Assignment 5
Foundations of Machine Learning

IIT-Hyderabad
Aug-Dec 2021

Questions: Programming

CS21MTECH14007
ADITI BAGORA

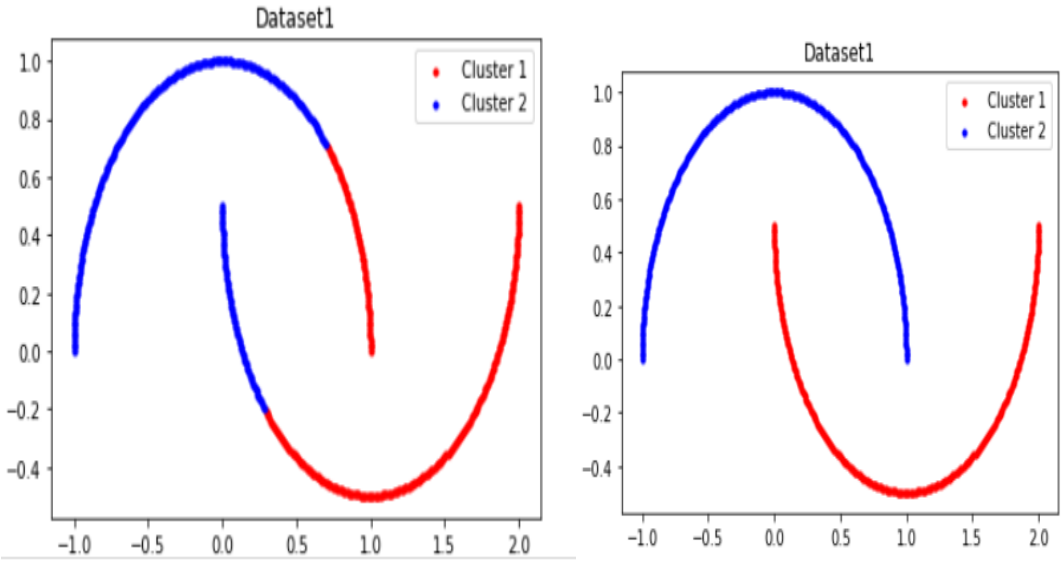
Questions: Programming

1. Clustering *(7 marks)*: DBSCAN, as we discussed in class, is a density-based clustering algorithm. In this problem, you need to implement your own DBSCAN algorithm. You can read more about it from paper that proposed this method [link].

   (a) Use the Kmeans clustering algorithm from sklearn and find the number of clusters in dataset1 shared with you. Plot the data points with different colors for different clusters. [*1 mark*]

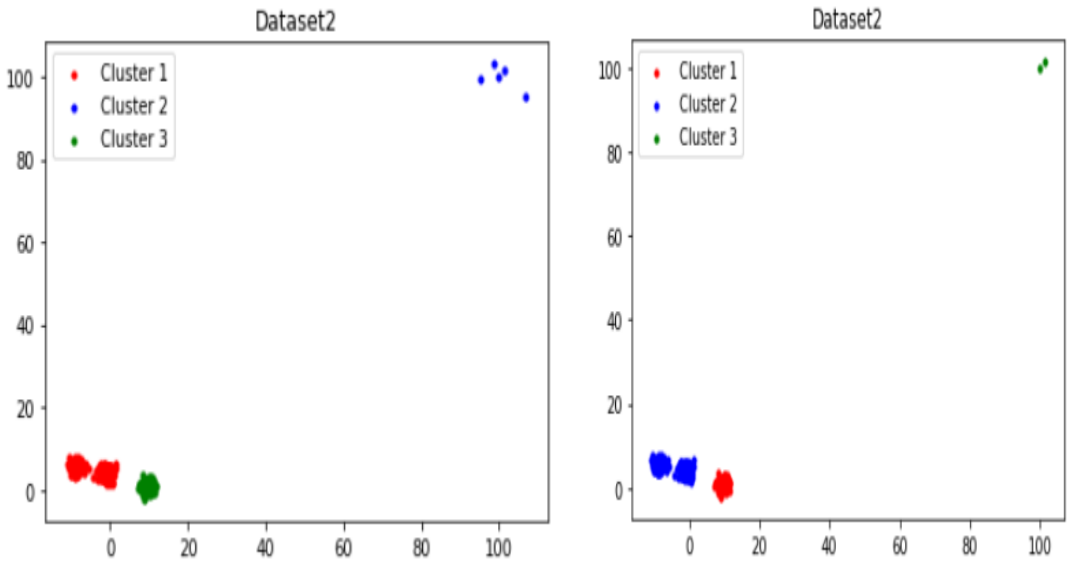
   (b) Implement your own DBSCAN algorithm on the same dataset and plot the data points. [*3 marks*]

   (c) What differences do you see between the DBSCAN and *k*-means methods, and why? [*1 mark*]



We can see that DBSCAN is able to cluster the samples better than K-means because of the bias of k-means towards clustering spherical structures even after using K-means ++ it is unable to cluster properly whereas DBSCAN can easily cluster since it works based of density connectivity an since both the structures are not at all connected it can cluster them easily.

   (d) Consider the dataset2 (also shared with you) with three clusters. Use (a) and (b) for dataset2, and compare the performance. List your observations clearly, and make conclusions on pros and cons of DBSCAN and *k*-means. [*2 marks*]



If we look at specifically a and b we can see that in first case DBSCAN works better as based on density connectivity it can easily cluster properly whereas in K-means even if the centroid are initialized properly the section where both the curves are closer than other points it is most likely to cluster them as done for dataset 1.But if we look at b we can see that K-means is able to cluster properly even the cluster which is less densely populated it is able to cluster but in DBSCAN the less densely populated cluster became more smaller as DBSCAN is treating other points in that cluster as noise hence for DBSCAN to work properly density of clusters matters it might happen that a cluster that is sparse and less densely populated it can be amused as noise or outlier by DBSCAN

K-Means

Pros:

- Perform on large dataset better compared to DBSCAN
- Can perform on varying density datasets better

Cons:

- Clusters formed are more or less spherical or convex in shape and must have same feature size.
- Sensitive to K
- Sensitive to outliers and noise

DBSCAN

Pros:

- Can cluster arbitrary shapes and one cluster surrounded by other
- Robust to outlier detection

Cons:

- Sensitive to min-points and epsilon
- Difficult to cluster having different density and spare with varying density