CLUSTERING ALGORITHMS(Problem 2)

Nitin K (2017csb1093)

30.10.2020
CSE FINAL YEAR, UG
IIT ROPAR

INTRODUCTION

This a report for two clustering algorithms(EM, Denclue). We are going to analyse how these algorithms perform on two different dataset(**Iris**, **Spiral**).

Algorithms Details:

I have Taken All the Algorithms from the books. That is why I am providing the reference only as the algorithm has already been discussed in the class.

EM(Expectation Maximization)

Reference book for the Algorithm is **Data Mining and Analysis: Fundamental Concepts and Algorithms, Mohammed J. Zaki and Wagner Meira Jr.**

Denclue

Reference book for the Algorithm is **Data Mining and Analysis: Fundamental Concepts and Algorithms, Mohammed J. Zaki and Wagner Meira Jr.**

DATA

DataSet Descriptions

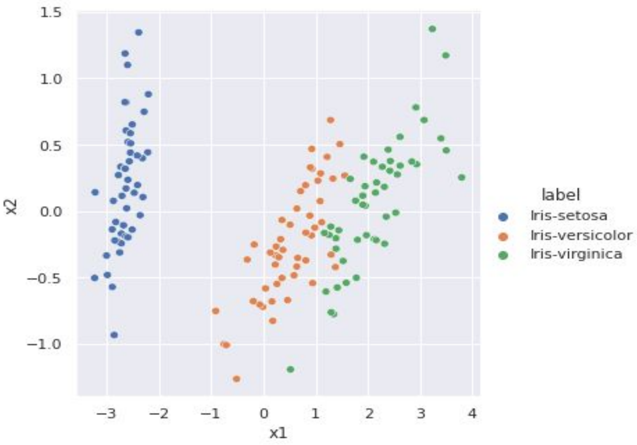
Here We have Provided two Different DataSet:

- **Iris DataSet (https://archive.ics.uci.edu/ml/datasets/iris)**

	sepal_length	sepal_width	petal_length	petal_width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

After PCA transformation of Iris Data on two principal components using sklearn.decomposition.PCA

	x1	x2
count	1.500000e+02	1.500000e+02
mean	3.552714e-16	-5.617729e-16
std	2.055442e+00	4.921825e-01
min	-3.225200e+00	-1.262492e+00
25%	-2.530159e+00	-3.235986e-01
50%	5.533290e-01	-3.251102e-02
75%	1.549463e+00	3.288601e-01
max	3.794687e+00	1.370524e+00



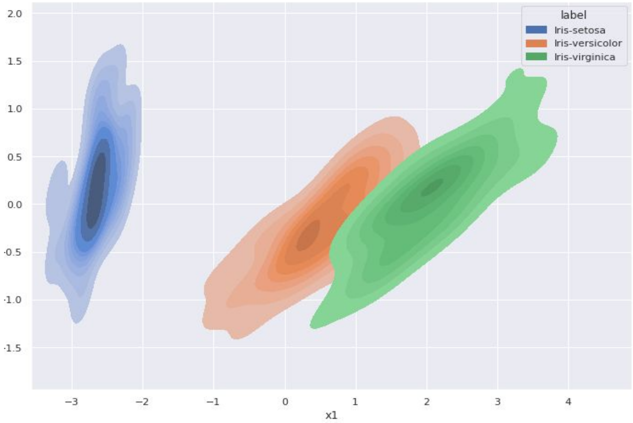
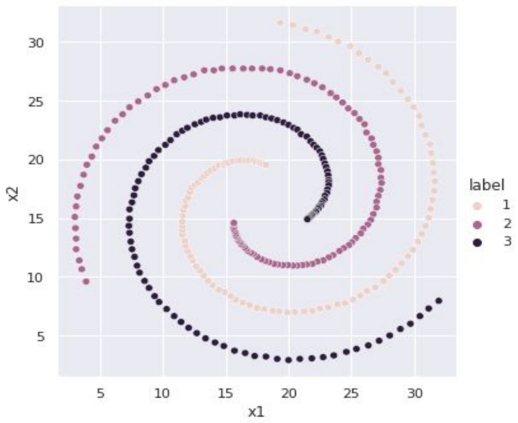


Fig:1.1(2D scatter Plot)

Fig: 1.2(2d Density Plot)

- **Spiral DataSet(https://github.com/milaan9/Clustering-Datasets)**

```
              x1          x2       label
count  312.000000  312.000000  312.000000
mean    18.408173   16.344712    2.016026
std      7.299923    6.867232    0.815682
min      3.000000    2.900000    1.000000
25%     12.912500   11.337500    1.000000
50%     18.325000   16.050000    2.000000
75%     23.400000   21.362500    3.000000
max     31.950000   31.650000    3.000000
```



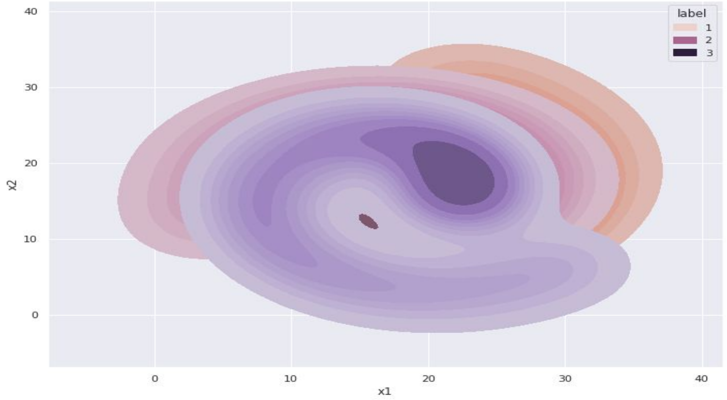


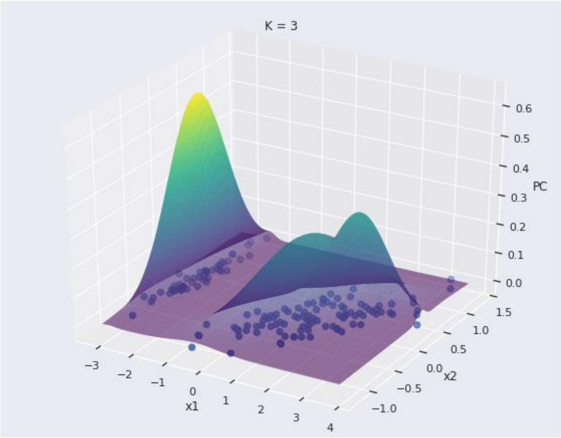
Fig:2.1(Scatter Plot)

Fig:2.2(2d Density Plot)

RESULTS

1. EM (Expectation Maximization)

Estimation Density Plot(For K = 3, Epsilon = 0.001)



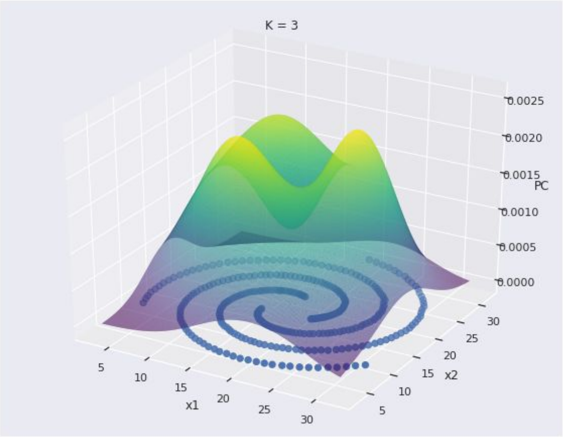
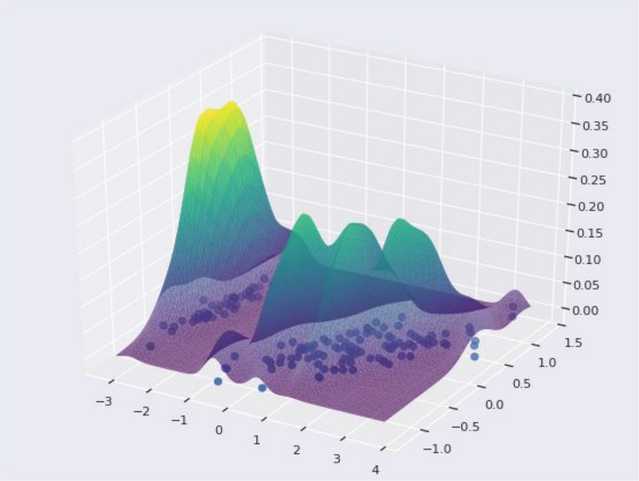


Fig:3.1(Iris DataSet)

Fig:3.2(Spiral DataSet)

2. Denclue

Estimation Density Plot



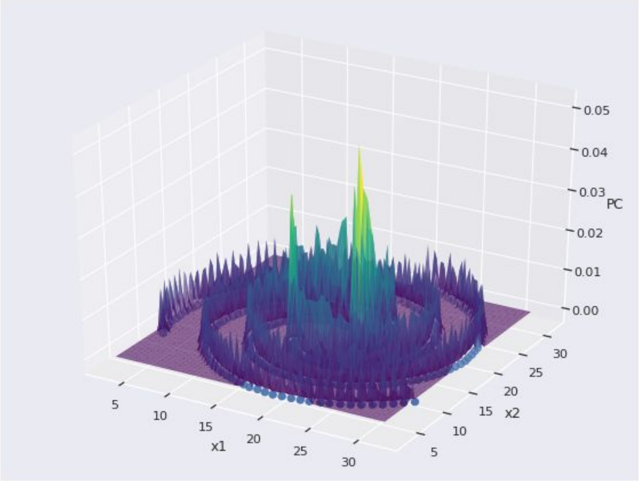


Fig: 4.1(Iris DataSet)

Fig:4.2(Spiral DataSet)

Expectation Maximization already considers that dataset is **GMM**. The algorithm uses expectation and maximization steps to find the local maximum likelihood parameter of the model. From the **Fig:3.1 and Fig:3.2** we can see that the the algorithm gives good result on Iris DataSet and bad result on Spiral DataSet because of it's GMM consideration.From **Fig:1.2 and Fig:2.2** we can see that Iris dataset is GMM but Spiral dataset is not. Hence the algorithm is not able to make good estimation on the spiral dataset.

Denclue is a clustering method based on density distribution functions. Clusters can be determined mathematically by identifying density attractors which is a local maxima of overall density function. The function is used to find density attractors(local maxima).From **Fig: 4.2** we can see that the algorithm gives good estimation on Spiral dataset. But on the iris dataset from **Fig:4.1** we can see that one cluster is well separated but it is hard to separate the other two clusters.Because it has many peaks near to each other which seem like merged structures.

The algorithm performs really well on an arbitrarily shaped dataset by using mathematical basis.

CONCLUSION

EM algorithms perform best on a dataset which are **GMM**. whereas **Denclue** performs best on an arbitrarily shaped dataset. However **Denclue** performs well on all types of dataset because here we do not have to calculate **expectation and maximization it is just a mathematical basis. which uses a hill climbing approach to find attractors.**

The only problem with Denclue is that it is really expensive in respect of runtime.Where as **EM** is faster.