Homework 5

Jarett Smith, Balin Allred, Charlie Deaton, and Max Thompson

2023-11-06

Questions

1. In the following, we have five data points and three cluster centers:

Table 1: Points

p1	0	0	6	2	5
p2	0	6	6	2	1
p3	2	4	6	0	3
p4	1	3	7	0	6
p5	3	3	4	0	0

Table 2: Clusters

c1	0	0	1	1	1
c2	0	1	1	1	0
c3	1	1	1	0	0

a. Please for each data point find the closest cluster center, based on Euclidean distance.

```
# Cluster 1
ed1<- sqrt(sum((p1-c1)^2))
# Cluster 2
ed2<- sqrt(sum((p1-c2)^2))
# Cluster 3
ed3<- sqrt(sum((p1-c3)^2))
ed1;ed2;ed3
```

```
## [1] 6.480741
```

```
## [1] 7.211103
```

```
## [1] 7.483315
```

```
# Result: Point 1 will be grouped with Cluster 1
```

```r
# Cluster 1
ed1<- sqrt(sum((p2-c1)^2))
# Cluster 2
ed2<- sqrt(sum((p2-c2)^2))
# Cluster 3
ed3<- sqrt(sum((p2-c3)^2))
ed1;ed2;ed3
```

```
## [1] 7.874008
```

```
## [1] 7.211103
```

```
## [1] 7.483315
```

```r
# Result: Point 2 will be grouped with Cluster 2
```

```r
# Cluster 1
ed1<- sqrt(sum((p3-c1)^2))
# Cluster 2
ed2<- sqrt(sum((p3-c2)^2))
# Cluster 3
ed3<- sqrt(sum((p3-c3)^2))
ed1;ed2;ed3
```

```
## [1] 7.071068
```

```
## [1] 6.928203
```

```
## [1] 6.63325
```

```r
# Result: Point 3 will be grouped with Cluster 3
```

```r
# Cluster 1
ed1<- sqrt(sum((p4-c1)^2))
# Cluster 2
ed2<- sqrt(sum((p4-c2)^2))
# Cluster 3
ed3<- sqrt(sum((p4-c3)^2))
ed1;ed2;ed3
```

```
## [1] 8.485281
```

```
## [1] 8.831761
```

```
## [1] 8.717798
```

```r
# Result: Point 4 will be grouped with Cluster 1
# Cluster 1
ed1<- sqrt(sum((p5-c1)^2))
# Cluster 2
ed2<- sqrt(sum((p5-c2)^2))
# Cluster 3
ed3<- sqrt(sum((p5-c3)^2))
ed1;ed2;ed3
```

```
## [1] 5.385165
```

```
## [1] 4.795832
```

```
## [1] 4.123106
```

```r
# Result: Point 5 will be grouped with Cluster 3
```

b. Please for each cluster center find the associated points, based on the above results.

```r
Q1<- data.frame(point = c(1,2,3,4,5), "grouped with cluster" = c(1,2,3,1,3))
kable(Q1)
```

point	grouped.with.cluster
1	1
2	2
3	3
4	1
5	3

c. Please for each cluster compute the new center, based on the above results.

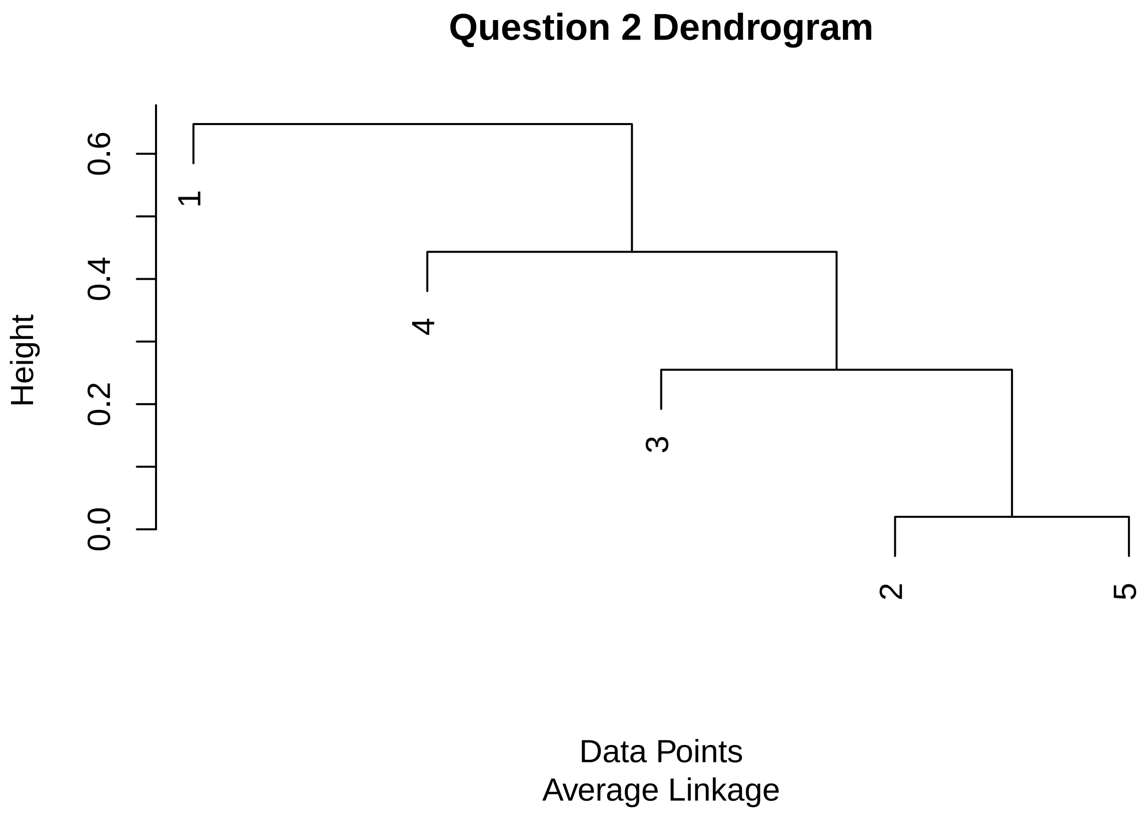
```r
# Cluster 1 -> Includes point 1 and point 4
newc1<- (p1+p4)/2
# Cluster 2 -> Includes point 2
newc2<- (p2)
# Cluster 3 -> Inlcudes point 3 and point 5
newc3<- (p3+p5)/2
kable(data.frame(cluster = c(1,2,3), coordinates = unname(rbind(newc1, newc2, newc3))))
```

cluster	coordinates.1	coordinates.2	coordinates.3	coordinates.4	coordinates.5
1	0.5	1.5	6.5	1	5.5
2	0.0	6.0	6.0	2	1.0
3	2.5	3.5	5.0	0	1.5

2. Use following similarities to perform average-linkage hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged

```r
similarity_matrix <- matrix(c(1.00, 0.10, 0.41, 0.55, 0.35,
                              0.10, 1.00, 0.64, 0.47, 0.98,
                              0.41, 0.64, 1.00, 0.44, 0.85,
                              0.55, 0.47, 0.44, 1.00, 0.76,
                              0.35, 0.98, 0.85, 0.76, 1.00),
                            nrow = 5, byrow = TRUE)
distance_matrix <- 1 - similarity_matrix

hc <- hclust(as.dist(distance_matrix), method = "average")
plot(hc, main = "Question 2 Dendrogram", xlab = "Data Points", sub = "Average Linkage")
```



3. Explain one advantage and one disadvantage of DBSCAN over the K-means clustering algorithm

DBSCAN is a good option for datasets with irregular shaped clusters. However, it reliant on user input for certain parameters. If the parameters aren't set correctly, it can cause issues with incorrect recognition of clusters.

4. Please Confirm:

Can confirm there is a github repo setup and registered with invites sent.

1. Plan to collect the data – We have been using the NFLverse which allows us access to NFL play by play data in R.
2. Plan to analyze – can confirm, plan on doing initial EDA and then feature engineering
3. Plan to improve – can confirm, the feature engineering process should increase model performance
4. Can confirm that we have met to work on project proposal as well.