CSIS734-01 Data Mining & Predictive Analytics
Garth Mortensen, mort0052@stthomas.edu

Graduate Program in Software
CSIS 734-01: Data Mining & Predictive Analytics
Assignment #9 (100 points)
Due Date: May 12th, 2018

1) What are the Silhouette coefficients for the following 7 points and their clustering results?

Data	**Cluster**	**Silhouette Coefficient**
1	1	
2	1	
5	1	
11	1	
12	2	
20	3	
21	3	

Definitions

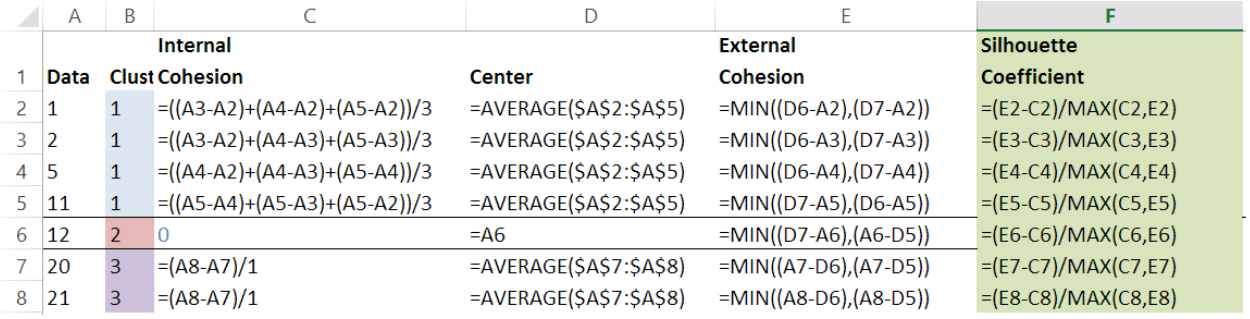
Internal cohesion
Find the distance of each point to all others in the same cluster, then average them.

Center
Find the center by averaging their locations

External cohesion
Find the distance of each point to the centers of the other clusters,
you take minimum to penalize it

Silhouette coefficient
Sq = (Eq - Iq) / Max(Eq, Iq)
1 is the best, -1 is the worse

Equations:

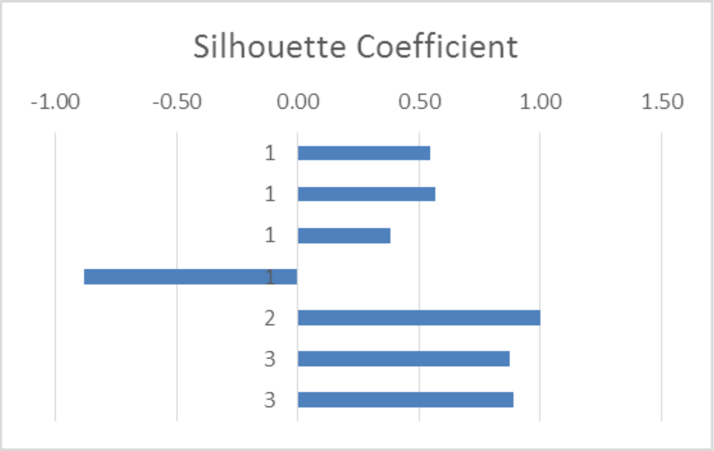


	A	B	C	D	E	F
1	Data	Clust	Internal Cohesion	Center	External Cohesion	Silhouette Coefficient
2	1	1	=((A3-A2)+(A4-A2)+(A5-A2))/3	=AVERAGE($A$2:$A$5)	=MIN((D6-A2),(D7-A2))	=(E2-C2)/MAX(C2,E2)
3	2	1	=((A3-A2)+(A4-A3)+(A5-A3))/3	=AVERAGE($A$2:$A$5)	=MIN((D6-A3),(D7-A3))	=(E3-C3)/MAX(C3,E3)
4	5	1	=((A4-A2)+(A4-A3)+(A5-A4))/3	=AVERAGE($A$2:$A$5)	=MIN((D6-A4),(D7-A4))	=(E4-C4)/MAX(C4,E4)
5	11	1	=((A5-A4)+(A5-A3)+(A5-A2))/3	=AVERAGE($A$2:$A$5)	=MIN((D7-A5),(D6-A5))	=(E5-C5)/MAX(C5,E5)
6	12	2	0	=A6	=MIN((D7-A6),(A6-D5))	=(E6-C6)/MAX(C6,E6)
7	20	3	=(A8-A7)/1	=AVERAGE($A$7:$A$8)	=MIN((A7-D6),(A7-D5))	=(E7-C7)/MAX(C7,E7)
8	21	3	=(A8-A7)/1	=AVERAGE($A$7:$A$8)	=MIN((A8-D6),(A8-D5))	=(E8-C8)/MAX(C8,E8)

Answers:

Data	Cluster	Internal Cohesion	Center	External Cohesion	Silhouette Coefficient
1	1	5.00	4.75	11.00	0.55
2	1	4.33	4.75	10.00	0.57
5	1	4.33	4.75	7.00	0.38
11	1	8.33	4.75	1.00	-0.88
12	2	0.00	12.00	7.25	1.00
20	3	1.00	20.50	8.00	0.88
21	3	1.00	20.50	9.00	0.89

Visual:



2) Use the following distance table to draw two dendrograms: one use single-link method and another one uses complete-link method.

	A	**B**	**C**	**D**	**E**
A	0	1	4	5	6
B	1	0	3	4	5
C	4	3	0	1	2
D	5	4	1	0	1
E	6	5	2	1	0

Distance table

	A	**B**	**C**	**D**	**E**
A	0	1	4	5	6
B	1	0	3	4	5
C	4	3	0	1	2
D	5	4	1	0	1
E	6	5	2	1	0

Single-link

min of min distances among clusters

1 ((A, B), (C), (D), (E))   A-B has shortest distance 1

2 ((A, B), (C, D), (E))   C-D has shortest distance 1   *e.g. AB->C min(4,3) = 3*

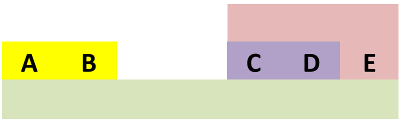
3 (A, B, C, D), (E)

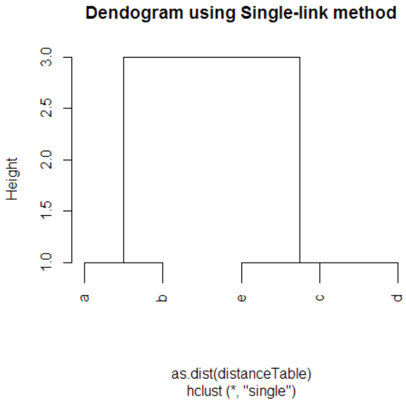
d(A, B), (C, D) = 3,   d(A, B), (E) = 5,   d(C, D), (E) = 1

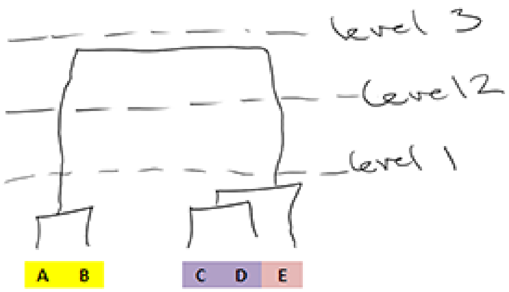
CD-E has shortest distance 1

4 (A, B), (C, D, E)

d(A, B), (C, D, E) = 3   AB-CDE has shortest distance 3







Distance table

	A	B	C	D	E
A	0	1	4	5	6
B	1	0	3	4	5
C	4	3	0	1	2
D	5	4	1	0	1
E	6	5	2	1	0

Complete-link

min of max distances among clusters

1 ((A, B), (C), (D), (E)) A-B has shortest distance 1

2 ((A, B), (C, D), (E)) C-D has shortest distance 1 *e.g. AB->C max(4,3) = 4*

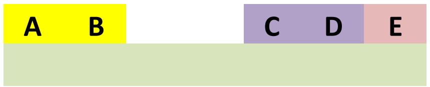
3 (A, B, C, D), (E)

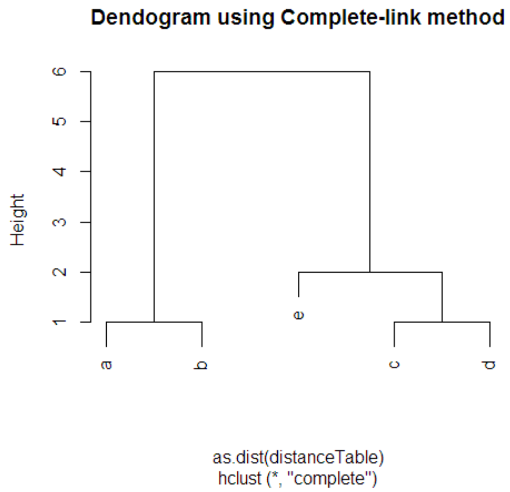
d(A, B), (C, D) = 5, d(A, B), (E) = 6, d(C, D), (E) = 2

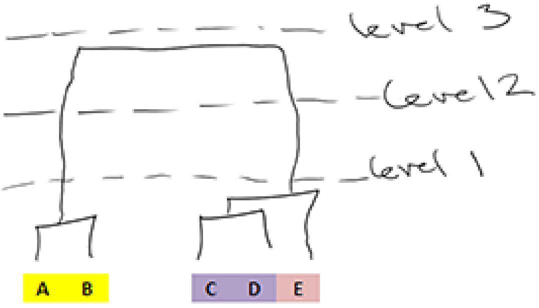
CD-E has the shortest distance 2

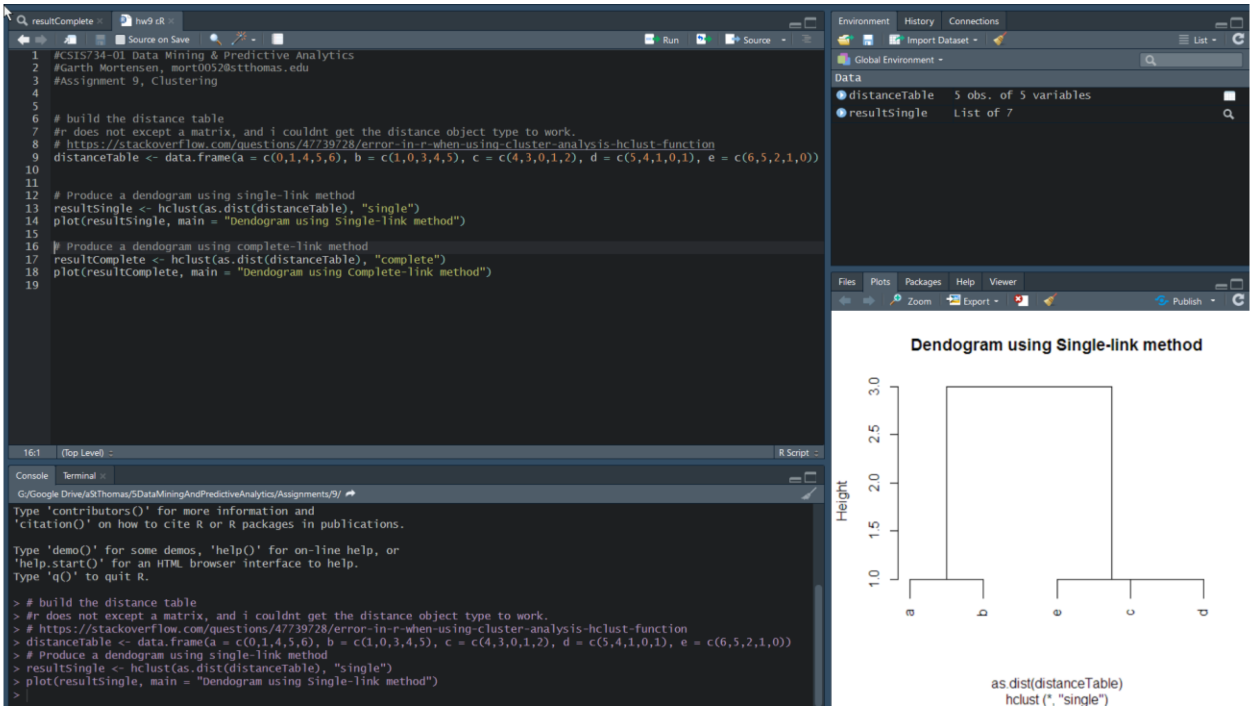
4 (A, B), (C, D, E)

d(A, B), (C, D, E) = 6 AB-CDE has the shortest distance 6









```r
#CSIS734-01 Data Mining & Predictive Analytics
#Garth Mortensen, mort0052@stthomas.edu
#Assignment 9, Clustering


# build the distance table
#r does not except a matrix, and i couldnt get the distance object type to work.
# https://stackoverflow.com/questions/47739728/error-in-r-when-using-cluster-analysis-hclust-function
distanceTable <- data.frame(a = c(0,1,4,5,6), b = c(1,0,3,4,5), c = c(4,3,0,1,2), d = c(5,4,1,0,1), e = c(6,5,2,1,0))


# Produce a dendogram using single-link method
resultSingle <- hclust(as.dist(distanceTable), "single")
plot(resultSingle, main = "Dendogram using Single-link method")

# Produce a dendogram using complete-link method
resultComplete <- hclust(as.dist(distanceTable), "complete")
plot(resultComplete, main = "Dendogram using Complete-link method")

```

Data

```
distanceTable    5 obs. of 5 variables
resultSingle     List of 7
```

G:/Google Drive/aStThomas/5DataMiningAndPredictiveAnalytics/Assignments/9/

```
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> # build the distance table
> #r does not except a matrix, and i couldnt get the distance object type to work.
> # https://stackoverflow.com/questions/47739728/error-in-r-when-using-cluster-analysis-hclust-function
> distanceTable <- data.frame(a = c(0,1,4,5,6), b = c(1,0,3,4,5), c = c(4,3,0,1,2), d = c(5,4,1,0,1), e = c(6,5,2,1,0))
> # Produce a dendogram using single-link method
> resultSingle <- hclust(as.dist(distanceTable), "single")
> plot(resultSingle, main = "Dendogram using Single-link method")
> 
```