Report for Programming questions-sm21mtech14001:

Question 1:

(c)

Difference between KMeans and DBSCAN Clustering:

With the above dataset, Kmeans gives us wrong clustering whereas DBSCAN gives us the correct clustering. In KMeans, even two points which are close to each other belong to two different clusters and the clustering mainly depends on the hyper parameter we are choosing. Kmeans tries to create same sized cluster no matter how the data is scattered and it does not work well for non-globular structures. DBSCAN overcomes the disadvantage of Kmeans by working with the density of points. Since, DBSCAN works with density, it can easily model non-globular structures.

(d)

Hence, for the dataset2, the number of clusters identified by the K-means and DBSCAN is **2** and **3** respectively. Intuitively from the scatter plot, it is easy to say that the number of clusters is 3. Hence the DBSCAN algorithm performs better than the K-means.

List of advantages of KMeans and DBSCAN:

* Execution time required for K-means is much lesser than that of DBSCAN.

* The number of clusters need not be given before the execution with respect to DBSCAN.

List of disadvantages of KMeans and DBSCAN:

* In the K-means, it is necessary to mention the number of clusters hidden in the dataset before executing.

* DBSCAN doesn’t work well over clusters with different densities.

* DBSCAN needs a careful selection of its parameters.

Question 2:

Qn.2a) Effect of increase in iterations:

Here, we could see that the experiments with iterations 1000 and 2000 produces same results. But minimum iterations of 250 did not produce optimal clustering. This could be because of the lack of convergence of the solution. The result for 1000 and 2000 iterations are the same because of cost function had already converged after certain number of iterations.

Qn.2b)

It is observed that, for some datasets, different runs of t-SNE algorithm with the same hyperparameters produce different results. The reason is because of the **non-convexity** of the cost function in the t-SNE. As the cost function is not convex, different initializations can give us different results. Hence only local convergence is guaranteed which is not very robust.