations in the test data. The Pred.Grant.Status variable should be 0, 1, or NA. There should be a row for each observation in the test data (even if they have missing data). Your score on this portion is worth 10 points, and it will be based on the number of predictions you get correct (total number not percentage).

2. In this problem, you are asked to do cluster data using finite mixture models and K-means on simulated data and deal with some practical issues, such as how to choose K and whether to do standardization or not before running the algorithm.

   (a) **(2 points)** Simulate 2-dimensional data from four clusters, each of which is specified by a Gaussian distribution with the same covariance matrix and different means. Using the code demonstrated in class to generate the data, though change the seed and make the variance of the means equal to 1 (the means are randomly generated).

   (b) **(10 points)** Using Gaussian mixture model clustering, plot K (ranging from 2 to 6) against BIC value for two different covariance structures in the `mclust` package. Find how many points are in wrong clusters (you can also use the RAND index). Dont forget to record the sample covariance matrix of the simulated data at this point.

   (c) **(5 points)** Using K-means clustering, plot K (ranging from 2 to 6) against within cluster sum of squares. Choose an appropriate K from the plot and argue why do you choose this particular K. Run the function K-means with chosen K and plot the clustering result. Write down how many points are in wrong clusters (you can also use the RAND index). Dont forget to record the sample covariance matrix of the simulated data at this point.

(d) **(5 points)** Now choose the number of clusters using the gap statistic with K-means. If the results differ from what you found in (b), run the function K-means with chosen K and plot the clustering result. Which method do you prefer?

(e) **(3 points)** Standardize the data. Now plot K against within cluster sum of squares from K-means on the standardized data and choose an appropriate K from the plot. Is the K you choose the same as the one you chose in parts (b) and (c)? Run the function K-means with all chosen K's (if they differ) on the standardized data. Compare the clustering results with the one you obtained in parts (b) and (c).

3. (Exercise 14.2 from "Elements of Statistical Learning")

   Consider a mixture model density in p-dimensional feature space,

$$g(x) = \sum_{k=1}^{K} \pi_k g_k(x)$$

   where $g_k = N(\mu_k, \mathbf{L} \cdot \sigma^2)$ and $\pi_k \geq 0 \; \forall k$ with $\sum_k \pi_k = 1$. Here, $\{\mu_k, \pi_k\}$ for $k = 1, 2, \ldots, K$ and σ^2 are the unknown parameters.

   Suppose we have data $x_1, x_2, \ldots, x_N \sim g(x)$ and we wish to fit the mixture model.

   (a) **(5 points)** Write down the log-likelihood of the data

   (b) **(5 points)** Derive an EM algorithm for computing the maximum likelihood estimates (see Section 8.1).

   (c) **(5 points)** Show that if σ has a known value in the mixture model and we take $\sigma \to 0$, then in a sense this EM algorithm coincides with K-means clustering.

4. **(20 points)** (Question 12.3 from "Modern Multivariate Statistical Techniques" by Izenman) Cluster the `primate.scapulae.txt` data using single, average, and complete-linkage agglomerative clustering methods. Find the five-cluster solutions for al three methods, which allows comparison with the true primate classifications. Find the misclassification rate for all three methods. Show that the orderings of the methods in terms of their misclassification rates (you can also use the RAND index). (Note: "Modern Multivariate Statistical Techniques" is electronically available through the library.)