Daniel Grewal
31505290
SCC 403 Data Mining
21 November 2014

Data Mining

This report is a detailed account of the data mining techniques performed on a phone call dataset which was collected over a period of 10 days. I preprocessed the data and performed a cluster analysis using various clustering techniques. My objective was to look for any interesting patterns amongst the data which could provide a useful insight into the sociocultural behaviour of this demographic. Through my analysis, I found various interesting relationships surrounding the amount of time that callers spend making calls at different periods of the day and then looked at the differences between the sexes in relation to time spent on the phone.

Introduction

This report is a detailed account of how I applied various data mining techniques to a dataset containing phone call data.

This report represents the first stage of the overall assignment and discusses methods for pre-processing the data followed by a cluster analysis. As part of the cluster analysis, I will also compare and highlight the advantages/disadvantages of using different cluster methods for different types of data. I have also developed much of my own code using the Matlab 2014Ra software. I will highlight this further into the report and also share this code (see Appendix).

The dataset given for this assignment is in raw form and each variable, as well as a brief description of the dataset is given in an accompanying PDF file (not included with this report). According to the assignment specifications, the dataset is made up of phone call data over a ten day period during 2006 on the island of Isla Del Sueno. Table 1.1 represents the meta-data of the dataset as well the description of that meta-data.

Literature Review

In preparation for this assignment, I studied the literature text, *Data Mining Concepts and Techniques, Third Edition,* written by Han, Kamber and Pei. The book was a fantastic study guide and it allowed me to understand further, the different methods which I have used throughout this assignment. There is great emphasis made on clustering analysis and the different types of clustering techniques used most frequently throughout academia and industry. This was a sufficient text for the assignment. Although the book is not a practical guide, it does teach the theoretical knowledge needed to competently perform data mining techniques. I would suggest that this text is definitely aimed at postgraduate level and above as it feels as if the authors discuss a lot of the text as if the reader already has a basic understanding of the different data mining concepts.

Variable Meta-Data & Description

Variable #	Meta-Data	Description
1	Caller	The numerical ID of each caller ranging from 0-399. There are 400 in total.
2	Receiver	The numerical ID of each receiver ranging from 0-400. There are 398 in total.
3	Duration	Duration of each call made, measured in seconds.
4	Time of day	The hour of day during which each call was made. Ranging from 0-24.
5	Day	The day which each call was made, ranging from 1-10.
6	Sex	The sex of each caller, 0 represents females and 1 represents males.
7	Age	Age of each caller, categorised into numerical representation ranging from 1-6.
8	Ethnic	Ethnicity of each caller,categorised into numerical representation ranging from 1-6.
9	Occupation	The occupation of each caller categorised numerically into groups ranging 1-20.
10	Country	The country that each caller was in when the call was made. Each country is represented numerically between 1-12.
11	Latitude	The geographical latitude that each call was made.
12	Longitude	The geographical longitude that each call was made.
13	Distance	The distance between the caller's location and the receiver's location.

Table 1.1

Data Visualisation

To get a clearer understanding of the data that I would be working with, I decided to perform a statistical analysis. This analysis included basic summary statistics such as finding the mean, median, range and interquartile ranges of the data. I performed summary statistics on each variable and also created box plots and histograms of each variable so that I could visualise the distribution of the data. *Figure 1.1* shows the box plot produced for the *Duration* variable. Looking at the box plot, one could assume that the spread of the data is fairly even and it seems to follow a normal distribution. *Figure 1.2* shows the accompanying histogram that was produced and one can see that the variable does have a normal distribution. The red line indicates what a perfect normal distribution would look like and it is clear that the data is very close to being perfectly aligned with this line.

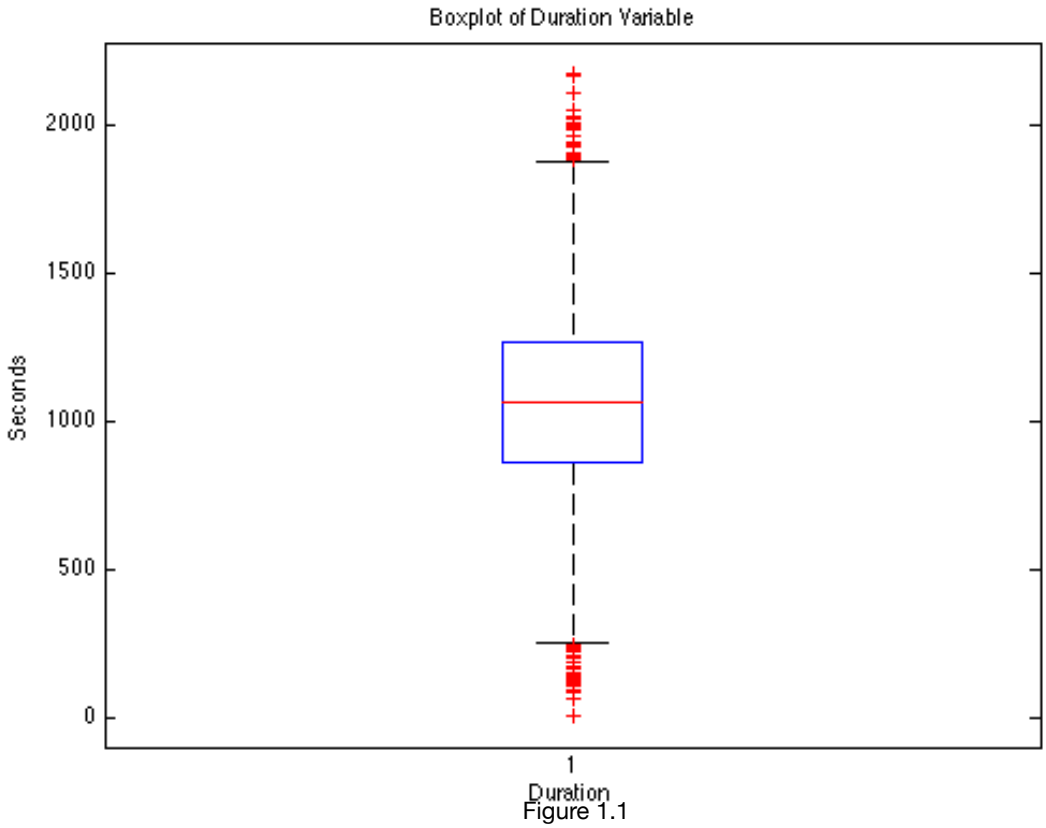

Figure 1.1

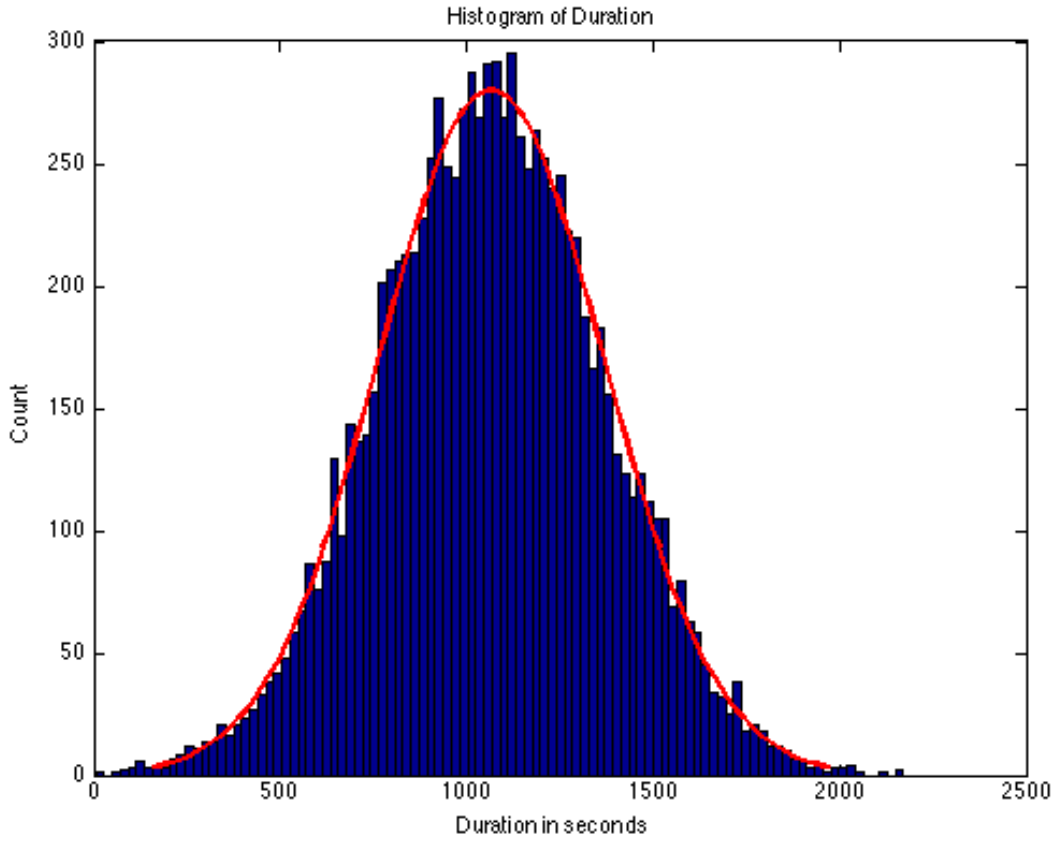

Figure 1.2

Data Pre-Processing

This section describes the various stages of the pre-processing phase. I will describe the exploratory analysis that I conducted as well as the feature selection/extraction methods used to gain a better insight into the data before moving onto cluster analysis.

Exploratory Analysis

The dataset in its rawest form is not very descriptive. To better visualise the data, I created a new dataset using the original '.csv' file in Matlab. A dataset specified in Matlab, is a special kind of data array which allows you to easily organise the data as well as give variable names.

Once I had created this new dataset and called it 'calls_dataset', I decided to add new variables. I created new categorised variables based on the information I was going to work with further to uncover patterns or relationships in the data. By creating the new variables, I could easily create subsets of the data whereby I view information relating to each categorised variable.

Description of Newly Created Variables

Original Variable	New Variable	Description
Caller	CallerCat	So that I could get information from the dataset based on each caller.
Receiver	ReceiverCat	So that I could get information from the dataset based on each receiver.
Sex	SexCat	A nominal variable coded as 0 or 1.
Age	AgeCat	An ordinal variable categorised into groups, the larger the number, the older the group.
Time of day	ToDCat	An ordinal variable categorised into groups. The larger the number, the later the hour in the day.

Table 1.2

Proposed Approach

This next section will describe my proposed approach to the assignment including the relationships between variables that I would like to explore further.

I decided to focus on the *Sex*, *Time of Day* and *Duration* variables within the dataset and look to discover patterns between the different sexes by observing their phone call behaviour.

I did not want to concentrate on too many relationships between different variables as I was more concerned with implementing the different methods correctly to show my understanding of the data mining practice. I have listed the objectives below:

- Do females make more calls than males?
- Do females spend more time on the phone during mid-day and early evening?
- Use PCA as a feature extraction method
- Perform 3 different types of cluster analysis and compare

Data Cleaning

As part of the pre-processing phase, I had to search the data for missing values or highly unusual attributes/observations.

I found that there were no missing values in the dataset however there were 4 highly unusual observations belonging to one of the variables. The *Duration* variable had 4 observations which were negative values. Considering that duration was measuring the length of a phone call between *Caller* and *Receiver*, I knew that these values could not exist in the real world.

To deal with these values, I simply removed them from the dataset. The reason for this, is that 4 observations out of approximately 10,000 is an extremely small value and is highly unlikely to have any effect on the outcomes of my analysis. However had there been a larger proportion of missing values or unusual observations, say a minimum of 5%, there would have been various techniques I could have applied to clean this data. I could have used the mean or median values of the variable observations and replaced the missing values with these measures of central tendency values. I could have applied Bayesian formalism or regression to better guesstimate the missing values or highly unusual observations in the dataset. Note that using measures of central tendency bias the data - the filled-in value may not be correct (Han, Kamber & Pei, 2012).

One of the variables which would feature heavily throughout my analysis was the *Duration* variable. The variable was measured in seconds which was not very intuitive. I decided to clean this variable by re-coding it as *DurationMins* where it would be measured in minutes. To do this, I simply divided the original variable by 60 (number of seconds which make up 1 minute).

Feature Selection

For this assignment, I did not need to use any specific feature selection methods as I had already defined which variables were going to be important to me when analysing the data. For the purposes of this assignment, I simply created subsets of the original dataset for me to work with when looking at the individual relationships. The main subset of data which I created was labelled ‘x’ and contained the following variables: *Sex*; *ToD* (Time of Day) and *DurationMins*. I then grouped this data according to the *Sex* and *ToD* variables so that I could look at all the combinations of the the two sexes and the hours in the day, and visualise the sum of minutes spent on the phone over the time period (shown in *figure 1.6*).

Feature Extraction

Feature Extraction is an important part of the data mining process. Feature extraction involves the various methods which Data Scientists use to extract information from datasets. This is especially important for very large and complex datasets. For this assignment, I focused on Principal Component Analysis (PCA) when extracting features from the dataset.

Principal Component Analysis is a dimensionality reduction method whereby data is projected orthogonally onto a linear space whilst retaining most the variance of the original data.

When you think about linear regression, points are projected vertically onto the regression line, this is known as the sum of least linear squares (LLS). With PCA, points are projected orthogonally on to the line, also known as the eigenvector, I think of this term as the sum of the orthogonal linear squares. *Figure 1.3* shows a set of data points which have been scaled using a feature scaling method (discussed later). Here, there are clearly what seems to be, two distinct groups of data points. These are represented in a two-dimensional space. *Figure 1.4* shows the same set of data points but once PCA has been applied. Note how the direction of the data points have changed. This is because PCA tries to retain as much variance as possible within the data itself. PCA computes the covariance matrix of the data and then uses the eigenvectors and eigenvalues from this matrix to project the new points. An eigenvector is a line which represents the direction of the data with the most variance and an eigenvalue is a point which is being projected on to this eigenvector. An eigenvector is, itself, a matrix of non-zero values. PCA uses the singular value decomposition algorithm (which will not be discussed here and is beyond the scope of this assignment) to compute eigenvectors and eigenvalues from the covariance matrix.

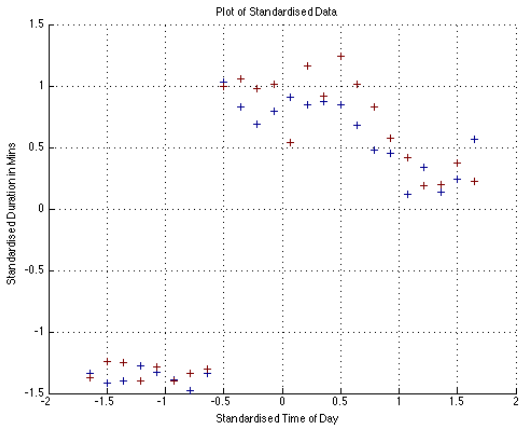

figure 1.3

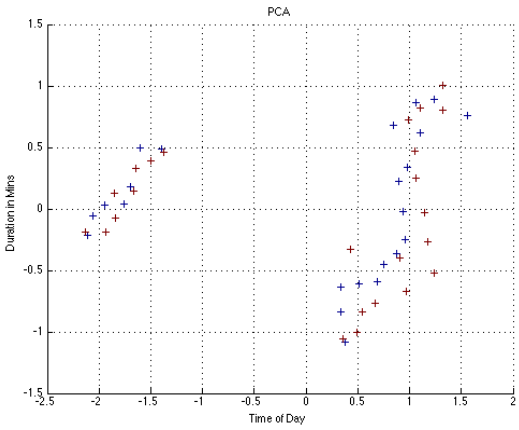

figure 1.4

Once computed, you take the first *k* eigenvectors and multiply by the original data and this gives you your new principal components for which the data is represented. A single principal component represents an axis on the bi-plot in *figure 1.5*, it can be seen how each variable contributes to each principal component. The red points are difficult to make out but they are focused to the right hand-side of the plot with a few more points on the far left hand-side of the plot. What this plot shows is that both variables contribute positively to the first principal component. *Duration* is the largest coefficient but *Time of Day* is only slightly less large. Only *Time of Day* has a positive contribution on the second principal component and *Duration* has a negative contribution. This indicates that the second component does in fact distinguish between calls which had a value for for the *Time of Day* variable and a low value for the *Duration* variable.

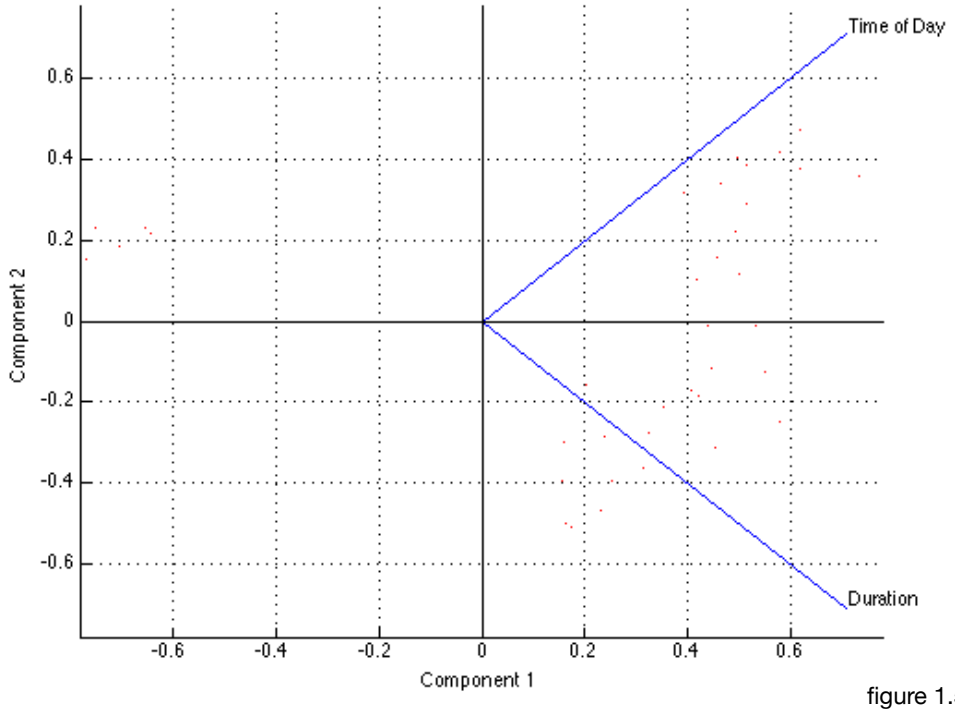

figure 1.5

To finalise my explanation of PCA, I have briefly described the steps to computing the procedure below. I also developed my own PCA algorithm (see Appendix). Another important phase of the process is to perform mean normalisation on the data. Also, if the features are of different scales, scale the features using standardisation.

Step 1.
Perform mean normalisation (x^i - u_j)

$$u_j = \frac{1}{n}\sum_{i=1}^{n} x_j^i$$

Step 2.
Compute the covariance matrix

$$sigma = \frac{1}{n}\sum_{i=1}^{n}(x^i)(x^i)^T$$

Step 3.
Take first *k* eigenvectors using singular value decomposition on sigma which then forms matrix U
***k* must be less than or equal to x^i**
Step 4.
Multiply U matrix by original data x

Another method in feature extraction, which was partially mentioned in the PCA explanation, is feature scaling. Very often in data mining, data of different scales is being visualised. Take *figure 1.6* for example. *Time of Day* is on an interval between 0 and 24 and the *Duration in Mins* is on a scale from 0 to 7000.

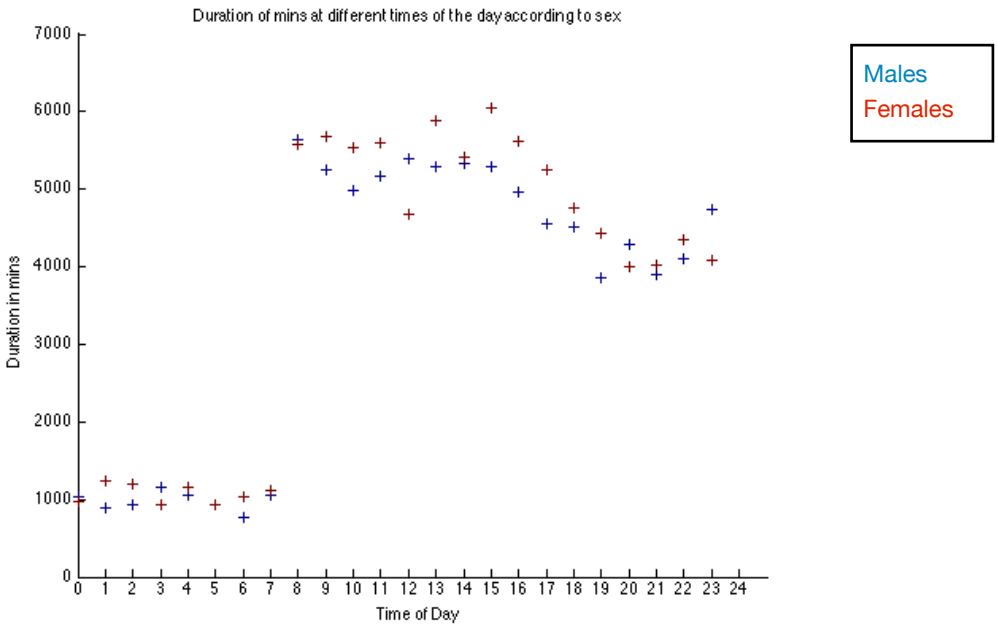

figure 1.6

This graph shows the sum of duration of calls made at different hours of the day by the different sexes. To perform certain data mining techniques, it is important to scale the data so that both axes lie on more similar scales. For instance, the PCA method requires that data is scaled before computing.

There are two common methods for feature scaling, one known as normalisation and the other as standardisation. I have developed algorithms for both in Matlab (see Appendix).

The equation for both of these concepts are very similar however there is a subtle difference between the two. Firstly, normalisation scales the features so that all values lie somewhere between -1 and 1. This is done by finding the difference between the current value and the minimum value of the data, then dividing by the range of the data. Standardisation, is a similar formula except you find the difference between the current value and the mean of the data and rather than dividing by the range, you instead divide by the standard deviation of the values within that data range. This measures how standardised the values are, i.e. how many standard deviations they are away from the mean. *Figure 1.3* shows data that has been standardised in preparation for PCA.

Normalisation	Standardisation
Step 1. Find the min and max values of x	Step 1. Find the mean and sd values of x
Step 2.	Step 2.
$\frac{x^i - min(x)}{max(x) - min(x)}$	$\frac{x^i - u_j}{sd}$

Now that feature scaling is explained, I want to move back to the topic of PCA. I have already discussed scaling the features and performing PCA onto the data itself. The results shown in *figures 1.3* to *1.5*. The final part I wish to discuss is how to choose *k* eigenvectors to get the principal components. First and foremost, *k* must always be less than or equal to the number of observations in the data, hence the term dimensionality reduction. The data that I have focused on is 2-dimensional, whereby I have 48 observations of the different combinations of *Sex* and *ToD* and I am analysing the sum of minutes spent on the phone by each sex at different time periods. This means that to reduce the data, I can only take a maximum of the first 2 eigenvectors computed from the covariance matrix.

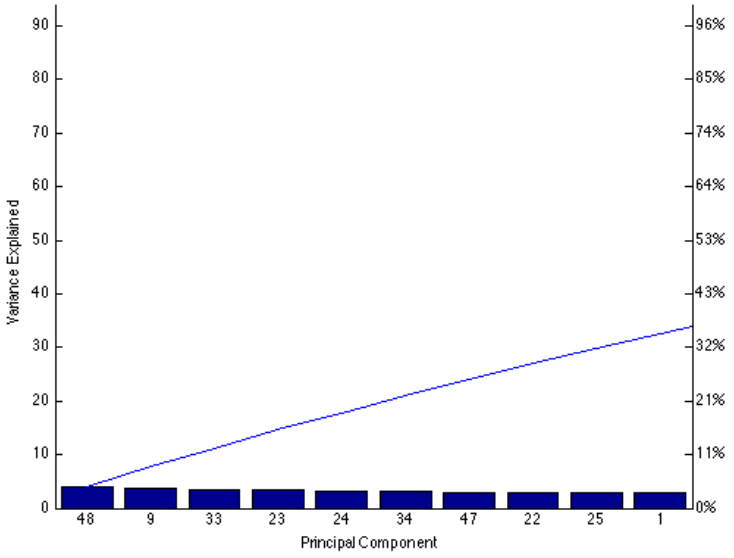

figure 1.7

Figure 1.7 shows how much variance is explained by each component. Only the first 10 components are shown on the plot and in total, only approximately **36%** of the variance is explained from these components. Ideally, you'd like to have at least two thirds of the variance explained by the first component in this instance but this is clearly not the case. Therefore, by reducing the dimensionality of the data, I will almost certainly lose a fair share of the variation of the original data. With this, I decided not to use the PCA data when performing clustering analysis.

Cluster Analysis

In this section, I shall describe the cluster analysis phase of the assignment and also analyse and compare the different clustering methods that I used and then produce the results.

K-Means Clustering

K-means clustering is one the oldest and most simple methods of clustering in existence. It is also one of the most widely used. For k-means to work, the number of clusters to compute must already be known. With *k* number of clusters already known, a centroid-based partitioning technique uses the centroid of a cluster, C_i, to represent that cluster (Han, Kamber and Pei, 2012). The centroid is also the centre of the cluster. two important concepts of k-means are the intracluster similarity and the intercluster similarity. The objective is to minimise the distance between each point within a cluster (intracluster) and maximise the distance between points of different clusters (intercluster). For the purpose of the assignment, I have developed my own k-means algorithm in Matlab and the code can be found in the Appendix as well as on the Matlab File Exchange.

```
K-Means
Input
        k: # of clusters
        d: training set containing n objects
Algorithm
        Randomly initiate centroid
        Repeat:
        for i = 1 to n
                c^i = index (from 1 to k) of
                cluster centroid closest to c^i
        for k = 1 to k
                u_k = mean of points
```

As I have developed my own k-means algorithm, I shall discuss in detail how this algorithm works. Firstly, once *k* is known, it is important to randomly initialise the first *k* centroids. There are many ways of doing this, including using the first *n* objects on the dataset that are equal to *k*. However, I have instead chosen to create a function whereby Matlab randomly chooses a set of data points as the centroids to begin with. Then, the algorithm enters a loop where each object in the dataset is assigned to a cluster, *k*, using a distance measurement, in this case, the Euclidean distance. Once an object has been assigned to the nearest cluster based on the Euclidean distance, the centroid is re-computed using the mean of the data points belonging to that cluster. This process repeats until convergence. Conceptually, convergence in this case is where the centroids of the clusters do not change from one step to the next.

K-means is a good performing algorithm on small to medium-sized datasets and as it is a distance-based algorithm, other distance measures can be used such as Seuclidean, Mahalanobis and Cosine. However, there are several drawbacks. K-means never terminates at global optimum and instead terminates at the local optimum. Different ways of

computing random initialisation often determine where the algorithm will terminate. Therefore, it is useful, if you have a small number of clusters (less than 100), to perform numerous iterations and choose the best performing. This is obviously time consuming and computational expensive, depending on the number of iterations. Also, k-means performs best when the data is well-separated, if there is not much sparsity in the data, it is difficult for the algorithm to maximise intercluster similarity.

Agglomerative Clustering

Agglomerative, or hierarchical clustering is another clustering method whereby data objects are grouped into a hierarchy or "tree" of clusters (Han, Kamber and Pei, 2012). Unlike k-means, you do not need to know beforehand the number of clusters to compute, however agglomerative clustering allows you to specify the maximum number of clusters to compute or it can naturally compute the number of clusters using "linkage". This algorithm first begins by assigning each object as it's own cluster and then merges the clusters in an iterative approach until all clusters satisfy the hierarchical conditions. To decide which cluster to merge with, it uses a distance measure to calculate which cluster is closest. It then merges with the closest cluster. For this assignment, I considered two approaches to agglomerative clustering which shall be discussed in the Results section.

Agglomerative Clustering

Step 1.
Assign each object to a separate cluster

Step 2.
Evaluate all pair-wise distances between clusters and compute distance matrix

Step 3.
Merge clusters with the shortest pair-wise distance

Step 4.
Update distance matrix and re-evaluate distances

Step 5.
Repeat steps 2-4 until distance matrix is left with a single element.

Agglomerative clustering is advantageous in the sense that it naturally produces smaller clusters than other clustering methods which may lead to better discoveries in the data. However one of it's disadvantages is that by using different distance measurements, you are likely to get different results. This leads to multiple experiments and comparing the results to support the veracity of the original results.

Subtractive Clustering

Although I have not used this method as part of the assignment, I will describe it in brief detail and highlight it's usefulness in the clustering process.

Subtractive Clustering is a enhanced development of Mountain clustering. In fact subtractive clustering is most often applied to find out the number of clusters with the data

before moving on to apply another form of clustering such as k-means. one of the advantages of subtractive clustering is that it's algorithm is recursive and thus satisfies an on-line method i.e. automatically updates in real-time as new data are added. The way the algorithm works, is that it initially treats all data points as a cluster. It then uses a likelihood function to calculate whether or not that data point is likely to be a cluster based on the density of points surrounding said data point.

Subtractive Clustering

Step 1.
Select the data point with the highest potential to be the first cluster centre

Step 2.
Remove all data points within the vicinity of the cluster, in order to determine the next data cluster and it's centre

Step 3.
Iterate until all data is within vicinity of a dat cluster

Gaussian Mixture Model

Although the Gaussian Mixture Model was not a specified clustering method within the assignment, as it was an area of clustering I had an interest in, I decided to apply it to the data and discover its usefulness and compare it to the other clustering algorithms used. Gaussian Mixture Models are said to perform better than k-means when the clusters are of different sizes, as is the case with my data.

The Gaussian Mixture Model (GMM) is actually more of a probability distribution function which is often used in clustering analysis. Each cluster is associated with some probability that some data object belongs to that cluster. A general assumption of mixture models is that a set of observed objects is a mixture of instances from multiple probabilistic clusters. A GMM generates each observed object independently following two steps (Han, Kamber and Pei, 2012):

1. Choosing a probabilistic cluster according to the probabilities of the clusters.

2. Choosing a sample according to the probability density function of the chosen cluster.

The objective of probabilistic model-based clustering is to infer a set of probabilistic clusters to generate the dataset using the generation process described above. Still, an important feature of mixture models is measuring the likelihood that a set of probabilistic clusters and their probabilities will generate an observed dataset. If we take C to be the set of, k, probabilistic clusters, each with their own probability distribution functions, f_k, and their probabilities w_k, then for an object, o, the probability that the object is generated by cluster C_j ($1 <= j <= k$), is given by the equation below.

$$P(o|C) = \sum_{j=1}^{k} w_j f_j(o)$$

Another thing to note about GMM is that it is a similar method to k-means in that it iterates until convergence is complete, terminates at local optimum but is considered a soft clustering method rather than a hard clustering method.

Results

This section highlights the results from the cluster analysis and compares each clustering method based on these findings.

K-Means Results

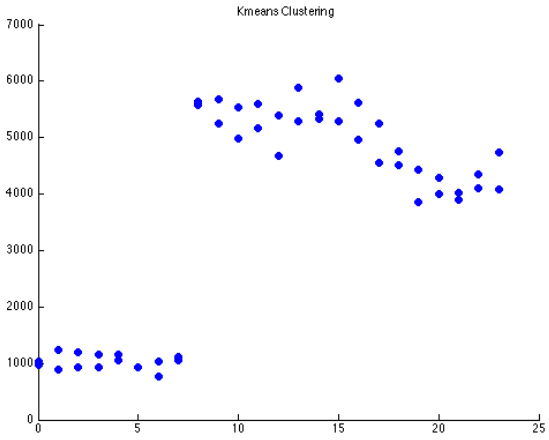

figure 1.8

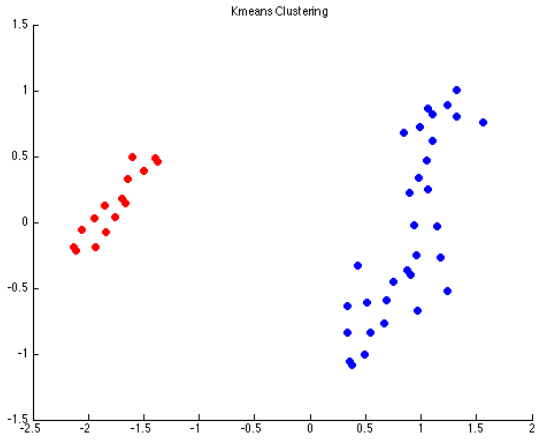

figure 1.9

An interesting comparison between *figures 1.8* and *1.9* show the results of k-means clustering on the data. *Figure 1.8* shows the results of clustering on the original dataset, as proposed during the discussion of PCA as the PCA algorithm did not retain the variance in the data. However, when applying clustering to the original data, after a total of 20 iterations, it still failed to separate the data even though the data itself appears well separated in the plot. *Figure 1.9* shows k-means applied to the data after PCA was applied. It took just one iteration to separate the data into the specified two clusters. The larger of the two clusters, the blue group, show that a more time was spent on the phone during the later hours of the day than the early hours of the day. However, I have concerns over these results as the direction of the data, once PCA is applied changes dramatically and this indicates to me that not a lot of the variance was retained, as already discussed. Therefore my results for k-means are inconclusive in the sense that the algorithm failed to find meaningful clusters for the original data points. My own interpretation as to why k-means failed in this instance could be due to the different cluster sizes.

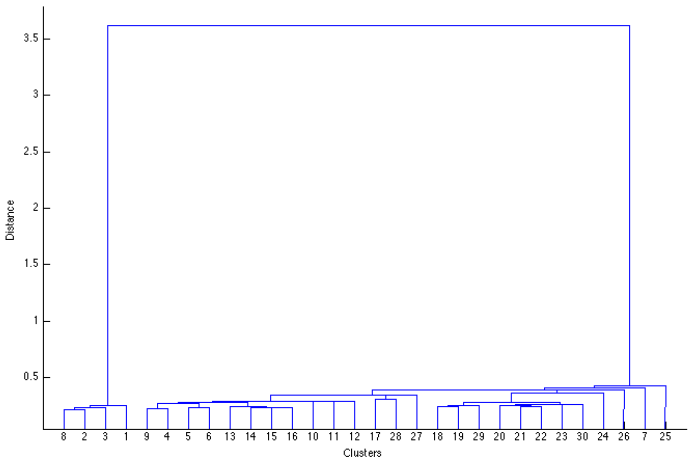

figure 2.0

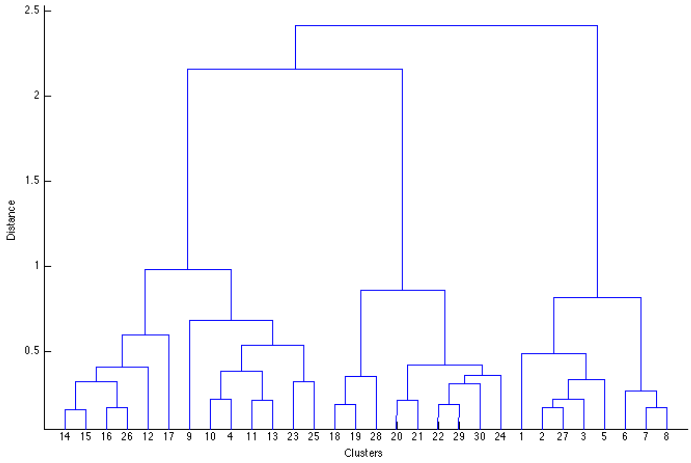

figure 2.1

Agglomerative Clustering Results

Before applying the clustering method, I first normalised the data and then computed the pair-wise distance. I used the proximity measures to decide how the data is to be clustered using the linkage function and the results are found below.

Figure 2.0 shows agglomerative clustering using the Euclidean distance and *figure 2.1* shows agglomerative clustering using Mahalanobis distance. There is a large distance between the hierarchical group of clusters in *figure 2.0* and this dendrogram is suggesting a cluster of two groups would be a good fit of the data. Also note that the distance between the lower group of clusters is extremely small and this is evidence that the dendrogram has done a good job in grouping datas points with are in close proximity to each other. I then plotted these clusters as a scatter graph, as can be seen in *figure 2.2*. The points that I was interested in were those which had a substantial distance between the sexes. Some of those points are represented by the blue line on the graph. I discuss these points further in *table 1.3*.

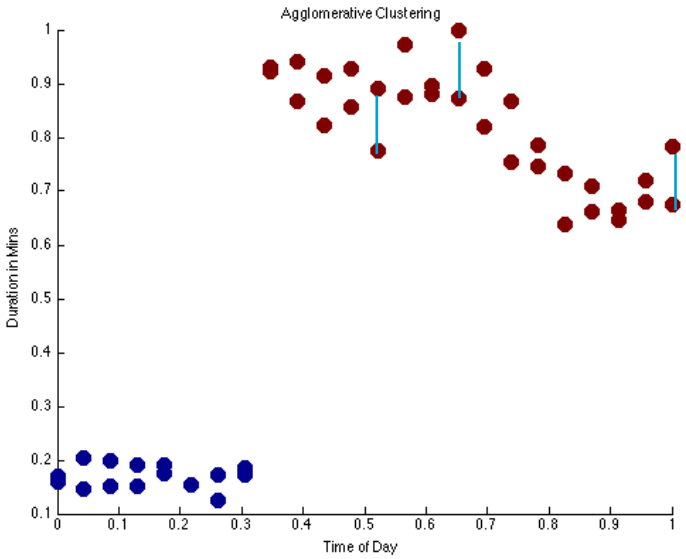

figure 2.2

The second dendrogram appears to show a cleaner divergence of the data and it is clearer where merging between the clusters took place.

However also note that the distance between lower groups in the hierarchy appears much greater than those in the previous dendrogram. This dendrogram also appears to suggest two clusters, although it is possible to suggest 3 based on the distance between the top level of the hierarchy and the level just below, specified at Distance, **2.15** on the graph.

First however, I wanted to verify that the clusters of both dendrograms represented the data by comparing the correlation coefficients. I computed the the *cophenetic correlation coefficients* for both dendrograms. In fact, *figure 2.0* represents the data better as it produced a correlation coefficient of **0.918** as opposed to the Mahalanobis coefficient of **0.888**. The closer to 1, the better the data is represented by the chosen clusters.

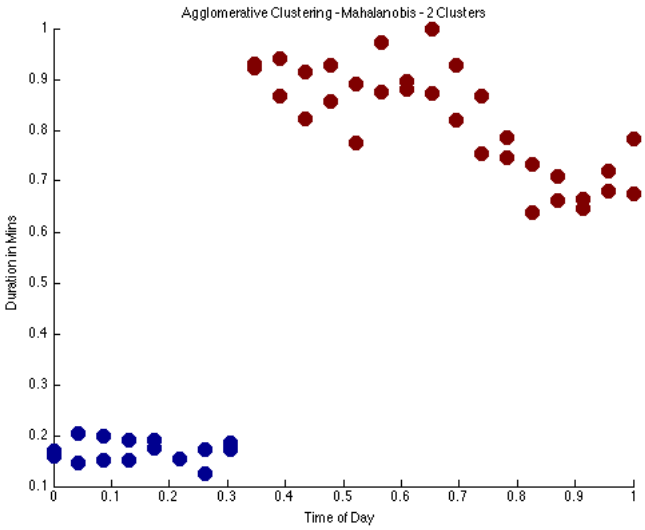

figure 2.3

I plotted both a 2-cluster and 3-cluster representation of the dendrogram measured with Mahalanobis distance. *Figure 2.3* shows the 2-cluster representation and it is clear that this is identical to the clustering grouped using the Euclidean distance as shown in *figure 2.2*. The 3-cluster representation is shown in *figure 2.4*. This cluster has produced quite an interesting grouping as it has nearly managed to distinguish between early hours, afternoon hours and evening hours. There are two points which are mis-represented by the clusters, these have been circled. These points represent the same time period, 5pm-6pm, but different sexes.

Table 1.3 shows a total of 11 different hours of the day where one sex spent a substantial more amount of time on the phone than the other. The results of the analysis show that females are responsible for much of the duration of calls made throughout the day. Males spend most of their phone time between the hours of 12pm and 1pm (a typical lunch hour in a standard working day) and between the hours of 11pm and 12am.

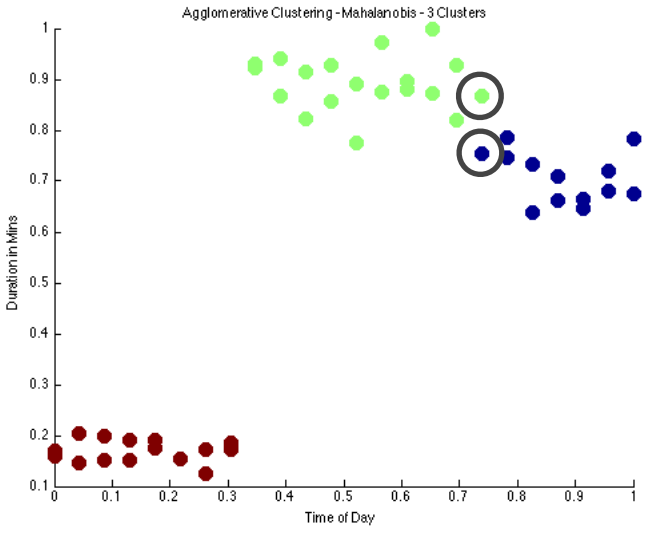

figure 2.4

Extreme Measurements Found in Agglomerative Clustering

Data Observation	Sex	Time of Day	Difference in Sum of Minutes Spent on The Phone
13	Male	12pm	711
24	Male	11pm	662
26	Female	1am	194
34	Female	9am	441
35	Female	10am	338
36	Female	11am	432
38	Female	1pm	603
40	Female	3pm	768
41	Female	4pm	652
42	Female	5pm	683
44	Female	7pm	579

table 1.3

Gaussian Mixture Model Results

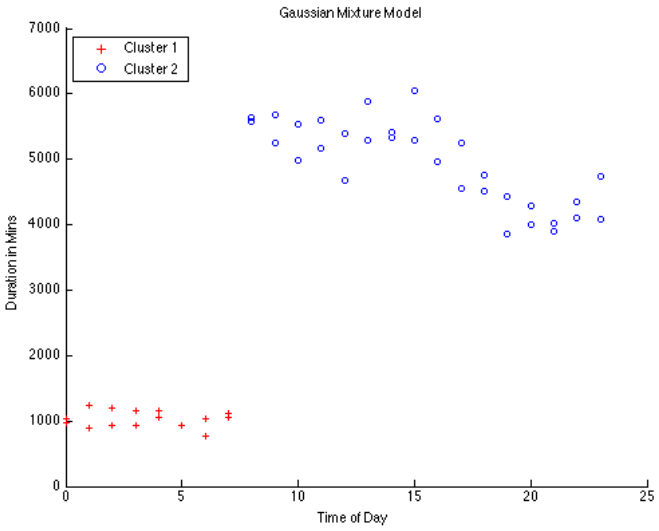

figure 2.5

The Gaussian Mixture Model produced a better result than the k-means clustering algorithm. One of the advantages of using a GMM over k-means is that it works better with clusters of different sizes. *Figure 2.5* shows the clear clustering of the specified two groups where the posterior probability of each data object determines the cluster that it will belong to. As I pre-determined two clusters, my hope was that it would produce similar results as the agglomerative clustering did using Euclidean distance. However, I also wanted to analyse the results when specifying three clusters, as in the previous case, agglomerative clustering using Mahalanobis distance suggested that the data could be split into early hours, afternoon and

evening. The results produced by a GMM with three clusters is less successful. As can be seen from *figure 2.6*, the model seemed to cluster four points which tell us nothing about the data itself.

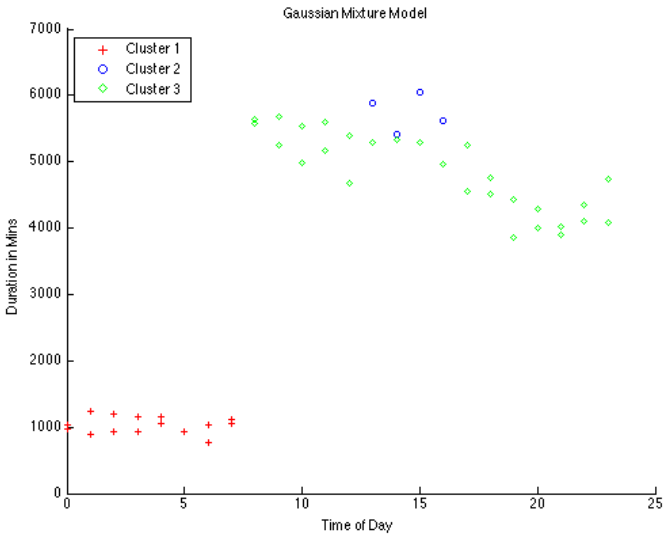

figure 2.6

As each cluster corresponds to one of the bivariate normal components in the mixture distribution, each data object is assigned based on the posterior probability that it came from one of the components. The highest posterior probability determines the cluster. I plotted these probabilities on a graph which can be seen in *figure 2.7*. Here the colour of the point shows the probability that it belongs to a cluster based on the component posterior probability, as determined by the colour map.

Notice how the colours of the blue points show that based on the component posterior probability, that they all belong to one cluster rather than two as shown in *figure 2.6*.

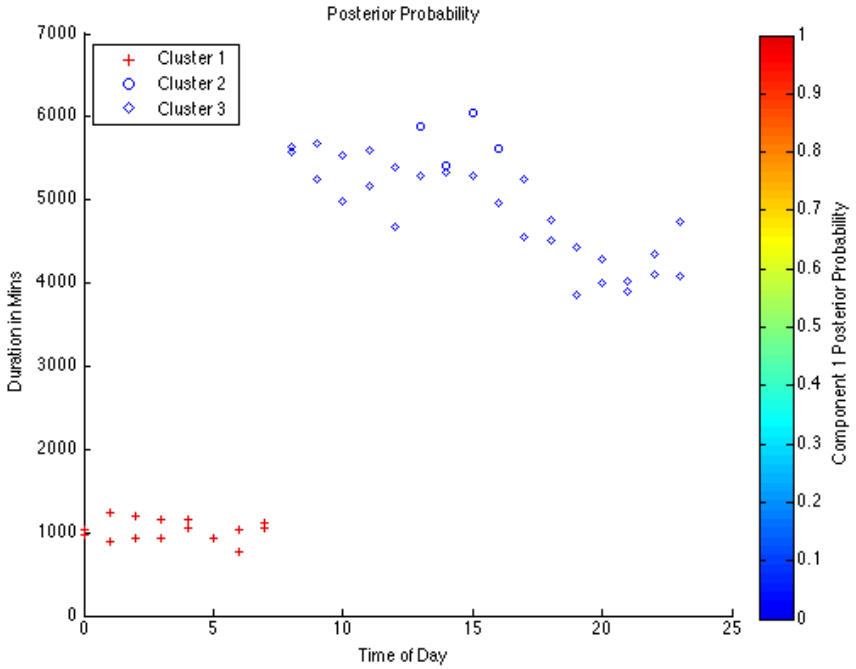

figure 2.7

The results of using a GMM are better than the k-means algorithm however the agglomerative clustering method performs much better on the data. This may be down to the fact that although the clusters are of different sizes, they are well separated and GMM works better on less sparse data.

Conclusion

To conclude my analysis, I will first compare my outcome to the proposed approach that was specified earlier in the report.

I was successful in implementing and comparing three different methods for clustering. I compared the methods in the Cluster Analysis chapter and then compared the results produced by each in the Results chapter. Overall, I believe that agglomerative clustering was the best approach used concerning the data that I was working with. My initial thoughts were that k-means would be the best performer however this turned out to be the worst. I decided that this was down to the fact that the data played in to the disadvantages of a k-means algorithm in the sense that the clusters were of different sizes and it is an algorithm notoriously known for terminating at the local optimum.

I was also successful in applying Principal Component Analysis to analyse whether or not it would be beneficial to reduce the dimensionality of the data. I concluded that using PCA, I would have lost a lot variance within the data, therefore losing the possibility to explore interesting points and relationships between the sexes. I progressed by clustering on the original data points.

The results produced from agglomerative clustering confirmed that females spend more time on the phone than males. One of the most interesting discoveries was that males spend a much higher amount of time on the phone between 12pm and 1pm, than their female counterparts. With no knowledge of the demographic or cultural society of Isla Del Sueno, this could be an indicator that this region conforms with the traditional working environment whereby males are the main “bread-winners” in the household and that a large proportion of females may be stay-at-home mothers or wives. However to confirm this, I would have to do more analysis on the data.

Another outcome of the results was that there is much less time spent making phone calls in the early hours of the day than the ‘prime’ hours of the day. This could conclude that the region which the dataset focuses on, has a population that does not conform to much night-time activity. However, this is inconclusive as the data is only a representation of the population and more analysis would be needed.

If I could improve this report further or do anything differently, I would spend more time expanding on the variables I had focused on by introducing the *Occupation* and *Age* variables into my analysis. I would then be able to analyse which age groups or occupation groups spend the most time on the phone during different hours of the day rather than generalising the analysis to the sex of the callers.

Appendix

The appendix contains all code produced by myself during this assignment. I have also included the algorithms which I developed to help perform the analysis. This code has been attached separately to this document.