# MA308: Statistical Calculation and Software

## Assignment 4 (Dec 07, 2021 - Dec 29, 2021)

**4.1** [48 points] For the *"planets"* dataset from *HSAUR3* package,

- **(a)** [8 points] Apply single linkage, complete linkage and average linkage hierarchical clustering to the *planets* data.
- **(b)** [8 points] Construct two three-dimensional drop-line scatterplot of the *planets* data in which the points are labelled with a suitable cluster label, use K-means (K=3 and K=5) method for clustering.
- **(c)** [8 points] Write a *R* function to fit a parametric model based on two-component normal mixture model for the *eccen* variable in the *planet* data. (Hint: refer to the "Mixture distribution estimation" section in Chapter 6)
- **(d)** [8 points] In fact, package *mclust* offers high-level functionality for estimating mixture models, apply *Mclust* to estimate normal mixture model for the *eccen* variable in the *planet* data .
- **(e)** [8 points] Implement principal component analysis on the *planet* data, find out the coefficients for the first two principal components and the principal component scores for each planet.
- **(f)** [8 points] Apply K-means (K=3) clustering to the first two principal components of the *planet* data. Compare with the K-means (K=3) clustering result in (b).

**4.2** [52 points] For the *"Default"* dataset from *ISLR* pacakge, we consider how to predict *default* for any given value of *student*, *balance* and *income*.

- **(a)** [12 points] Split the sample set into a training set (70%) and a validation set (30%). Fit a multiple logistic regression model (*default* $\sim$ *student* + *balance* + *income*) using only the training observations. Obtain a prediction of default

status for each individual in the validation set by computing the posterior probability of default for that individual, and classifying the individual to the *default* category if the posterior probability is greater than 0.5. Compute the confusion matrix and the validation set error, which is the fraction of the observations in the validation set that are misclassified.

**(b)** [7 points] Apply Classical Decision Tree on the training data in (a). Use the *plotcp()* function to plot the cross-validated error against the complexity parameter and choose the most appropriate tree size.

**(c)** [7 points] Apply Conditional Inference Tree on the training data. Obtain a prediction in the validation set and compute the confusion matrix

**(d)** [12 points] Write down the algorithm for a random forest involves sampling cases and variables to create a large number of decision trees. Implement random forest algorighm based on traditional decision trees and conditional inference trees respectively. Use the random forest models built to classify the validation sample, compute the confusion matrix and compare the predictive accuracy of the two models.

**(e)** [14 points] Fit a support vector machine classifier to the *Default* dataset with $gamma = 1$ and $cost = 1$. Compare the *sensitivity, specificity, positive predictive power and negative predictive power* of the svm, Conditional Inference Tree, random forest and logistic regression classifiers.