Assignment 3

Problem

1) (**4 marks**) Consider the following Union territory capitals in India i.e. Leh, Pondicherry, New Delhi, Chandigarh, Srinagar, Daman, Port Blair, Kavaratti. Take the dissimilarity measure as the aerial distance between two cities. Plot the dissimilarity matrix among these cities. On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering using average linkage and single-linkage. Be sure to indicate on the plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram. In R, hclust function can be considered for hierarchical clustering. Choose the number of clusters based on the elbow method.
Can you apply k-means clustering in this dataset? Explain.

2) (**4 marks**) Generate 3 sets of 100 observations each from bi-variate normal distribution with means $\mu_1 = (0, 0)$, $\mu_2 = (2, 2)$ and $\mu = (-2, 2)$ with Variance-covariance matrix given by $\lambda_1 I$, $\lambda_2 I$ and $\lambda_3 I$ with $\lambda_1 = 0.1$, $\lambda_2 = 0.2$ and $\lambda_3 = 0.3$. Combine the data and plot the data. Perform K-means clustering with K=1,2,3,4,5,6 on the dataset after scaling and without scaling. Choose K based on the elbow method in both the cases.

3) (**6 marks**) The family of Exponential distribution $E(\alpha)$ has pdf

$$f(x|\alpha) = \begin{cases} \alpha \exp(-\alpha x) & \text{if } x \geq 0 \\ 0, & x < 0 \end{cases},$$

with $\alpha > 0$.
a) Consider the 100 observations given in the second column of "Prob1.csv". Let, these observations be independent realisations of a random variable Y which follows Exponential distribution with unknown α.
Write a function in R/python which computes the log-likelihood for exponential distribution based on the generated data and then optimise the function to find the MLE. Also, find the asymptotic variance of the MLE. In R, use the function optim with the method "L-BFGS-B" for optimisation using gradient descent. Check the argument hessian in the optim function for variance. The inverse of the observed information matrix is the estimate of variance.
b) Use bootstrapping to find the 95% confidence interval of the MLE.

In this question, all the computations need to be performed using the program.

4) (**8 marks**) Take the ptitanic dataset available in R. Randomly sample 800 observations from the dataset. Fit a classification tree with maximum tree depth as 7 and split based on gini index. Prune the tree with complexity parameter 0.02. Which are the two most important variables? Now, fit a random Forest to the same dataset with 250 trees and number of randomly selected predictor at each split to be 2. Impute the missing covariates by its median. Report the variable importance. Also, report the specificity and sensitivity

considering the positive class in the response to be the class named "died".
The ptitanic dataset needed for Q8 is in the rpart.plot package of R. You need to import the package then use the command data("ptitanic") to get the data.

5) (**8 marks**) In this exercise, predict the number of applications received using the other variables in the College data set available in the ISLR2 package in R.
(a) Split the data set into a training set and a test set in the ratio 1:1.
(b) Fit a linear model using least squares on the training set, and report the test error obtained.
(c) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Choose the λ where the cross validation error is minimum. Report the test error obtained.
(d) Fit a lasso model on the training set, with λ chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates.