**IDS 572 Business Data Mining**

**Homework 9**

**Submitted by:**

Bhalerao, Amit (abhale2)

Joshi, Madhura (mjosh4)

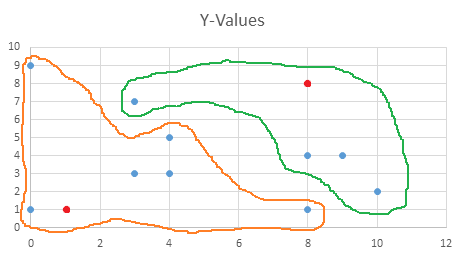
Singh, Vivek (vsingh34**)**

**Problem 1:** Consider the below points in a 2-D space:

1. Beginning with centroids at c1 = (1,1) and c2 = (8,8) and using 2-means clustering algorithm, the iterations are as shown below:
2. Allocate the points to centroids, then find the new centroids.

|  |  |
| --- | --- |
| Points | Cluster (Iteration 1) |
| (8, 4) | c2 |
| (3, 3) | c1 |
| (4, 5) | c1 |
| (0, 1) | c1 |
| (10, 2) | c2 |
| (3, 7) | c2 |
| (0, 9) | C1 |
| (8, 1) | C2 |
| (4, 3) | c1 |
| (9, 4) | c2 |

Per the centroid allocation in this table, the clusters are as below:

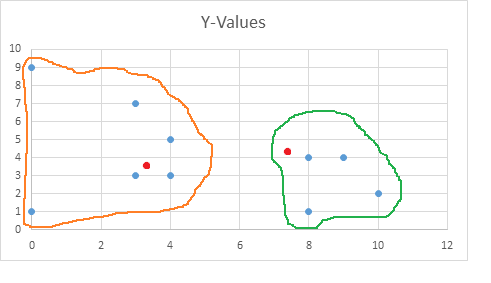


New centroids are c1 = (3.17, 3.67) and c2 = (7.5, 4.25)

1. Again, using the new centroids from step (i) above, the allocation for iteration 2 is:

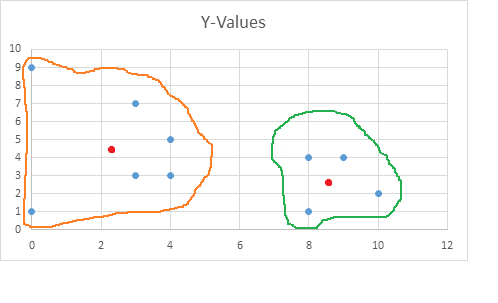
|  |  |
| --- | --- |
| Points | Cluster (Iteration 2) |
| (8, 4) | c2 |
| (3, 3) | c1 |
| (4, 5) | c1 |
| (0, 1) | c1 |
| (10, 2) | c2 |
| (3, 7) | c1 |
| (0, 9) | c1 |
| (8, 1) | c2 |
| (4, 3) | c1 |
| (9, 4) | c2 |

Per the centroid allocation in this table, the clusters are as below:



New centroids are c1 = (2.33, 4.67) and c2 = (8.75, 2.75)

At the end of the second iteration, using the new centroids from step (ii), the clusters will be as below:



1. Suppose we are interested in a binary (yes, no) output. Suppose outputs for the points above are yes, yes, no, no, yes, yes, no, no, yes, yes respectively. Consider the point (5,3)
2. What are the three closest points in our data set?

Using Manhattan distance, the three closest points to the point (5, 3) are:

(4, 3), (3, 3) and (4, 5)

1. Using the K-nearest neighbors approach, what would be the predicted output for (5,3) using K = 3 neighbors?

Based on the 3 nearest neighbors of (5, 3) as depicted in the table below, the predicted output for (5, 3) would be majority of the outputs of the 3 neighbors.

|  |  |
| --- | --- |
| Point | Output |
| (4, 3) | Yes |
| (3, 3) | Yes |
| (4, 5) | No |

So, the predicted output for (5, 3) will be **Yes**.

**Problem 2:** Distance Matrix is as given below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dist | A | B | C | D | E | F |
| A | 0.00 |  |  |  |  |  |
| B | 0.71 | 0.00 |  |  |  |  |
| C | 5.66 | 4.95 | 0.00 |  |  |  |
| D | 3.61 | 2.92 | 2.24 | 0.00 |  |  |
| E | 4.24 | 3.54 | 1.41 | 1.00 | 0.00 |  |
| F | 3.20 | 2.50 | 2.50 | 0.50 | 1.12 | 0.00 |

Using Single-link agglomerative clustering,

Step 1: Merge D and F

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dist | A | B | C | D, F | E |
| A | 0.00 |  |  |  |  |
| B | 0.71 | 0.00 |  |  |  |
| C | 5.66 | 4.95 | 0.00 |  |  |
| D, F | 3.20 | 2.50 | 2.24 | 0.00 |  |
| E | 4.24 | 3.54 | 1.41 | 1.00 | 0.00 |

Step 2: Merge A and B

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dist | A, B | C | D, F | E |
| A, B | 0.00 |  |  |  |
| C | 4.95 | 0.00 |  |  |
| D, F | 2.50 | 2.24 | 0.00 |  |
| E | 3.54 | 1.41 | 1.00 | 0.00 |

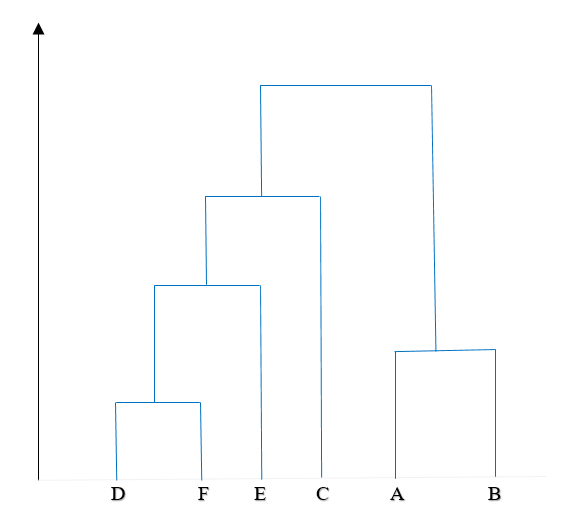
Step 3: Merge E with (D, F)

|  |  |  |  |
| --- | --- | --- | --- |
| Dist | A, B | C | D, F, E |
| A, B | 0.00 |  |  |
| C | 4.95 | 0.00 |  |
| D, F, E | 2.50 | 1.41 | 0.00 |

Step 4: Merge C with (D, F, E)

|  |  |  |
| --- | --- | --- |
| Dist | A, B | D, F, E, C |
| A, B | 0.00 |  |
| D, F, E, C | 2.50 | 0.00 |

Step 5: Merge (A, B) with (D, F, E, C) to complete the dendogram.



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dist | A | B | C | D | E | F |
| A | 0.00 |  |  |  |  |  |
| B | 0.71 | 0.00 |  |  |  |  |
| C | 5.66 | 4.95 | 0.00 |  |  |  |
| D | 3.61 | 2.92 | 2.24 | 0.00 |  |  |
| E | 4.24 | 3.54 | 1.41 | 1.00 | 0.00 |  |
| F | 3.20 | 2.50 | 2.50 | 0.50 | 1.12 | 0.00 |

Using Complete-link agglomerative clustering

Step 1: Merge D and F

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dist | A | B | C | D, F | E |
| A | 0.00 |  |  |  |  |
| B | 0.71 | 0.00 |  |  |  |
| C | 5.66 | 4.95 | 0.00 |  |  |
| D, F | 3.61 | 2.92 | 2.50 | 0.00 |  |
| E | 4.24 | 3.54 | 1.41 | 1.12 | 0.00 |

Step 2: Merge A and B

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dist | A, B | C | D, F | E |
| A, B | 0.00 |  |  |  |
| C | 5.66 | 0.00 |  |  |
| D, F | 3.61 | 2.50 | 0.00 |  |
| E | 4.24 | 1.41 | 1.12 | 0.00 |

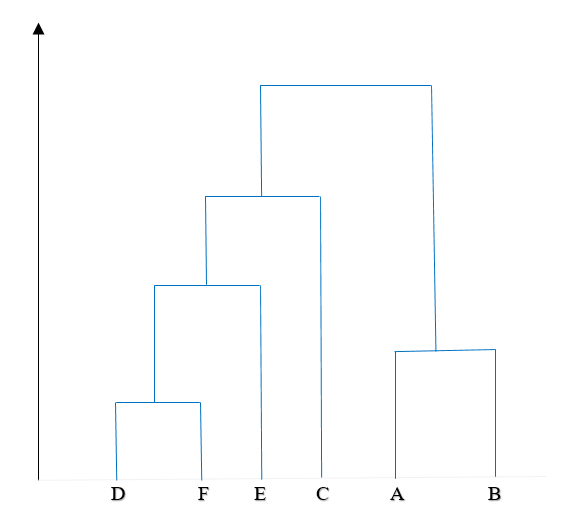
Step 3: Merge (D, F) and E

|  |  |  |  |
| --- | --- | --- | --- |
| Dist | A, B | C | D, F, E |
| A, B | 0.00 |  |  |
| C | 5.66 | 0.00 |  |
| D, F, E | 4.24 | 2.50 | 0.00 |

Step 4: Merge (D, F, E) with C

|  |  |  |
| --- | --- | --- |
| Dist | A, B | D, F, E, C |
| A, B | 0.00 |  |
| D, F, E, C | 5.66 | 0.00 |

Step 5: Merge (A, B) with (D, F, E, C) to complete the dendogram.



**Problem 3:** K-means clustering on ‘prospects’ data set:

We imported the prospects dataset into R and performed the tasks as below.

> prospects <- read.csv("F:/UIC Fall 2016/IDS 572 Data Mining for Business/prospects.csv")

> summary(prospects)

ID AGE INCOME SEX MARRIED OWNHOME LOC

1.07E+10 : 1 Min. :18.00 Min. : 15.00 : 106 Min. :0.0000 Min. :0.0000 E :1039

100107557: 1 1st Qu.:38.00 1st Qu.: 35.00 F:2161 1st Qu.:0.0000 1st Qu.:0.0000 F :1021

100115636: 1 Median :44.00 Median : 50.00 M:2434 Median :1.0000 Median :0.0000 B : 872

100305277: 1 Mean :44.23 Mean : 47.69 Mean :0.5785 Mean :0.3277 G : 453

100415298: 1 3rd Qu.:50.00 3rd Qu.: 61.00 3rd Qu.:1.0000 3rd Qu.:1.0000 H : 445

100430775: 1 Max. :75.00 Max. :116.00 Max. :1.0000 Max. :1.0000 A : 313

(Other) :4695 NA's :106 NA's :106 NA's :106 NA's :106 (Other): 558

CLIMATE FICOGTE700

Min. :10.00 Min. :0.0000

1st Qu.:20.00 1st Qu.:0.0000

Median :20.00 Median :0.0000

Mean :20.06 Mean :0.4135

3rd Qu.:20.00 3rd Qu.:1.0000

Max. :30.00 Max. :1.0000

NA's :106

The summary of the data set shows that various fields contain null values. So, data cleaning needs to be done on the data set.

First, we excluded the fields LOCATION and ID from the dataset.

> prospects <- prospects[, -1]

> prospects <- prospects[, -6]

Per the summary, the continuous variables AGE and INCOME contain missing values which are filled in using the means,

> sum(is.na(prospects$AGE))

[1] 106

> prospects$AGE[is.na(prospects$AGE)] = mean(prospects$AGE, na.rm=TRUE)

> sum(is.na(prospects$AGE))

[1] 0

Similarly,

> sum(is.na(prospects$INCOME))

[1] 106

> prospects$INCOME[is.na(prospects$INCOME)] = mean(prospects$INCOME, na.rm=TRUE)

> sum(is.na(prospects$INCOME))

[1] 0

To impute the missing values for the other fields, mice command is used.

> library(mice)

> md.pattern(prospects)

AGE INCOME SEX CLIMATE MARRIED OWNHOME FICOGTE700

4595 1 1 1 1 1 1 1 0

106 1 1 1 1 0 0 0 3

0 0 0 0 106 106 106 318

> prospects\_new = mice(prospects, m=5, maxit=50, meth='pmm', seed=500)

To retrieve the completed dataset, the complete command was used as below:

> prospects\_complete = complete(prospects\_new, 1)

> View(prospects\_complete)

> summary(prospects\_complete)

AGE INCOME SEX MARRIED OWNHOME CLIMATE FICOGTE700

Min. :18.00 Min. : 15.00 : 106 Min. :0.0000 Min. :0.000 Min. :10.00 Min. :0.0000

1st Qu.:38.00 1st Qu.: 35.00 F:2161 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:20.00 1st Qu.:0.0000

Median :44.00 Median : 49.00 M:2434 Median :1.0000 Median :0.000 Median :20.00 Median :0.0000

Mean :44.23 Mean : 47.69 Mean :0.5807 Mean :0.327 Mean :20.06 Mean :0.4108

3rd Qu.:50.00 3rd Qu.: 60.00 3rd Qu.:1.0000 3rd Qu.:1.000 3rd Qu.:20.00 3rd Qu.:1.0000

Max. :75.00 Max. :116.00 Max. :1.0000 Max. :1.000 Max. :30.00 Max. :1.0000

The field SEX was converted into numerical values:

> prospects\_complete$SEXCODE[prospects\_complete$SEX=="M"] <- "1"

> prospects\_complete$SEXCODE[prospects\_complete$SEX=="F"] <- "2"

> prospects\_complete$SEXCODE <- factor(prospects\_complete$SEXCODE)

> View(prospects\_complete)

> prospects\_complete <- prospects\_complete[,-3]

The SEX column was removed as SEXCODE column (having the numeric values) will now be used for

clustering. The missing values in the SEXCODE column were imputed using the mice command.

> View(prospects\_complete)

> summary(prospects\_complete)

AGE INCOME MARRIED OWNHOME CLIMATE FICOGTE700 SEXCODE

Min. :18.00 Min. : 15.00 Min. :0.0000 Min. :0.000 Min. :10.00 Min. :0.0000 1:2490

1st Qu.:38.00 1st Qu.: 35.00 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:20.00 1st Qu.:0.0000 2:2211

Median :44.00 Median : 49.00 Median :1.0000 Median :0.000 Median :20.00 Median :0.0000

Mean :44.23 Mean : 47.69 Mean :0.5807 Mean :0.327 Mean :20.06 Mean :0.4108

3rd Qu.:50.00 3rd Qu.: 60.00 3rd Qu.:1.0000 3rd Qu.:1.000 3rd Qu.:20.00 3rd Qu.:1.0000

Max. :75.00 Max. :116.00 Max. :1.0000 Max. :1.000 Max. :30.00 Max. :1.0000

The dataset is now prepared for clustering.

(a) Use the K-means method to cluster the prospects dataset.

The below command was run to create the clusters:

> km = kmeans(prospects\_complete, 4, nstart = 100)

How many points are in each cluster? What are cluster means and variances?

> km$size

[1] 1196 1131 1198 1176

> aggregate(prospects\_complete,by=list(km$cluster),FUN=mean)

Group.1 AGE INCOME MARRIED OWNHOME CLIMATE FICOGTE700 SEXCODE

1 1 48.37458 65.23746 0.7140468 0.4230769 19.94147 0.6521739 NA

2 2 34.54642 55.47657 0.3032714 0.3244916 20.92838 0.3996463 NA

3 3 51.41632 43.49125 0.7479132 0.2762938 18.99833 0.3764608 NA

4 4 42.02126 26.64711 0.5416667 0.2831633 20.41667 0.2108844 NA

> aggregate(prospects\_complete,by=list(km$cluster),FUN=var)

Group.1 AGE INCOME MARRIED OWNHOME CLIMATE FICOGTE700 SEXCODE

1 1 58.19430 48.04566 0.2043548 0.2442871 28.53214 0.2270329 0.2275395

2 2 40.94541 56.25498 0.2114849 0.2193908 40.28779 0.2401415 0.2424747

3 3 63.54353 43.53303 0.1886966 0.2001226 36.75689 0.2349342 0.2498072

4 4 81.07785 43.66855 0.2484752 0.2031546 43.31560 0.1665538 0.2404747

(b) For each of the four clusters, briefly describe the characteristics of members of that cluster.

Group 1. These are married people, fall in above average age range, are very rich and own houses. They seem to be well settled families with long and good credit score and educated hence earning good.

Group 2. These are young group of people, may be recently graduated and hence still to get married. They are well educated hence have high income. They are a mix who own houses and rent and have started to build a good credit score.

Group 3. These are elderly, retired families who have mid income range and renting houses. They have an average credit score, may be have stopped using credit cards and have an older credit history.

Group 4. These are very low earning, mid age, mix of married and unmarried population. Due to low income, they almost have no credit history and very low credit score. They also seem to be renting houses.

(c) What is the best value of k for this data set?

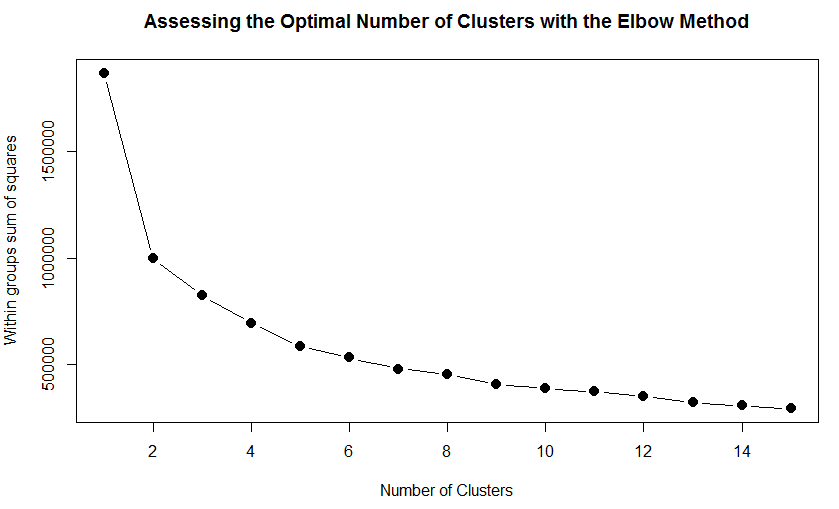
> mydata = prospects\_complete

> wss = (nrow(mydata)-1)\*sum(apply(mydata,2,var))

> for (i in 2:15) wss[i] = sum(kmeans(mydata, centers=i)$withinss)

> plot(1:15, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares",

main="Assessing the Optimal Number of Clusters with the Elbow Method", pch=20, cex=2)



As depicted in the graph above, we can say that after 7 clusters the observed difference in the within-cluster dissimilarity is not substantial.

Consequently, we can say that the optimal number of clusters to be used is k=7.