301115 Advanced Statistical Methods

Assignment 1: Case Study

Spring 2019

Word/page limit: 2000 words, 12 pages

In this assignment there are 3 questions (split into various parts). For each question you should draw appropriate plots and summary tables needed to present the results at each stage of the analysis. In addition to your plots and summary tables, for each question you should write at least two paragraphs (of at least 150 words each) describing the analysis you have just performed. At least one paragraph should be written for a statistician, with a focus on the choice of methods used in the analysis. You can assume this statistician is familiar with all of the methods used in weeks 1-6 of this unit, so you do not need to explain how the methods work. There should however be enough detail so that the statistician could reproduce your analysis without having to read your R code. At least one paragraph should be written for a non-statistician subject-matter expert, with a focus on summarising and interpreting the results and providing conclusions. You should avoid the use of overly technical terms in these non-statistician paragraphs.

Your assignment can be submitted as either a PDF or a Word document. It is advisable that your assignment includes all of the code used to perform your analysis in an appendix (not within the body of the report), so that partial credit can be awarded in case of error. Do not use raw R (text) output to present your results, instead present all results using appropriate plots, summary tables and the text of your report. Submission is due by the end of Monday for week 11. (30th Sep 2019). Submission is by the vUWS online system.

Assignment 1

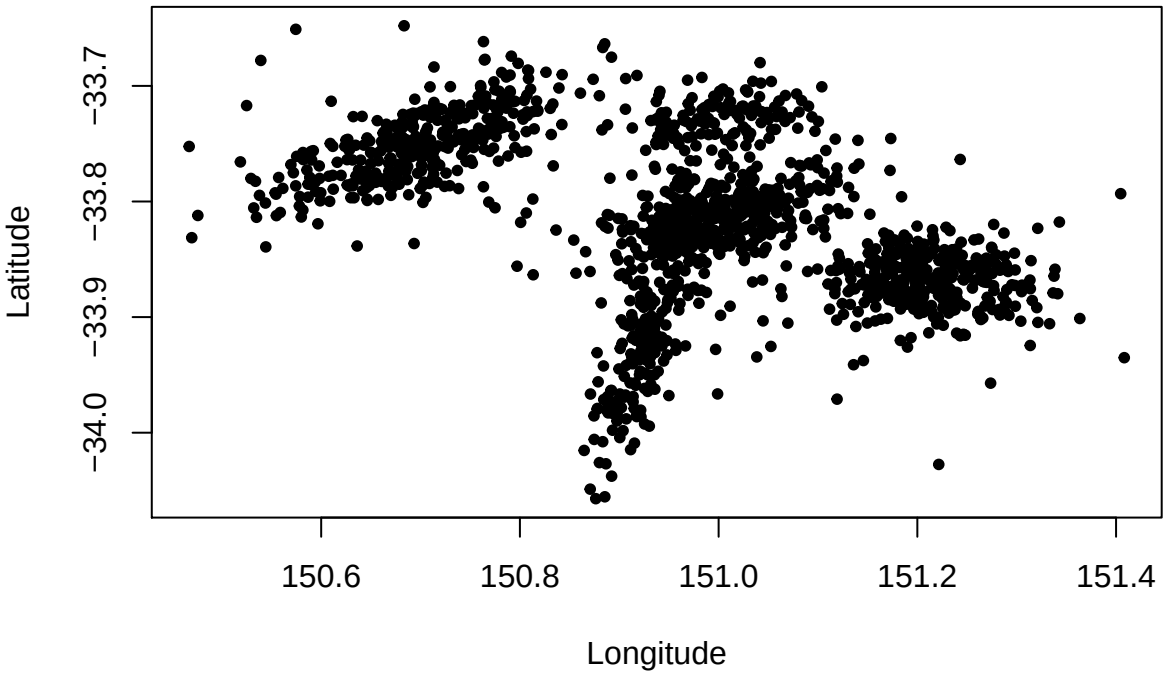
You are conducting an analysis of car crashes for an insurance company. The dataset `assignment1.csv` contains a sample of the times and locations of car crashes. Additionally, the colour (light or dark) of the vehicle is recorded. The following table shows the first 6 rows.

Colour	Hour	Minutes	Longitude	Latitude
dark	16	7	151.0476	-33.80303
light	15	36	151.0911	-33.76662
light	15	49	150.7100	-33.74679
dark	16	32	151.2432	-33.76362
light	8	44	150.8819	-33.88775
light	7	54	150.9440	-33.81858

1. It is desired to understand the relationship between *time of day* and car colour.
   a. Construct a time-of-day variable from the Hour and Minutes variables.
   b. For the *light* and *dark* vehicles seperately, create kernel density estimates and plot them.
   c. Construct and plot a single function that uses KDEs and Bayes rule to predict the colour of a vehicle in a crash as a function of time of day.
2. It is desired to understand the relationship between time of day and the number of vehicle accidents. Ignoring vehicle colour, use a parametric bootstrap to determine whether a Normal mixture model with 2, 3, or 4 components is the best fit for the time of day data. Is this supported by AIC and BIC? (Use 100 or more bootstraps). For each model, produce a plot which overlays the density of each component in the fitted mixture model with some form of density estimate for the data.

(See page 2)

3. The following plot shows the location data for the car crashes. It is desired to understand the relationship between geographical location and density of car crashes. It is thought that some crashes occur within traffic 'hotspots'.



a. Compute a two-dimensional kernel density estimate for this data, and contour plot the result.
b. Fit a five (5) component two-dimensional Normal mixture to the data and plot the result. (Warning: this takes up to 10 minutes computer time on a Macbook pro)

The most likely component for each observation can be computed from the posterior probabilities `cl1 = apply(fit$posterior,1,which.max)`. This can be considered a form of clustering.

An alternate method is *k-means*. Given a data matrix `X` (Long and Lat in this case) 5 clusters can be found using `cl2 = kmeans(X,5)$cluster`. Whilst this *may* produce similar clusters the labels will probably not be the same.

c. Using the code above (or similar) find a cluster labeling for the data points, using **both** a 5-component Normal mixture model and *k-means*. Plot the location data using cluster labels from each method (you can use separate plots). Produce a table of cross-cluster memberships `table(cl1,cl2)`. Using the plots and the cross-cluster memberships, comment on which clusters are most similar between the individual clusters identified by each method. Comment on any qualitative differences between the clusters produced by the two methods.