CS 545: Machine Learning, Spring 2020

Name: Sree Vandana Nadipalli

Algorithm #1: K-Means clustering

In this experiment, standard version of k-means algorithm is implemented. The initial centers(centroids) of 'k' clusters are randomly chosen from the dataset.

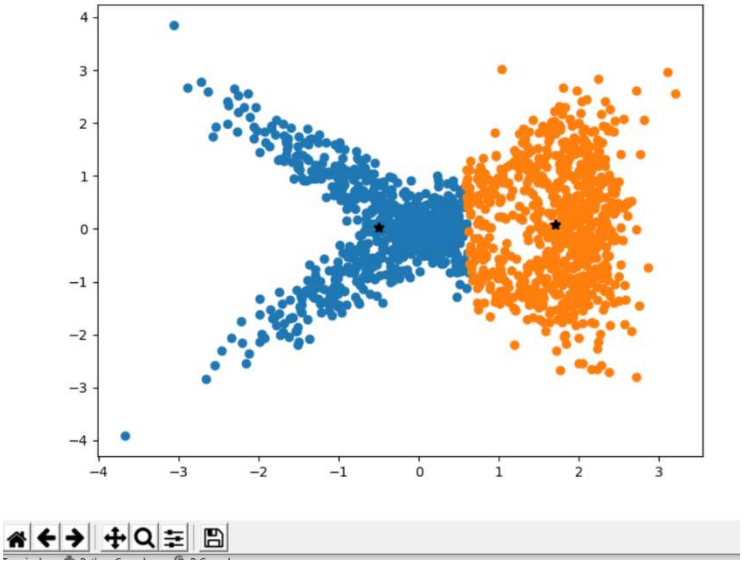
For a fixed 'k' value, This algorithm runs for 'r' number of times. For each of this 'r' we randomly chosen 'k' number of points (initial cluster centers). The sum-square-error is calculated for each 'r' and finally we select one model that gives minimum sum-square-error, for that 'k'. (here r value is taken as 10)

After selecting initial centroids, the distance between each data point and the 'k' centroids is calculated. And the data point is placed in that cluster which has minimum distance from the centroid. After placing all data points in respective clusters, the centroid of each cluster is re-calculated, and centroid values are updated. And then again, we repeat the process of adding data points to respective clusters (based on distance between the data point and the cluster centroid). We repeat this process until the centroids converge or till it reaches maximum number of iterations.

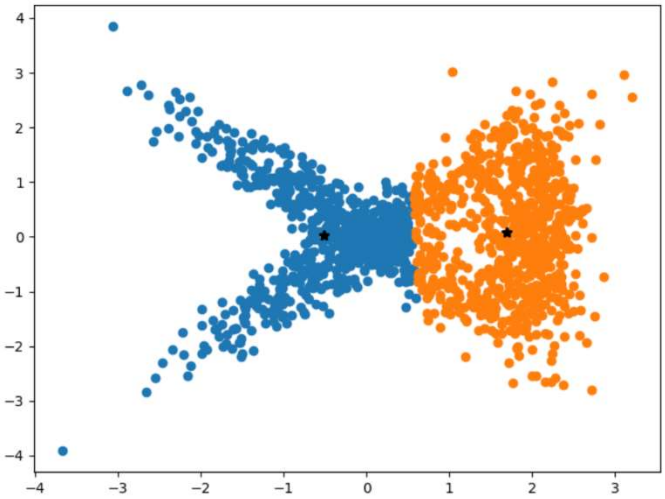
Observations:

1. **Number of Clusters (k) = 2:**

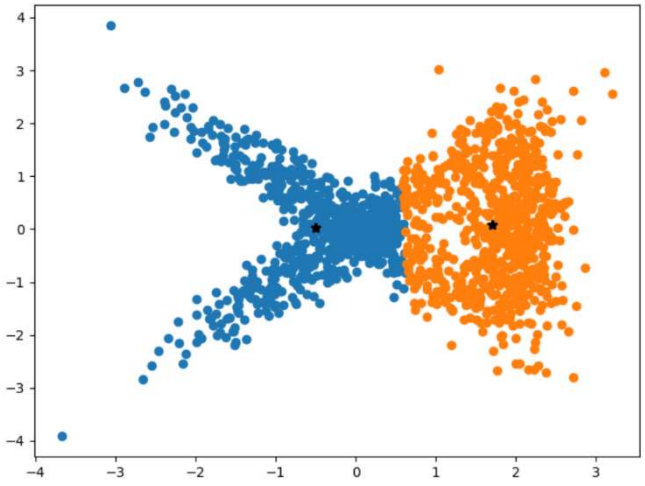
r = 1



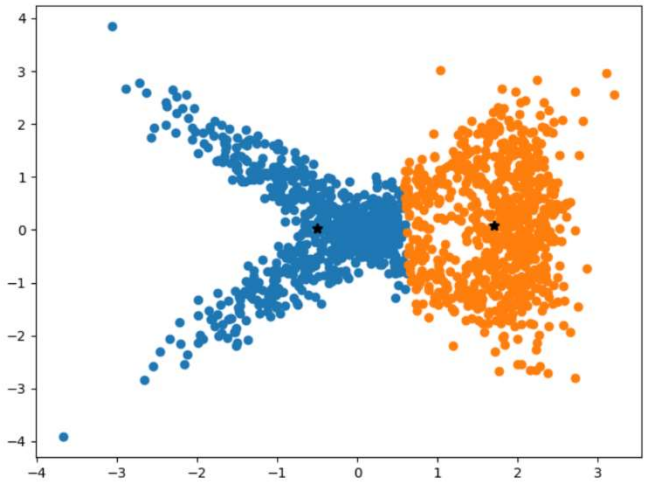
r = 2



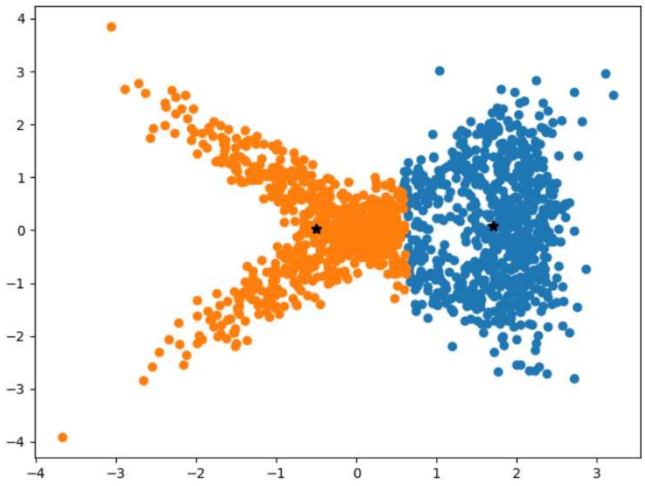
r = 3



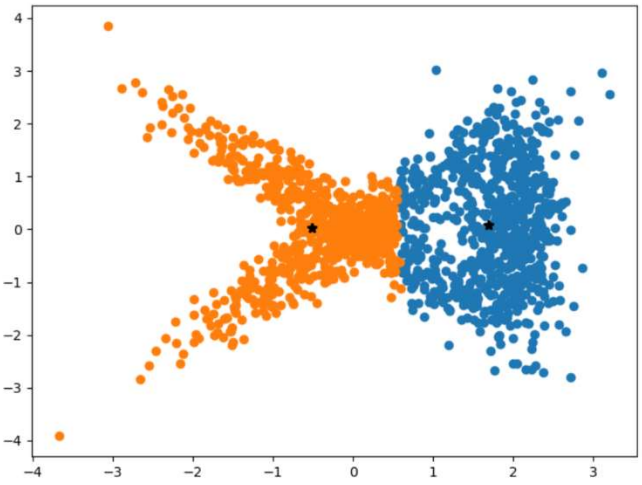
r = 4



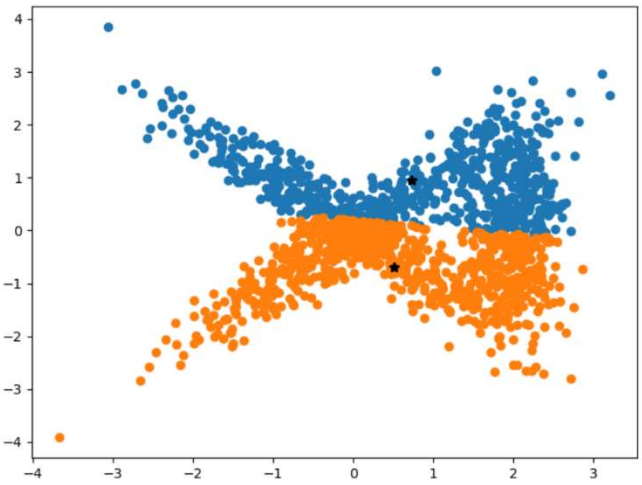
r = 5



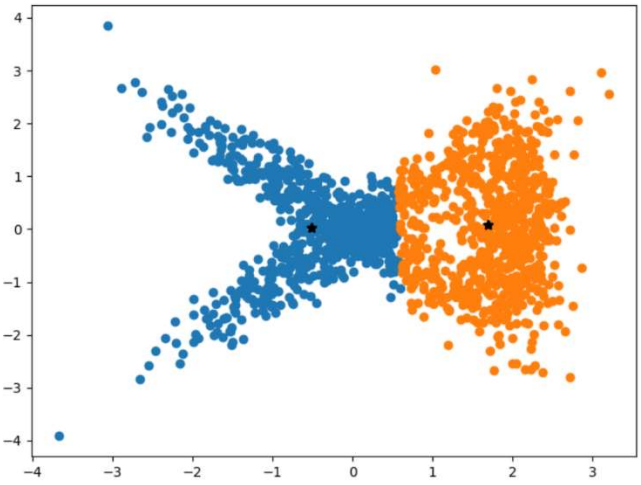
r = 6



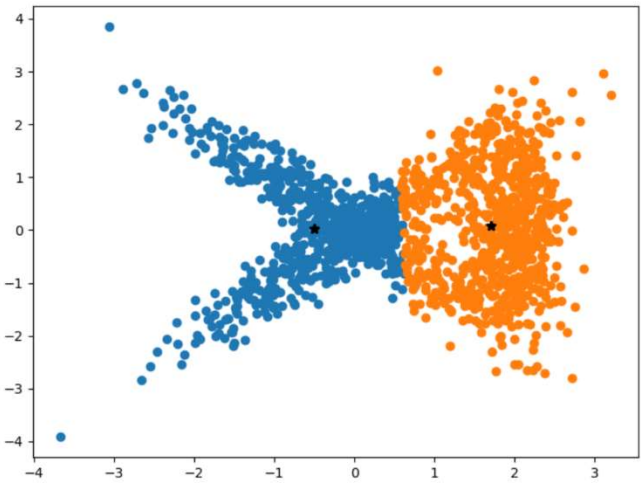
r = 7



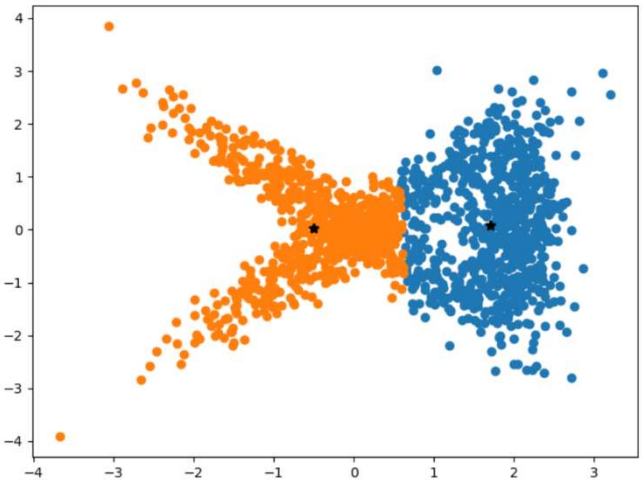
r = 8



r = 9



r = 10



Below is the pic showing sum square error at each ‘r’ with different initial centroids

```
C:\Users\SreeV\anaconda3\envs\tensor\pythonw.exe "C:/(
k =  2
centroids converged, breaking the loop
r =  1   sum-square-error =  2168.2788438938423
centroids converged, breaking the loop
r =  2   sum-square-error =  2171.0019696814525
centroids converged, breaking the loop
r =  3   sum-square-error =  2168.2788438938423
centroids converged, breaking the loop
r =  4   sum-square-error =  2168.2788438938423
centroids converged, breaking the loop
r =  5   sum-square-error =  2168.2788438938355
centroids converged, breaking the loop
r =  6   sum-square-error =  2171.001969681462
centroids converged, breaking the loop
r =  7   sum-square-error =  2811.3702170236525
centroids converged, breaking the loop
r =  8   sum-square-error =  2171.0019696814525
centroids converged, breaking the loop
r =  9   sum-square-error =  2168.2788438938423
centroids converged, breaking the loop
r =  10   sum-square-error =  2168.2788438938355
errors =  [2168.2788438938423, 2171.0019696814525, 21(
min error when r = 5
error =  2168.2788438938355
```

Sum-square-error at each ‘r’

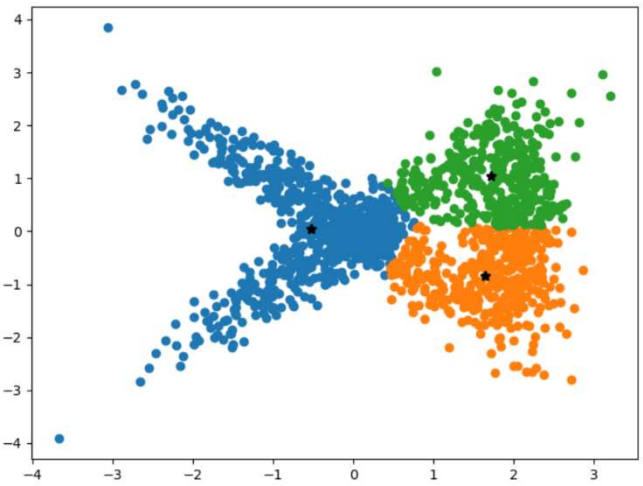
errors = [2168.2788438938423, 2171.0019696814525, 2168.2788438938423, 2168.2788438938423, 2168.2788438938355, 2171.001969681462, 2811.3702170236525, 2171.0019696814525, 2168.2788438938423, 2168.2788438938355]

min error when r = 5

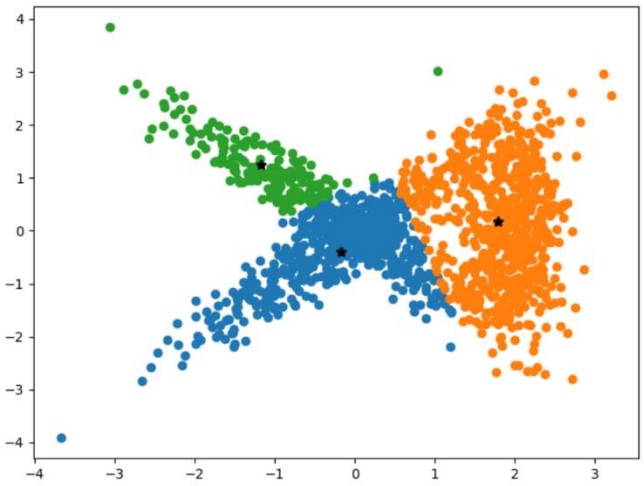
error = 2168.278

2. Number of Clusters K = 3

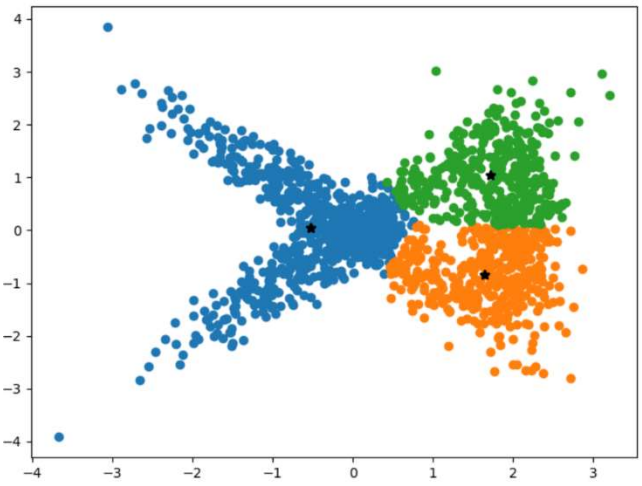
r = 1



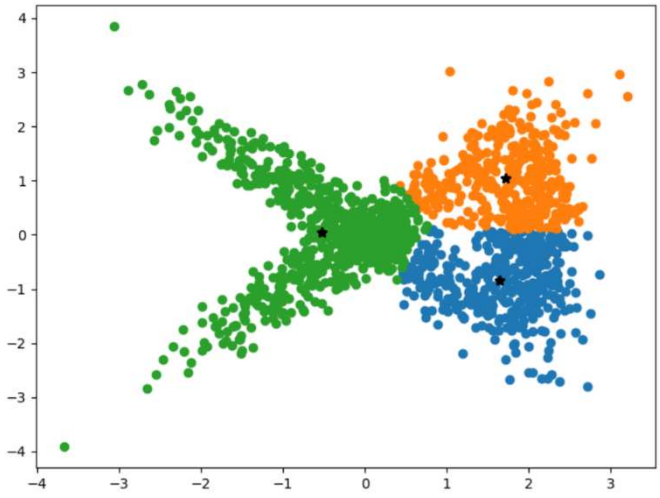
r = 2



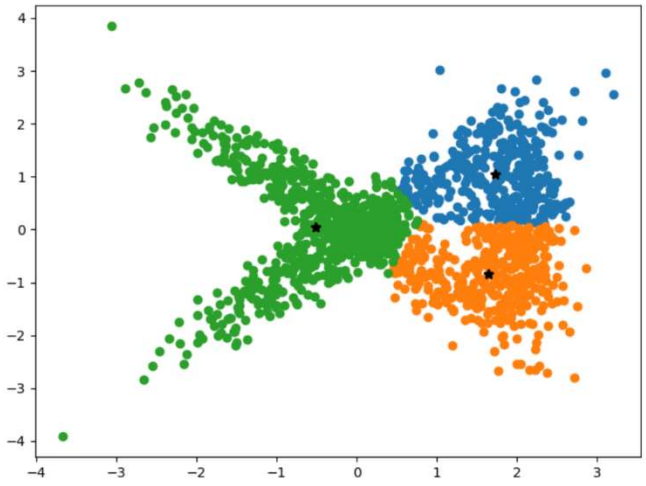
r = 3



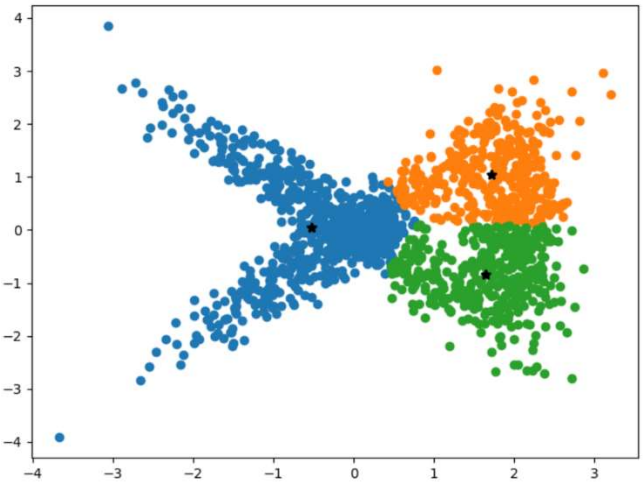
r = 4



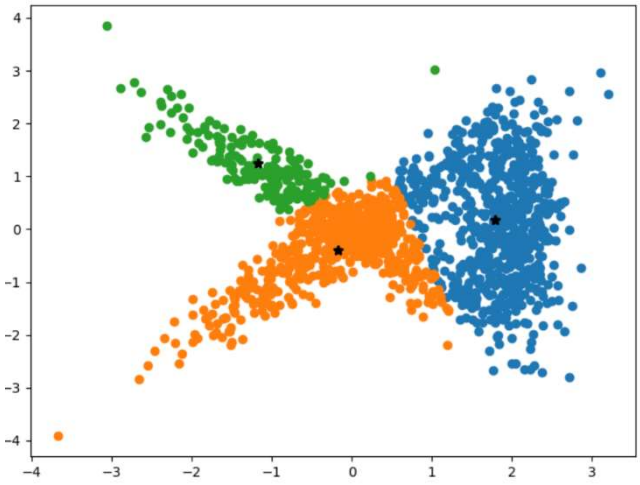
r = 5



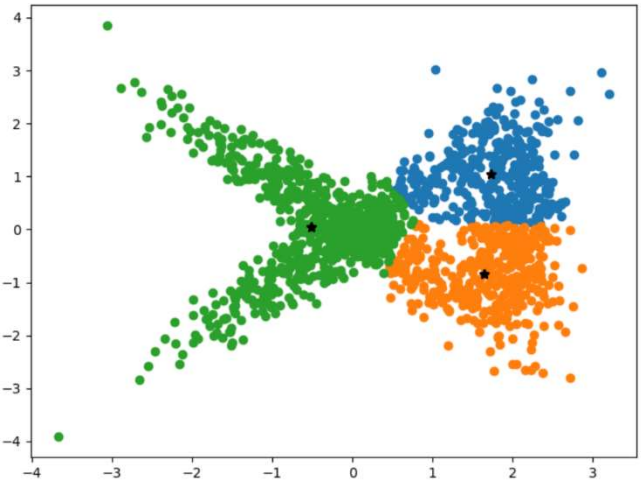
r = 6



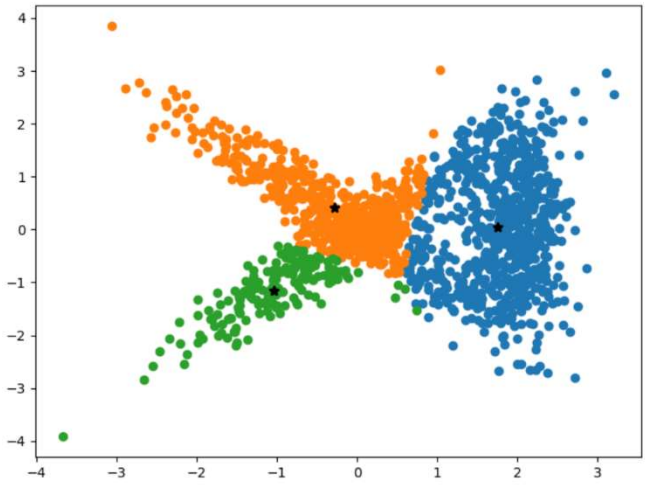
r = 7



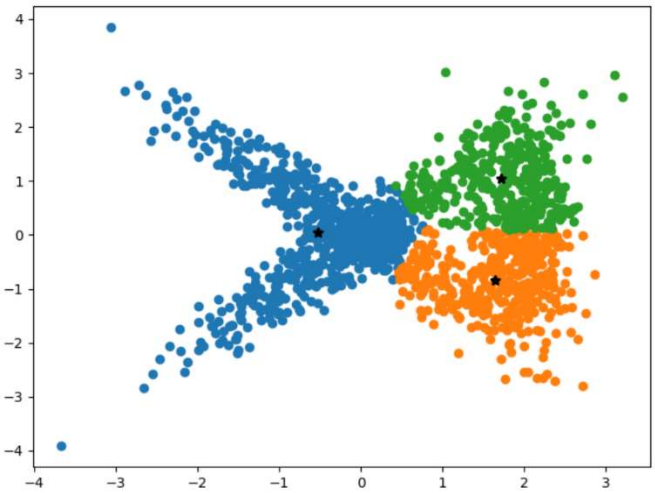
r =8



r = 9



r = 10



```
C:\Users\SreeV\anaconda3\envs\tensor\pythonw.exe "C:/U
k =  3
centroids converged, breaking the loop
r =  1   sum-square-error =  1478.4278229881143
centroids converged, breaking the loop
r =  2   sum-square-error =  1912.939497722299
centroids converged, breaking the loop
r =  3   sum-square-error =  1478.4278229881143
centroids converged, breaking the loop
r =  4   sum-square-error =  1478.427822988117
centroids converged, breaking the loop
r =  5   sum-square-error =  1479.5970650234522
centroids converged, breaking the loop
r =  6   sum-square-error =  1478.4278229881152
centroids converged, breaking the loop
r =  7   sum-square-error =  1912.9394977223055
centroids converged, breaking the loop
r =  8   sum-square-error =  1479.5970650234522
centroids converged, breaking the loop
r =  9   sum-square-error =  1603.7708155559499
centroids converged, breaking the loop
r =  10   sum-square-error =  1478.4278229881143
errors =  [1478.4278229881143, 1912.939497722299, 1478
min error when r = 1
error =  1478.4278229881143
```

Sum-square-error at each 'r'

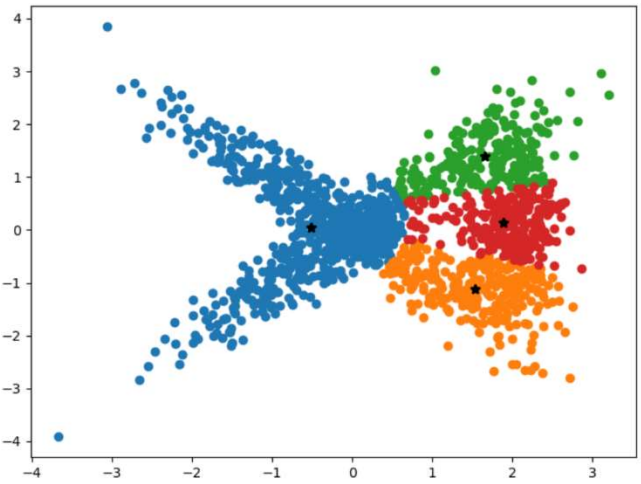
errors = [1478.4278229881143, 1912.939497722299, 1478.4278229881143, 1478.427822988117, 1479.5970650234522, 1478.4278229881152, 1912.9394977223055, 1479.5970650234522, 1603.7708155559499, 1478.4278229881143]

min error when r = 1

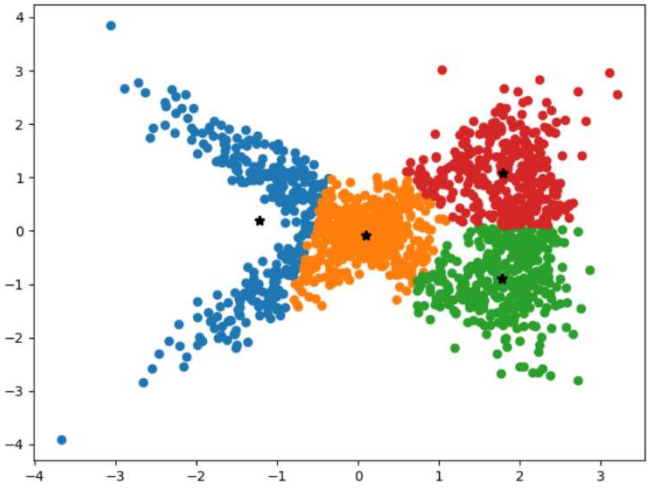
error = 1478.427

3. Number of clusters K = 4

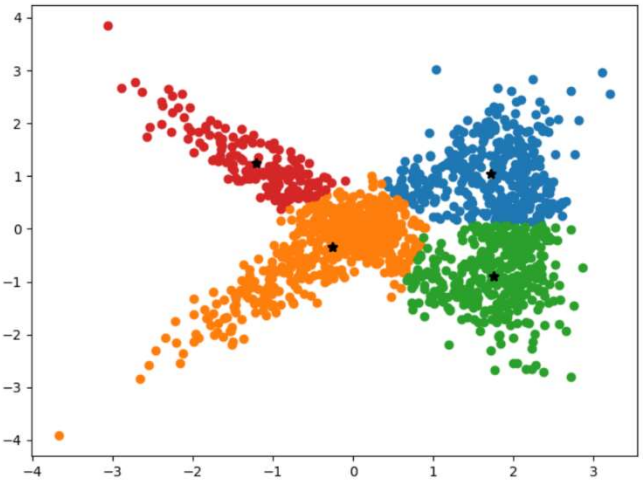
r = 1



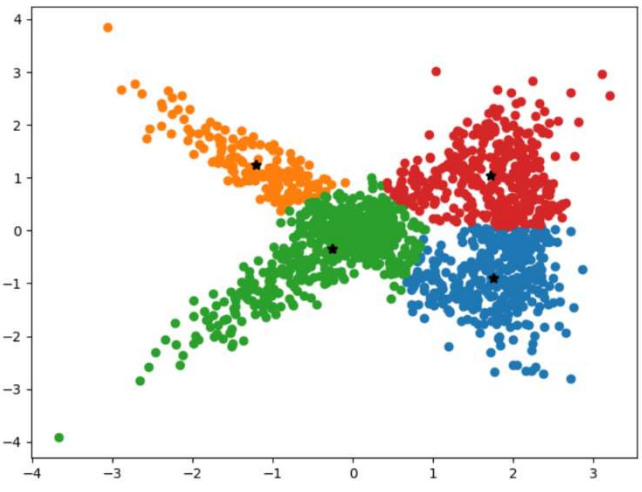
r = 2



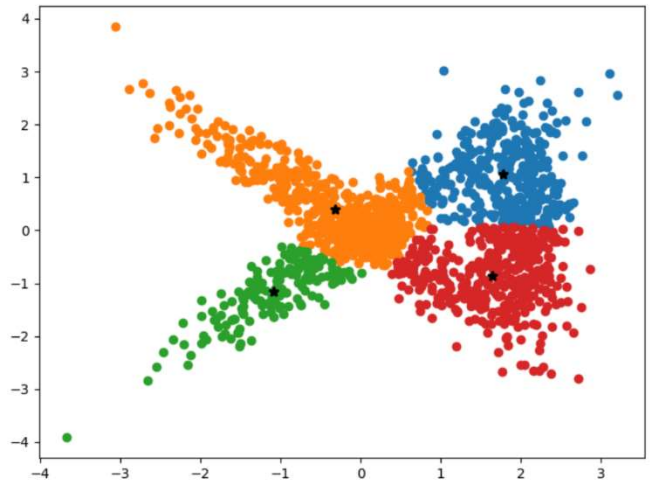
r = 3



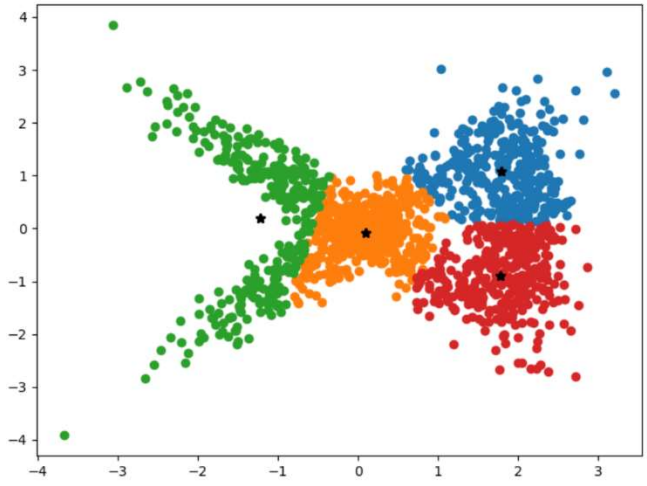
r =4



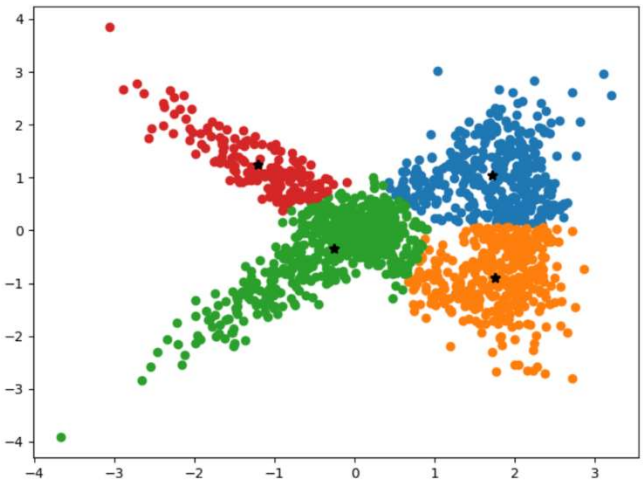
r = 5



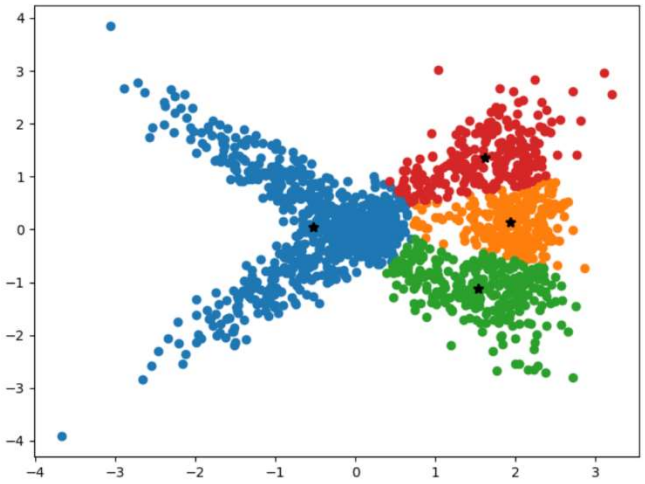
r = 6



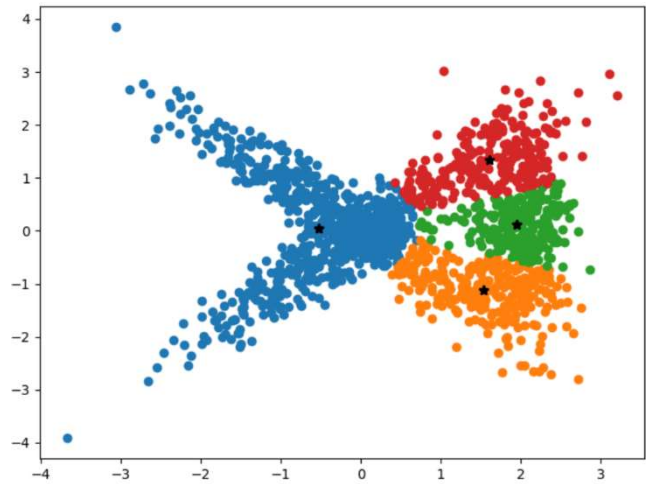
r = 7



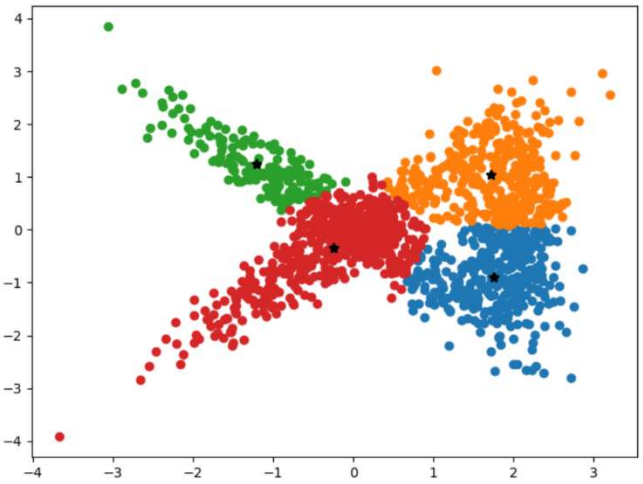
r = 8



r = 9



r = 10



```
C:\Users\SreeV\anaconda3\envs\tensor\pythonw.exe "C:/U
k =  4
centroids converged, breaking the loop
r =  1   sum-square-error =  1312.6298925288215
centroids converged, breaking the loop
r =  2   sum-square-error =  1295.4239593113969
centroids converged, breaking the loop
r =  3   sum-square-error =  1420.470317927069
centroids converged, breaking the loop
r =  4   sum-square-error =  1420.470317927068
centroids converged, breaking the loop
r =  5   sum-square-error =  810.7170368991299
centroids converged, breaking the loop
r =  6   sum-square-error =  1295.423959311399
centroids converged, breaking the loop
r =  7   sum-square-error =  1420.470317927065
centroids converged, breaking the loop
r =  8   sum-square-error =  1325.900823091777
centroids converged, breaking the loop
r =  9   sum-square-error =  1333.5799736968975
centroids converged, breaking the loop
r =  10   sum-square-error =  1422.1776450881312
errors =  [1312.6298925288215, 1295.4239593113969, 142
min error when r = 5
error =  810.7170368991299
```

Sum-square-error at each 'r'

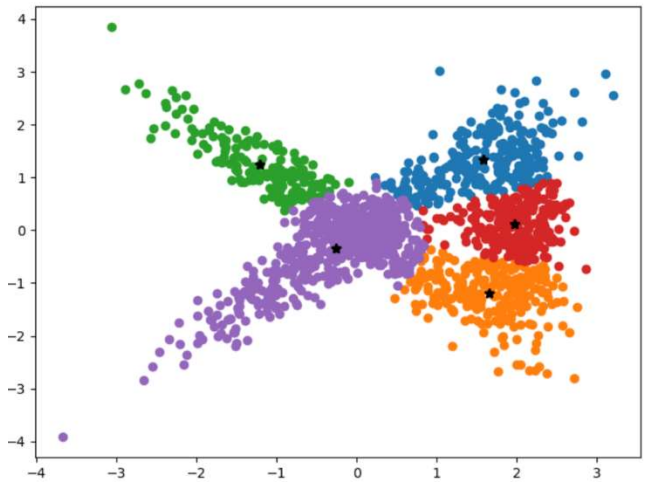
errors = [1312.6298925288215, 1295.4239593113969, 1420.470317927069, 1420.470317927068, 810.7170368991299, 1295.423959311399, 1420.470317927065, 1325.900823091777, 1333.5799736968975, 1422.1776450881312]

min error when r = 5

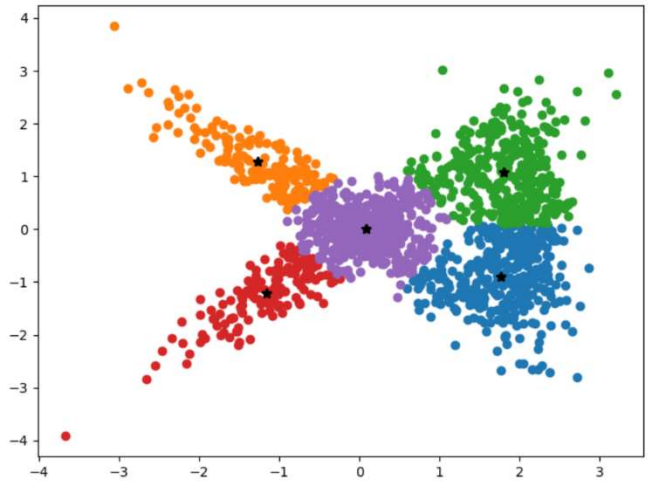
error = 810.717

4. Number of Clusters k = 5

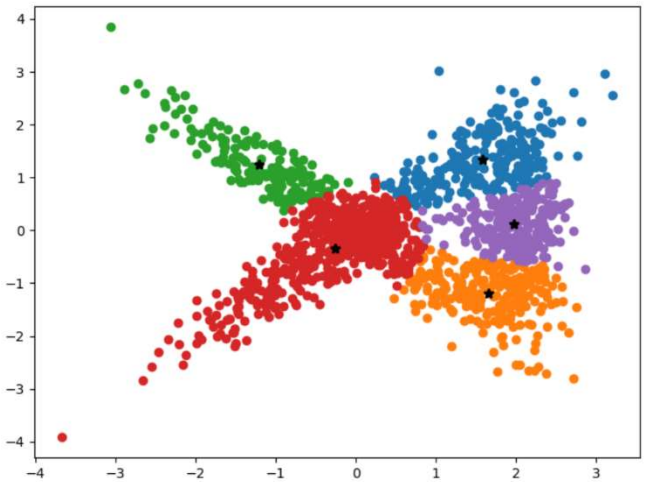
r = 1



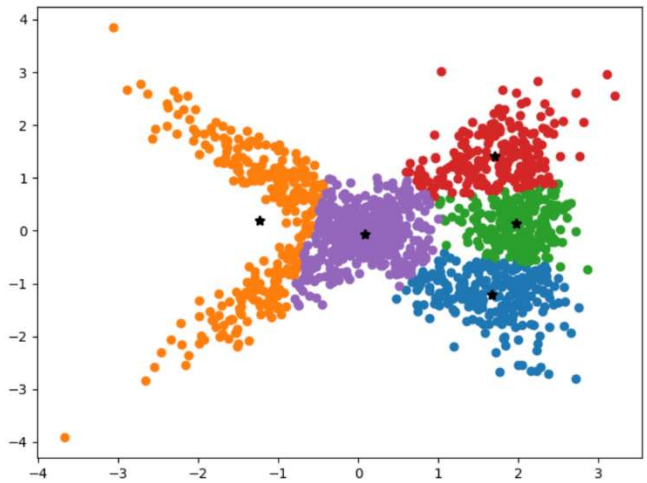
r = 2



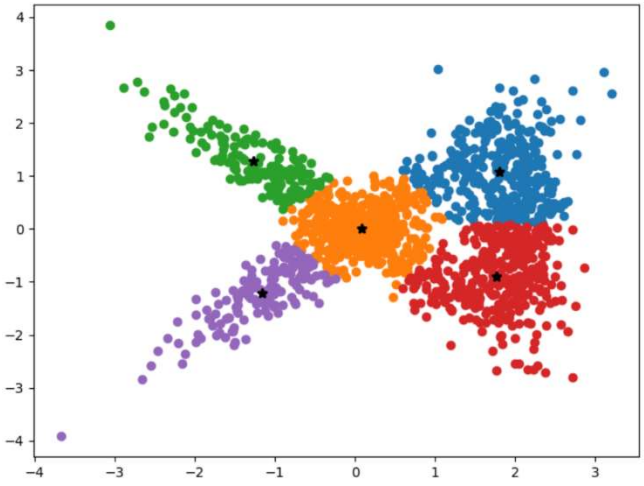
r = 3



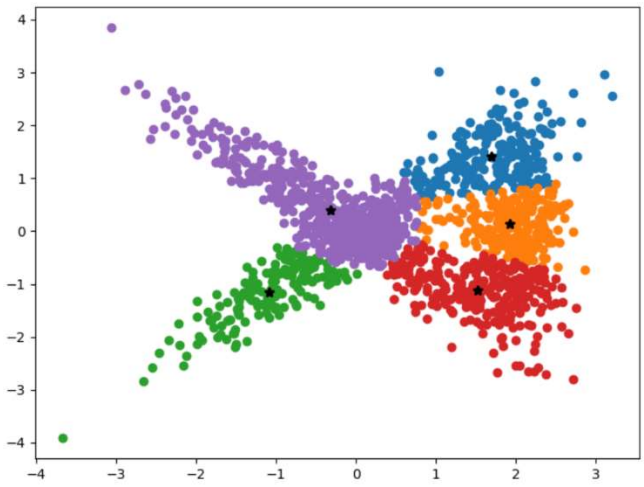
r = 4



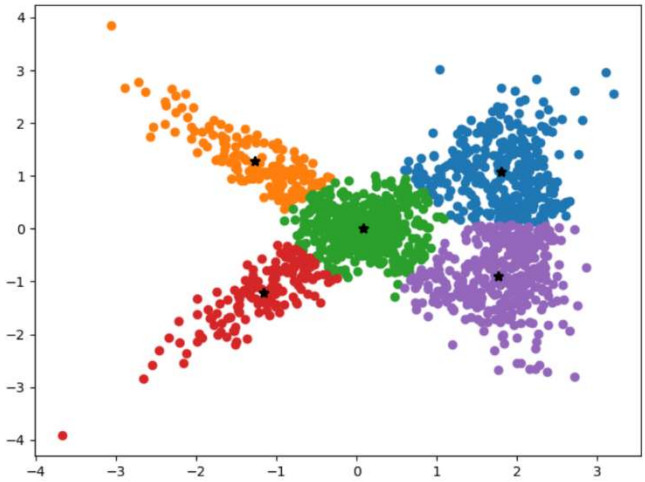
r = 5



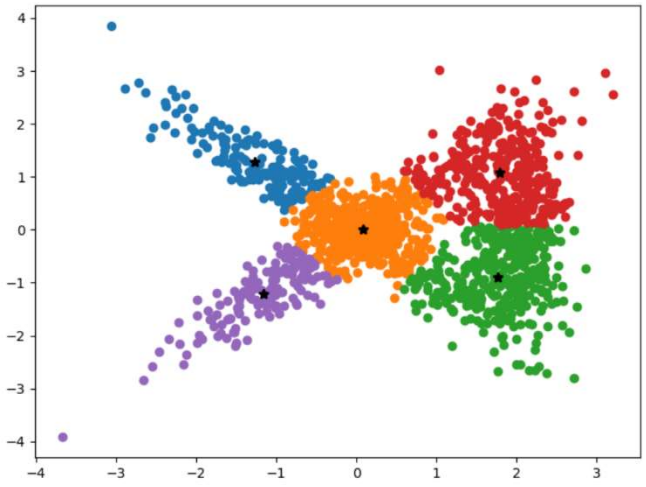
r = 6



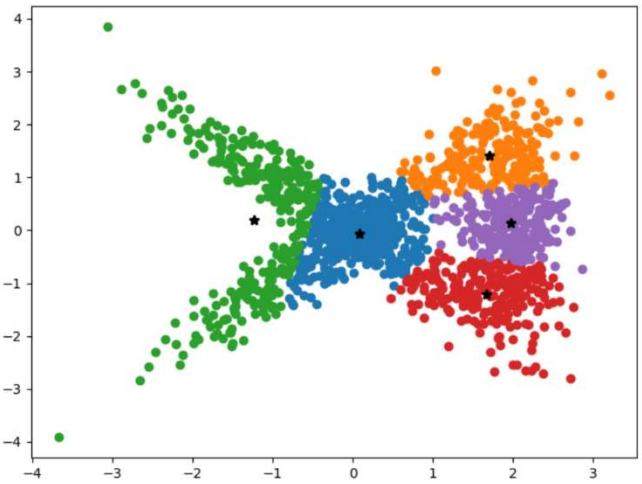
r = 7



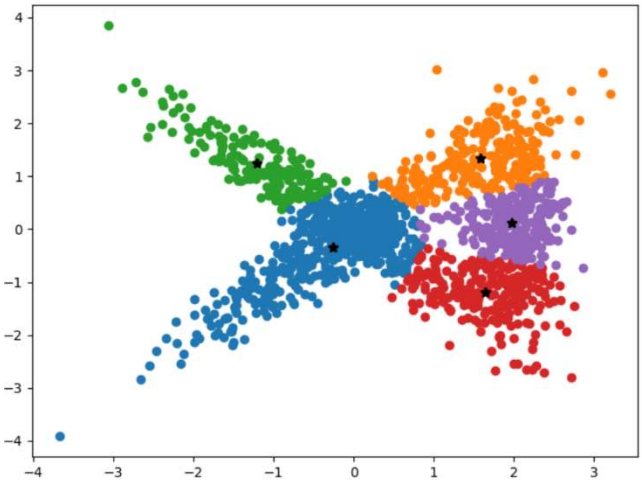
r = 8



r = 9



r = 10



```
C:\Users\SreeV\anaconda3\envs\tensor\pythonw.exe "(
k =  5
centroids converged, breaking the loop
r =  1   sum-square-error =  1289.7559492275616
centroids converged, breaking the loop
r =  2   sum-square-error =  780.2356428010877
centroids converged, breaking the loop
r =  3   sum-square-error =  1289.7559492275618
centroids converged, breaking the loop
r =  4   sum-square-error =  1147.5332491157528
centroids converged, breaking the loop
r =  5   sum-square-error =  780.2356428010888
centroids converged, breaking the loop
r =  6   sum-square-error =  638.2989423487171
centroids converged, breaking the loop
r =  7   sum-square-error =  780.2356428010888
centroids converged, breaking the loop
r =  8   sum-square-error =  779.9152903382292
centroids converged, breaking the loop
r =  9   sum-square-error =  1147.5332491157515
centroids converged, breaking the loop
r =  10   sum-square-error =  1288.4819711122955
errors =  [1289.7559492275616, 780.2356428010877, 1
min error when r = 6
error =  638.2989423487171
```

Sum-square-error at each 'r'

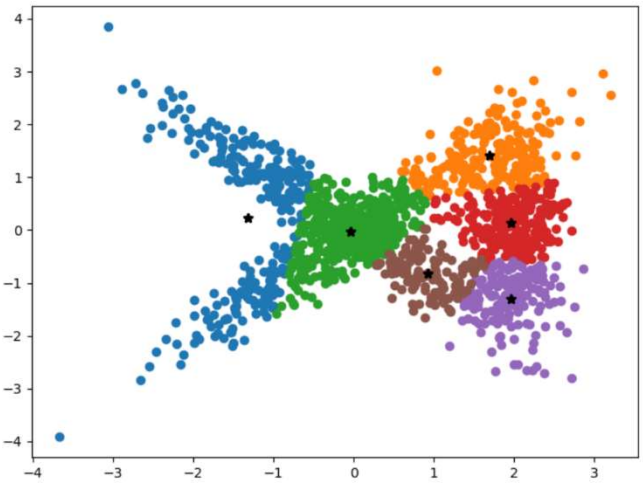
errors = [1289.7559492275616, 780.2356428010877, 1289.7559492275618, 1147.5332491157528, 780.2356428010888, 638.2989423487171, 780.2356428010888, 779.9152903382292, 1147.5332491157515, 1288.4819711122955]

min error when r = 6

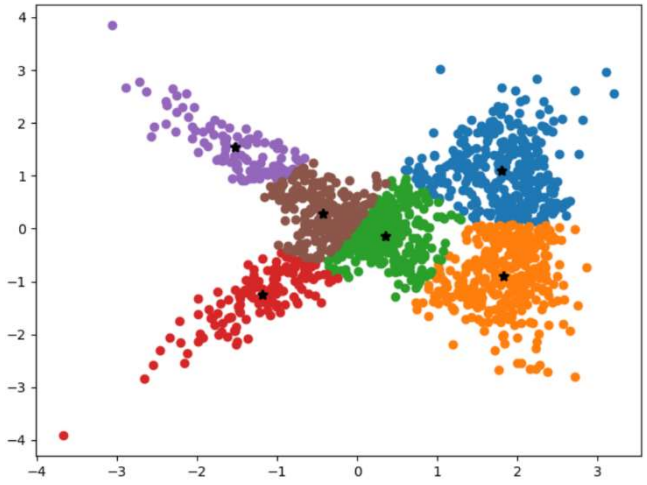
error = 638.298

5. Number of Cluster k = 6

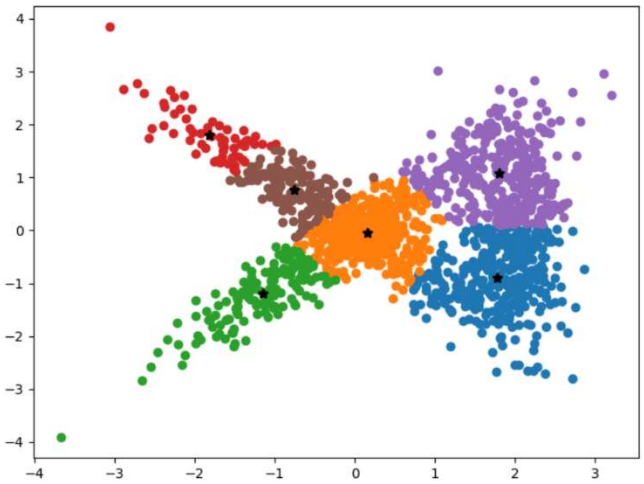
r = 1



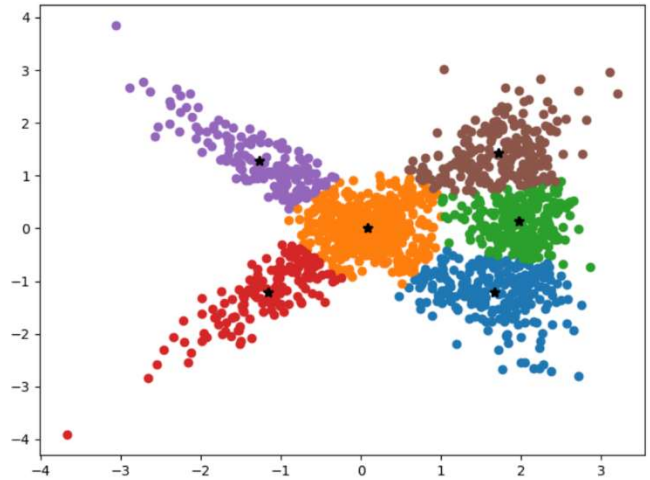
r = 2



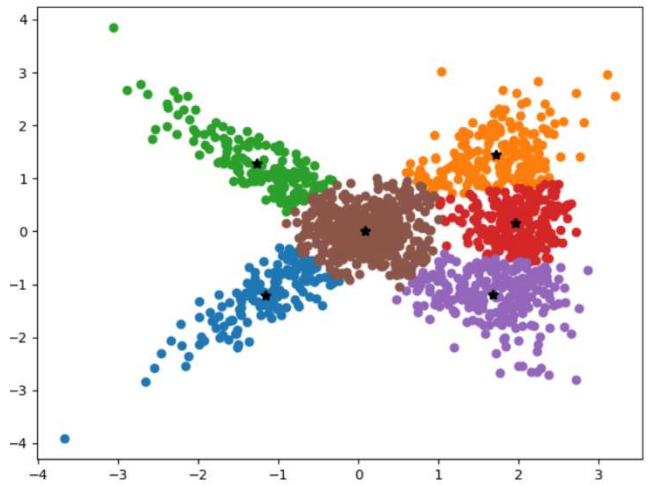
r = 3



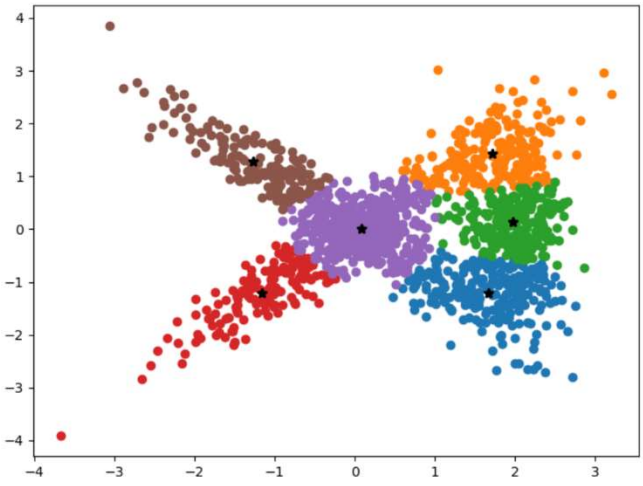
r = 4



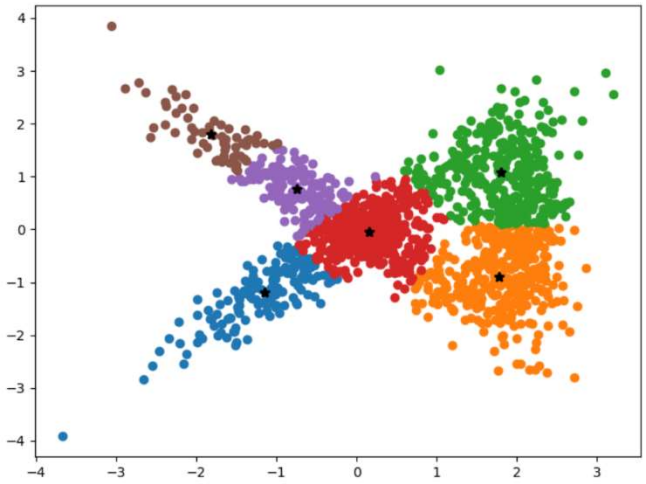
r = 5



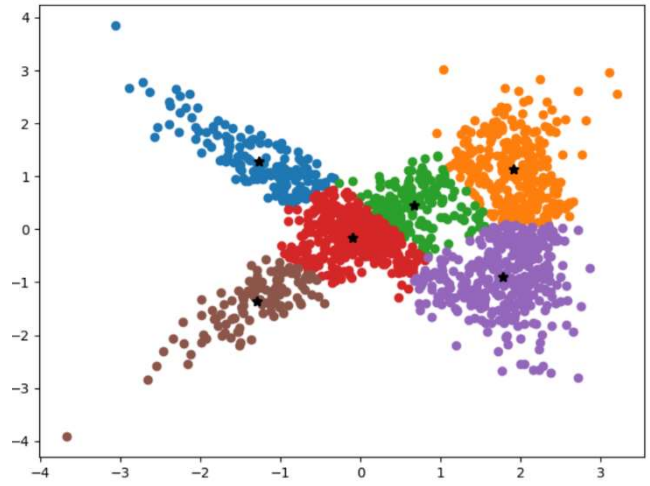
r = 6



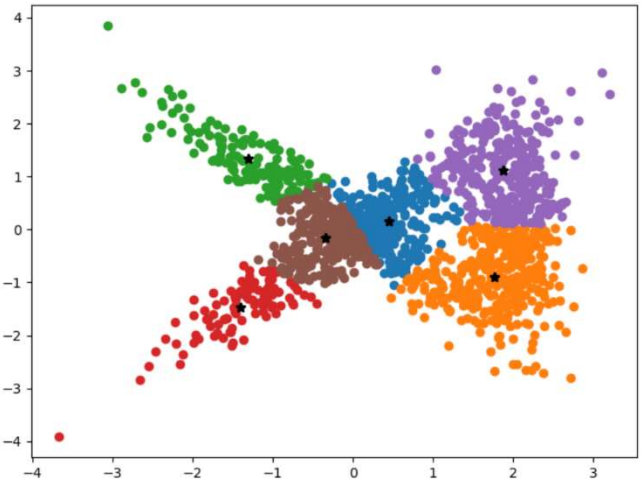
r = 7



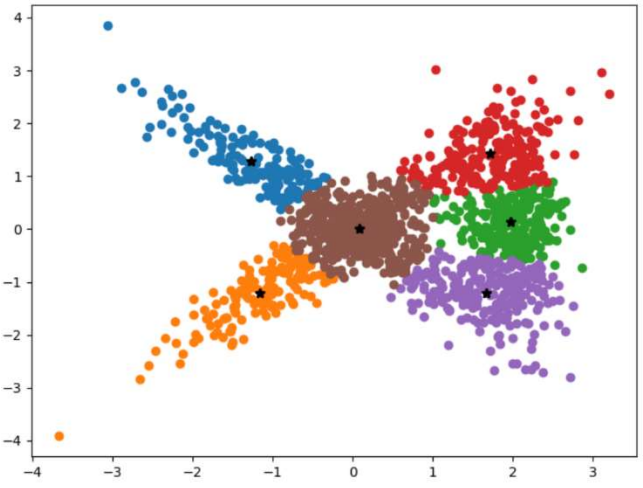
r = 8



r = 9



r = 10



```
C:\Users\SreeV\anaconda3\envs\tensor\pythonw.exe "C:/
k =  6
centroids converged, breaking the loop
r =  1   sum-square-error =  1136.0792369997837
centroids converged, breaking the loop
r =  2   sum-square-error =  747.5879388754727
centroids converged, breaking the loop
r =  3   sum-square-error =  774.854901155855
centroids converged, breaking the loop
r =  4   sum-square-error =  627.8969606785537
centroids converged, breaking the loop
r =  5   sum-square-error =  634.9732787162732
centroids converged, breaking the loop
r =  6   sum-square-error =  627.8969606785549
centroids converged, breaking the loop
r =  7   sum-square-error =  774.8738001854558
centroids converged, breaking the loop
r =  8   sum-square-error =  643.696857472202
centroids converged, breaking the loop
r =  9   sum-square-error =  672.842896386114
centroids converged, breaking the loop
r =  10   sum-square-error =  627.8969606785549
errors =  [1136.0792369997837, 747.5879388754727, 774
min error when r = 4
error =  627.8969606785537
```

Sum-square-error at each 'r'

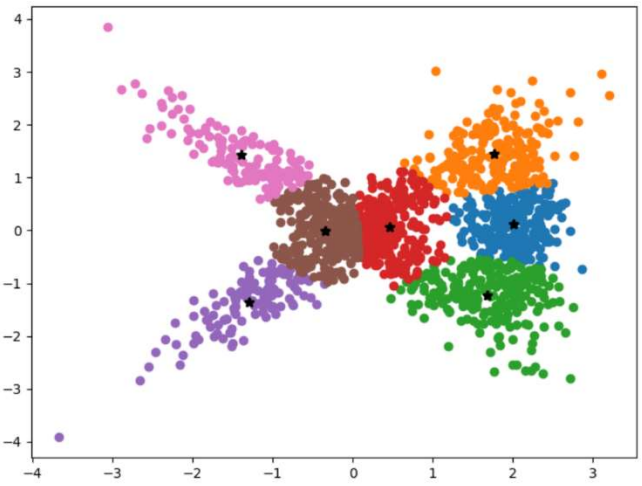
errors = [1136.0792369997837, 747.5879388754727, 774.854901155855, 627.8969606785537, 634.9732787162732, 627.8969606785549, 774.8738001854558, 643.696857472202, 672.842896386114, 627.8969606785549]

min error when r = 4

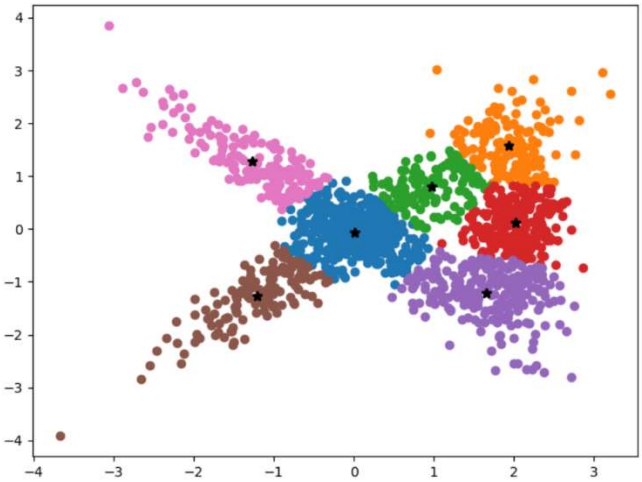
error = 627.896

6. Number of clusters k = 7

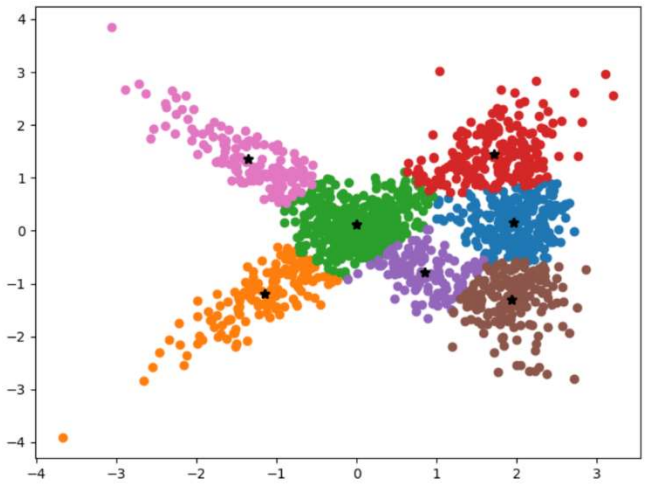
r = 1



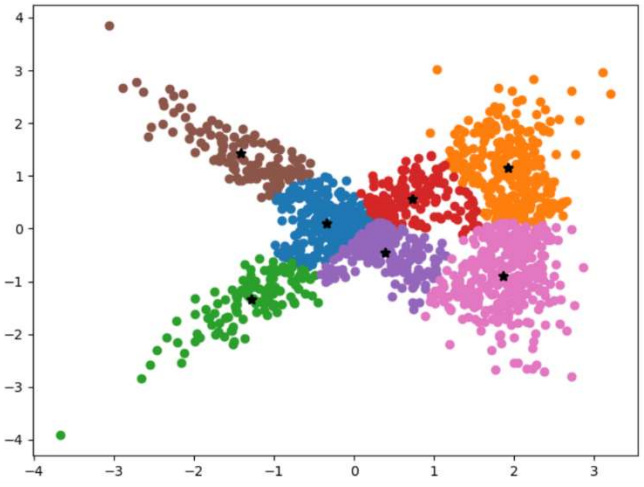
r = 2



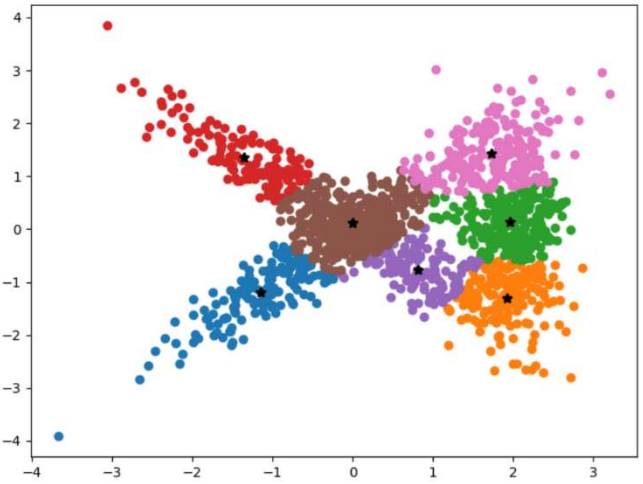
r = 3



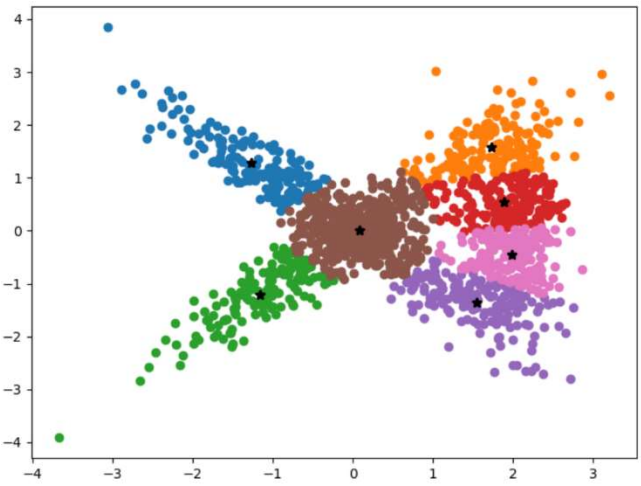
r = 4



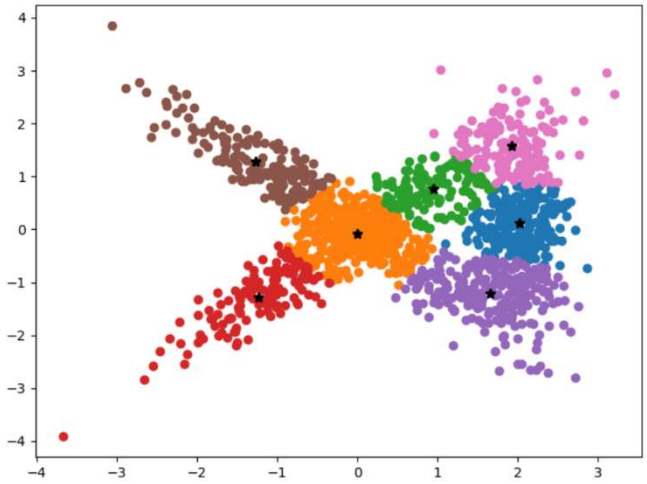
r = 5



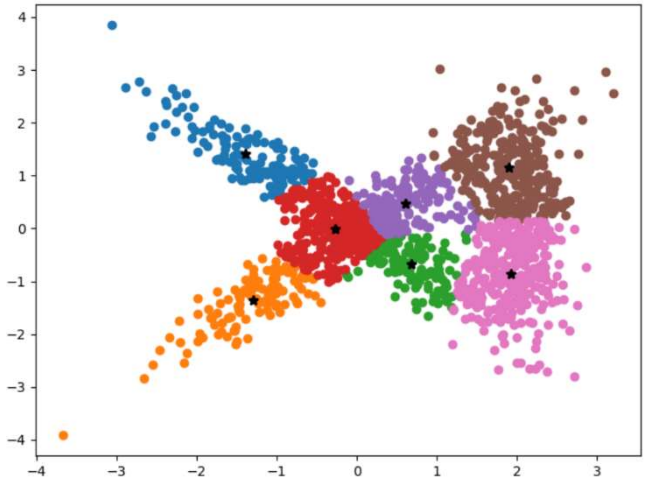
r = 6



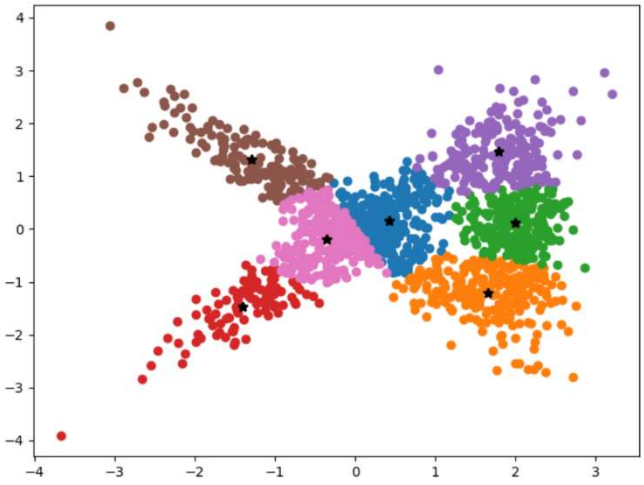
r = 7



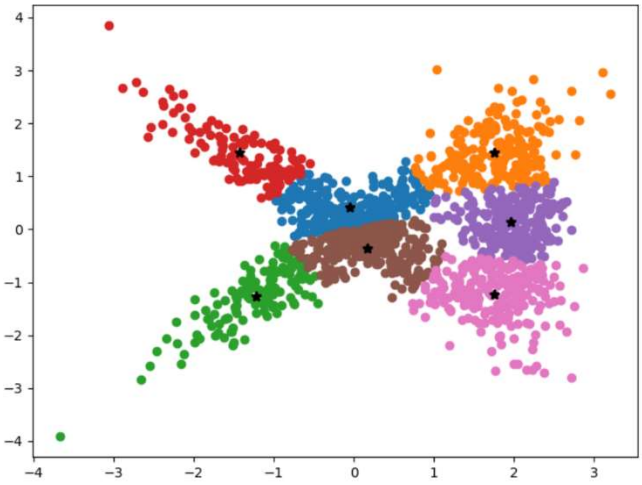
r = 8



r = 9



r = 10



```
C:\Users\SreeV\anaconda3\envs\tensor\pythonw.exe '
k =  7
centroids converged, breaking the loop
r =  1   sum-square-error =  545.9540564031288
centroids converged, breaking the loop
r =  2   sum-square-error =  501.2878001639498
centroids converged, breaking the loop
r =  3   sum-square-error =  617.7514065874752
centroids converged, breaking the loop
r =  4   sum-square-error =  617.1514051268657
centroids converged, breaking the loop
r =  5   sum-square-error =  616.2788521530157
centroids converged, breaking the loop
r =  6   sum-square-error =  566.8367585057922
centroids converged, breaking the loop
r =  7   sum-square-error =  500.78099397912627
centroids converged, breaking the loop
r =  8   sum-square-error =  600.7586107151776
centroids converged, breaking the loop
r =  9   sum-square-error =  510.10714414141347
centroids converged, breaking the loop
r =  10   sum-square-error =  589.0977975929953
errors =  [545.9540564031288, 501.2878001639498, 6
min error when r = 7
error =  500.78099397912627
```

Sum-square-error at each 'r'

errors = [545.9540564031288, 501.2878001639498, 617.7514065874752, 617.1514051268657, 616.2788521530157, 566.8367585057922, 500.78099397912627, 600.7586107151776, 510.10714414141347, 589.0977975929953]

min error when r = 7

error = 500.78

I ran this algorithm for different 'K' values (k = 2 to 7). And for each 'k' value it is run for 'r' (10) times. Below table shows selected models sum-square-error for each 'k' value and 'r' value at which that model occurred.

K Value	**Minimum Sum-Square Error**	**Model occurred at 'r'**
K = 2	2168.278	5
K = 3	1478.427	1
K = 4	810.717	5
K = 5	638.298	6
K = 6	627.896	4
K = 7	500.78	7