k-Means και Clustering Evaluation στην R

1 Εισαγωγή

1.1 Εισαγωγή στον k-Means

Ο k-Means είναι ένας αλγόριθμος διαχωρισμού που χρησιμοποιείται για να ομαδοποιήσει τα δείγματα ενός σετ x1, x2, … σε ομάδες C1, C2, … με κέντρα m1, m2, … Ο σκοπός του αλγορίθμου είναι να βρει την ομαδοποίηση που ελαχιστοποιεί το άθροισμα των τετραγώνων των αποστάσεων κάθε σημείου από το κέντρο της ομάδας που ανήκει (Within cluster Sum of Squares – WSS ή SSE):

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - m_i\|^2$$

όπου K είναι ο αριθμός των ομάδων (clusters) και C_i είναι η i-οστή ομάδα που έχει κέντρο (centroid) το m_i. Ο αλγόριθμος επαναλαμβάνει διαδοχικά δύο βήματα: την ανάθεση σε κέντρα των ομάδων και την ενημέρωση των κέντρων. Με δεδομένα κάποια αρχικά κέντρα, αρχικά ο αλγόριθμος τοποθετεί κάθε σημείο στην ομάδα που βρίσκεται το κοντινότερο κέντρο. Στη συνέχεια, αφού έχουν τοποθετηθεί όλα τα σημεία σε ομάδες, για κάθε ομάδα υπολογίζεται το νέο κέντρο ως ο μέσος όρος όλων των σημείων της ομάδας.

Η πιο γνωστή παραλλαγή του αλγορίθμου είναι ο αλγόριθμος k-Medoids, ο οποίος χρησιμοποιεί σημεία από τα δεδομένα ως medoids των clusters (αντί για μέσους όρους σημείων όπως ο k-Means). Έτσι, επιτρέπει τη χρήση κατηγορικών δεδομένων καθώς και διαφόρων συναρτήσεων απόστασης.

1.2 k-Means στην R

Για ομαδοποίηση με χρήση του k-Means στην R, χρησιμοποιούμε την εντολή kmeans:

```
> model = kmeans(data, centers = 3)
```

Η εντολή centers μπορεί να πάρει ως όρισμα είτε τα αρχικά κέντρα είτε τον αριθμό των κέντρων που θα θέλαμε να βρει ο αλγόριθμος οπότε τα κέντρα αρχικοποιούνται τυχαία.

Μπορούμε επίσης να εμφανίσουμε τα centroids και την κατανομή των σημείων σε clusters:

```
> model$centers
```

```
> model$cluster
```

1.3 k-Medoids στην R

Για να χρησιμοποιήσουμε τον αλγόριθμο k-Medoids χρειαζόμαστε τη βιβλιοθήκη cluster:

```
> library(cluster)
```

Για να εφαρμόσουμε τον αλγόριθμο για 3 clusters εκτελούμε την εντολή:

```
> model = pam(conferences, 3)
```

Μπορούμε να εμφανίσουμε τα medoids με την εντολή `model$medoids` και την κατανομή των σημείων σε clusters με την εντολή `model$clustering`. Αν το dataset έχει κατηγορικές μεταβλητές, μπορούμε επίσης να εμφανίσουμε τα medoids με τις μεταβλητές τους με την εντολή `data[model$id.med,]`.

1.4 Clustering Evaluation στην R

Για να αξιολογήσουμε ένα μοντέλο ομαδοποίησης χρησιμοποιούμε τη βιβλιοθήκη cluster:

```
> library(cluster)
```

Έχοντας το αποτέλεσμα του k-means (`model`) μπορούμε αρχικά να το λάβουμε το cohesion (within-cluster sum of squares) και το separation (between-cluster sum of squares):

```
> cohesion = model$tot.withinss
> separation = model$betweenss
```

Για τον υπολογισμό του silhouette εκτελούμε την εντολή:

```
> model_silhouette = silhouette(model$cluster, dist(data))
```

όπου `data` είναι τα αρχικά δεδομένα. Στη συνέχεια, μπορούμε να σχεδιάσουμε το silhouette plot με την εντολή `plot(model_silhouette)` ή να εμφανίσουμε το μέσο silhouette με την εντολή `mean(model_silhouette[, 3])`.

Για να κατασκευάσουμε ένα heatmap αρχικά ταξινομούμε τα δεδομένα με βάση το cluster στο οποίο ανήκουν:

```
> data_ord = data[order(model$cluster),]
```

Στη συνέχεια, σχεδιάζουμε το heatmap με την παρακάτω εντολή:

```
> heatmap(as.matrix(dist(data_ord)), Rowv = NA, Colv = NA, col =
          heat.colors(256), revC = TRUE)
```

2 Κατασκευή Μοντέλου k-Means

Για την κατασκευή ενός μοντέλου k-Means θα χρησιμοποιήσουμε ως εφαρμογή τα δεδομένα εκπαίδευσης του παρακάτω πίνακα για ένα πρόβλημα ομαδοποίησης.

	X	**Y**
x1	7	1
x2	3	4
x3	1	5
x4	5	8
x5	1	3
x6	7	8
x7	8	2
x8	5	9

Θα απαντήσουμε στα παρακάτω ερωτήματα:

α) Σχεδιάστε τα δεδομένα.

β) Εφαρμόστε τον αλγόριθμο K-means ώστε να ομαδοποιηθούν τα σημεία σε 3 ομάδες. Θεωρήστε πως τα αρχικά κέντρα είναι τα x1, x2 και x3.

γ) Υπολογίστε το cohesion και το separation της τελικής ομαδοποίησης.

δ) Κατασκευάστε το μοντέλο στην R και επαναλάβετε τα ερωτήματα (β) και (γ).

ε) Σχεδιάστε εκ νέου τα δεδομένα με διαφορετικά χρώματα για κάθε cluster και στο ίδιο διάγραμμα σχεδιάστε επίσης τα centroids.

2.1 Κατασκευή Δεδομένων και Εισαγωγή Βιβλιοθηκών

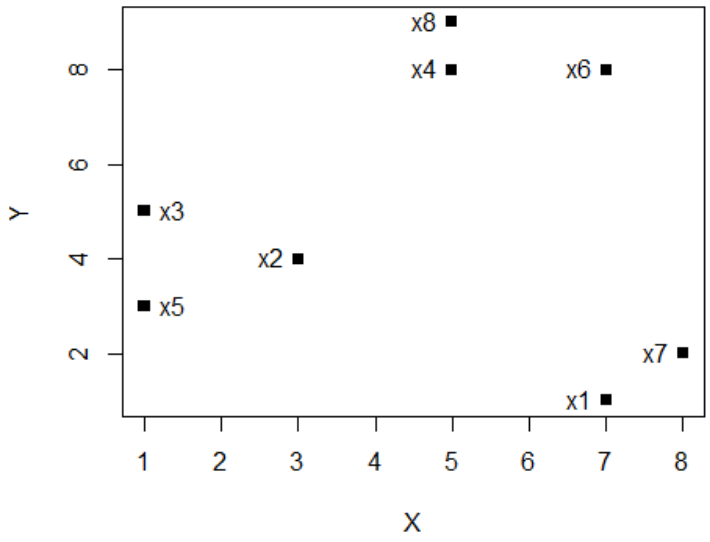
Αρχικά κατασκευάζουμε τα δεδομένα με τις παρακάτω εντολές:

```
> X = c(7, 3, 1, 5, 1, 7, 8, 5)
> Y = c(1, 4, 5, 8, 3, 8, 2, 9)
> rnames = c("x1", "x2", "x3", "x4", "x5", "x6", "x7", "x8")
> kdata = data.frame(X, Y, row.names = rnames)
```

Στη συνέχεια μπορούμε να τα σχεδιάσουμε με τις εντολές:

```
> plot(kdata, pch = 15)
> text(kdata, labels = row.names(kdata), pos = 2)
```

Τα δεδομένα φαίνονται στο σχήμα:



2.2 Υπολογισμός k-Means

Για το ερώτημα (β) εφαρμόζουμε τον k-Means με αρχικά κέντρα τα A1, A2, A3:

$$z_1 = \begin{bmatrix}7\\1\end{bmatrix} \qquad z_2 = \begin{bmatrix}3\\4\end{bmatrix} \qquad z_3 = \begin{bmatrix}1\\5\end{bmatrix}$$

1η επανάληψη

Υπολογίζουμε τις αποστάσεις όλων των σημείων από τα κέντρα:

	z1	**z2**	**z3**
x1	**0.00**	5.00	7.21
x2	5.00	**0.00**	2.24
x3	7.21	2.24	**0.00**
x4	7.28	**4.47**	5.00
x5	6.32	2.24	**2.00**
x6	7.00	**5.66**	6.71
x7	**1.41**	5.39	7.62
x8	8.25	**5.39**	5.66

Τα νέα κέντρα είναι:

$$z_1' = \frac{x1 + x7}{2} = \begin{bmatrix}7.5\\1.5\end{bmatrix} \qquad z_2' = \frac{x2 + x4 + x6 + x8}{4} = \begin{bmatrix}5\\7.25\end{bmatrix} \qquad z_3' = \frac{x3 + x5}{2} = \begin{bmatrix}1\\4\end{bmatrix}$$

Τα κέντρα έχουν αλλάξει οπότε συνεχίζουμε με την επόμενη επανάληψη του αλγορίθμου.

2η επανάληψη

Υπολογίζουμε τις αποστάσεις όλων των σημείων από τα κέντρα:

	z1′	z2′	z3′
x1	**0.71**	6.56	6.71
x2	5.15	3.82	**2.00**
x3	7.38	4.59	**1.00**
x4	6.96	**0.75**	5.66
x5	6.67	5.84	**1.00**
x6	6.52	**2.14**	7.21
x7	**0.71**	6.05	7.28
x8	7.91	**1.75**	6.40

Τα νέα κέντρα είναι:

$$z_1'' = \frac{x1 + x7}{2} = \begin{bmatrix} 7.5 \\ 1.5 \end{bmatrix} \qquad z_2'' = \frac{x4 + x6 + x8}{3} = \begin{bmatrix} 5.67 \\ 8.33 \end{bmatrix} \qquad z_3'' = \frac{x2 + x3 + x5}{3} = \begin{bmatrix} 1.67 \\ 4 \end{bmatrix}$$

Τα κέντρα έχουν αλλάξει οπότε συνεχίζουμε με την επόμενη επανάληψη του αλγορίθμου.

3η επανάληψη

Υπολογίζουμε τις αποστάσεις όλων των σημείων από τα κέντρα:

	z1′	z2′	z3′
x1	**0.71**	7.45	6.12
x2	5.15	5.09	**1.33**
x3	7.38	5.73	**1.20**
x4	6.96	**0.75**	5.21
x5	6.67	7.09	**1.20**
x6	6.52	**1.37**	6.67
x7	**0.71**	6.75	6.64
x8	7.91	**0.94**	6.01

Τα νέα κέντρα είναι:

$$z_1''' = \frac{x1 + x7}{2} = \begin{bmatrix} 7.5 \\ 1.5 \end{bmatrix} \qquad z_2''' = \frac{x4 + x6 + x8}{3} = \begin{bmatrix} 5.67 \\ 8.33 \end{bmatrix} \qquad z_3''' = \frac{x2 + x3 + x5}{3} = \begin{bmatrix} 1.67 \\ 4 \end{bmatrix}$$

Τα κέντρα έχουν σταθεροποιηθεί οπότε ο αλγόριθμος έχει συγκλίνει στις 3 επαναλήψεις.

2.3 Υπολογισμός Cohesion και Separation

Θεωρούμε τα clusters C_1, C_2, C_3 με τα centroids m_1, m_2, m_3 όπως υπολογίστηκαν παραπάνω.

Για το (γ), υπολογίζουμε το Cohesion (Within cluster Sum of Squares) για κάθε cluster:

$$WSS(C_1) = \sum_{x \in C_1} \|x - m_1\|^2 = \|x_1 - m_1\|^2 + \|x_7 - m_1\|^2$$

$$= \sqrt{(7-7.5)^2 + (1-1.5)^2}^2 + \sqrt{(8-7.5)^2 + (2-1.5)^2}^2$$

$$= 0.25 + 0.25 + 0.25 + 0.25 = 1$$

$$WSS(C_2) = \sum_{x \in C_2} \|x - m_2\|^2 = \|x_4 - m_2\|^2 + \|x_6 - m_2\|^2 + \|x_8 - m_2\|^2$$

$$= \sqrt{(5-5.67)^2 + (8-8.33)^2}^2 + \sqrt{(7-5.67)^2 + (8-8.33)^2}^2$$
$$+ \sqrt{(5-5.67)^2 + (9-8.33)^2}^2$$

$$= 0.4489 + 0.1089 + 1.7689 + 0.1089 + 0.4489 + 0.4489 = 3.3334$$

$$WSS(C_3) = \sum_{x \in C_3} \|x - m_3\|^2 = \|x_2 - m_2\|^2 + \|x_3 - m_2\|^2 + \|x_5 - m_2\|^2$$

$$= \sqrt{(3-1.67)^2 + (4-4)^2}^2 + \sqrt{(1-1.67)^2 + (5-4)^2}^2$$
$$+ \sqrt{(1-1.67)^2 + (3-4)^2}^2$$

$$= 1.7689 + 0 + 0.4489 + 1 + 0.4489 + 1 = 4.6667$$

Το συνολικό Cohesion για όλα τα clusters θα είναι:

$$WSS_{Total} = \sum_i WSS(C_i) = WSS(C_1) + WSS(C_2) + WSS(C_3) = 1 + 3.3334 + 4.6667 = 9$$

Για το Separation θα υπολογίσουμε αρχικά το μέσο όρο των δεδομένων:

$$m = \frac{x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8}{8} = \begin{bmatrix} 4.625 \\ 5 \end{bmatrix}$$

Το Separation (Between cluster Sum of Squares) θα είναι:

$$BSS_{Total} = \sum_i |C_i| \cdot \|m - m_1\|^2$$

$$= 2 \cdot \sqrt{(4.625-7.5)^2 + (5-1.5)^2}^2 + 3 \cdot \sqrt{(4.625-5.67)^2 + (5-8.33)^2}^2$$
$$+ 3 \cdot \sqrt{(4.625-1.67)^2 + (5-4)^2}^2$$

$$= 41.031 + 36.589 + 29.255 = 106.875$$

2.3 Κατασκευή Μοντέλου με την R

Για το ερώτημα (δ) εφαρμόζουμε τον k-Means για το dataset:

```
> model = kmeans(kdata, centers = kdata[1:3,])
```

Εμφανίζουμε αν θέλουμε τα centroids και την κατανομή των σημείων σε clusters:

```
> model$centers
```

```
> model$cluster
```

Υπολογίζουμε εκ νέου το ερώτημα (γ):

```
> cohesion = model$tot.withinss
```

```
> separation = model$betweenss
```

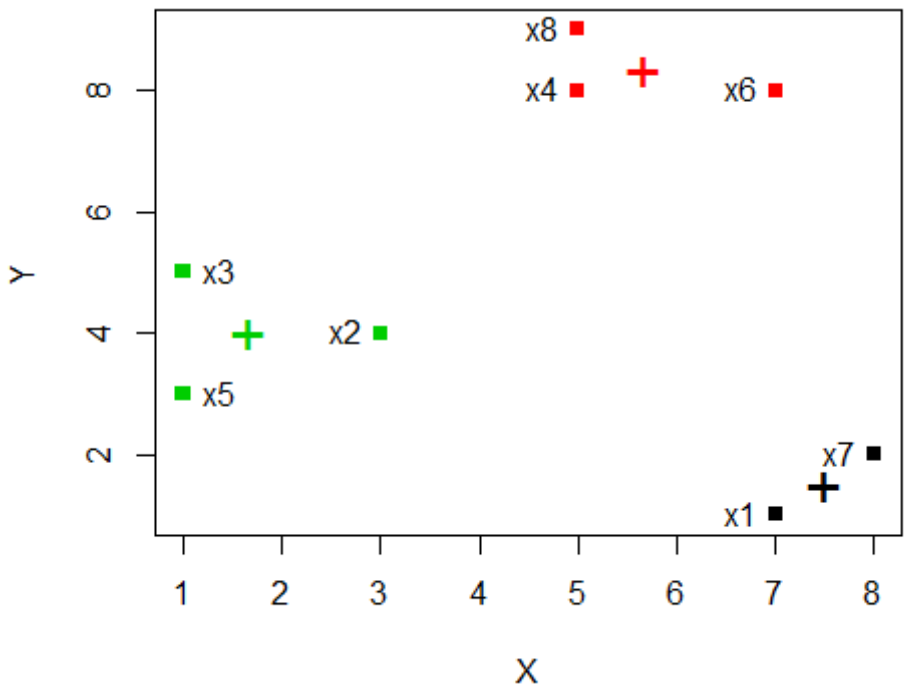
Μπορούμε τέλος να σχεδιάσουμε τα δεδομένα με τα clusters (ερώτημα (ε)) με τις εντολές:

```
> plot(kdata, col = model$cluster, pch = 15)
```

```
> text(kdata, labels = row.names(kdata), pos = 2)
```

```
> points(model$centers, col = 1:length(model$centers), pch = "+", cex = 2)
```

Οπότε προκύπτει η ομαδοποίηση που φαίνεται στο σχήμα:



3 Εφαρμογή με k-Means και Μετρικές Αξιολόγησης

Για την κατασκευή ενός μοντέλου k-Means θα χρησιμοποιήσουμε ως εφαρμογή τα δεδομένα εκπαίδευσης του αρχείου που δίνεται (cdata.txt) για ένα πρόβλημα ομαδοποίησης.

Ένα summary των δεδομένων με την R είναι το παρακάτω:

```
       X1                X2                Y
 Min.   :-4.720   Min.   :-3.974   Min.   :1
 1st Qu.:-1.237   1st Qu.:-1.059   1st Qu.:1
 Median : 1.757   Median : 2.096   Median :2
 Mean   : 1.322   Mean   : 1.532   Mean   :2
 3rd Qu.: 3.592   3rd Qu.: 3.675   3rd Qu.:3
 Max.   : 6.077   Max.   : 6.689   Max.   :3
```

Αρχικά, κάνουμε import τη βιβλιοθήκη cluster, εισάγουμε τα δεδομένα και τα διαχωρίζουμε:

```
> library(cluster)

> cdata = read.csv("cdata.txt")

> target = cdata[, 3]

> cdata = cdata[, 1:2]
```

Στη συνέχεια, θα απαντήσουμε στα παρακάτω ερωτήματα:

α) Σχεδιάστε τις τιμές του συνόλου δεδομένων με διαφορετικά χρώματα για κάθε κατηγορία.

β) Εφαρμόζοντας τον αλγόριθμο k-Means επιλέξτε τον ελάχιστο αριθμό των clusters έτσι ώστε να περιγράφουν ικανοποιητικά τα δεδομένα με βάση το SSE.

γ) Εφαρμόστε τον k-Means για 3 clusters.

δ) Υπολογίστε το cohesion και το separation για την ομαδοποίηση που προέκυψε.

ε) Σχεδιάστε εκ νέου τα δεδομένα με διαφορετικά χρώματα για κάθε cluster και στο ίδιο διάγραμμα σχεδιάστε επίσης τα centroids.

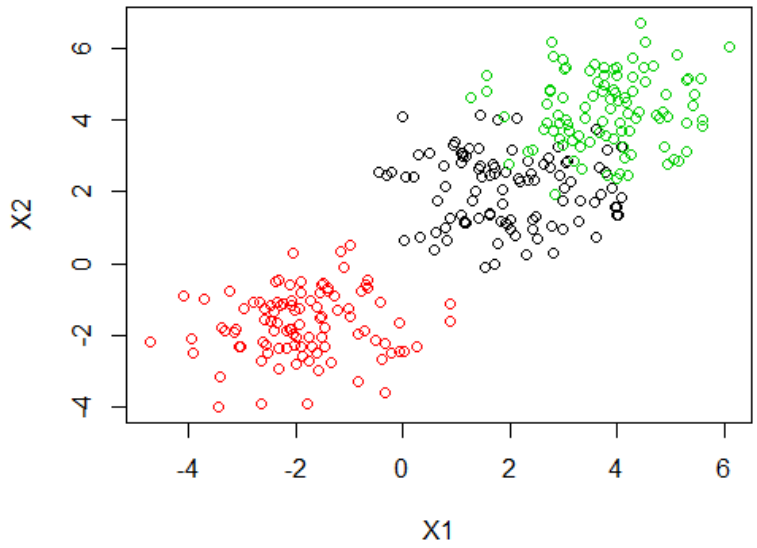
στ) Υπολογίστε το silhouette και σχεδιάστε το silhouette plot για την ομαδοποίηση που προέκυψε.

ζ) Σχεδιάστε το heatmap για την ομαδοποίηση που προέκυψε.

3.1 Επιλογή Αριθμού Clusters με τον k-Means

Αρχικά, μπορούμε να σχεδιάσουμε τα δεδομένα με την παρακάτω εντολή (ερώτημα (α)):

```
> plot(cdata, col = target)
```



Για να βρούμε ένα πλήθος clusters που να περιγράφουν τα δεδομένα (ερώτημα (β)), υπολογίζουμε το SSE για τους διαχωρισμούς σε 1, 2, …, 10 clusters με τις παρακάτω εντολές:

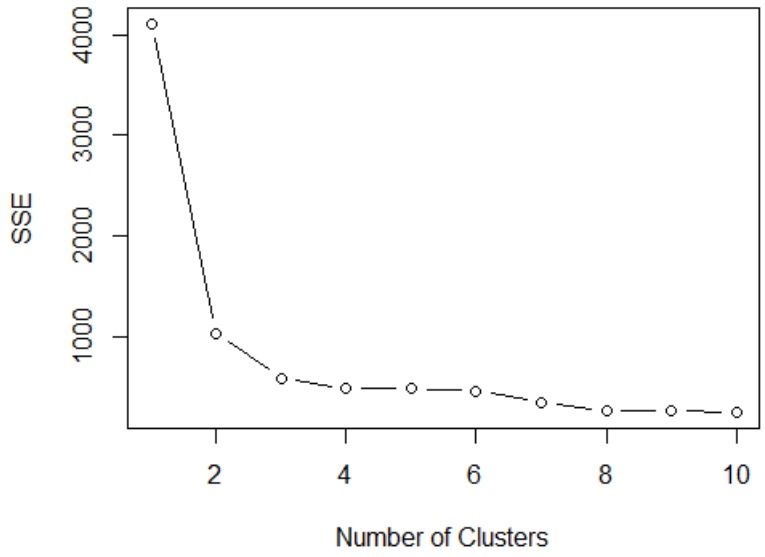
```
SSE <- (nrow(cdata) - 1) * sum(apply(cdata, 2, var))

for (i in 2:10)

  SSE[i] <- kmeans(cdata, centers = i)$tot.withinss
```

Στη συνέχεια μπορούμε να σχεδιάσουμε το SSE με την παρακάτω εντολή:

```
plot(1:10, SSE, type="b", xlab="Number of Clusters", ylab="SSE")
```



3.2 Κατασκευή Μοντέλου k-Means

Για το ερώτημα (γ) εφαρμόζουμε τον k-Means για το dataset:

```
> model = kmeans(cdata, centers = 3)
```

Εμφανίζουμε αν θέλουμε τα centroids και την κατανομή των σημείων σε clusters:

```
> model$centers
```

```
> model$cluster
```

Υπολογίζουμε το cohesion και το separation (ερώτημα (δ)):

```
> cohesion = model$tot.withinss
```

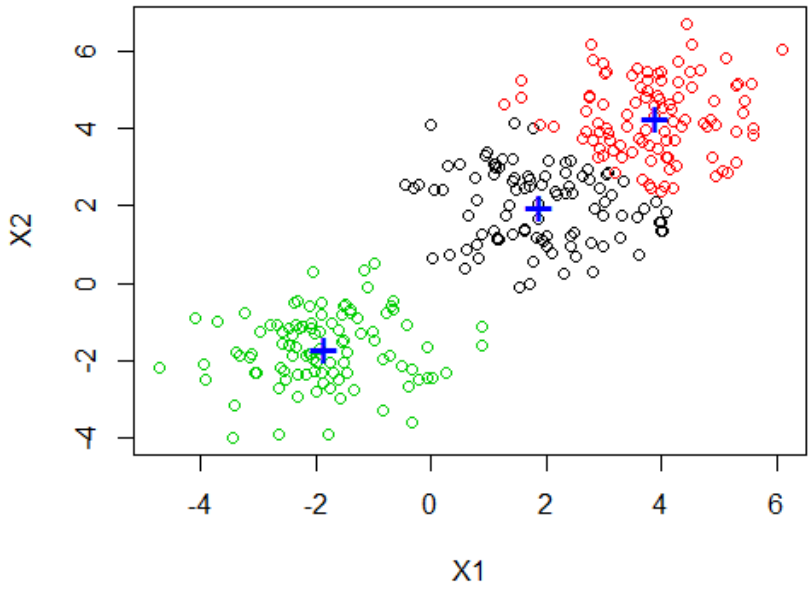
```
> separation = model$betweenss
```

Μπορούμε τέλος να σχεδιάσουμε τα δεδομένα με τα clusters (ερώτημα (ε)) με τις εντολές:

```
> plot(cdata, col = model$cluster)
```

```
> points(model$centers, col = 4, pch = "+", cex = 2)
```

Οπότε προκύπτει η ομαδοποίηση που φαίνεται στο σχήμα:



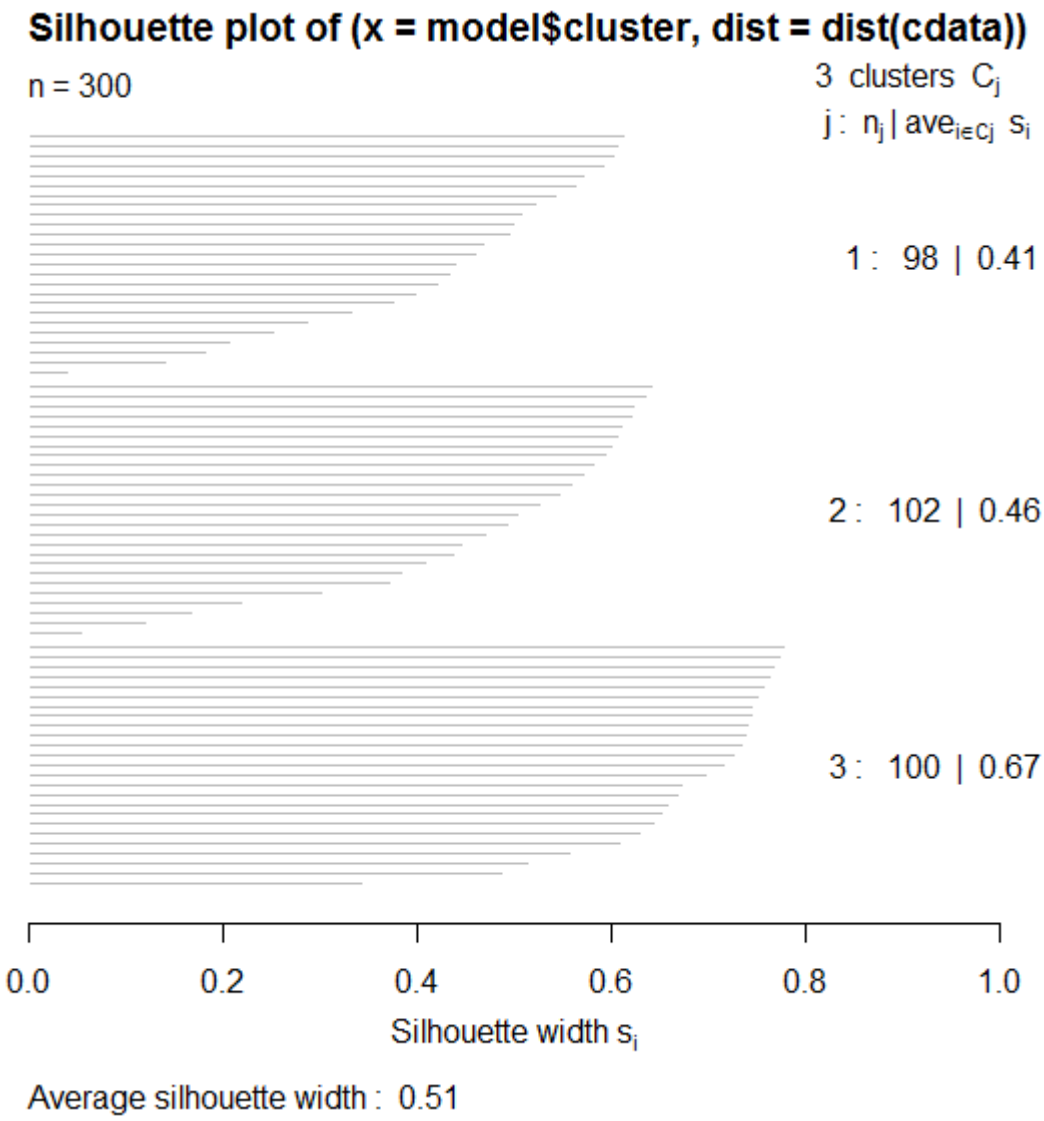
3.3 Υπολογισμός Silhouette και Κατασκευή Silhouette Plot

Για τον υπολογισμό του silhouette (ερώτημα (στ)) εκτελούμε την εντολή:

```
> model_silhouette = silhouette(model$cluster, dist(cdata))
```

Στη συνέχεια μπορούμε να σχεδιάσουμε το silhouette plot με την παρακάτω εντολή:

```
> plot(model_silhouette)
```



Μπορούμε επίσης να εμφανίσουμε το μέσο silhouette με την εντολή:

```
> mean_silhouette = mean(model_silhouette[, 3])
```

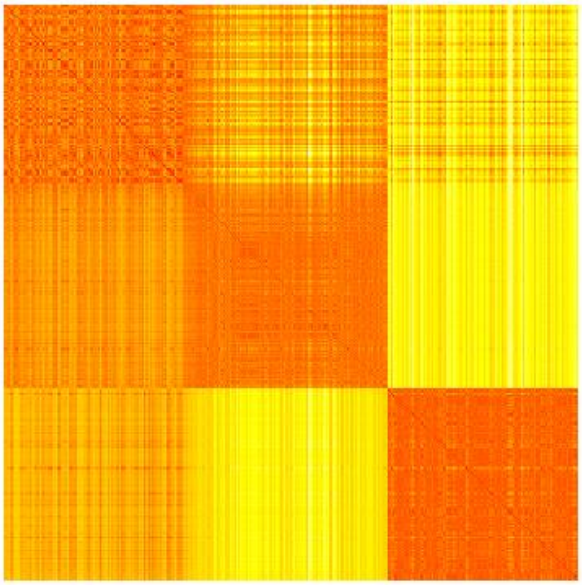
3.4 Κατασκευή Heatmap

Για να κατασκευάσουμε το heatmap (ερώτημα (ζ)) αρχικά ταξινομούμε τα δεδομένα με βάση το cluster στο οποίο ανήκουν:

```
> cdata_ord = cdata[order(model$cluster),]
```

Στη συνέχεια, σχεδιάζουμε το heatmap με την παρακάτω εντολή:

```
> heatmap(as.matrix(dist(cdata_ord)), Rowv = NA, Colv = NA,
          col = heat.colors(256), revC = TRUE)
```



4 Εφαρμογή με k-Medoids

Για την κατασκευή ενός μοντέλου k-Medoids θα χρησιμοποιήσουμε ως εφαρμογή τα δεδομένα εκπαίδευσης του παρακάτω πίνακα για ένα πρόβλημα ομαδοποίησης.

Rank	Topic
High	SE
Low	SE
High	ML
Low	DM
Low	ML
High	SE

Θα απαντήσουμε στα παρακάτω ερωτήματα:

α) Εφαρμόστε τον αλγόριθμο k-Medoids ώστε να ομαδοποιηθούν τα δεδομένα σε 3 ομάδες. Εμφανίστε τα κέντρα των 3 ομάδων.

β) Σχεδιάστε τα δεδομένα με διαφορετικά χρώματα και σύμβολα για κάθε medoid.

4.1 Κατασκευή Δεδομένων και Εισαγωγή Βιβλιοθηκών

Εισάγουμε τη βιβλιοθήκη cluster και κατασκευάζουμε τα δεδομένα με τις εντολές:

```
> library(cluster)
> Rank = c("High", "Low", "High", "Low", "Low", "High")
> Topic = c("SE", "SE", "ML", "DM", "ML", "SE")
> conferences = data.frame(Rank, Topic)
```

4.2 Κατασκευή Μοντέλου k-Medoids

Για το ερώτημα (α) εφαρμόζουμε τον k-Medoids για το dataset:

```
> model = pam(conferences, 3)
```

Εμφανίζουμε τα medoids και την κατανομή των σημείων σε clusters:

```
> model$medoids
```

```
> model$clustering
```

Μπορούμε επίσης να εμφανίσουμε τα κέντρα των ομάδων με τις μεταβλητές τους:

```
> conferences[model$id.med,]
```

Στη συνέχεια, για το ερώτημα (β) σχεδιάζουμε τα δεδομένα με τις παρακάτω εντολές:

```
> L1 = levels(conferences$Rank)
```

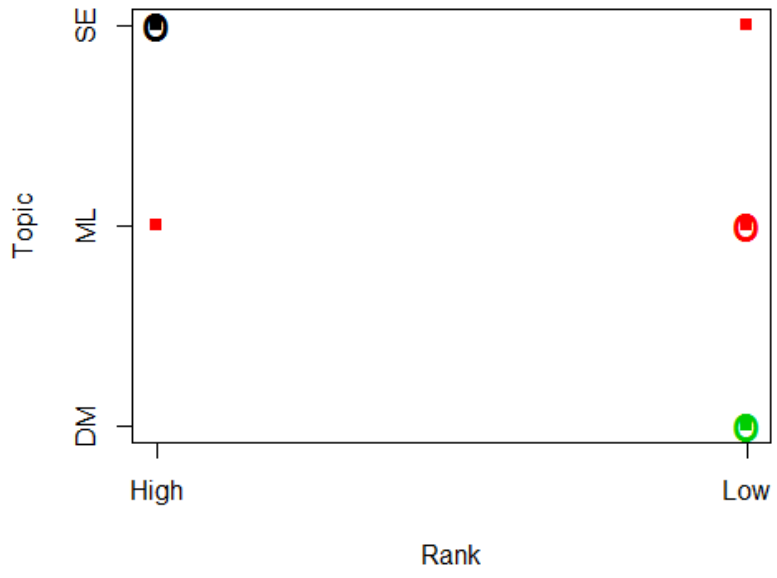
```
> L2 = levels(conferences$Topic)
```

```
> plot(model$data, xaxt = "n", yaxt = "n", pch = 15, col = model$cluster)
```

```
> axis(1, at = 1:length(L1), labels = L1)
```

```
> axis(2, at = 1:length(L2), labels = L2)
```

```
> points(conferences[model$id.med,], col = 1:3, pch = "o", cex = 2)
```



5 Πρόβλημα για Εξάσκηση

Για την κατασκευή ενός μοντέλου k-Means δίνονται τα δεδομένα του αρχείου kdata.txt για ένα πρόβλημα ομαδοποίησης.

Ένα summary των δεδομένων με την R είναι το παρακάτω:

```
       X1                  X2
 Min.   :-6.45171   Min.   :-1.9740
 1st Qu.:-3.14300   1st Qu.: 0.9636
 Median : 0.06884   Median : 9.3382
 Mean   :-0.01162   Mean   : 6.8853
 3rd Qu.: 3.05494   3rd Qu.:10.5456
 Max.   : 6.07726   Max.   :12.6890
```

Αρχικά, εισάγετε τα δεδομένα:

```
> sdata = read.csv("sdata.txt")
```

Στη συνέχεια, απαντήστε στα παρακάτω ερωτήματα:

α) Ομαδοποιήστε τα δεδομένα χρησιμοποιώντας τον k- Means με αρχικά κέντρα τα (-4, 10), (0, 0), και (4, 10).

β) Σχεδιάστε την ομαδοποίηση που προέκυψε και υπολογίστε τις μετρικές cohesion, separation και silhouette.

γ) Ομαδοποιήστε τα δεδομένα χρησιμοποιώντας τον k- Means με αρχικά κέντρα τα (-2, 0), (2, 0), και (0, 10).

δ) Σχεδιάστε τη νέα ομαδοποίηση που προκύπτει και υπολογίστε εκ νέου τις μετρικές cohesion, separation και silhouette. Συγκρίνετε τα δύο μοντέλα.