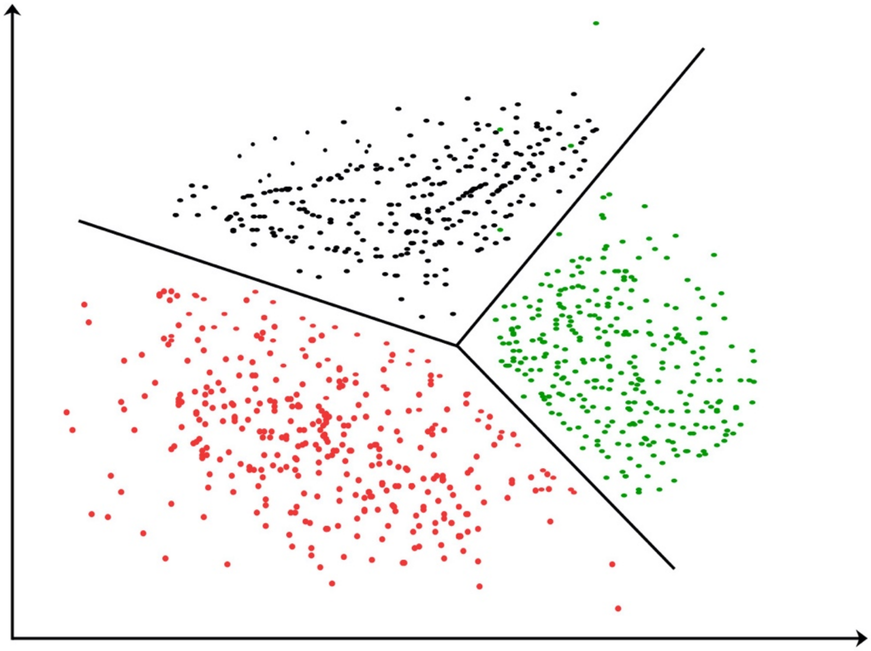
2nd Project - Clustering



DS 310

Daehoon Gwak

dpg5313@psu.edu

1. **Implement three clustering algorithms to find clusters of genes that exhibit similar expression profiles: K-means, Hierarchical Agglomerative clustering and density-based clustering. Compare these three methods and discuss their pros and cons.**

   >> For this project, there are three clustering algorithms: K-means, Hierarchical Agglomerative, and Density-Based clustering. Each clustering has pros and cons. First, K-means initializes by specifying the initial centroids then partition into K clusters. Compared to other algorithms, it is easy to implement and efficient. However, it cannot handle non-spherical shape clusters and has problems when the data contain outliers. Secondly, Hierarchical clustering has benefit when it comes to the output of the dendrogram. However, intermediate decision cannot be undone. Lastly, Density-Based clustering uses DBSCAN. It is useful since it detects arbitrary-shape clusters. However, it has major parameters and is very sensitive to the parameters so sometimes it is not useful.

2. **For each of the above tasks, you are required to validate your clustering results using the following methods using Rand Index, F1, and purity.**

   a. K-means

   i. K = 4

   1. Rand index = 0.8723930792334701
   2. F1 = 0.7276912660798917
   3. Purity = 0.775974025974026

   ii. K = 5

   1. Rand index = 0.8470535978679301
   2. F1 = 0.6335579992905286
   3. Purity = 0.7467532467532467

   iii. K = 6

   1. Rand index = 0.9065950336308642
   2. F1 = 0.7567746199603437
   3. Purity = 0.9285714285714286

   iv. K = 7

   1. Rand index = 0.8778501628664495
   2. F1 = 0.6608327949726904
   3. Purity = 0.8896103896103896

   b. Hierarchical Agglomerative

   i. K = 4

   1. Rand index = 0.7924827615381361
   2. F1 = 0.5663646408839779
   3. Purity = 0.6103896103896104

   ii. K = 5

   1. Rand index = 0.8352299166631414
   2. F1 = 0.6094455028577158
   3. Purity = 0.75

iii. **K = 6**

1. Rand index = 0.8390371843140573
2. F1 = 0.5901109555100722
3. Purity = 0.7694805194805194

iv. **K = 7**

1. Rand index = 0.8395871229747451
2. F1 = 0.5666285714285715
3. Purity = 0.7727272727272727

c. Density-Based

i. **eps = 3, min_samples = 3**

1. Rand index = 0.7906425821735268
2. Clustering F1 = 0.21407019215499445
3. Clustering Purity = 0.8928571428571429

ii. **eps = 3, min_samples = 4**

1. Rand index = 0.7433267058674224
2. Clustering F1 = 0.2480946774893116
3. Clustering Purity = 0.7954545454545454

iii. **eps = 3, min_samples = 5**

1. Rand index = 0.5697364524726088
2. Clustering F1 = 0.3307231690465223
3. Clustering Purity = 0.5681818181818182

iv. **eps = 3, min_samples = 6**

1. Rand index = 0.21951013156224883
2. Clustering F1 = 0.35999722492021646
3. Clustering Purity = 0.3246753246753247

v. **eps = 4, min_samples = 3**

1. Rand index = 0.8083040737763865
2. Clustering F1 = 0.5733653438779833
3. Clustering Purity = 0.7272727272727273

vi. **eps = 4, min_samples = 4**

1. Rand index = 0.8342357967765134
2. Clustering F1 = 0.47603128969713177
3. Clustering Purity = 0.8798701298701299

vii. **eps = 4, min_samples = 5**

1. Rand index = 0.8268116248572275
2. Clustering F1 = 0.3705412054120541
3. Clustering Purity = 0.9383116883116883

viii. **eps = 4, min_samples = 6**

1. Rand index = 0.7888658572697661
2. Clustering F1 = 0.3041962916492402
3. Clustering Purity = 0.8538961038961039

ix. **eps = 5, min_samples = 3**
1. Rand index = 0.5537882313126613
2. Clustering F1 = 0.45802075840098655
3. Clustering Purity = 0.5

x. **eps = 5, min_samples = 4**
1. Rand index = 0.5537882313126613
2. Clustering F1 = 0.45802075840098655
3. Clustering Purity = 0.5

xi. **eps = 5, min_samples = 5**
1. Rand index = 0.7312703583061889
2. Clustering F1 = 0.5421126608281976
3. Clustering Purity = 0.6461038961038961

xii. **eps = 5, min_samples = 6**
1. Rand index = 0.7893311899826558
2. Clustering F1 = 0.5469432314410481
3. Clustering Purity = 0.6525974025974026

xiii. **eps = 6, min_samples = 3**
1. Rand index = 0.21951013156224883
2. Clustering F1 = 0.35999722492021646
3. Clustering Purity = 0.3246753246753247

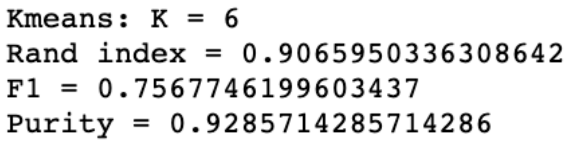
xiv. **eps = 6, min_samples = 4**
1. Rand index = 0.21951013156224883
2. Clustering F1 = 0.35999722492021646
3. Clustering Purity = 0.3246753246753247

xv. **eps = 6, min_samples = 5**
1. Rand index = 0.21951013156224883
2. Clustering F1 = 0.35999722492021646
3. Clustering Purity = 0.3246753246753247

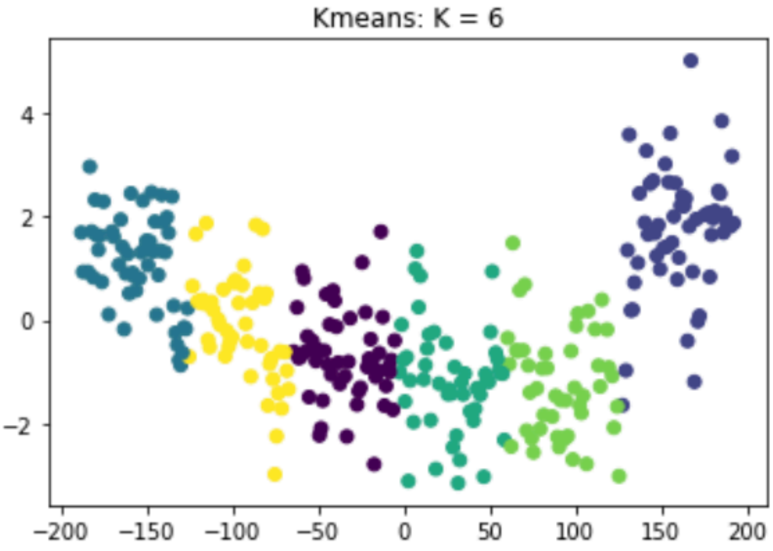
xvi. **eps = 6, min_samples = 6**
1. Rand index = 0.21951013156224883
2. Clustering F1 = 0.35999722492021646
3. Clustering Purity = 0.3246753246753247

3. Analyze the results and judge which parameters are the best for different clustering algorithms. Report these values (with plots) and reasons.

a. K-means: For K-means clustering algorithm, it is best when K equals to 6.



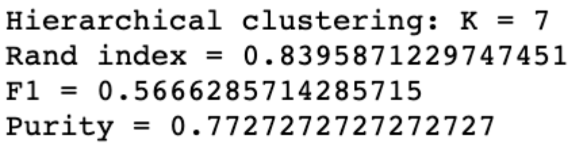
```
Kmeans: K = 6
Rand index = 0.9065950336308642
F1 = 0.7567746199603437
Purity = 0.9285714285714286
```



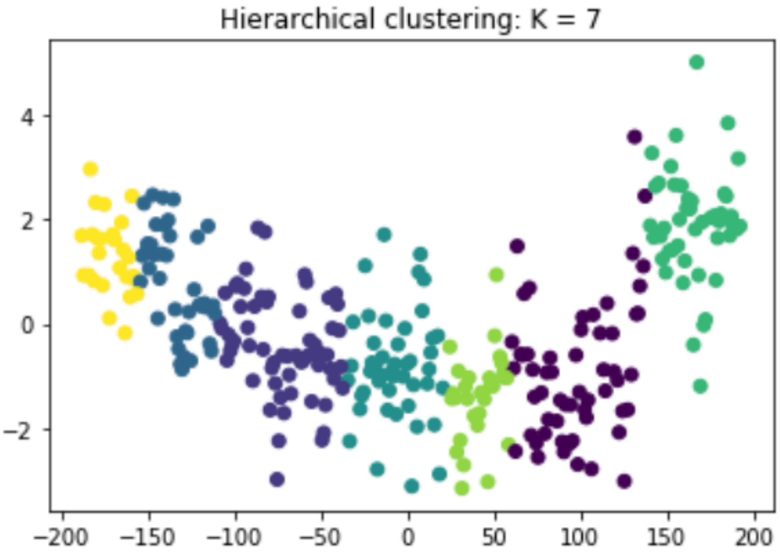
<Figure 1>

As the ‘Figure 1’ shows, the Rand index is ‘0.906595’, F1 is ‘0.75677’, and Purity is ‘0.92857’, which are the highest among other K values.

b. Hierarchical: For Hierarchical clustering algorithm, it is best when K equals to 7.



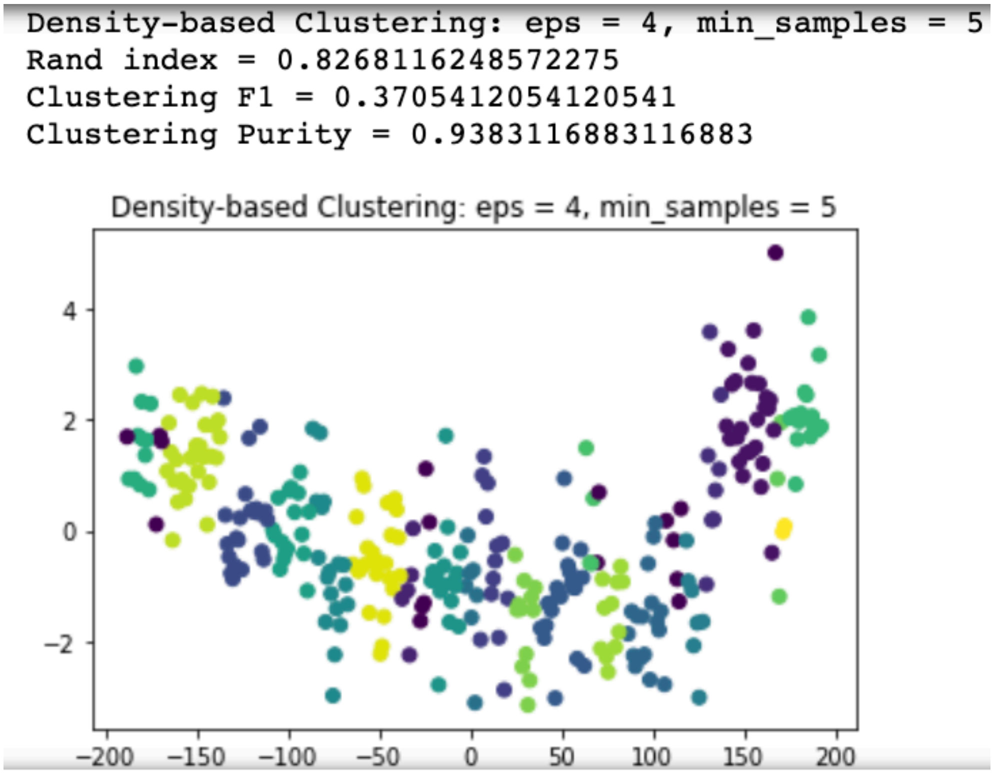
```
Hierarchical clustering: K = 7
Rand index = 0.8395871229747451
F1 = 0.5666285714285715
Purity = 0.7727272727272727
```



<Figure 2>

With same reason as K-means, all three values(Rand index, F1, Purity) are the highest among other K values.

c. Density-Based: For Density-Based clustering algorithm, it is best when eps value is 4 and minimum samples are 5.



```
Density-based Clustering: eps = 4, min_samples = 5
Rand index = 0.8268116248572275
Clustering F1 = 0.3705412054120541
Clustering Purity = 0.9383116883116883
```

<Figure 3>

With same reason as above algorithms, all three values(Rand index, F1, Purity) are the highest among other eps and minimum sample values.