Question #2 Report

Implementation of common soft assignment clustering algorithms

Abhinav Khanna(2017CSB1061)

CS524: Data mining

INTRODUCTION

In the area of clustering algorithms using soft assignment, 2 major algorithms were discussed: EM algorithm and Denclue. The purpose of this project is to implement these algorithms on various datasets, analyze their space and time complexity, and conclude which algorithm is better for which dataset. The scope of this project is limited by the hardware capabilities, such as RAM, number of cores/processors, and CPU clock cycles, availability and usage of GPUs et al.

THEORETICAL FOUNDATION

EM algorithm is a Gaussian Mixture model based representative clustering algorithm. In this, for a given K, we find the probability of a point lying in the ith cluster using the multivariate Gaussian distribution. The probability density function of a point is the sum of pdfs over all clusters. It consists of 2 parts: Expectation step, and Maximization step. In the expectation step, we find the posterior probability for each pair of data points and cluster. Then, in the maximization step, we find the new means, covariance matrix and mixture parameters. The expectation-maximization steps continue until we get convergence in all the parameters.

Denclue is also a GMM based density clustering method. In this, we find the attractors for each of the data points. If the pdf of the generated point is greater than an input parameter zi, the generated point is counted as an attractor. Once we have all the attractors, we form density based clusters by clustering all points which are density reachable.

IMPLEMENTATION

- EM algorithm

Data Preparation

In this implementation, I have taken the first two principal components as my dataset(for both Iris and 2d spiral). The dataset is represented as a list of lists.

Clustering implementation

As inferred from the previous question, I have taken K = 3 number of clusters. Initially, I have chosen K random means, K covariance matrix which are 2X2 identity matrices, and mixture parameters as 1/K for each of the clusters. Expectation and maximization steps are performed repeatedly until convergence. An important observation to be made here is that while updating the covariance matrix, the mean to be used is the existing mean and not the updated mean.

- Denclue

Data Preparation

In this implementation, I have taken the first two principal components as my dataset(for both Iris and 2d spiral). The dataset is represented as a list of lists.

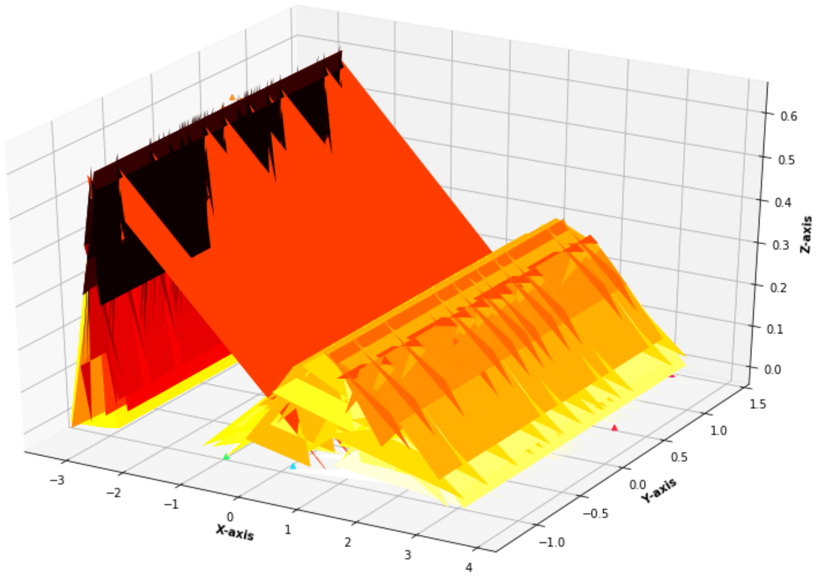
Clustering implementation

For each data point, we find the attractor. The attractor is found by updating the given point, using the formula discussed in class, until the difference between the new point and the existing point is less than an input parameter epsilon. Then, the same steps as discussed earlier are followed.

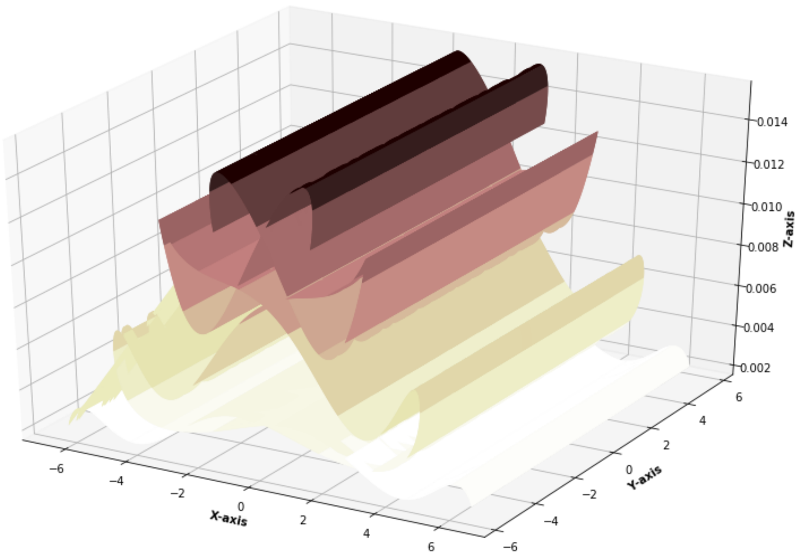
OBSERVATIONS AND OUTPUT

- EM algorithm

  Iris dataset (Time taken: 0.0578 sec)

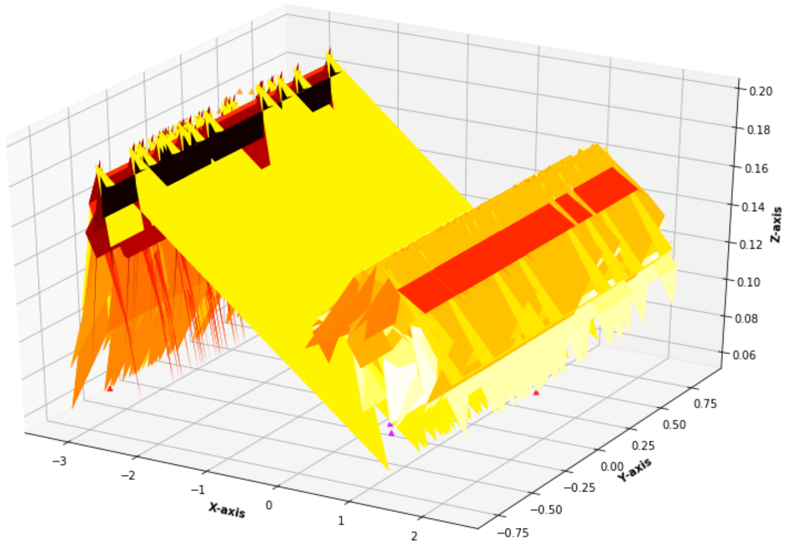


2d spiral dataset (Time taken: 0.09 sec)

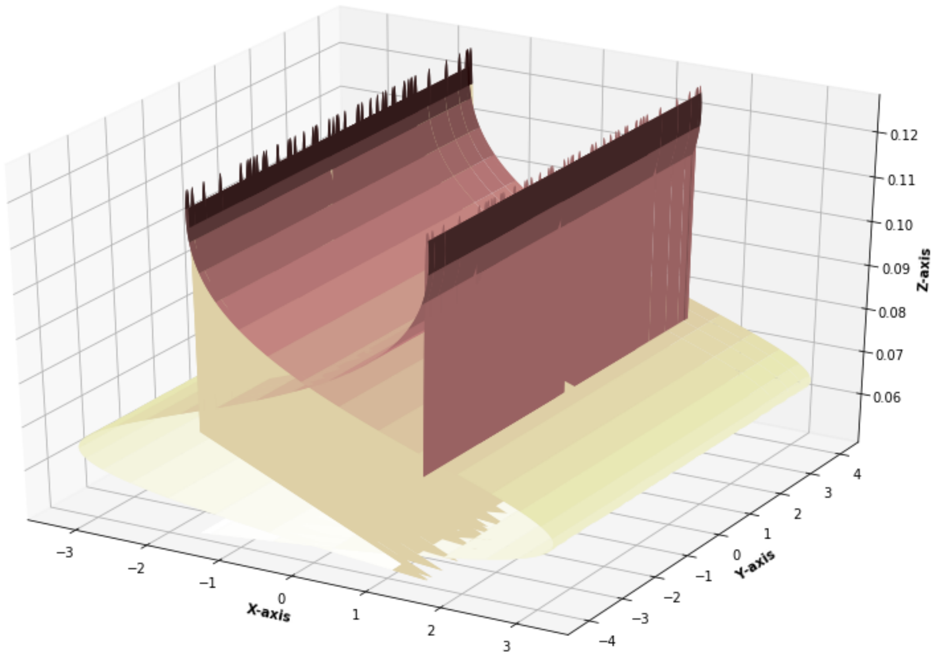


- Denclue

Iris dataset (Time taken: 0.16 sec + 0.002 sec)



2d spiral (Time taken: 2.74 sec + 0.013 sec)



Note: Time taken is shown as a sum of time taken to find attractors + time taken to assign clusters.

CONCLUSION

From the above outputs and observations, we can again see that the EM algorithm is better suited to describe the Iris dataset, while the Denclue algorithm is better for 2d spiral. It is also clear that the EM algorithm is much faster than the Denclue algorithm, the reason being having to find attractors for each point, which increase the complexity to O(n^2)

REFERENCES

1. Data Mining and Analysis: Fundamental Concepts and Algorithms: Textbook by Mohammed J. Zaki and Wagner Meira
2. Class notes