Programming Assignment 3

Sneha Reddy Aenugu, EE11B059

November 6, 2015

1 Clustering

- For each of the below data, the visualizations are provided in Figures 1 - 7.
- K means is run on the R15 cluster with the number of clusters set as 15 and k value set as 8. Out of 599 instances of the data, 114 are incorrectly classified, giving rise to 19% error in clustering. The graph of k vs cluster purity is given in Figure 8.

$$ClusterPurity = percentage\ of\ instances\ classified\ correctly = 80.97\% \quad (1)$$

- DBSCAN is run on the Jain dataset with minpoints set as 20 and epsilon set as 0.1. 3 instances out of the total 372 instances are incorrectly classified, giving rise to 0.8% error in the clustering.

$$ClusterPurity = percentage\ of\ instances\ classified\ correctly = 99.2\% \quad (2)$$

  With decrease in minpoints the cluster purity decreases. Increase in epsilon as well as its decrease from the said optimal point decreases the cluster purity.

- Comparison of DBSCAN and heirarchical clustering for different datasets is shown in Table 1.

2 Decision trees

- The J48 Decision tree algorithm is highly accurate and is giving a precision, recall and f-measure for Minobjects set as 5.

Dataset	**DBSCAN**	**Heirarchical**
Path-based	98	Ward linkage - 75.6
Spiral	100	Single linkage - 100
Flames	99.2	Ward linkage - 99.2

Table 1: Comparision of DBSCAN and Heirarchicals

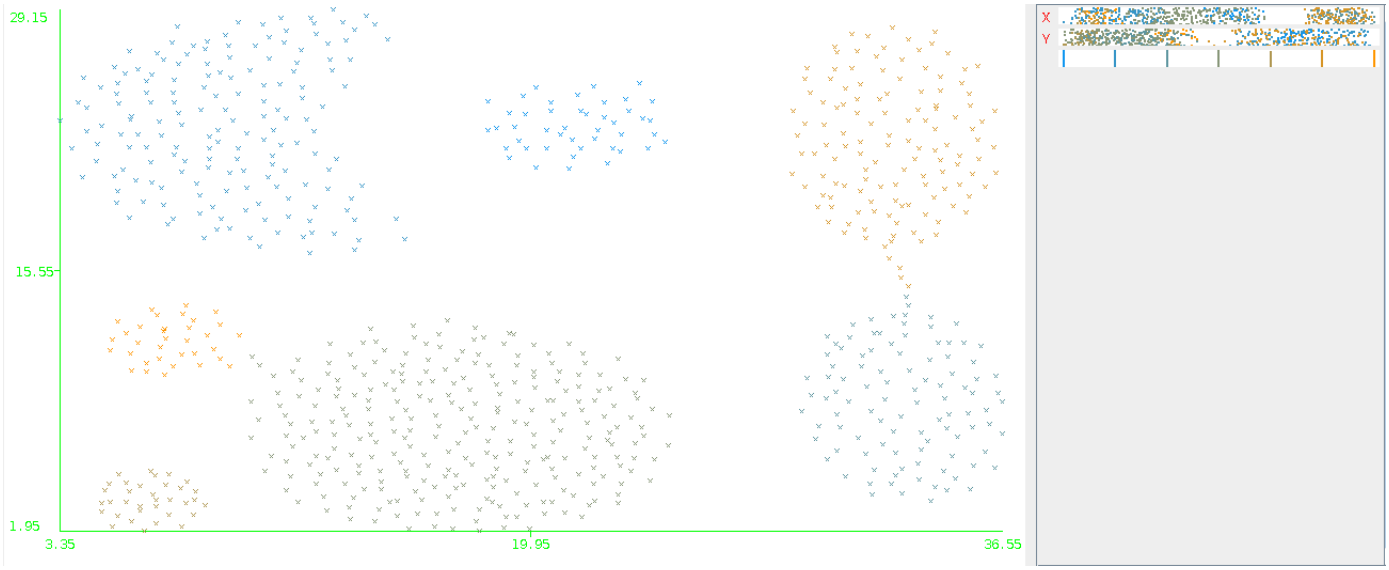

Figure 1: Aggregation

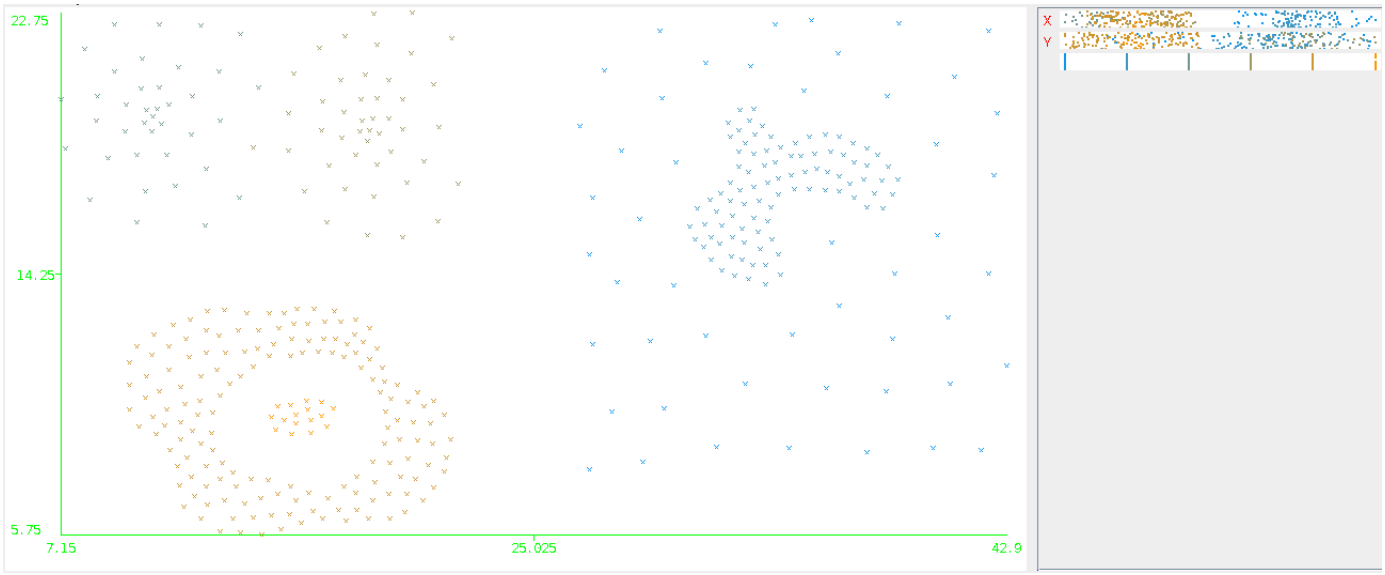

Figure 2: Compound

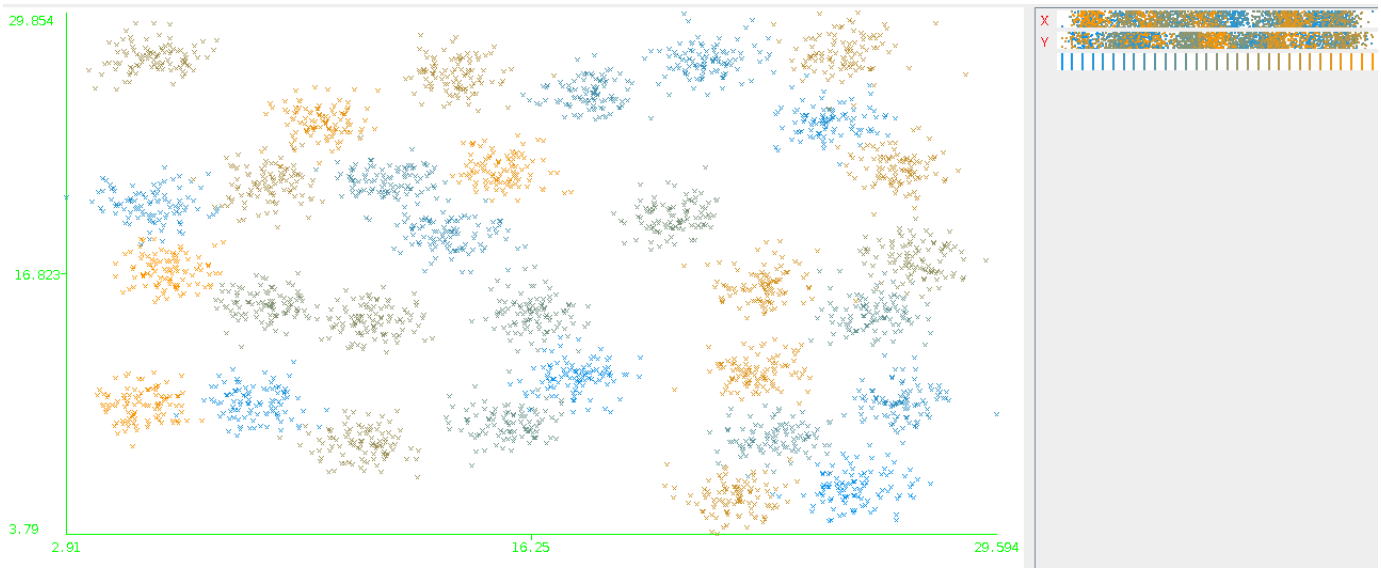

Figure 3: D31

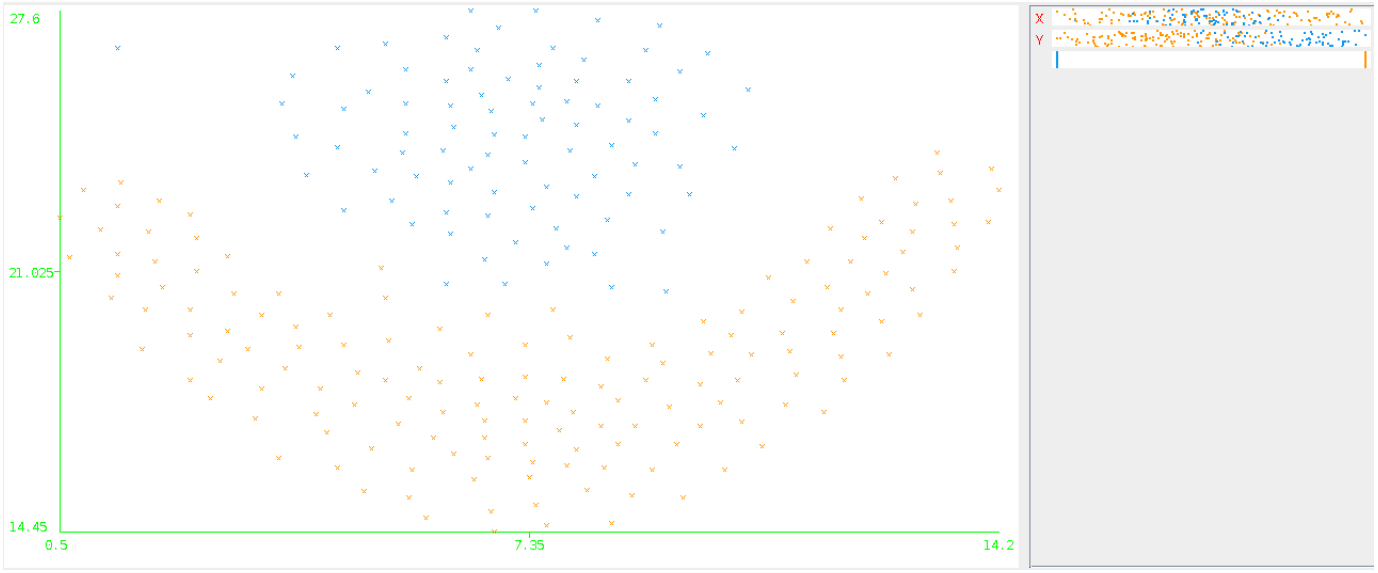

Figure 4: Flames

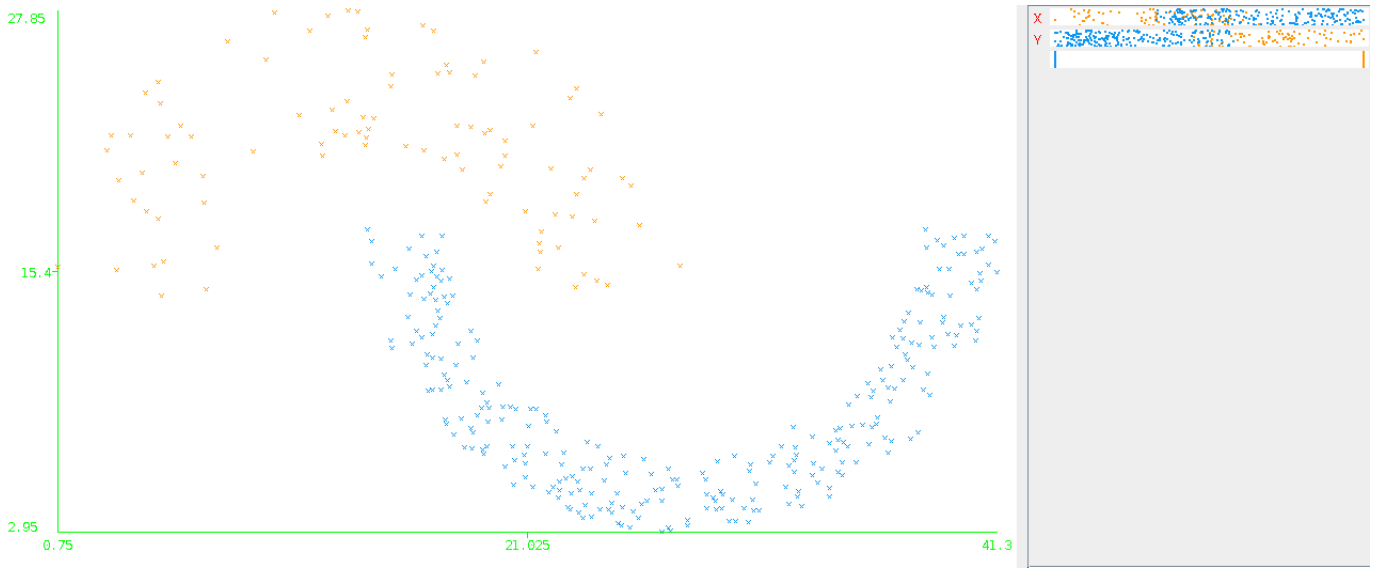

Figure 5: Jain

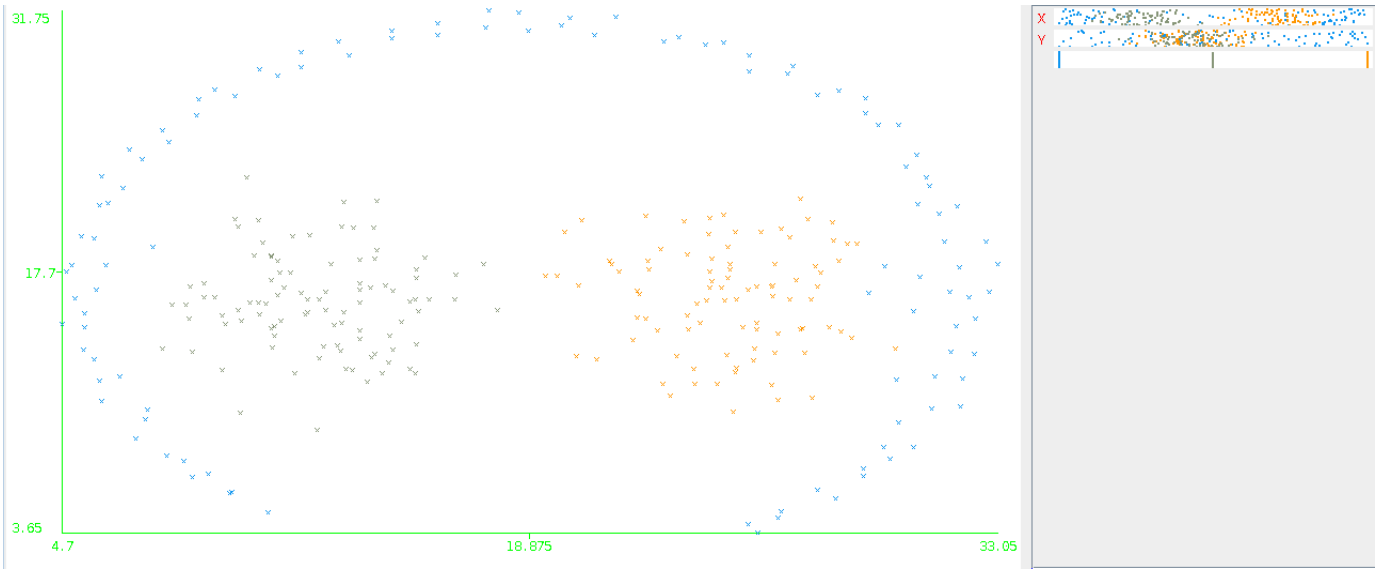

Figure 6: Path-based

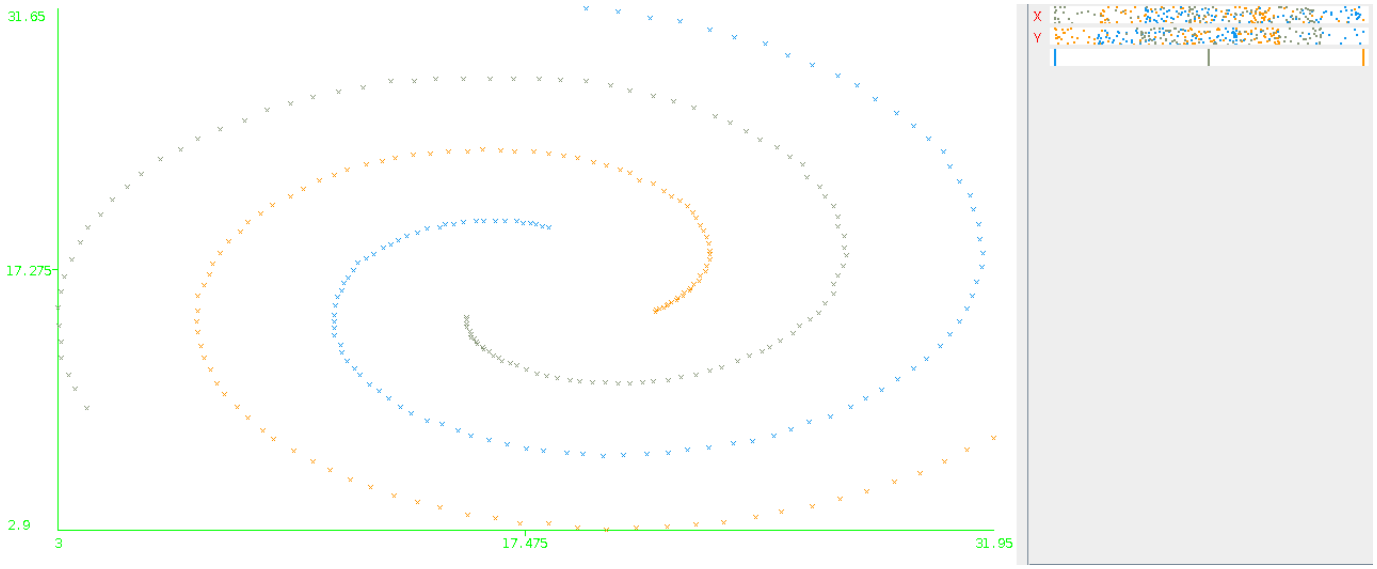

Figure 7: Spiral

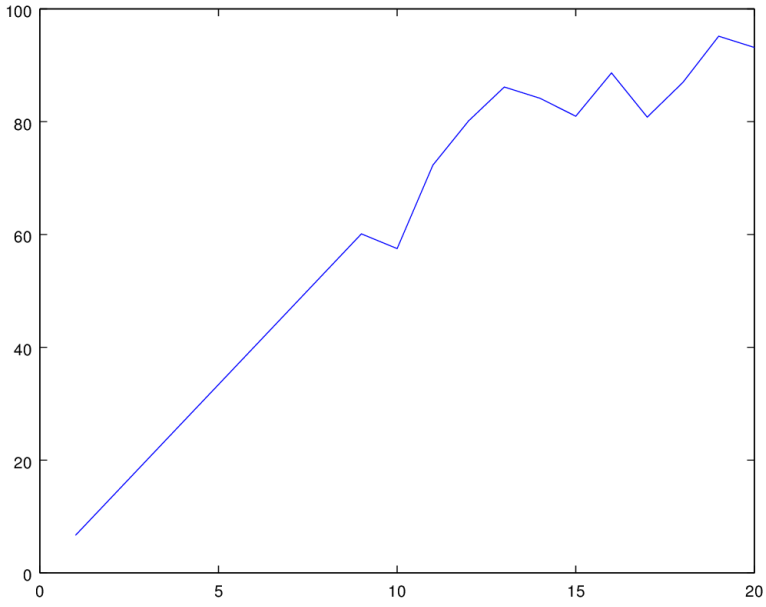

Figure 8: k vs Cluster purity

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

5 = p: p (256.0)
5 = a: e (400.0)
5 = l: e (400.0)
5 = n
|   10 = e
|   |   20 = k: e (48.0)
|   |   20 = n: e (52.0)
|   |   20 = u: e (0.0)
|   |   20 = h: e (48.0)
|   |   20 = w
|   |   |   8 = n
|   |   |   |   7 = c: p (32.0)
|   |   |   |   7 = w
|   |   |   |   |   21 = s: e (0.0)
|   |   |   |   |   21 = n: e (0.0)
|   |   |   |   |   21 = a: e (0.0)
|   |   |   |   |   21 = v: e (48.0)
|   |   |   |   |   21 = y: e (0.0)
|   |   |   |   |   21 = c: p (9.0)
|   |   |   8 = b: e (240.0)
|   |   20 = r: p (72.0)
|   |   20 = o: e (7.0)
|   |   20 = y: e (4.0)
|   |   20 = b: e (1.0)
|   10 = t: e (2496.0)
5 = f: p (1952.0)
5 = c: p (192.0)
5 = y: p (379.0)
5 = s: p (359.0)
5 = m: p (5.0)
```

Figure 9: Training data model

- As the MinObjects increase, precision of one class decreases and the recall of the other class decreases.
- The most important feature is the bruises.If there are bruises (f) then the mushroomn is poisonous and if there are no bruises (t) mushroom is edible.
- The Decision tree learnt by the model is given in the figure 9.