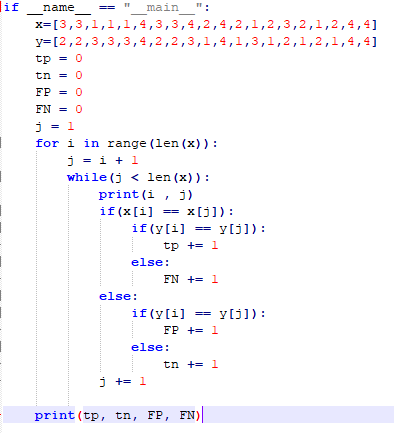
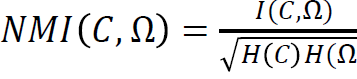
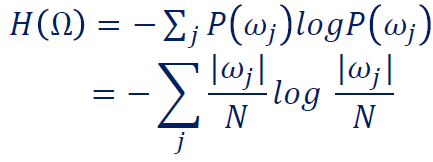
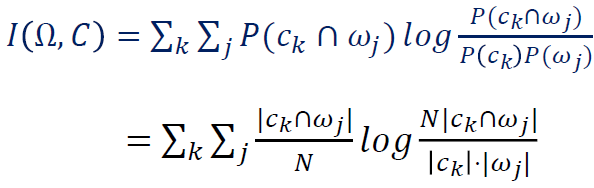
Hun Lee 604958834

1.



TP = 32 TN = 141 FP = 9 FN = 8



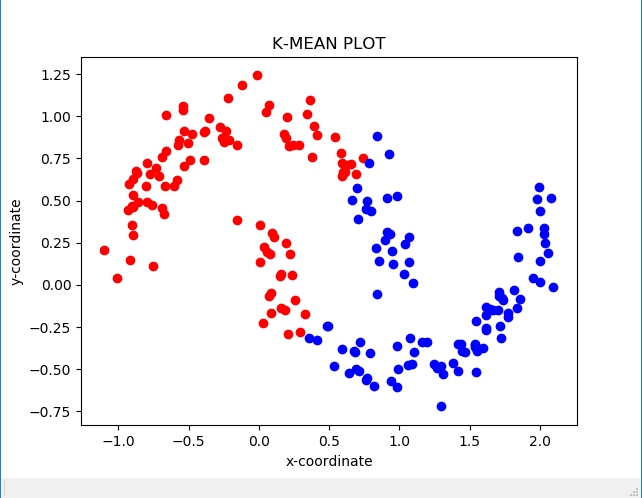
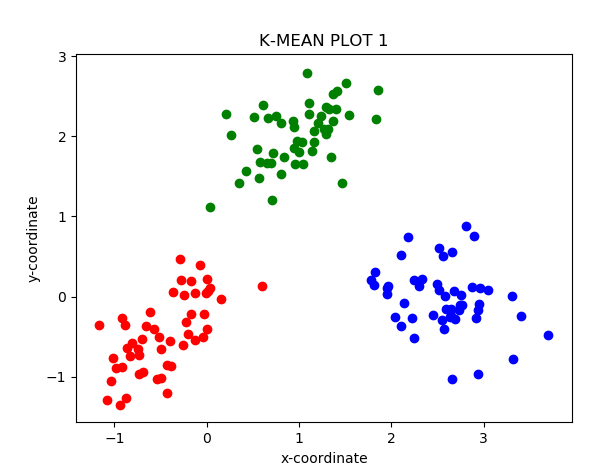


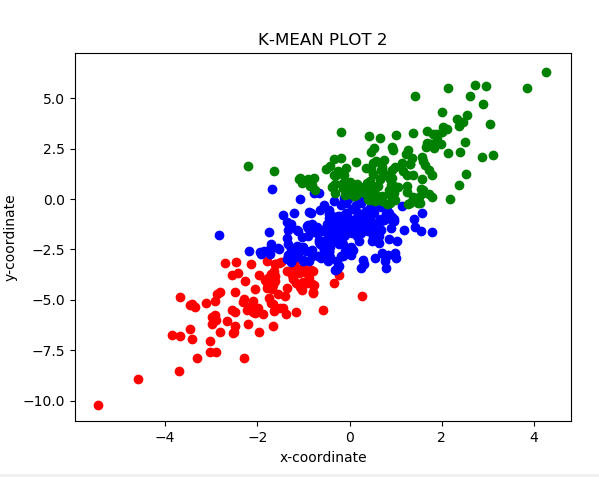
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ＧＴ／Ｏ | １ | ２ | ３ | ４ |  |
| １ | ０ | １ | ４ | ０ | ５ |
| ２ | ５ | ０ | ０ | ０ | ５ |
| ３ | ０ | ５ | ０ | ０ | ５ |
| ４ | ０ | ０ | １ | ４ | ５ |
|  | ５ | ６ | ５ | ４ | ２０ |

Evaluation:

This test did a good job clustering the data. Purity of 0.9 means maximum of points that are matched with their ground truth level is 90%. Precision and recall values are also high with the value of 0.8. this stands for number of labels matched right is substantially higher that the number of labels matched mistakenly. Combining these two values, we obtain F-measure of 0.79 which indicates a very high level of performance. The model gave us NMI value of 0.815. this means there is a high dependency between the points in each clusters. In other words, given a point, we are more likely to make a right guess where this point belongs to. This also is a indication of a good test performance.

２．



Data set 1 : Purity is :1.0 NMI :1.0

Evaluation:

K-means does a wonderful job with 1st data set. Purity of 1 means that all data points matches its ground truth label. NMI of 1 means all the data in each cluster are highly correlated and existence of one can tell the label of another.

Data set 2 : Purity is :0.764 NMI :0.046851365954979234

Evaluation:

K means does a moderate job with the 2nd data set with the purity of 0.764. this means that 76.4% of the points in the data set matched their ground truth label. As you can also see NMI is very low. This means the dependency among the points within the same cluster are very low. This can be seen on the plot. If you look at the borderline of the blue cluster, you can see a lot of points that belong to red and green clusters. Although those points are close to the blue cluster, they belong to other cluster groups. This is not surprising given .0468 NMI.

Data set 3 : Purity is :0.76 NMI :0.14502499937722388

Evaluation:

The plot has purity of 0.76. This tells us how well these two clusters match to the ground truth label of the points. The value 0.76 is fairly high and tells us the quality of this cluster is fine. As you can see, the end part of the clusters are colored differently from the rest of points. This can be explained by moderate to low level of independency between the points specified by NMI. This means that inner parts of the red and blue clusters had higher correlation with the clusters they belong to.

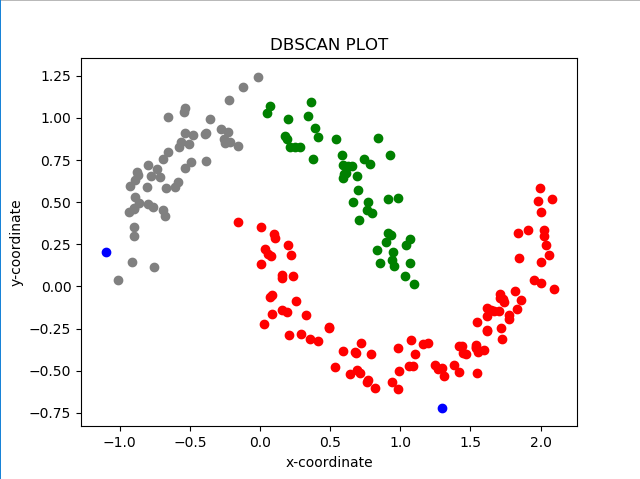
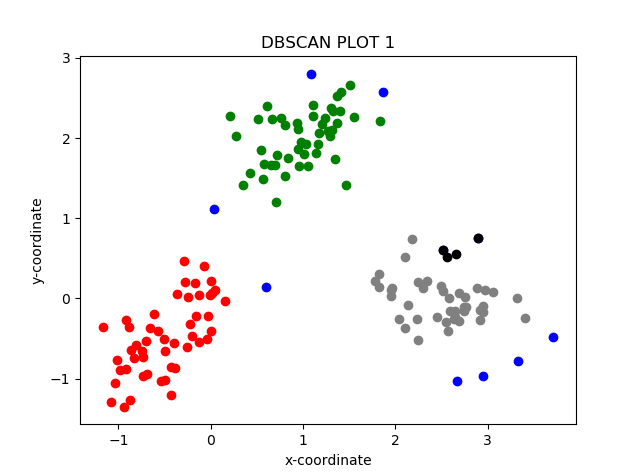
Strength: K-mean algorithm is very efficient.

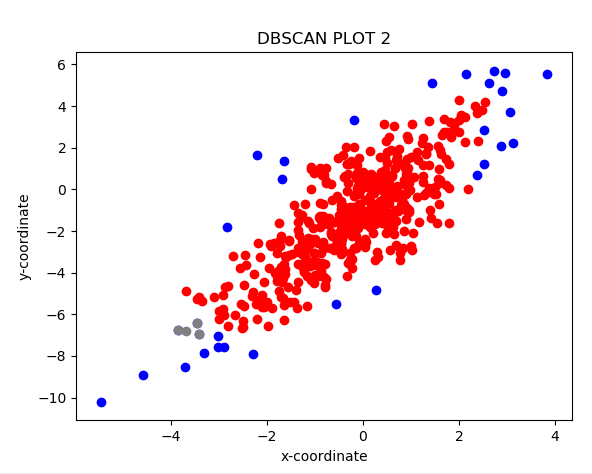
Weakness:

1. algorithm terminates at a local optimal.
2. The algorithm is only applicable to objects in a continuous n-dimensional space.
3. The number for K has to be specified.
4. It is sensitive to noisy data and outliers.

It is inefficient in discovering clusters with non-convex shapes.

3.



Data set 1 : Purity is :0.94 NMI :0.9590647490609898

Evaluation:

DB scan does not do as well as the other models by a little bit. Purity of 0.94 means that 94 percent of the points match their ground truth label. This still is a very good performance. NMI is not 1 like models because as you can see, there is a small number of outliers and the dependency of the points within the same clusters are less strong.

Data set 2 : Purity is :0.714 NMI :0.011352036737124684

Evaluation:

DB scan does a moderate job with the 2nd data set as well with purity of 0.714. This says that 71.4 percent of the points in the data are matched with their ground truth label. As you can see on the plot, there does not seem like to be a high dependency among the points in the clusters. This can be seen on the low NMI value of 0.01135.

Data set 3 : Purity is :0.985 NMI :0.8173489274692755

Evaluation: DB scan does the best job with the 3rd data set. Its purity value suggests that 98.5% of the data points match their ground truth label. That is substantially better than performance of the other models. The reason why it is not 1 is because of some outliers as you can see on the plot. As you can see from NMI, the independency among points in each cluster is high. That means given a point, it is easier to guess the label of an unknown point. Not surprisingly, this leads to a better performance for DB scan.

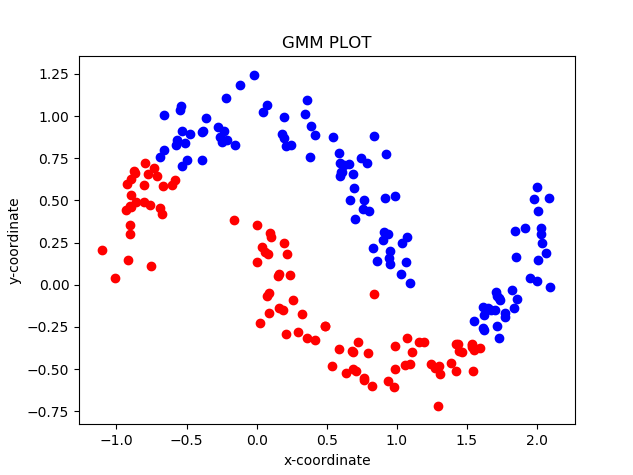
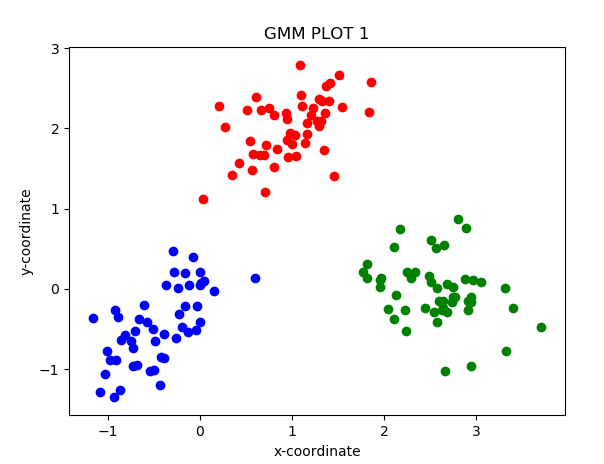
Strength:

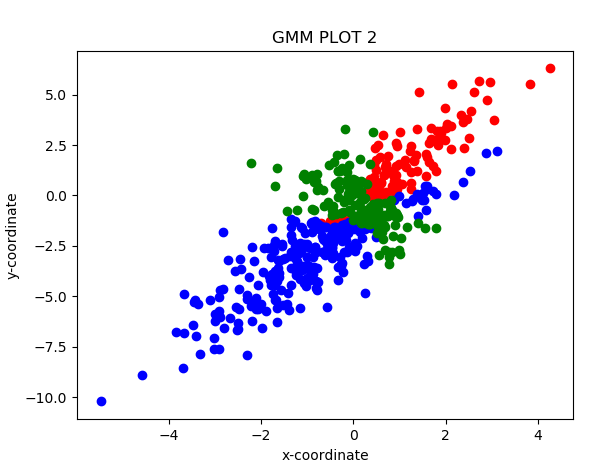
1. It can discover clusters of arbitrary shape
2. It can handle noise well
3. It is cost efficient with one scan

Weakness:

1. It needs density parameters as termination condition
2. It does not work too well when dealing with clusters of varying densities and higher dimensional data

4.



Data set 1 : Purity is :1.0 NMI :1.0

Evaluation:

GMM does a wonderful job with 1st data set. Purity of 1 means that all data points matches its ground truth label. NMI of 1 means all the data in each cluster are highly correlated and existence of one can tell the label of another.

Data set 2 : Purity is :0.764 NMI :0.07560366608611019

Evaluation:

GMM also does a moderate job with the 2nd data set with purity of 0.764. From this we can say 76.4% of the points in the data are matched with their ground truth label. Also, NMI value for this model is slightly higher than those of other models. This means that given a point, there is a more chance that this model would make an accurate decision. In fact it performs as well as K-means and better than DB scan.

Data set 3 : Purity is :0.69 NMI :0.07594783950403936

Evaluation:

The lowest value for purity says this model performs the worst among the other models for this given clustering task with purity of 0.69. This mean 69% of the data points match their ground truth label. As you can see from the plot, there is less dependency among the points within the same clusters with NMI of 0.0759. Thus it is hard to tell the label of a point given the label of another point. Also, lack of dependency is a reason why we can get clusters that divide the plot into two arbitrary shape.

Strength:

1. It can deal with different densities and sizes of clusters.
2. Clusters can be characterized by a small number of parameters.
3. The results may satisfy the statistical assumptions of the generative models.

Weakness:

1. It converges to local optimal
2. It is computationally expensive if the number of distributions is large
3. It is hard to estimate the number of clusters
4. It can only deal with spherical clusters