Programming Assignment-03
Intro to Machine Learning -CS5011

Sunil Kumar J S(NA13B031)

November 16, 2016

1

For all graphs first column is plotted on y axis and second column on x axis.The legend of all graphs are matched with above figure

1.1 Visualization of aggregation dataset

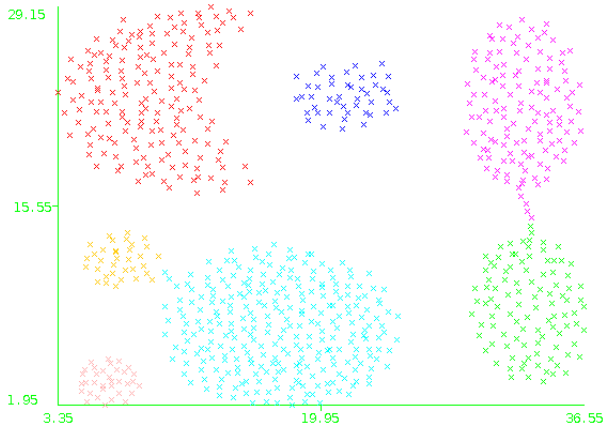

Figure 1: visualisation of aggregation dataset

Analysis:

- **K-means:Rating(3/5)**kmeans wouldn't give best because:
  - count of observations for each class variable aren't roughly same
  - all classes do not have same variance
  - clusters are non globular

- it might be difficult to distinguish between classes 4,5 and 7 as they are located very closely and there is high mismatch between count of observations

- **Heirarchical Clustering:**
  - Heirarchical structure isn't present.It seems bad idea to run heirarchical clustering as it would take more running time also.
  - **Singlelink Clustering:Rating(2.5/5)**single link heirarchical clustering will give bad prediction between classes (6 and 3) and (4,5 and 7). This is because single link considers minimum pairwise distance to cluster.
  - **Completelink Clustering:Rating(3.5/5)**
- **DBSCAN:Rating(4/5)**
  - DBSCAN performs well on arbitary shaped cluster and is robust toward outliers. In this sense, with appropritae choice of minpoints and epsilon we could distinguish between classes (6 and 3) and (4,5,7) better.

1.2 Compound dataset

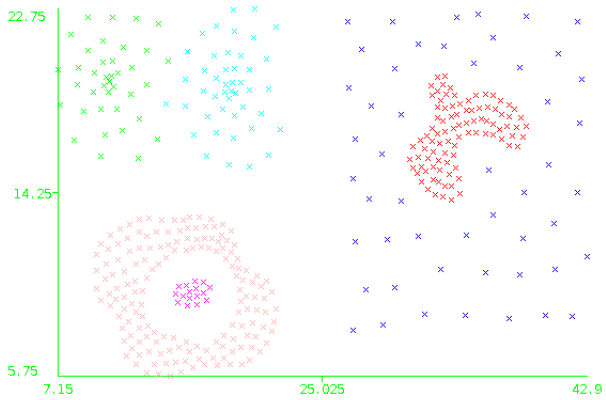

Figure 2: visualization of compound dataset

Analysis:

- **K-means:Rating(1.5/5)**kmeans wouldn't give best because:
  - it will fails to recognise heirarchial structure and clusters are non globular here. it wouldn't be able to distinguish between classes (2 and (6,5))
  - all classes doesn't have same variacnce
- **Heirarchical Clustering:**
  - Heirarchical clustering is expected to be winner here becuase of heirarchical structure present in data.
  - **Singlelink Clustering:Rating(3.5/5)**single link heirarchical clustering will give good prediction between classes (5 and 6)
  - **Completelink Clustering:Rating(3.5/5)** complete link heirarchical clustering might give better results than single linked because it would be able to give better prediction between classes (5 and 6)
- **DBSCAN:Rating(2.5/5)**
  - DBSCAN doesn't perform well because density varies significantly and is sparse for some clusters. DBSCAN would certainly give bad performace on classes 3 and 4.

1.3 visualization of D31

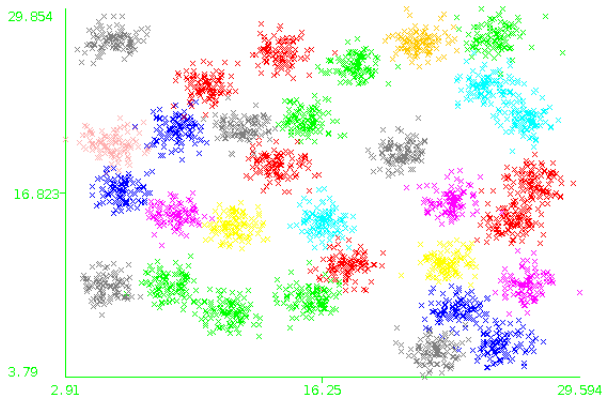

Figure 3:

Analysis:

- **K-means:Rating(4/5)**kmeans is expected to perform quite good here because
  - class clusters look approximately globular
  - count of observations in each class is roughly equal
- **Heirarchical Clustering:**
  - Heirarchical structure isn't present.It seems bad idea to run heirarchical clustering as it would take more running time also.
  - **Singlelink Clustering:Rating(2/5)**no hierarchical structure is present. so no need of single link as it is comutatioanlly expensive .
  - **Completelink Clustering:Rating(2/5)** no hierarchical structure is present. so no need of single link as it is comutatioanlly expensive .
- **DBSCAN:Rating(4/5)**
  - DBSCAN gives moderate results as clusters are packed together

1.4 Visualisation of flame

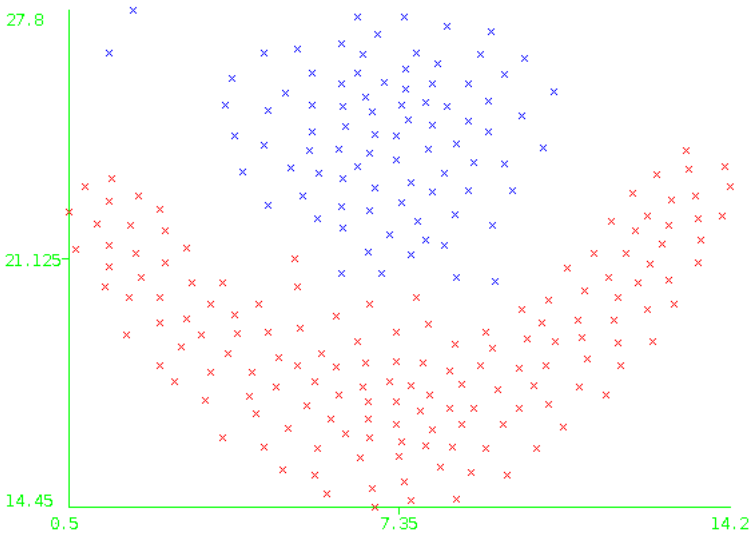

Figure 4: alphas(hyperparameters) are taken to be equal to 10 in this graph

Analysis:

- **K-means:Rating(1.5/5)**kmeans would give poor results because:
  - class clusters not even roughly spherical
  - From visulasiation it seems distance based algorithms fail here.
- **Heirarchical Clustering:**
  - Heirarchical clustering is expected to be perform poor as visual inspection says distance based algorithms fail.Also there is is no heirarchical structure
- **DBSCAN:Rating(5/5)**
  - DBSCAN is winner here as classes appear to get seperated based on density. Also graph isn't sparse.

1.5 Visualisation of jain

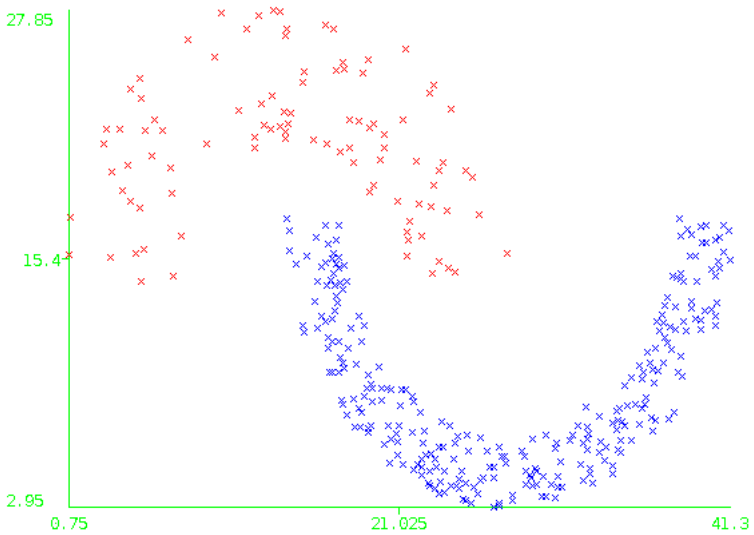

Figure 5: alphas(hyperparameters) are taken to be equal to 10 in this graph

Analysis:

- **K-means:Rating(1.5/5)**kmeans is expected to perform poor because:
  - clusters aren't globular in nature
  - Kmeans uses distance metric.
- **Heirarchical Clustering:**
  - Heirarchical clustering is expected to be perform poorly on this dataset computing based on distance fails here. This is applicable for both singlelink and Completelink clustering.
- **DBSCAN:Rating(4.5/5)**
  - DBSCAN seems clear winner in this case,although sparse nature of class 2 raise some doubts.With appropriate choice of minpoints and epsilon DBSCAN would give good results

1.6 Visualisation of pathbased

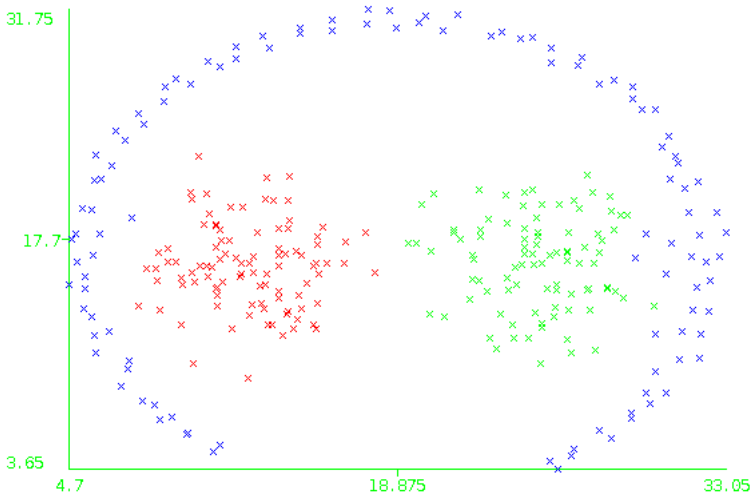

Figure 6: alphas(hyperparameters) are taken to be equal to 10 in this graph

Analysis:

- **K-means:Rating(1.5/5)**kmeans would perform poor because
  - class clusters aren't spherical
  - variance of each class isn't approximately same
- **Heirarchical Clustering:**
  - Heirarchical clustering is expected to be winner here becuase of heirarchical structure present in data.
  - **Singlelink Clustering:Rating(3.5/5)** performs good becuase of heirarchical structure
  - **Completelink Clustering:Rating(3.5/5)** complete link heirarchical performs good becuase of heirarchical structure
- **DBSCAN:Rating(2.5/5)**
  - DBSCAN is also expected to do good as graph isn't sparse and density appears approximately constant. however density difference raise some doubts

1.7 Visualisation of R15

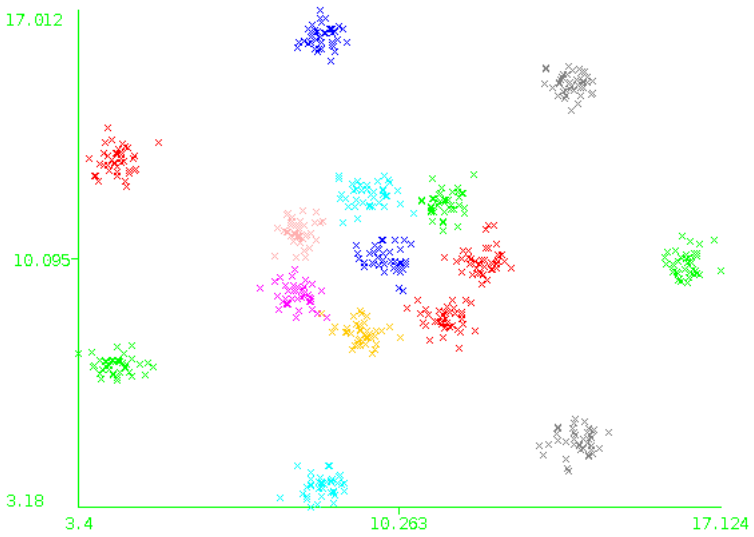

Figure 7: alphas(hyperparameters) are taken to be equal to 10 in this graph

Analysis:

- **K-means:Rating(1.5/5)**kmeans is expected perform good because heirarchical structure is present
- **Heirarchical Clustering:**
  - Heirarchical clustering is expected to be winner here becuase of heirarchical structure present in data.
  - **Singlelink Clustering:Rating(3.5/5)**performs good because heirarchical structure
  - **Completelink Clustering:Rating(3.5/5)** complete link heirarchical performs good because heirarchical structure
- **DBSCAN:Rating(3.5/5)**
  - DBSCAN is also expected to perform good because clusters are well seperated and density is same

1.8 Visualisation of Spiral

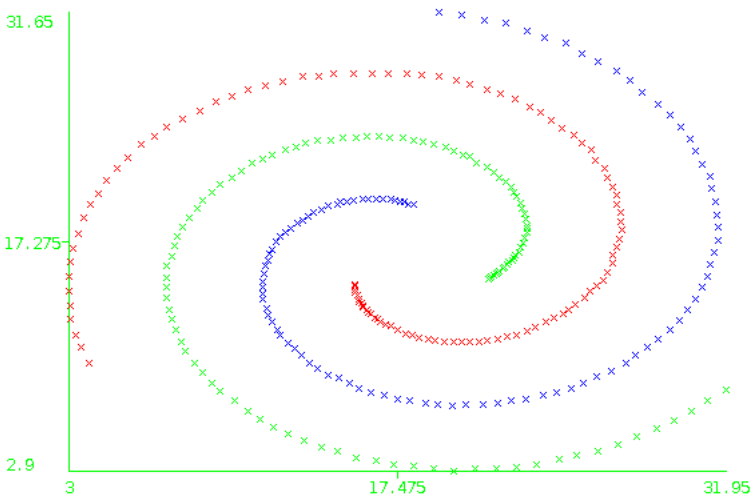

Figure 8: Visualisation of spiral dataset

Analysis:

- **K-means:Rating(1.5/5)**kmeans would give poor results because:
  - shapes of clusters aren't globular, not even roughly.Also all distance based algorithms are expected to fail in this case
- **Heirarchical Clustering:Rating 1.5/5**
  - Heirarchical clustering also wouldn't work well as there is no heirachical structure and heirarchical structure uses distance method. Distance method wouldn't work well in this case.Both Single link and Complete link performs poorly in this case
- **DBSCAN:Rating(4.5/5)**
  - DBSCAN is clear winner here. DBSCAN clusters arbitary clusters really well and density seems constant

2 K-means with R15

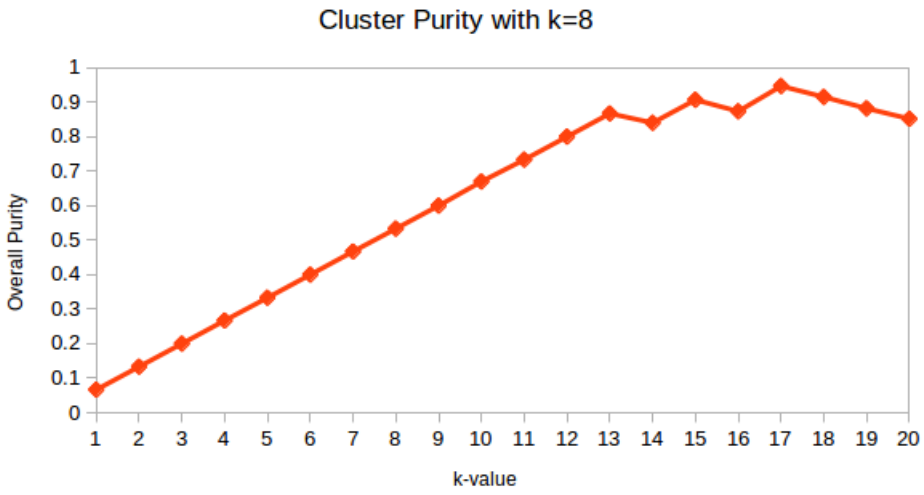

Figure 9: cluster purity variation with value of 'k'

This is similar to Knee-plot taught in class and based on gradient change we select k=15 as optimal value of k For the k=15,K-means classified points as shown below

3 DBSCAN with Jain

for e=0.1 and minPoints=25

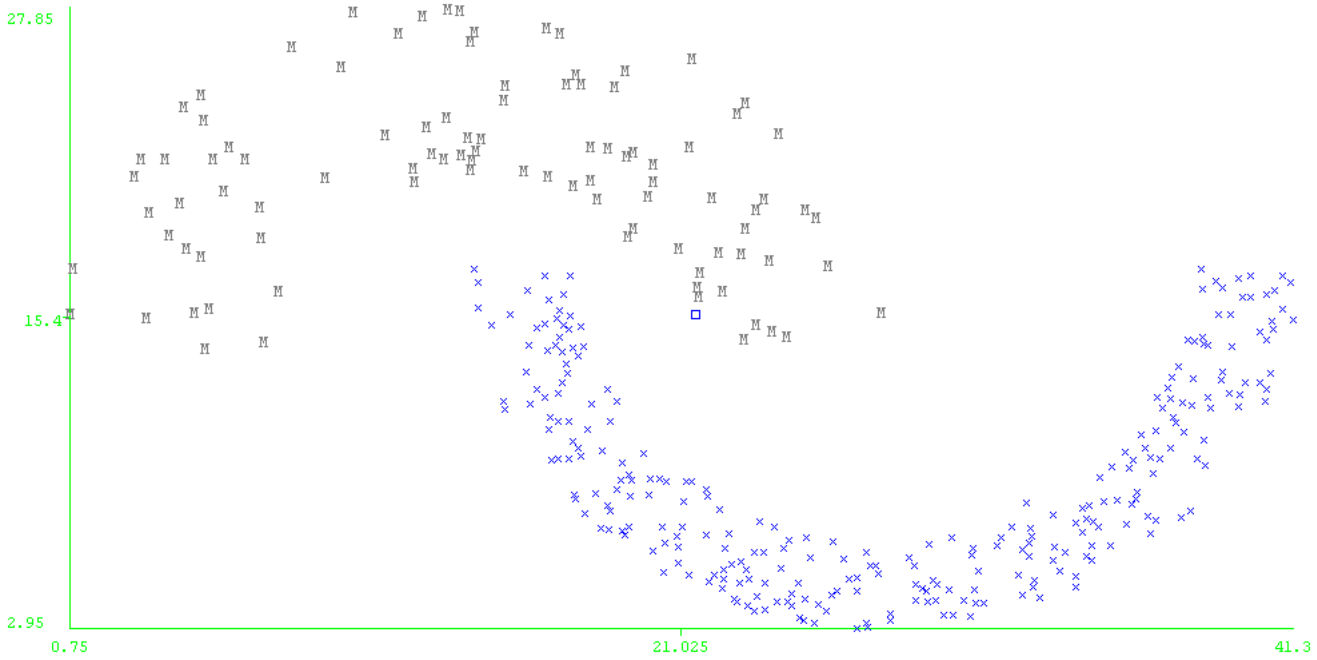

Figure 10: Resulting labels when e=0.12 and minpoints in neighbourhood is 25

DBSCAN fails in correctly classifying because the density difference between two clusters is large.when we try different values of epsilon and minPoints either it classfies all points to single cluster or detects class 2 clusters as NOISE.

for e=0.15 and minPoints=25

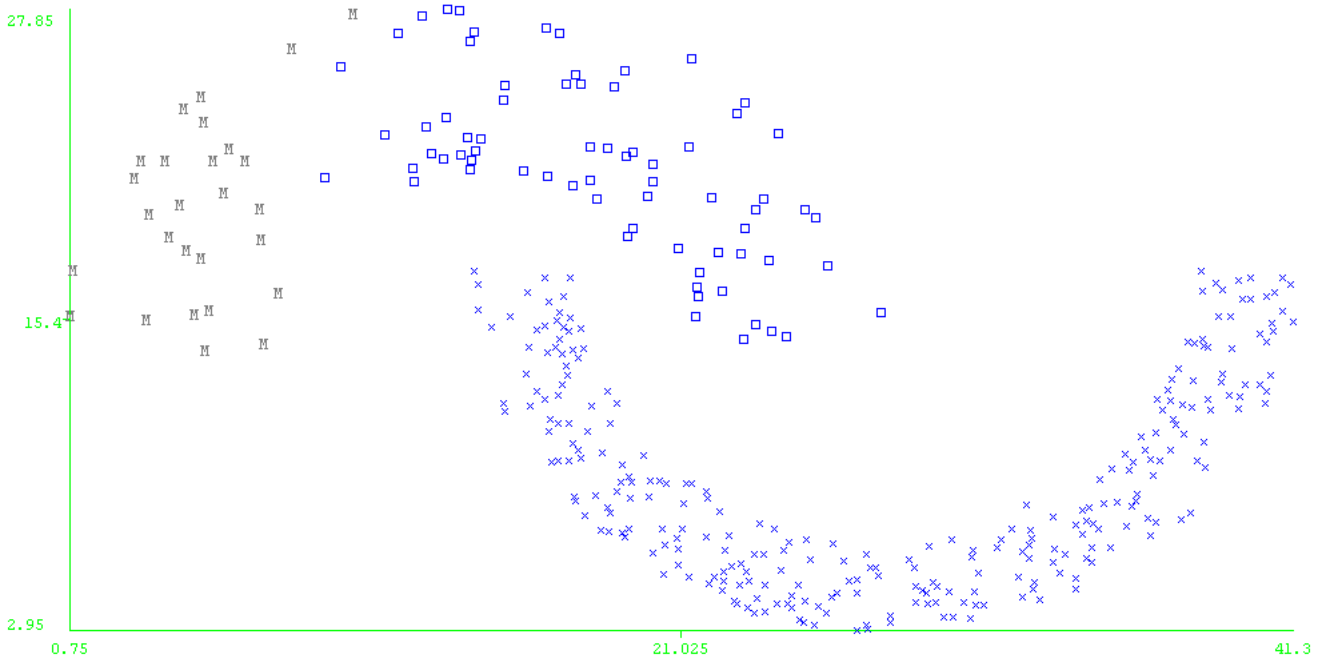

Figure 11: Resulting labels when e=0.16 and minpoints in neighbourhood is 25

4 DBSCAN vs Heirarchical Clustering

4.1 performance on path-based dataset

4.1.1 DBSCAN

DBSCAN correctly classifies inner clusters but fails to classify outer clusters as two inner clusters have similar density while outer has different density.DBSCAN classifies 111 datapoints as noise.

Plot of resulting labels obtained after running DBSCAN on pathbased parameters for DBSCAN were epsilon= 0.1 minPoints=20

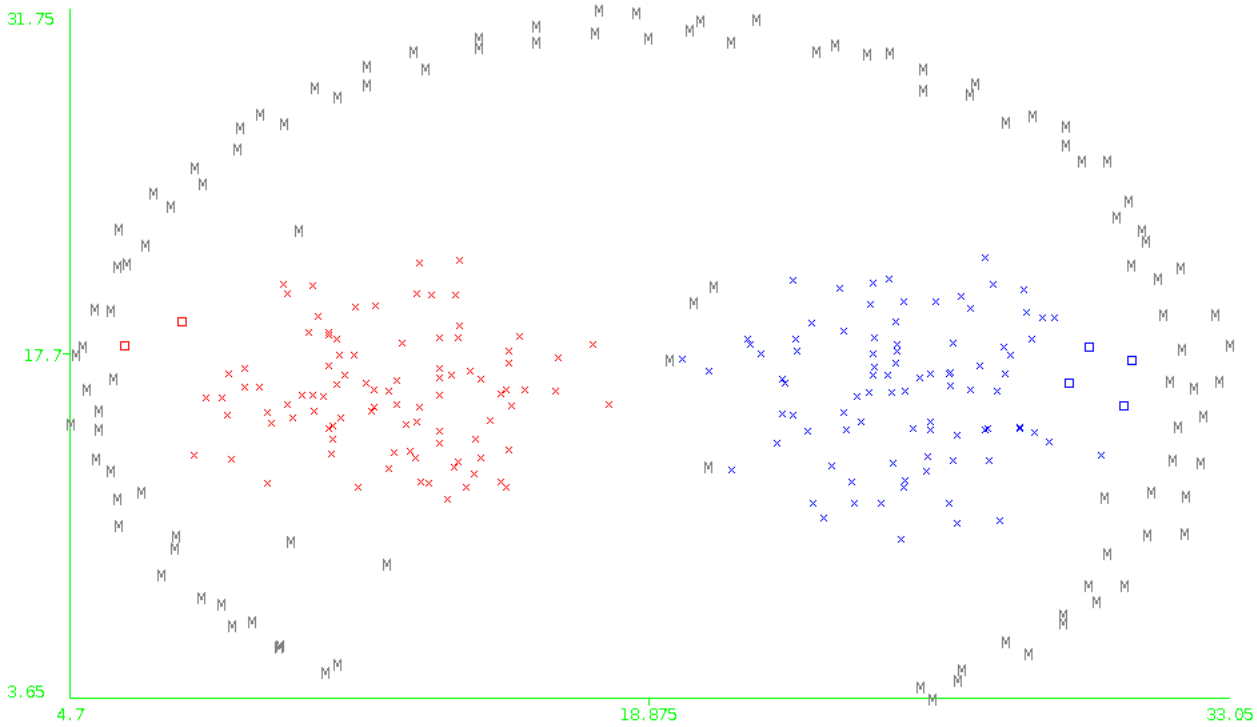

Figure 12: Resulting labels when e=0.1 and minpoints in neighbourhood is 20

4.1.2 Heirarchical

Ward linkage gave better performanceamong Heirarchical Clustering. Plot of resulting labels obtained after running Wardslinkage on pathbased Purity of class 2 and class 3 is 100%. Purity of Class 1 is 32.72%. Best results were obtained from Wards Linkage.

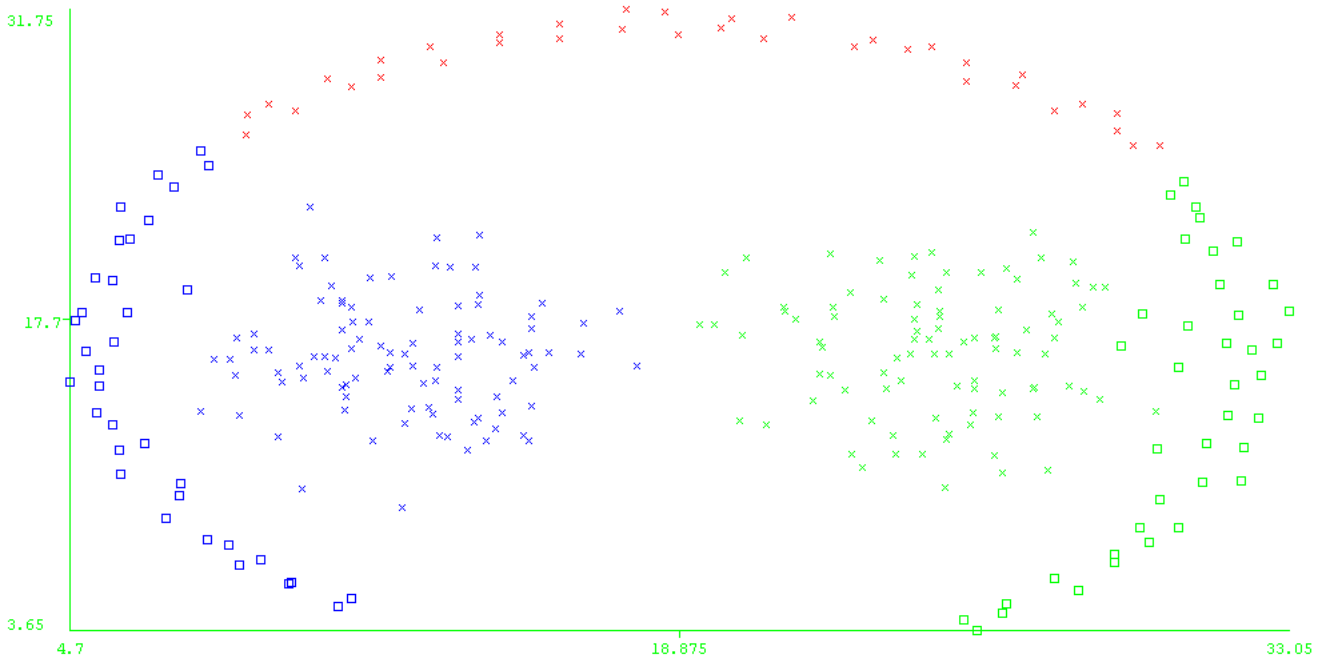

Figure 13: Resulting labels ward linkage was run on path-based

4.2 Performance on spiral dataset

4.2.1 DBSCAN

DBSCAN classifies all points correctly in this case as density difference is same.

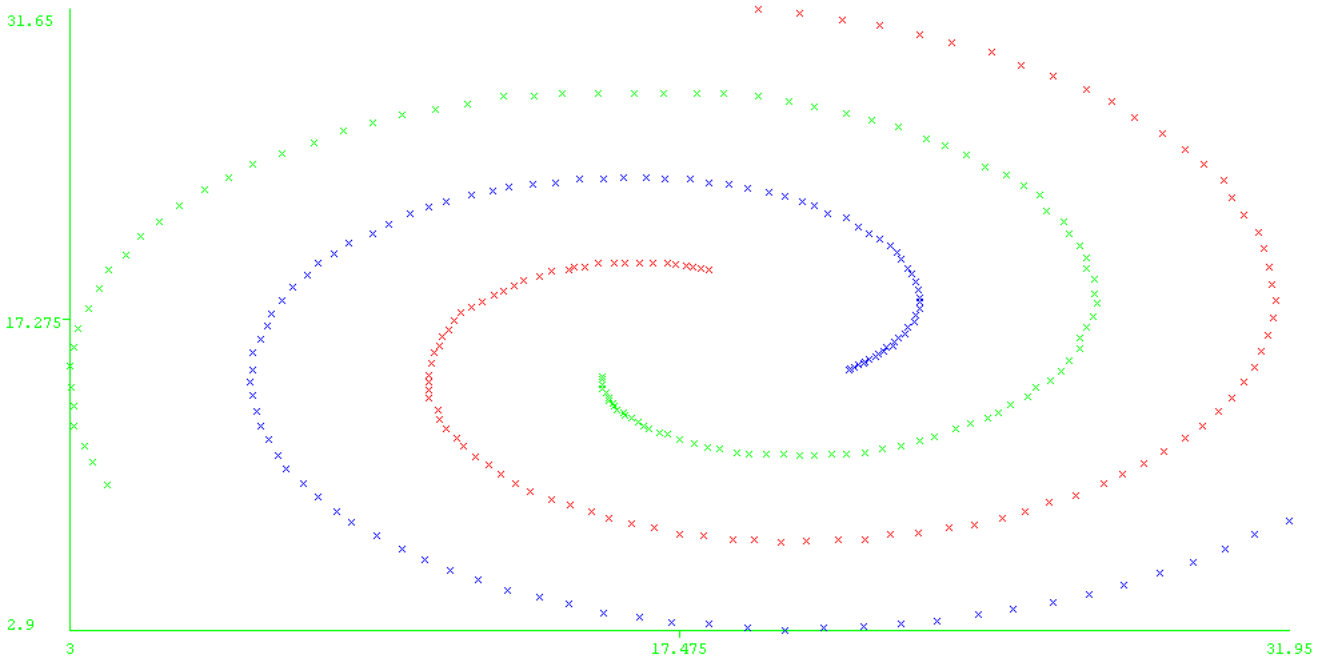

Figure 14: Resulting labels ward linkage was run on path-based

4.2.2 Hierarchical

Single linkage gave best performance among hierarchical purity of clustering is 100%

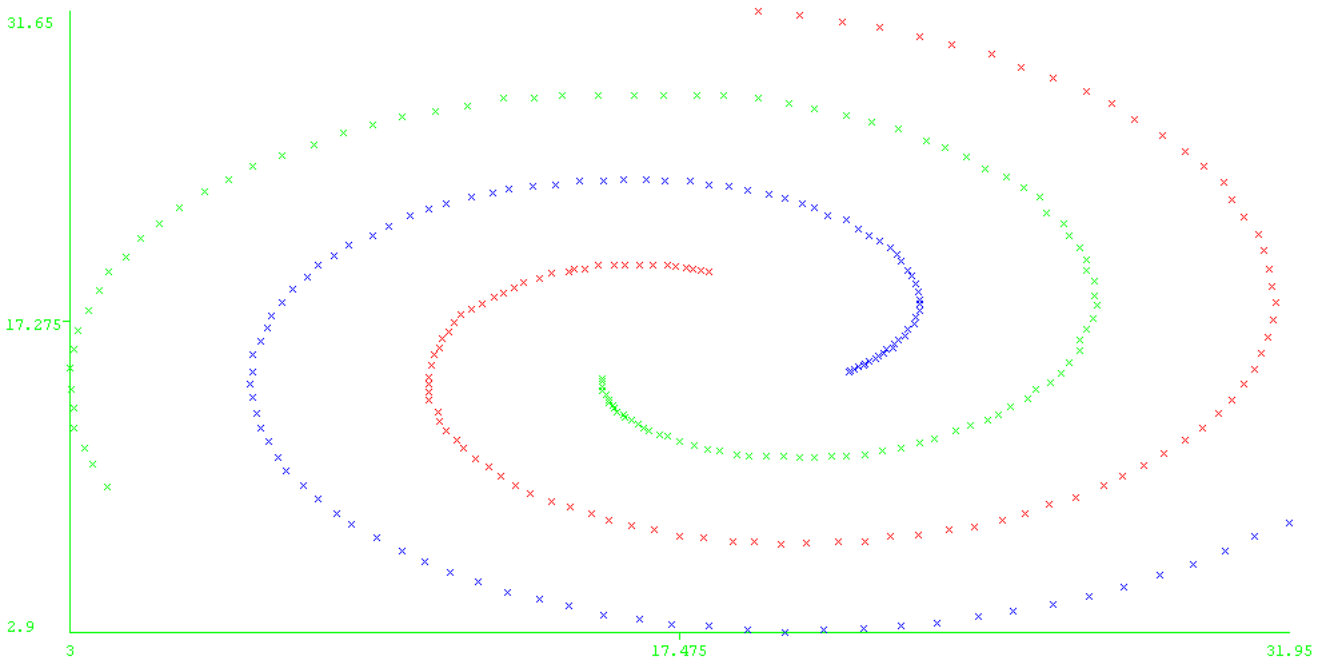

Figure 15: Resulting labels single linkage was run on path-based

4.3 performance on flames dataset

4.3.1 DBSCAN

density difference between two classes are same so it classifies most of the points correctly however some points are left as outliers

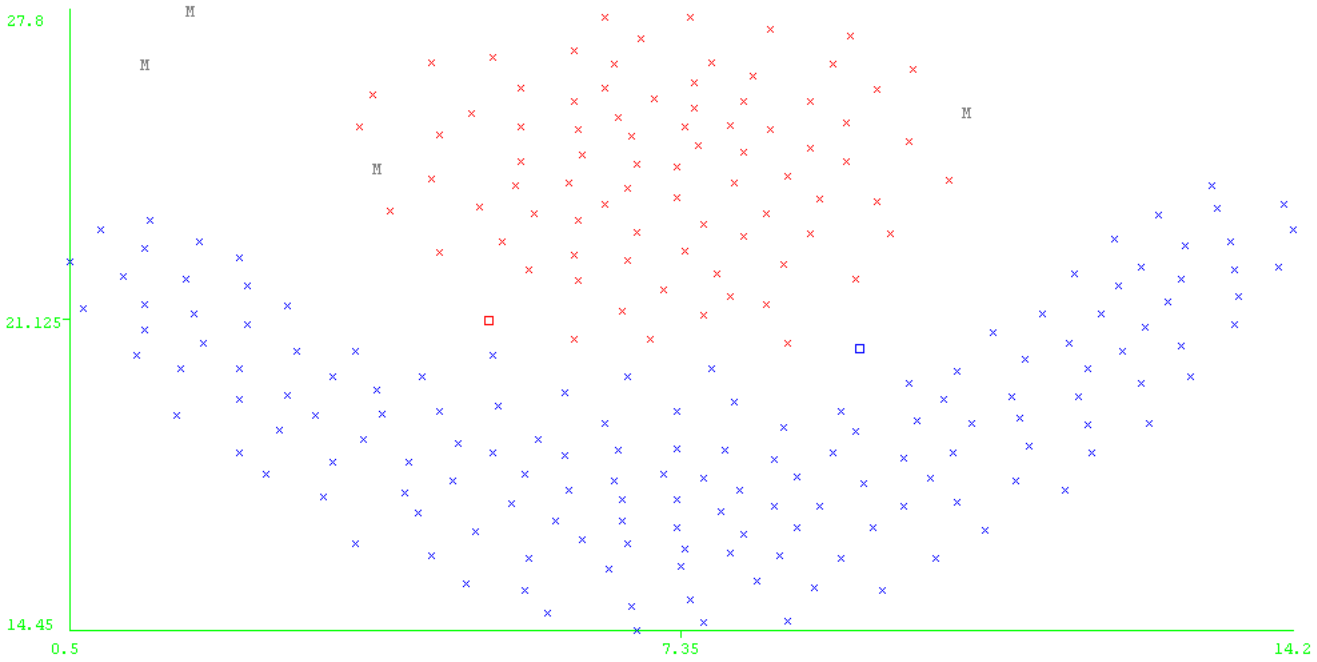

Figure 16: DBSCAN on flames

4.3.2 Heirarchical clustering

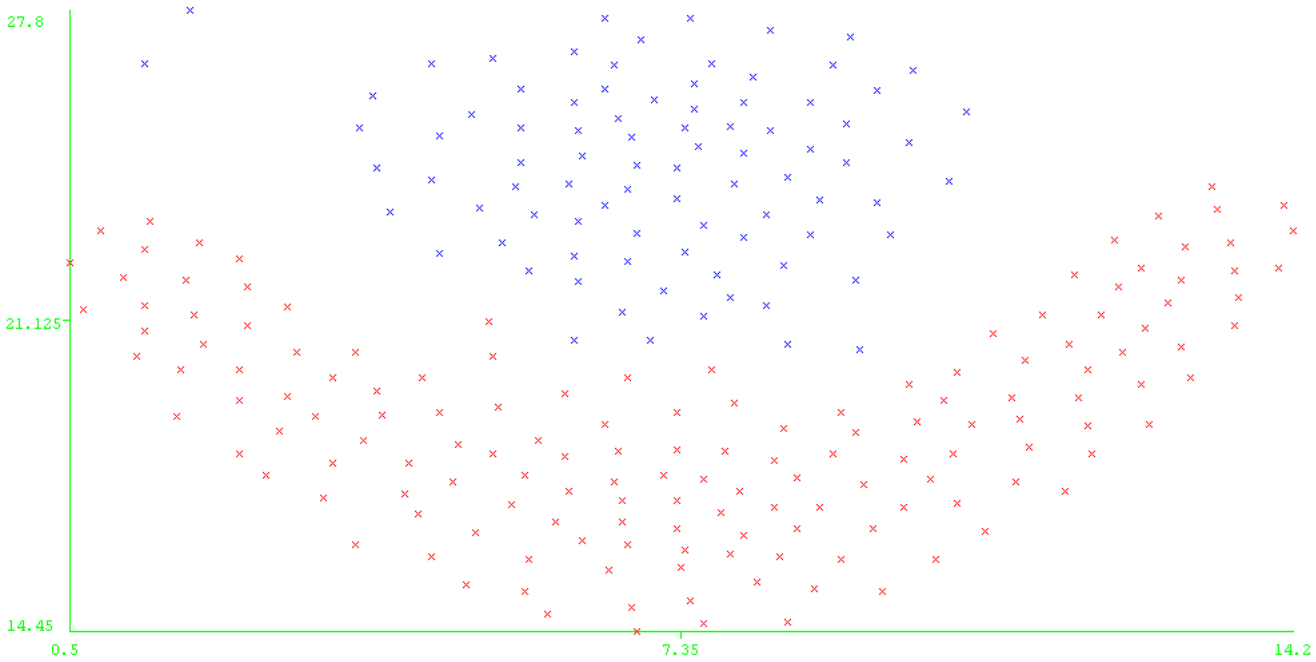

Figure 17: Resulting labels ward linkage was run on path-based

5 K-means with D31 dataset

we can't recover all classes with k=32 we should increase k to 38 to make it recognise all points

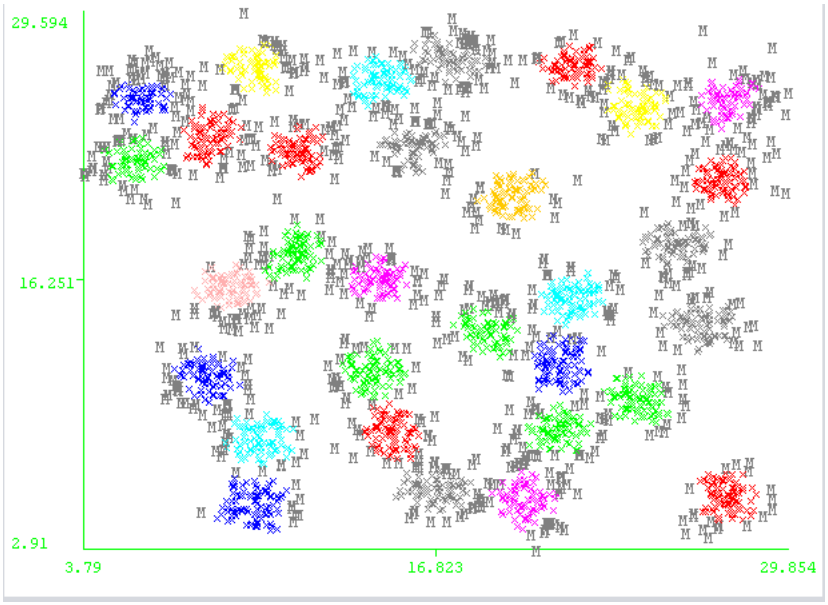

Figure 18: resulting graph when k=38