**BUSINESS DATA MINING**

**(IDS 572)**

**Solutions to Homework 9**

**Group Members**

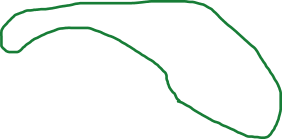
* Amey Pophali (apopha2@uic.edu)
* Karthik Varanasi (vvaran3@uic.edu)
* Mrinal Dhawan ([mdhawa3@uic.edu](mailto:mdhawa3@uic.edu))

**Solution 1**

The two centroids are highlighted in Red. Starting with centroids at c1(1,1) and c2(8,8) and using 2 mean clustering analysis,

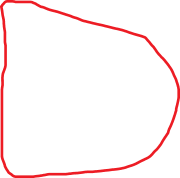
1. Allocating points to centroids and finding a new one

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



1. New centroids are c1(3.17,3.67) and c2(7.5,4.25). Using 2-means algorithm for the points above, clustering them into groups.

|  |  |
| --- | --- |
| Points | Cluster |
| (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 |



|  |  |
| --- | --- |
| Points |  |
| (8, 4) | Y |
| (3, 3) | Y |
| (4, 5) | N |
| (0, 1) | N |
| (10, 2) | Y |
| (3, 7) | Y |
| (0, 9) | N |
| (8, 1) | N |
| (4, 3) | Y |
| (9, 4) | Y |

Closet points to point (5,3) are (4,3), (4,5) and (3,3).

1. As per the data above, we have

(4,3) - Yes

(3,3) - Yes

(4,5) - No

Hence the predicted output for (5,3) would be yes.

**Solution 2**

1. Distance Matrix

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Distance | 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 |

**Solution 3**

Clearing and Preparing the data -

Reading the file -

p <- read.csv("D:/UIC Fall/Data Mining/HW/9/prospects.csv")

Changing relevant variables to Factor -

p$OWNHOME = factor(p$OWNHOME)

p$FICO..700 = factor(p$FICO..700)

Getting rid of the irrelevant columns ID and Location -

p = p[,-1]

p = p[,-6]

Data -

> str(p)

'data.frame': 4701 obs. of 7 variables:

$ AGE : int 37 46 45 38 34 69 46 28 37 46 ...

$ INCOME : int 57 71 65 50 44 60 42 63 59 57 ...

$ SEX : Factor w/ 3 levels "","F","M": 2 3 3 2 3 2 2 2 3 3 ...

$ MARRIED : int 0 1 1 0 0 0 1 0 1 1 ...

$ OWNHOME : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 2 1 2 ...

$ CLIMATE : int 20 20 20 10 20 30 20 20 20 20 ...

$ FICO..700: Factor w/ 2 levels "0","1": 1 1 2 1 1 1 2 2 2 2 ...

> summary(p)

AGE INCOME SEX MARRIED OWNHOME CLIMATE FICO..700

Min. :18.00 Min. : 15.00 : 106 Min. :0.0000 0 :3089 Min. :10.00 0 :2695

1st Qu.:38.00 1st Qu.: 35.00 F:2161 1st Qu.:0.0000 1 :1506 1st Qu.:20.00 1 :1900

Median :44.00 Median : 50.00 M:2434 Median :1.0000 NA's: 106 Median :20.00 NA's: 106

Mean :44.23 Mean : 47.69 Mean :0.5785 Mean :20.06

3rd Qu.:50.00 3rd Qu.: 61.00 3rd Qu.:1.0000 3rd Qu.:20.00

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

NA's :106 NA's :106 NA's :106

Checking for missing values -

> sum(is.na(p))

[1] 530

There are 530 missing values.

Replacing missing values by mean for Age and Income.

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

[1] 106

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

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

[1] 0

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

[1] 106

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

> sum(is.na(p$CLIMATE))

[1] 0

To attribute for the missing values of other variables, we use mice command.

library(mice)

md.pattern(p)

p1 = mice(p, m=5, maxit=50, meth='pmm', seed=500)

p\_complete = complete(p1, 1)

> library(mice)

> md.pattern(p)

AGE INCOME SEX CLIMATE MARRIED OWNHOME FICO..700

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

> library(mice)

> p1 = mice(p, m=5, maxit=50, meth='pmm', seed=500)

To retrieve the complete dataset, we use complete command -

p\_complete = complete(p1, 1)

summary(p\_complete)

str(p\_complete)

> summary(p\_complete)

AGE INCOME SEX MARRIED OWNHOME CLIMATE FICO..700

Min. :18.00 Min. : 15.00 : 106 Min. :0.0000 0:3159 Min. :10.00 0:2747

1st Qu.:38.00 1st Qu.: 35.00 F:2161 1st Qu.:0.0000 1:1542 1st Qu.:20.00 1:1954

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

Mean :44.23 Mean : 47.69 Mean :0.5831 Mean :20.06

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

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

The attribute sex was converted to numerical values -

p\_complete$SEX1[p\_complete$SEX=="M"] <- "1"

p\_complete$SEX1[p\_complete$SEX=="F"] <- "2"

p\_complete$SEX1 <- factor(p\_complete$SEX1)

View(p\_complete)

p\_complete <- p\_complete[,-3]

We get rid of the sex column as sex1 column, which has the numerical value, will now be used for clustering. The missing values in Sex1 column was attributed to using MICE command.

> summary(p\_complete)

AGE INCOME MARRIED OWNHOME CLIMATE FICO..700 SEX1

Min. :18.00 Min. : 15.00 Min. :0.0000 0:3159 Min. :10.00 0:2747 1 :2434

1st Qu.:38.00 1st Qu.: 35.00 1st Qu.:0.0000 1:1542 1st Qu.:20.00 1:1954 2 :2161

Median :44.00 Median : 49.00 Median :1.0000 Median :20.00 NA's: 106

Mean :44.23 Mean : 47.69 Mean :0.5831 Mean :20.06

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

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

The data is now prepared for clustering.

1. Run the clustering -

c = kmeans(p\_complete, 4, nstart = 100)

Points in the cluster and its means and variances.

> c$size

[1] 1190 1136 1179 1196

Mean

> aggregate(p\_complete,by=list(c$cluster),FUN=mean)

Group.1 AGE INCOME MARRIED OWNHOME CLIMATE FICO..700 SEX1

1 1 48.42353 65.24286 0.7142857 0.4218487 19.94958 0.6521008 1.348739

2 2 34.56866 55.53609 0.3054577 0.3265845 20.91549 0.4014085 1.411972

3 3 42.00763 26.68109 0.5402884 0.2824427 20.41561 0.2103478 1.599661

4 4 51.43876 43.49709 0.7491639 0.2767559 18.99666 0.3770903 1.477425

Variance

> aggregate(p\_complete,by=list(c$cluster),FUN=var)

Group.1 AGE INCOME MARRIED OWNHOME CLIMATE FICO..700 SEX1

1 1 57.95504 48.18067 0.2042533 0.2440975 28.59291 0.2270561 0.2273113

2 2 40.91158 56.29826 0.2123402 0.2201208 40.21840 0.2404914 0.2424645

3 3 80.95325 44.01026 0.2485877 0.2028409 43.20573 0.1662426 0.2402715

4 4 63.34812 43.58545 0.1880746 0.2003296 36.81673 0.2350898 0.2496991

The means and variances of the groups are shown in the table above.

Group 1

This group belongs to married people who fall in above average age range, own houses and are very rich. They are well off families with a very good credit score. They are educated and are earning well.

Group 2

This group belongs to young people. They are well educated and have high incomes. They might be recently graduated and are yet to get married. This group has started building a good credit score. Group contains a mix of individual who own houses and rent.

Group 3

This group is of elderly and retired people who have mid-range incomes. These people stay in rented houses. They have an average credit score. They have stopped using their credit cards and have older credit history.

Group 4

This group belongs to the individuals who have low earning and are mid aged. It is a mix of married and unmarried people. They have a low credit score or no credit history. They live in rented houses.

1. Best value of k for the Dataset

data = p\_complete

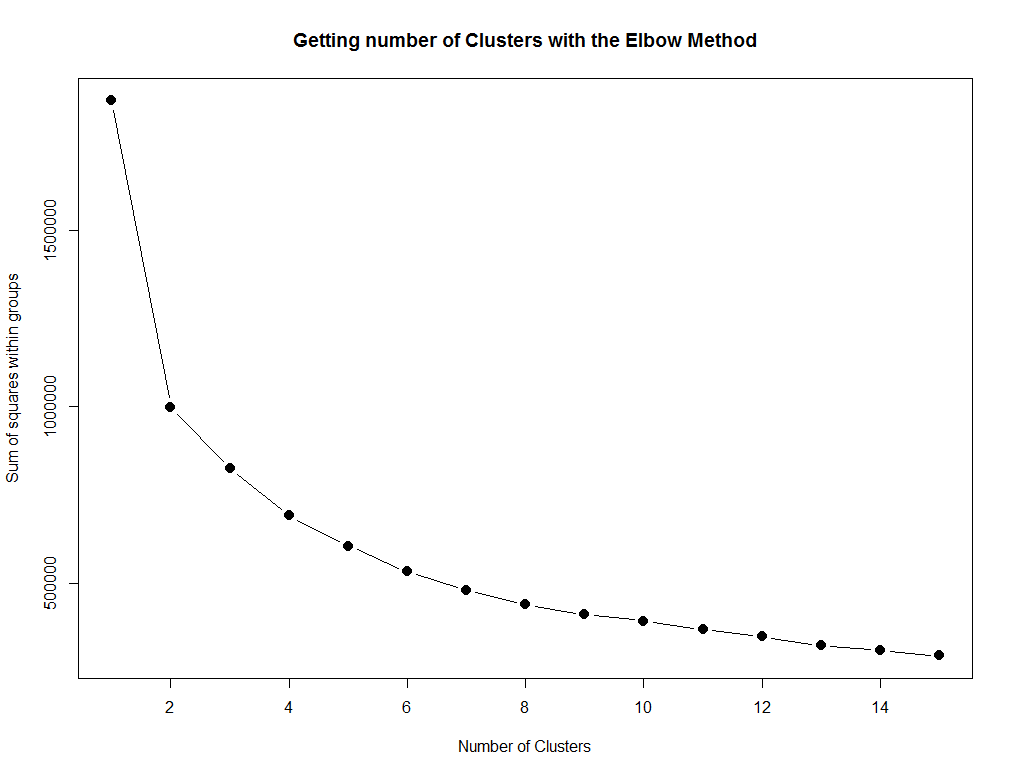
a = (nrow(data)-1)\*sum(apply(data,2,var))

for (i in 2:15) a[i] = sum(kmeans(data, centers=i)$withinss)

plot(1:15, a, type="b", xlab="Number of Clusters", ylab="Sum of squares within groups",

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

The plot -



As per the plot above, we can conclude that post 7 clusters the difference within cluster dissimilarity is not significant.

Hence we can say that optimal number of clusters to be used is 7.