Behavioural Segmentation

Objectives:

This assignment aims to achieve the following general learning objectives:

- To gain experience using model based clustering at an introductory level;
- To critically evaluate specific clustering techniques;

Plagiarism:

The Department of Statistics regards plagiarism as a serious offence. Your submission will be subject to plagiarism checks and appropriate action will be taken against offending parties. You may also refer to the the Library's website at *www.library.up.ac.za/plagiarism/index.htm* for more information.

Report,hand in and mark allocation:

Marks are awarded as indicated in each question.

Submit only the report. No additional files of any sort should be submitted. Do not submit files in any other format other than PDF. Failure to follow any of these instructions will result in a zero mark for the assignment.

Upload the report, PDF file, (named s99999999.pdf, where 99999999 is your student number). **Ensure that you name and student number is clearly indicated on the front page of your assignment.**

Suggested sources

1. Video as discussed in class
2. Class discussion
3. There are many sources on the internet that can also assist.

Software:

You can use any appropriate software. Use the mixtools package (normalmixEM) when using *R* for mixture modelling.

Data

Consider the data, *data asg2.csv* , on ClickUp. This file contains the duration (Time) of 250 calls received at a call centre. You want to model the length of time required to answer a random call received. Each of these calls can be classified as easy, specialized or difficult. The data contains 2 variables, the **type** of call and the **duration**. In the actual observed data only the **duration** will be available, but for the purposes of this assignment **type** is also given. In your modelling you should not use the **type**. Type will only be used in the evaluation of your model.

1: Question 1

Do a k-nearest neighbour, k-NN, **density** based clustering on the variable $Time$ using mode seeking. Use the following algorithm in answering the question.

- Estimate the density of the variable X using the k-NN density estimator below.

$$\hat{f}_{knn}(x_i) = \frac{1}{||x_i - x_k||^2}$$

  Thus the density at point x_i, $\hat{f}_{knn}(x_i)$, is the reciprocal of the squared distance to the $k-th$ nearest neighbour x_k.

- For each of the observations x_i
  - Define a pointer to the observation within the k-nearest neighbours of x_i with the highest k-NN-density.
  - Repeat the process by following pointers from the initial pointer until a pointer that points to itself is found. This will be taken as a local mode of $\hat{f}_{knn}$.
- Assign each point, x_i, that converged to the same mode to the same cluster.

1. Give the k-NN density estimate for the observed data. Overlay the graph of the density estimate on a histogram of the data.
2. Use the algorithm above to determine the relevant mode(s), comparing different values of k. How many clusters, **c**, does your cluster solution suggest.
3. Give the estimated means, variances and proportion of observations in each of the clusters.
4. Consider the attached paper on $k-nn$ mode seeking. Add to your solution the co-presence ensemble section. Compare your results to those obtained in (2).

(**20**)

2: Question 2

Estimate a mixture of Gaussians model with three components.

1. Give the estimated mixture model with all estimated parameters.

2. Overlay the graph of the mixture density estimate on a histogram of the data. (**15**)

3: Question 3

1. Compare the results of Questions 1 and 2 with respect to the classification accuracy of each of the models. You should compare the classification based on your models to the variable *Type* in the data. This is the only instance where you are allowed to use the variable *Type*. (**5**)

2. Use a co-presence measure to evaluate the degree of correspondence of the two solutions. (**10**)

Total [**50**]