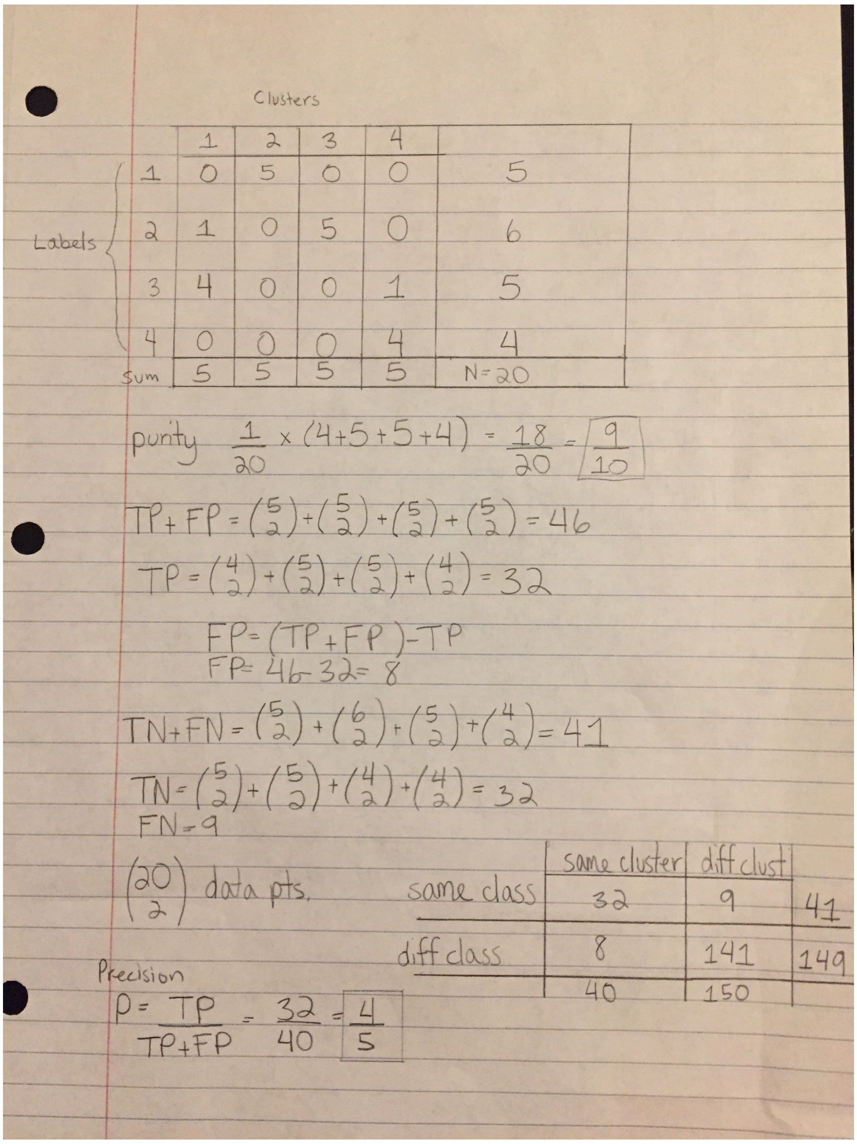
Homework 3

Daniel O'Laughlin

November 11, 2017

1 Problem 1



Clusters

		1	2	3	4	
Labels	1	0	5	0	0	5
	2	1	0	5	0	6
	3	4	0	0	1	5
	4	0	0	0	4	4
	Sum	5	5	5	5	N=20

purity $\frac{1}{20} \times (4+5+5+4) = \frac{18}{20} = \boxed{\frac{9}{10}}$

$TP + FP = \binom{5}{2} + \binom{5}{2} + \binom{5}{2} + \binom{5}{2} = 46$

$TP = \binom{4}{2} + \binom{5}{2} + \binom{5}{2} + \binom{4}{2} = 32$

$FP = (TP + FP) - TP$

$FP = 46 - 32 = 8$

$TN + FN = \binom{5}{2} + \binom{6}{2} + \binom{5}{2} + \binom{4}{2} = 41$

$TN = \binom{5}{2} + \binom{5}{2} + \binom{4}{2} + \binom{4}{2} = 32$

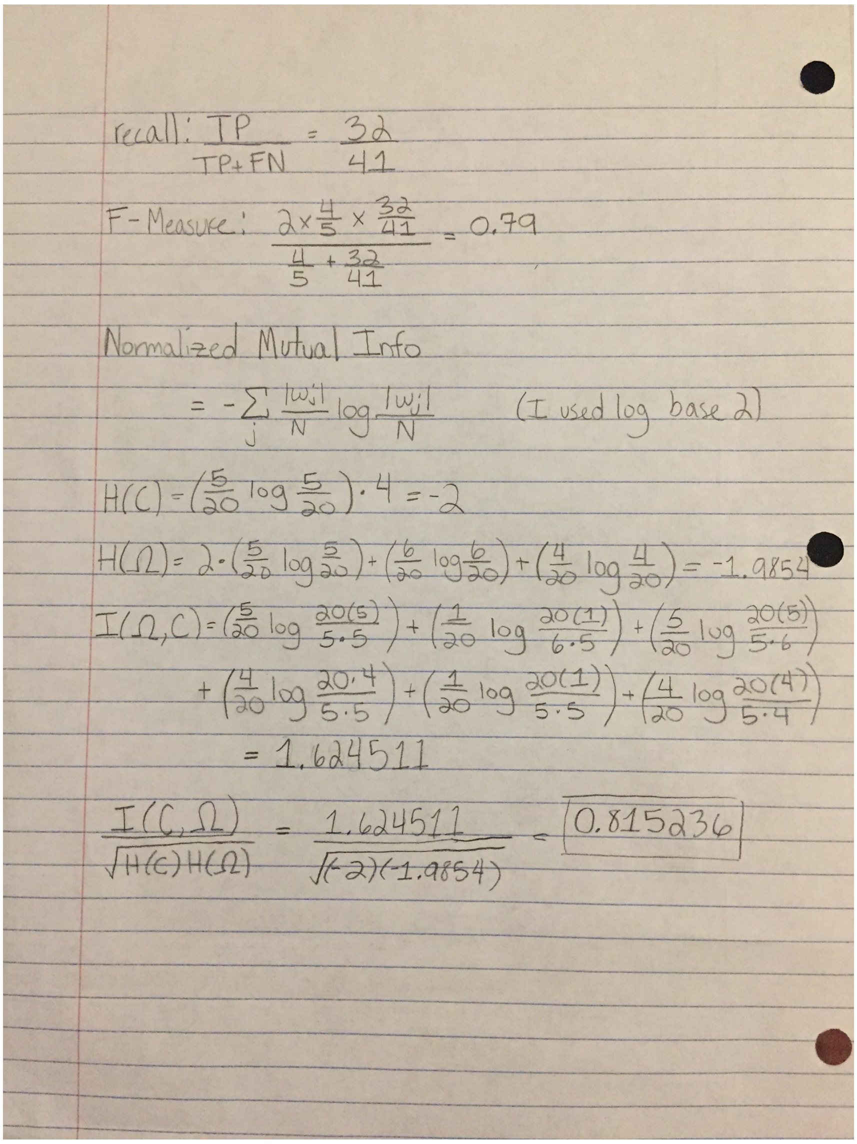
$FN = 9$

$\binom{20}{2}$ data pts.

	same cluster	diff clust	
same class	32	9	41
diff class	8	141	149
	40	150	

Precision

$P = \frac{TP}{TP+FP} = \frac{32}{40} = \boxed{\frac{4}{5}}$



recall: $\frac{TP}{TP+FN} = \frac{32}{41}$

F-Measure: $\frac{2 \times \frac{4}{5} \times \frac{32}{41}}{\frac{4}{5} + \frac{32}{41}} = 0.79$

Normalized Mutual Info

$$= -\sum_j \frac{|\omega_j|}{N} \log \frac{|\omega_j|}{N}$$ (I used log base 2)

$$H(C) = \left(\frac{5}{20} \log \frac{5}{20}\right) \cdot 4 = -2$$

$$H(\Omega) = 2 \cdot \left(\frac{5}{20} \log \frac{5}{20}\right) + \left(\frac{6}{20} \log \frac{6}{20}\right) + \left(\frac{4}{20} \log \frac{4}{20}\right) = -1.9854$$

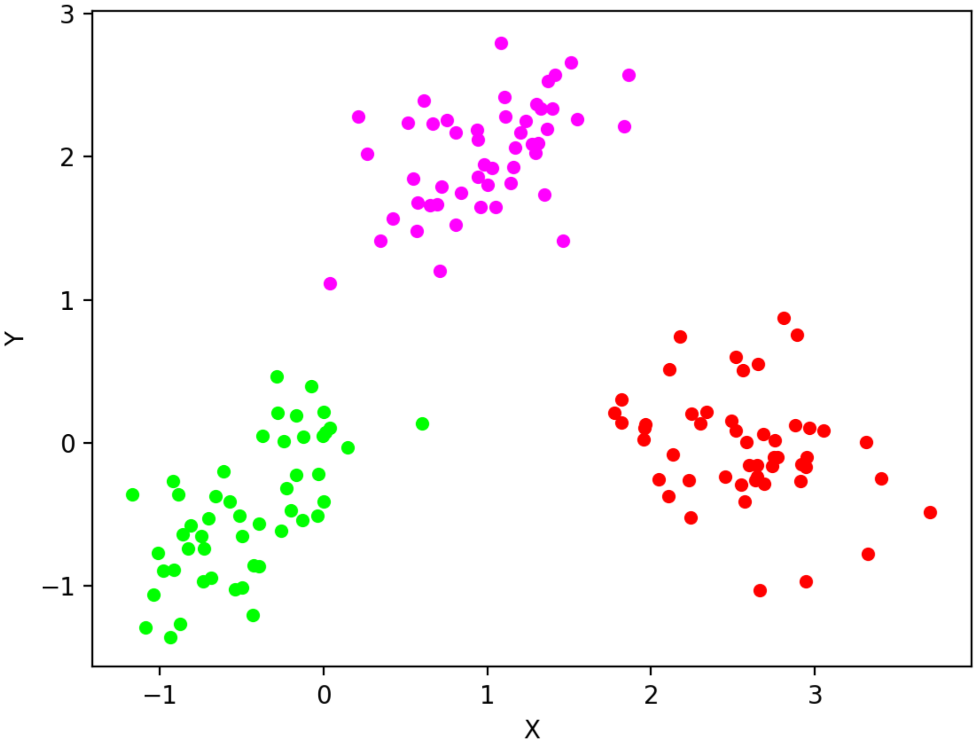
$$I(\Omega, C) = \left(\frac{5}{20} \log \frac{20(5)}{5 \cdot 5}\right) + \left(\frac{1}{20} \log \frac{20(1)}{6 \cdot 5}\right) + \left(\frac{5}{20} \log \frac{20(5)}{5 \cdot 6}\right)$$
$$+ \left(\frac{4}{20} \log \frac{20 \cdot 4}{5 \cdot 5}\right) + \left(\frac{1}{20} \log \frac{20(1)}{5 \cdot 5}\right) + \left(\frac{4}{20} \log \frac{20(4)}{5 \cdot 4}\right)$$
$$= 1.624511$$

$$\frac{I(C, \Omega)}{\sqrt{H(C) H(\Omega)}} = \frac{1.624511}{\sqrt{(-2)(-1.9854)}} = \boxed{0.815236}$$

2 Problem 2

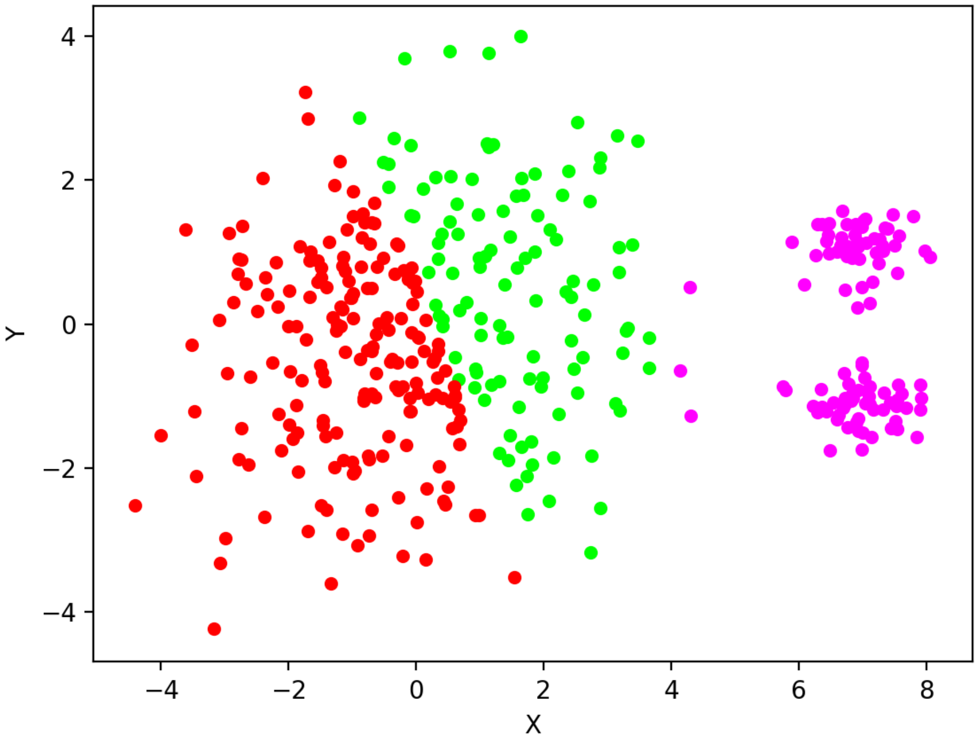
Part 1)
Implemented in code.

Part 2)



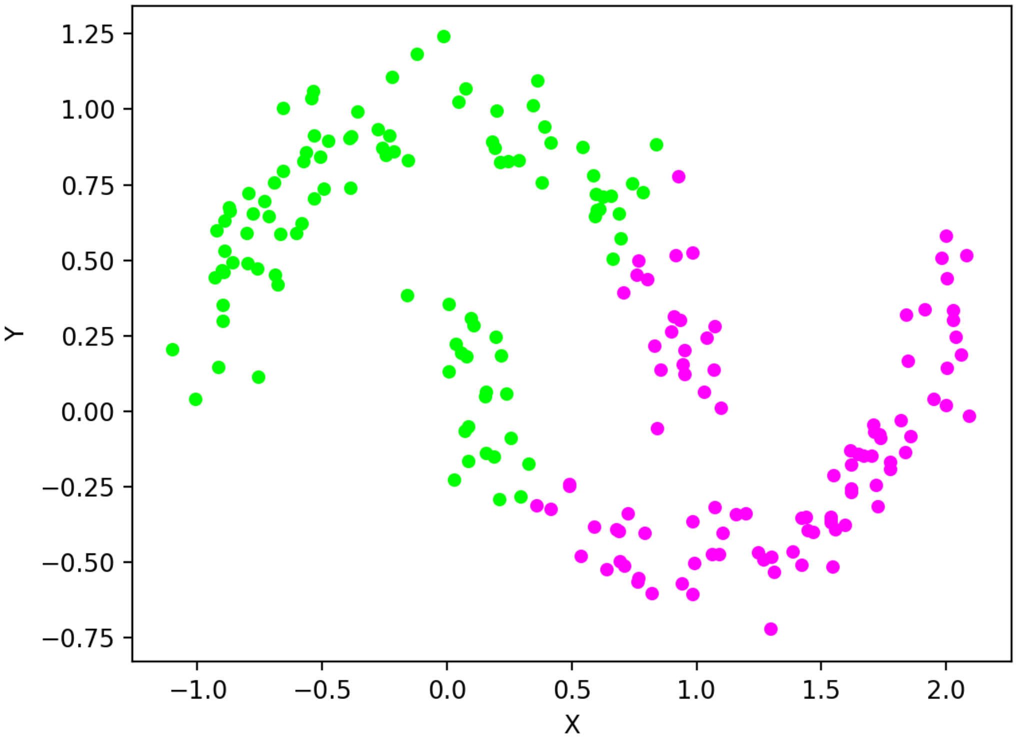
Data Set 1:
Purity is :1.0
NMI :1.0

By evaluating the purity and normalized mutual information values obtained for dataset 1, we see that he data was perfectly clustered.



Data Set 2:
Purity is :0.8675
NMI :0.464256439333

Purity is the measure of the extent to which clusters contain a single class and thus, a purity a higher purity is more desirable than a lower purity. Our purity of 0.8675 is thus not perfect, but still a relatively acceptable value. Mutual information is a measure of the information that two variable X and Y share. Thus, it tells us how much knowing one variable is able to reduce the uncertainty about another variable. Thus, a higher NMI is more desirable. Our NMI is 0.464 which is low and not ideal with regards to the quality of the resulting clusters.



Data Set 3:
Purity is :0.78
NMI :0.169704955284

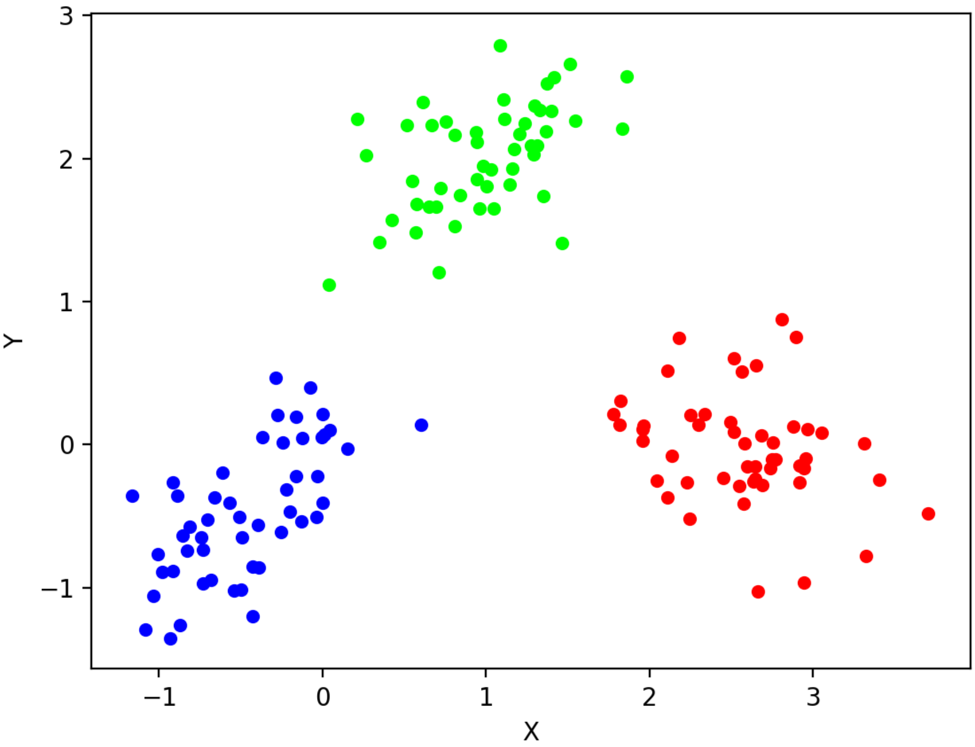
Our purity for dataset 3 is 0.78. This is lower than the purity for both datasets 1 and 2. Intuitively, we can see that the decided clusters contain a good deal of points belonging to different classes. Moreover, we have a very low NMI of 0.169 further diminishing the quality of our obtained cluster.

Part 3)
Some strengths of the K-means algorithm are the fact that it is easy to implement and generally has high speed performance relative to hierarchical clustering. Furthermore, it is very good at identifying tight, globular clusters. Some disadvantages are the fact that the number of clusters (K) must be determined beforehand, different random choice of cluster centers can produce different clustering results, it equally weights all attributes, outliers may heavily impact centroid location, and the performs worse on non-circular clusters.

3 Problem 3

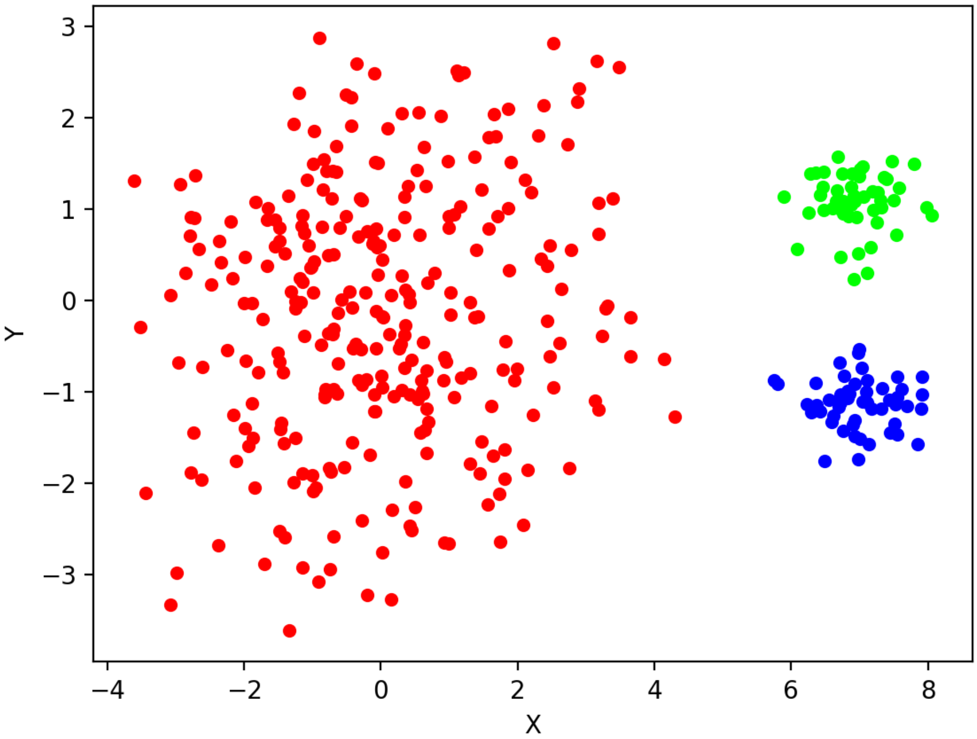
Part 1)
Implemented in code.

Part 2)



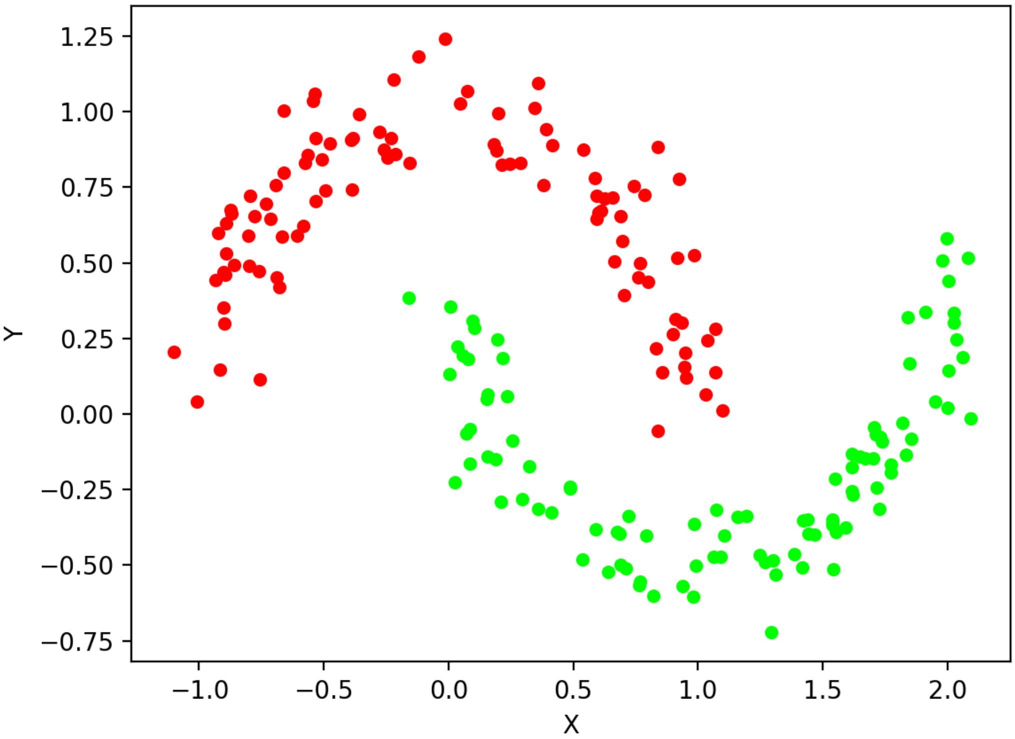
For dataset1
Purity is :1.0
NMI :1.0

We have both a purity and NMI of 1 and can thus see that we have clustered our data correctly.



For dataset2
Purity is :0.9625
NMI :1.0

We obtained a perfect NMI of 1 and a near perfect purity of 0.9625 indicating that our clusters are near perfect.



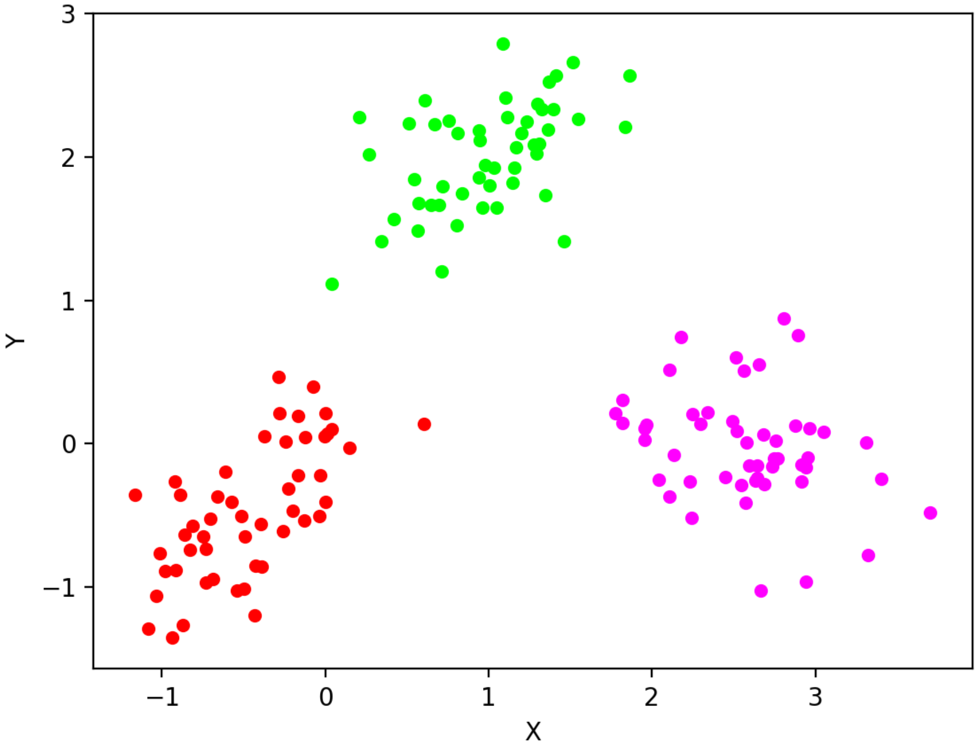
For dataset3
Purity is :1.0
NMI :1.0

We have purity of 1 and and NMI of 1 indicating that our data is perfectly clustered.

Part 3)
A good thing about DBSCAN is that it can handle clusters of many different shapes and sizes (ex. moon shaped cluster), thus it can find many clusters that an algorithm such as K-means would not find. However, a weakness of DBSCAN is that it does not perform well on data that has many dimensions or clusters of varying densities.

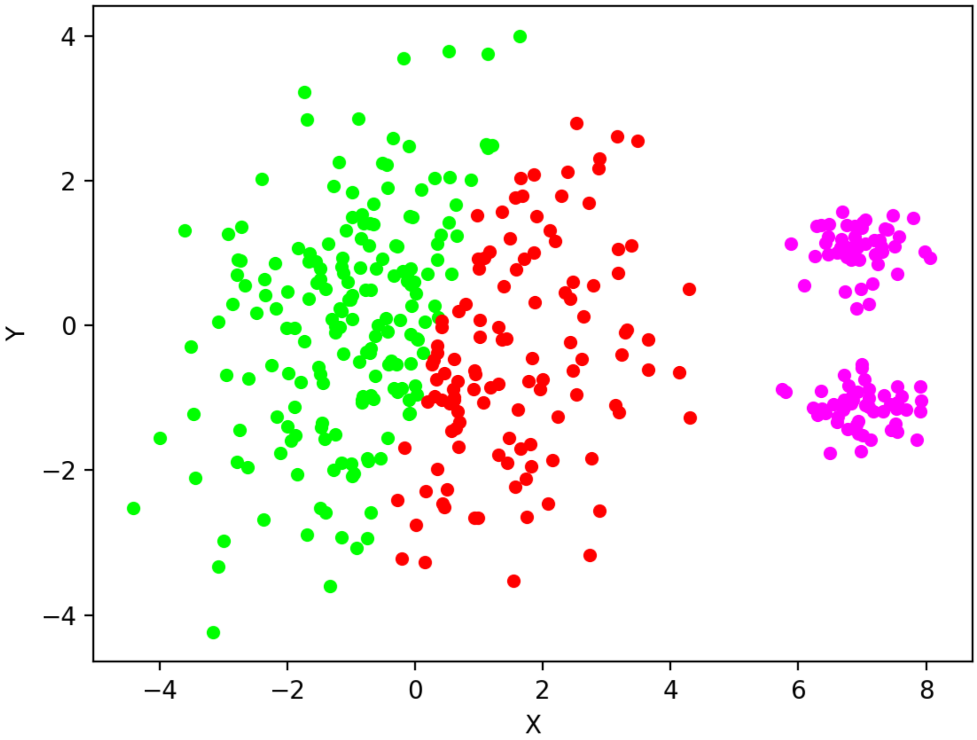
4 Problem 4

Part 1)
This was implemented in code.



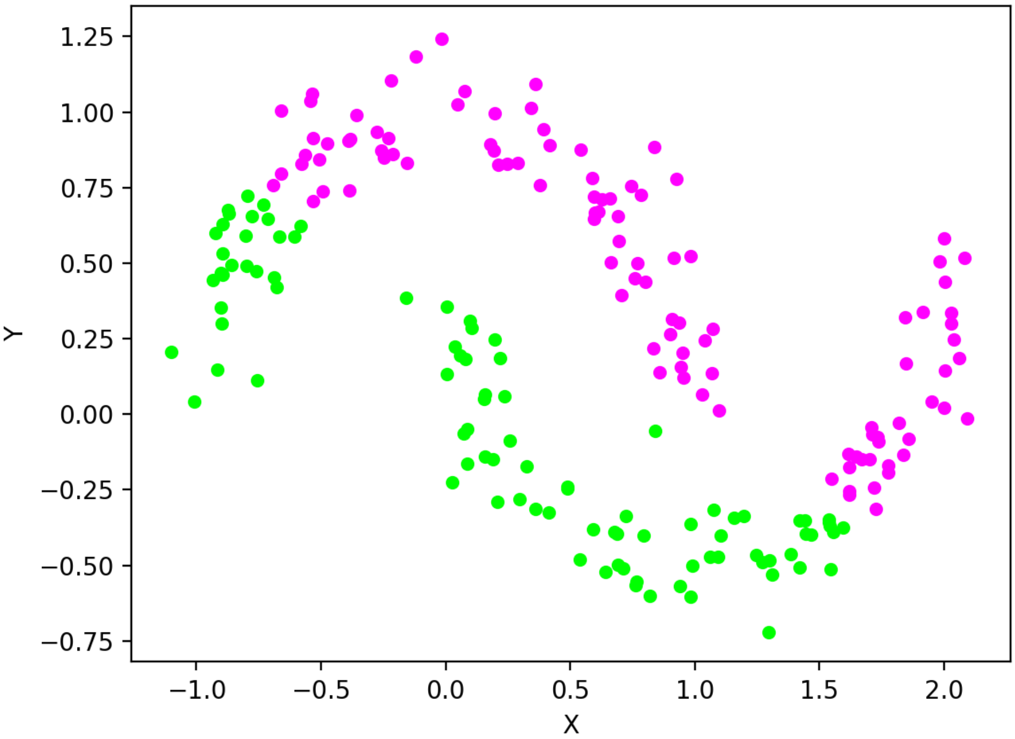
dataset 1:
Purity is :1.0
NMI :1.0

Both our purity and NMI are 1 indicating that our data is correctly clustered.



dataset2:
Purity is :0.875
NMI :0.552059635187

Our purity is 0.875 which is not unacceptably low and our NMI is 0.55. Thus, we can say that our clusters are relatively accurate but certainly not perfect.



dataset3:
Purity is :0.69
NMI :0.075947839504

Our purity is 0.69 which is relatively low, while our NMI is 0.076 which is very bad. Thus, we can infer that our data is very incorrectly clustered.

Part 3)
The strengths of GMM's are that they provide probabilistic cluster assignments and can handle clusters of various sizes and variances. Unlike K-means it can be used on non-spherical clusters. On the downside, the loss function for GMM is non-convex and thus optimization is not an overwhelmingly simple task. Furthermore, initialization can affect the algorithm and it can be prone to over-fitting