Clustering Project

Jordan Brown

Fairleigh Dickinson University

CSCI_3269_31: Introduction to Data Mining

Dr. Tamraparni Dasu

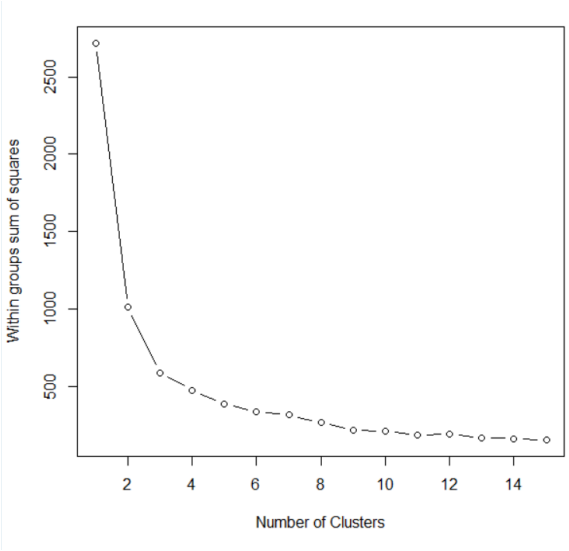
March 20, 2025

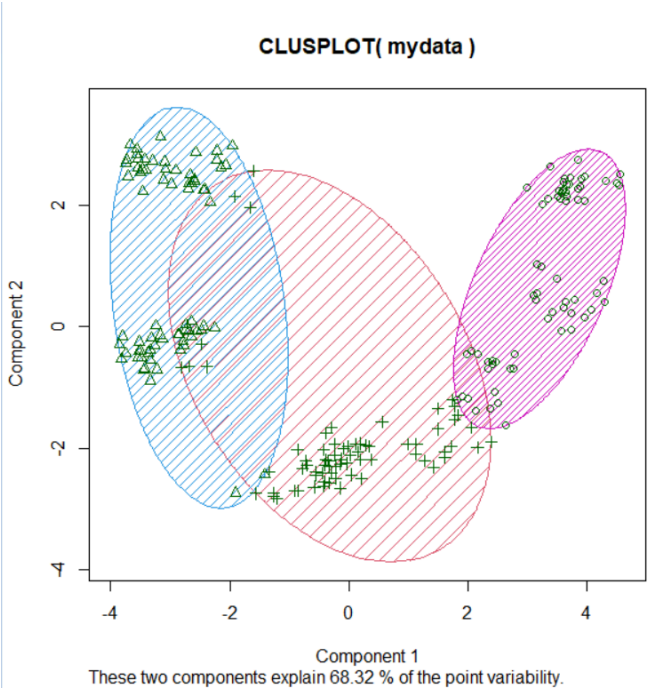
Clustering Project

Please randomize your data set by choosing a sample with replacement from the Seeds data and Iris data. The sample size should be the same as the original data set.

UG (150 points)

1. Run k-means on seeds data set **(50 pts)**
    a. With raw data
        i. Number of clusters= 3, cluster sizes=70 , cluster centers=(-3,0), (0,0), (1,3) , SSE for each cluster= C1:9, C2: 12, C3: 45 , SSB/TotalSS= 0.728





SSE for each cluster

C_1: (-3, 3), (-3, 0), (-3, 0)

C_2: (0, -2), (2, -2), (3, 0)

C_3: (2, -1), (4, 0), (4, 2), (4, 3)

$SSE = \left(\sqrt{(-3--3)^2+(3-0)^2}\right)^2 + \left(\sqrt{(-3--3)^2+(0-0)^2}\right)^2 \leftarrow C_1\ SSE = 9$

$\left(\sqrt{(0-0)^2+(-2-0)^2}\right)^2 + \left(\sqrt{(2-0)^2+(-2-0)^2}\right)^2 = 12 = C_2 SSE$

$\left(\sqrt{(2-1)^2+(-1-3)^2}\right) + \left(\sqrt{(4-1)^2+(3-0)^2}\right)^2 + \left(\sqrt{(4-1)^2+(3-2)^2}\right)^2 = 45 = C_3 SSE$

$SSE = 9+12+45 = 66$

$SSB = 9 \cdot 2 + 12 \cdot 2 + 45 \cdot 3 = 177$

$TSS = 66 + 177 = 243$

$\frac{SSB}{TSS} = \frac{177}{243} \approx 0.728$

ii. Compare purity of each cluster using species as the class label. R-command: table(cluster_id, species)

```
> table(input_data$type, fit$cluster)

            1  2  3
  Canadian  0 66  4
  Kama      2  6 62
  Rosa     65  0  5
```

Total = 70*3 = 210

P_C1 = 0/70 = 0

P_C2 = 66/70 = 0.94

P_C3 = 4/70 = 0.06

P_K1 = 2/70 = 0.03

P_K2 = 6/70 = 0.09

P_K3 = 62/70 = 0.89

P_R1 = 65/70 = 0.93

P_R2 = 0/70 = 0

P_R3 = 5/70 = 0.07

- Each cluster's purity is 1.

b. With standardized data

i. Number of clusters= 3, cluster sizes=70 , cluster centers=(-3,-1), (0,-1), (3,2), SSE for each cluster= C1:10, C2: 15, C3: 6 , SSB/TotalSS= 0.689

SSE for each cluster

C_1: $(-4,2)$, $(-3,-1)$, $(-3,-1)$

C_2: $(1,1)$, $(3,-2)$, $(3,-1)$

C_3: $(3,4)$, $(3,2)$, $(4,1)$, $(3,2)$

$$SSE = \left(\sqrt{(-4-3)^2+(2-1)^2}\right)^2 + \left(\sqrt{(-3-3)^2+(-1-1)^2}\right)^2 \leftarrow C_1 \; SSE = 10$$

$$\left(\sqrt{(1-0)^2+(1-1)^2}\right)^2 + \left(\sqrt{(3-0)^2+(-2-1)^2}\right)^2 = 15 = C_2 SSE$$

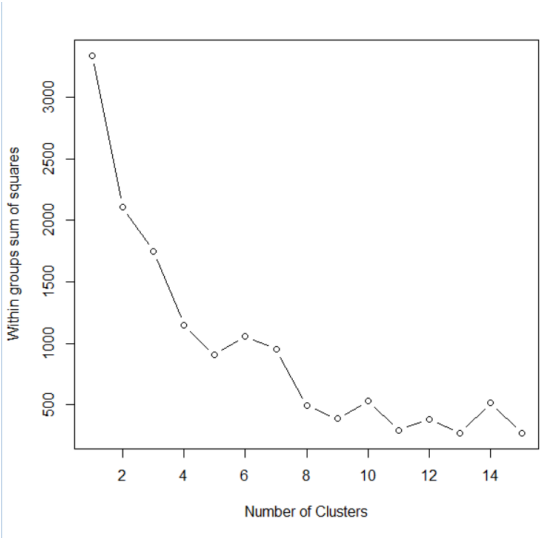
$$\left(\sqrt{(3-3)^2+(4-2)^2}\right) + \left(\sqrt{(3-3)^2+(2-2)^2}\right)^2 + \left(\sqrt{(4-3)^2+(1-2)^2}\right)^2 = 6 = C_3 SSE$$

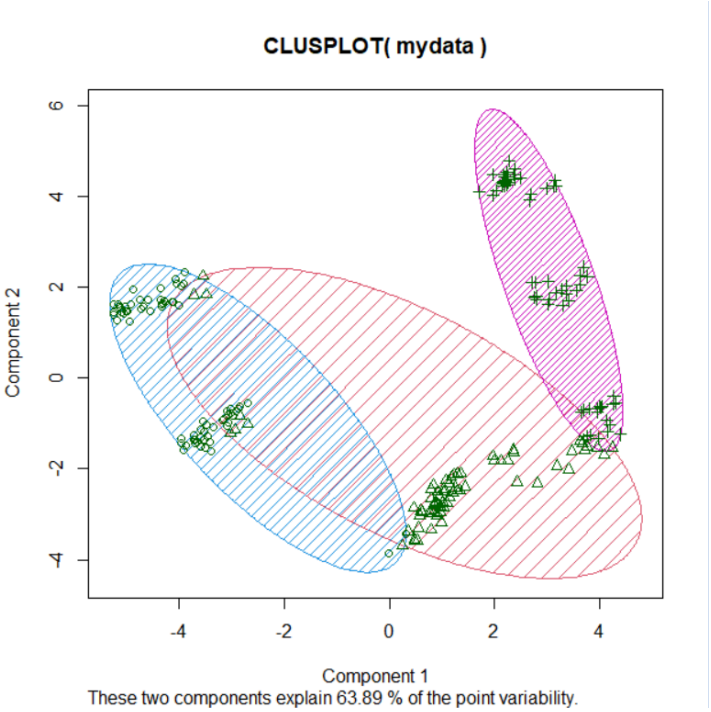
$$SSE = 10+15+6 = 31$$

$$SSB = 10 \cdot 2 + 15 \cdot 2 + 6 \cdot 3 = 68$$

$$TSS = 31 + 68 = 99$$

$$\frac{SSB}{TSS} = \frac{68}{99} = 0.689$$





ii. Compare purity of each cluster using species as the class label.

```
> table(input_data$type, fit$cluster)

           1  2  3
Canadian 66  4  0
Kama      6 62  2
Rosa      0  5 65
```

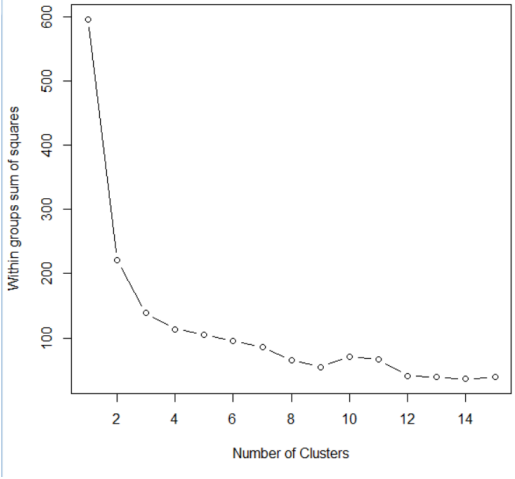
- Same purity as raw data for each cluster, just reordered.

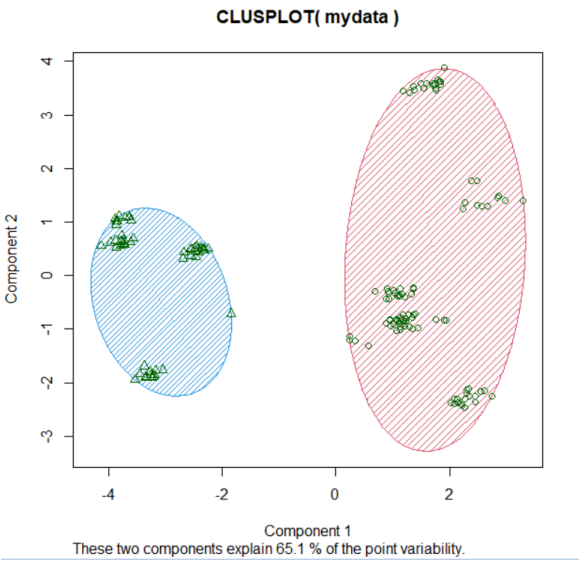
c. Compare the results for 2a and 2b

- By looking at the CLUSPLOT you can tell that the clusters are different based on their shapes. They are more stretched out after the data was standardized. The amount of clusters for both raw and standardized data are the same, which is 3 clusters. Their purity is the same for all 3 classes.

2. Run k-means on **Iris.** Which has better clusters based on purity, seeds or iris? Why? **(20 pts)**

- Iris has much better clusters based on purity because its clusters are more cohesive and separated more. The two clusters are completely separate from each other and don't overlap. They are very clear and are more pure than seeds.





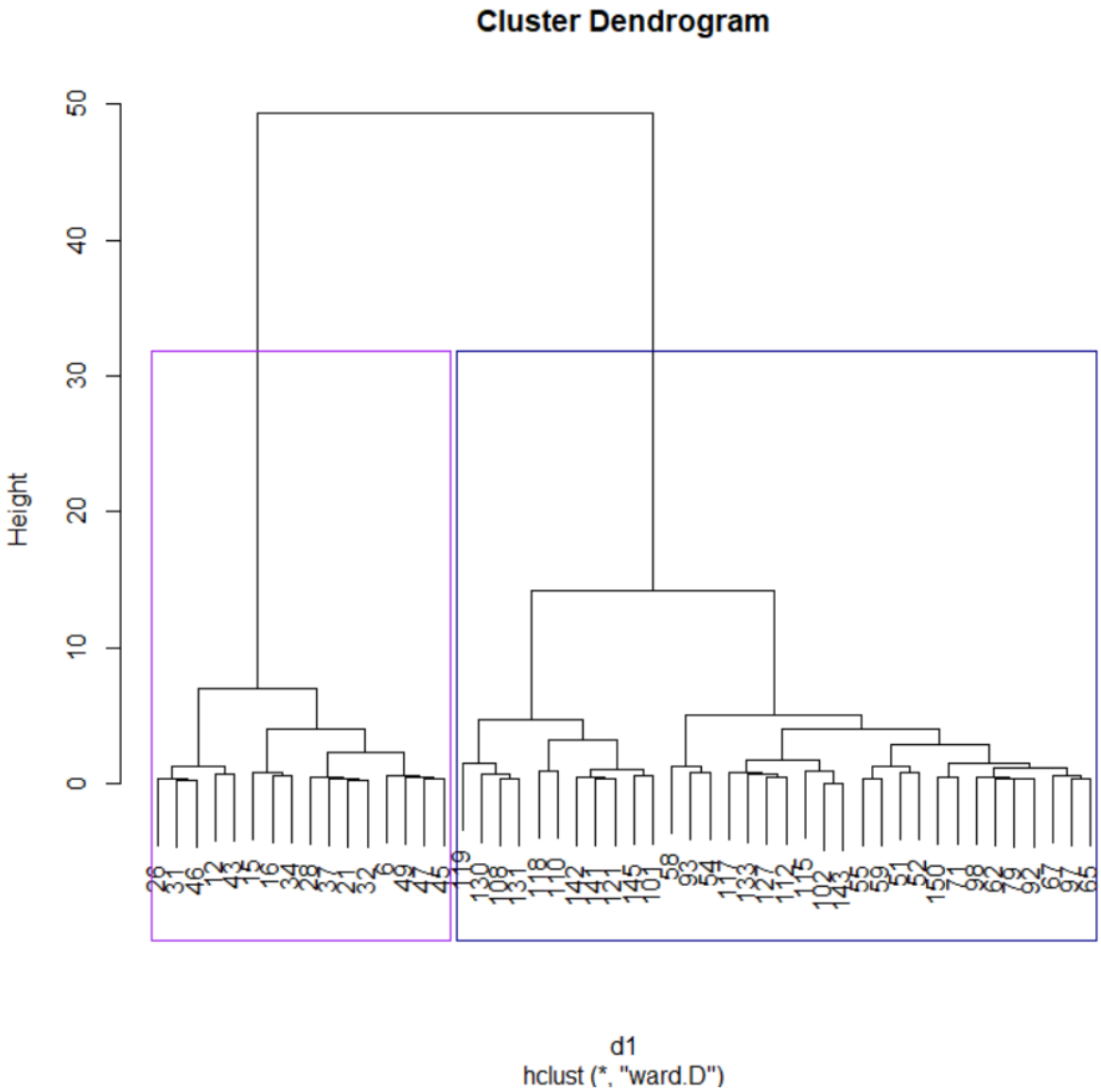
```
> table(input_data$type, fit$cluster)

             1  2
  setosa     0 50
  versicolor 50  0
  virginica  50  0
```

3. Run an agglomerative hierarchical algorithm on **Iris data. (40 pts)**

a. Show and explain how you chose the number of clusters

i. There are two clusters. Based on the cluster dendrogram, when you cut it horizontally at around height=30, there are two lines that go vertically through it, giving you two clusters.



b. Report number of clusters=2, cluster sizes= 15 & 33, cluster statistics (centroid= 2.628872, diameter= 2.81555) and cluster purity= 1.0, the weighted average of the purity of all clusters= 1.

- cluster purity
  - P_B1 = 0/50= 0
  - P_B2 = 50/50 = 1
  - P_VE1 =50/50 = 1
  - P_VE2 = 0/50= 0
  - P_VI1 =50/50 = 1
  - P_VI2 = 0/50= 0

- weighted average of the purity of all clusters
  - (50 + 50 + 50)/150 = 1.00

```
> table(input_data$type, fit$cluster)

             1  2
  setosa     0 50
  versicolor 50  0
  virginica  50  0
```

4. Run DBSCAN on **Iris data (40 pts)**

a. Show and explain how you chose MinPts and Eps

- I chose 15 for MinPts because the cluster number is 15. I chose 1.5 for Eps because the minimum cluster size is 1.

```
$n
[1] 50

$cluster.number
[1] 15

$cluster.size
 [1] 11  1  4  1  1  5  2  5  1  5  4  2  3  2  3

$min.cluster.size
[1] 1
```

```
>  res.fpc <- fpc::dbscan(scale(input_data[,in_attr]), eps = 1.5, MinPts = 15)
>  res.fpc
dbscan Pts=150 MinPts=15 eps=1.5
         1  2
border   5  3
seed    95 47
total  100 50
```

b. Report number of clusters=2, cluster sizes= 15 & 33, cluster statistics (centroid= 2.628872, diameter= 2.81555) and cluster purity= 1.0, the weighted average of the purity of all clusters= 0.67.

- cluster purity
    - P_B1 = 5/8 = 0.63
    - P_B2 = 3/8 = 0.38

    - P_S1 = 95/142 = 0.67
    - P_S2 = 47/142 = 0.33

- weighted average of the purity of all clusters
    - (5+95)/150 = 0.67