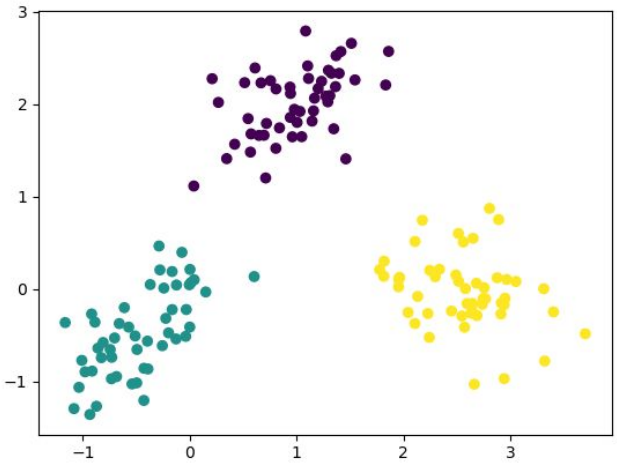
Earle Aguilar
804 501 476
11/10/2017

1. K-means
    a.

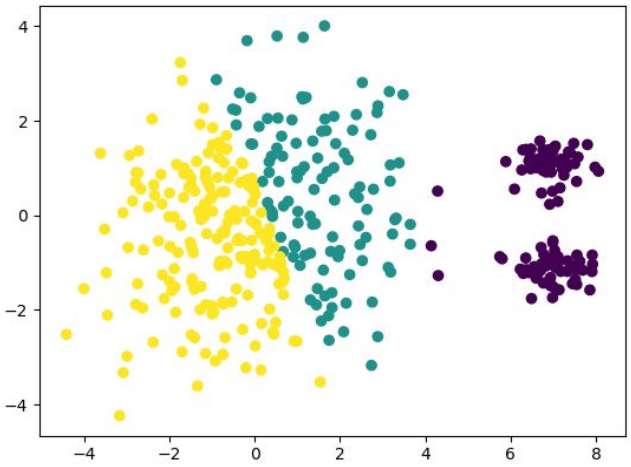
	Dataset 1	Dataset 2	Dataset 3
Purity	1.0	0.868	0.78
NMI	1.0	0.464	0.170

    b. Strengths
        i. Efficient
    c. Weaknesses
        i. Need to specify K in advance.
        ii. Sensitive to noisy data and outliers.
        iii. Can terminate at a local optima.

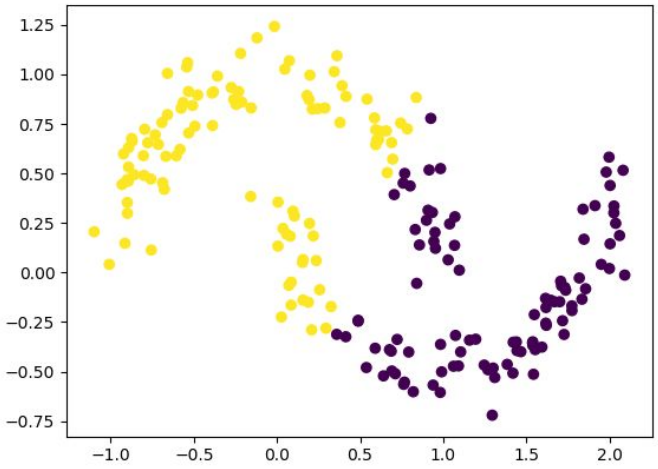


    d.

e.



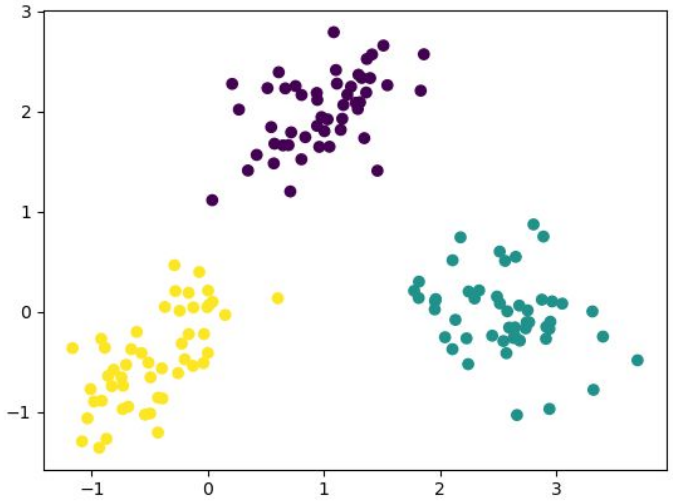
f.



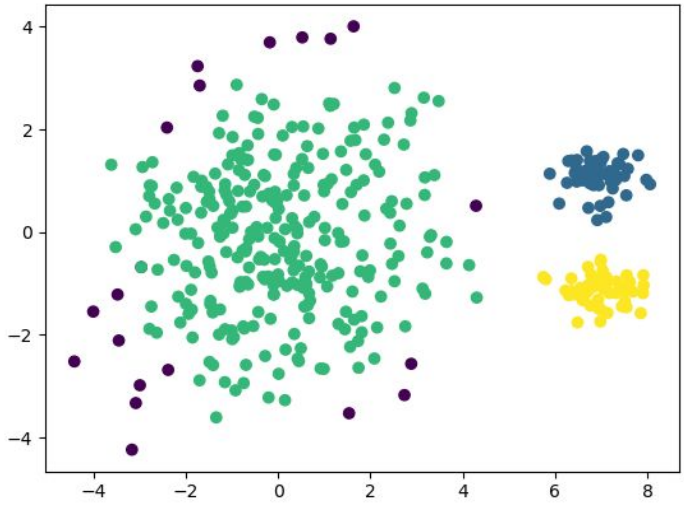
2. DBSCAN
   a.

	Dataset 1	Dataset 2	Dataset 3
Purity	1.0	0.953	1.0
NMI	1.0	1.0	1.0

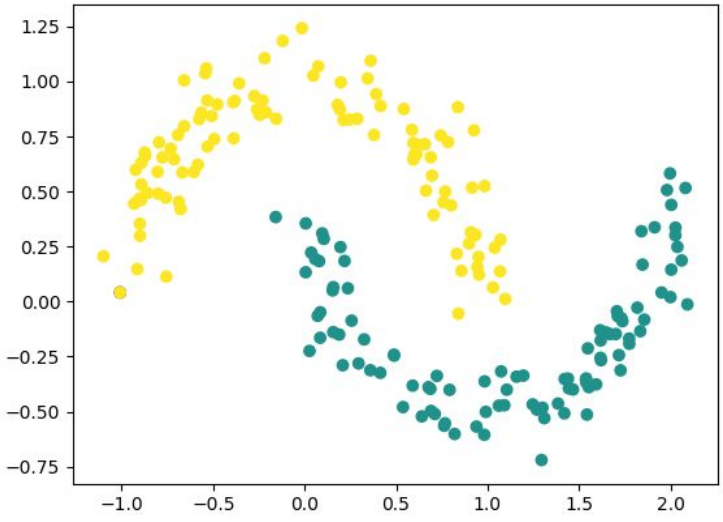
   b. Strengths
      i. Robust to noise and outliers.
      ii. Can handle clusters of different shapes and sizes.
   c. Weaknesses
      i. Sensitive to parameters.



d.



e.

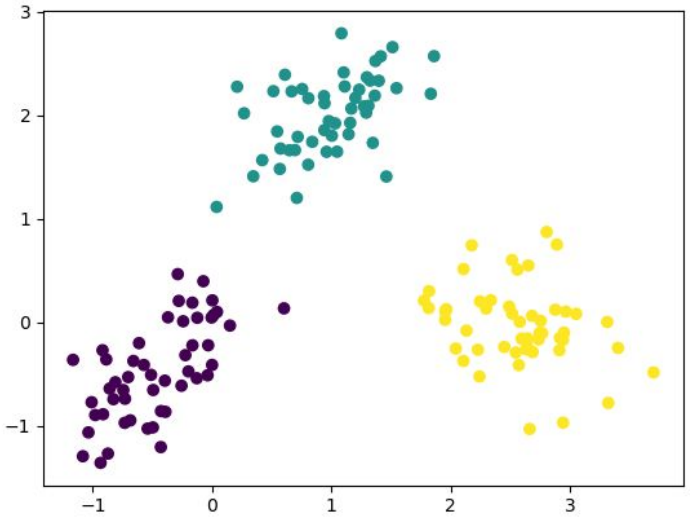


f.

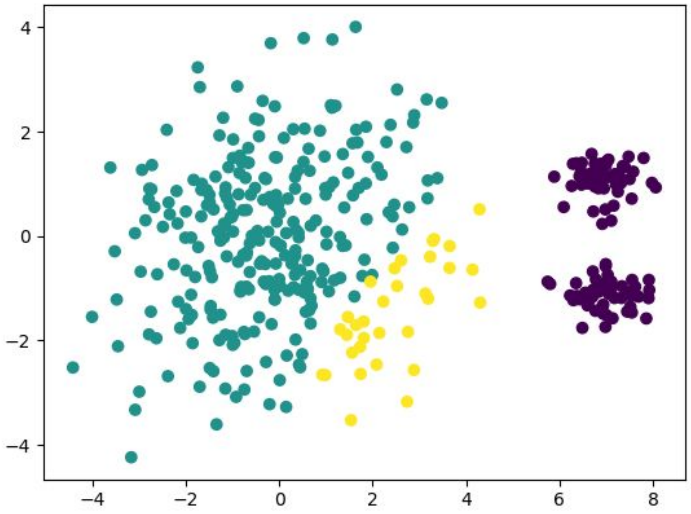
3. GMM
   a.

	Dataset 1	Dataset 2	Dataset 3
Purity	1.0	0.875	0.69
NMI	1.0	0.608	0.076

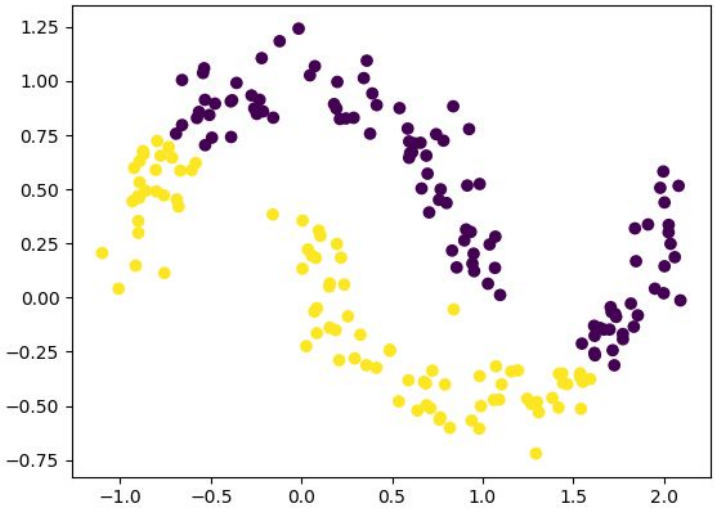
   b. Strengths
      i. Mixture models are more general than partitioning. Can have different densities and size of clusters
      ii. Clusters may be characterized by a small number of parameters.
   c. Weaknesses
      i. Converges to local optima.
      ii. Computationally expensive for large distributions
      iii. Hard to estimate number of clusters
      iv. Can only deal with spherical clusters.



d.



e.



f.

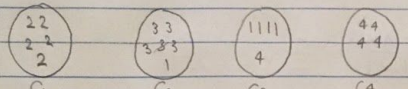
4. **Problem 1**

a.

Actual \ Predicted	C_1	C_2	C_3	C_4
C_1	0	1	4	0
C_2	5	0	0	0
C_3	0	5	0	0
C_4	0	0	1	4

$$\text{Purity} = \frac{1}{20}(4+5+5+4)$$

$$= \frac{18}{20} = 0.9$$



$$TP+FP = \binom{5}{2}+\binom{6}{2}+\binom{5}{2}+\binom{4}{2}$$

$$TP+FP = 41$$

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

$$\boxed{TP = 32}$$

$$\boxed{FP = 9}$$

$$\#\text{pairs} = 190$$

$$\text{Total Neg} = 190 - 41 = 149$$

	Same Cluster	Different Cluster
Same Class	32	8
Differen Class	9	141

$\boxed{FN = 8}$ and $TN = 149 - 8 = 141$

$\boxed{TN = 141}$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{32}{41}$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{32}{32+8}$$

$\boxed{\text{Precision} = 0.78}$

$\boxed{\text{Recall} = 0.8}$

$$\boxed{F = \frac{2(0.78)(0.8)}{0.8+0.78} = 0.79}$$

b.

	C1	C2	C3	C4	sum
1	0	1	4	0	5
2	5	0	0	0	5
3	0	5	0	0	5
4	0	0	1	4	5
sum	5	6	5	4	20

$$I(\Omega,C) = \sum_{k=1}^{4}\sum_{j=1}^{4} \frac{|c_k \cap w_j|}{N} \log\left(\frac{N|c_k \cap w_j|}{|c_k||w_j|}\right) =$$

$$= \frac{5}{20}\log(4) + \frac{1}{20}\log\left(\frac{2}{3}\right) + \frac{5}{20}\log\left(\frac{0}{3}\right) + \frac{4}{20}\log\left(\frac{0}{5}\right) + \frac{1}{20}\log\left(\frac{4}{5}\right) + \frac{4}{20}\log(4)$$

$$I(\Omega,C) = 1.623$$

$$H(\Omega) = 2 \quad , H(C) = 1.985$$

$$NMI(\Omega,C) = \frac{1.623}{\sqrt{2 \times 1.985}} = 0.8145$$