TITLE: Clustering Algorithms **AUTHOR**: Giacomo Garbarino **DATE**: 31/05/2021

HOMEWORK DESCRIPTION

This homework consist of implementing and testing K-Means, C-Means and GPC-Means clustering algorithms, WTA and alpha-cut defuzzifiers and RAND and Jaccard indeces.

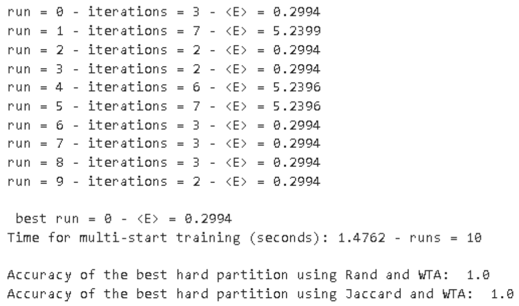
HOMEWORK APPROACH

Once the libraries are imported, an auxiliar function named *get_real_U* to convert datasets into membership matrices for accuracy measuring is defined. Then, three datasets are prepared: separate blobs, overlapping blobs and iris dataset. At the end of the implementation of each dataset, the membership matrices obtained using the *get_real_U* are added in a list used for the test phase. Starting from the implementation of K-Means already provided, C-Means and GPC-Means classes have almost the same structure (some adjustments are needed according the corresponding formulas). After that, WTA and alpha-cut defuzzifiers are implemented using simple user-defined functions which will be used together with other two functions for the implementation of RAND and Jaccard indeces: *co_association* to convert a matrix into its co-association version and *components* to return *n00* (number of data points in dataset Z both in different clusters in A and in different clusters in B), *n01* (number of data points in dataset Z both in different clusters in A and in the same clusters in B), *n10* (number of data points in dataset Z both in the same clusters in A and in different clusters in B), *n11* (number of data points in dataset Z both in the same clusters in A and in the same clusters in B). In particular, there are two versions of RAND and Jaccard indexes: one using WTA and one using alpha-cut. Then, the implementation of the test phase consists of defining three functions to run K-Means, C-Means and GPC-Means multiple times, a function to plotted the clustered data points, a function to measure accuracy using RAND and Jaccard with WTA rule. These functions are used with class methods of the clustering algorithms to prepare cells to get the results.

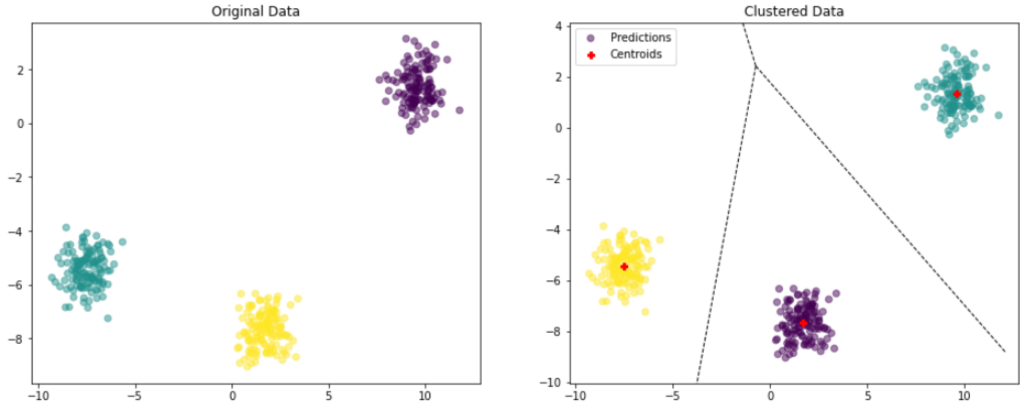
RESULTS

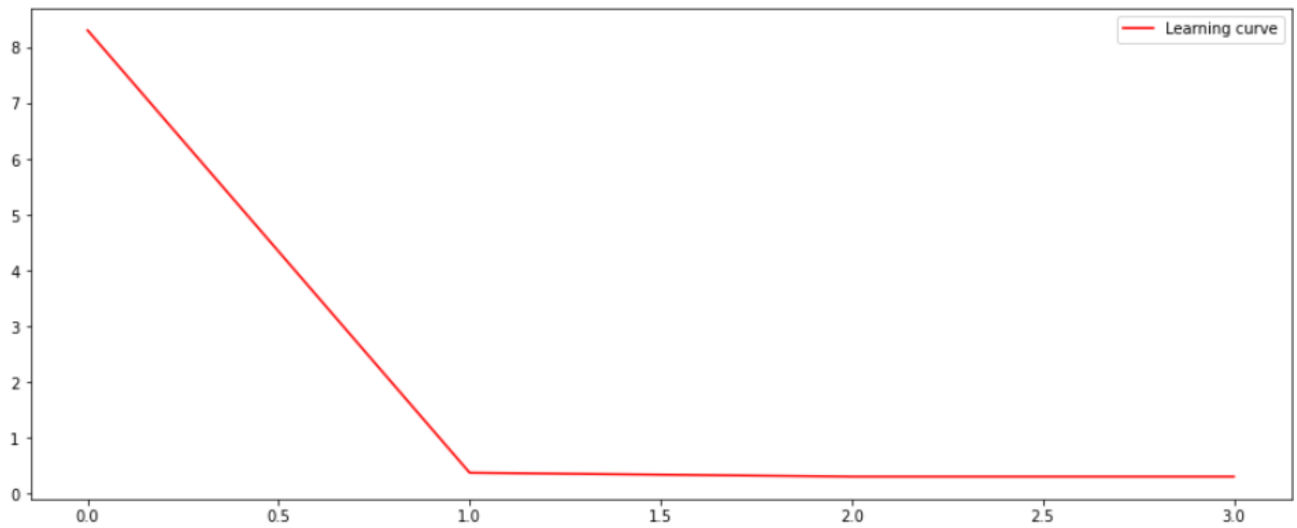
First dataset (400 points, 2 features, 3 clusters and standard deviation=0.70):

- K-Means:

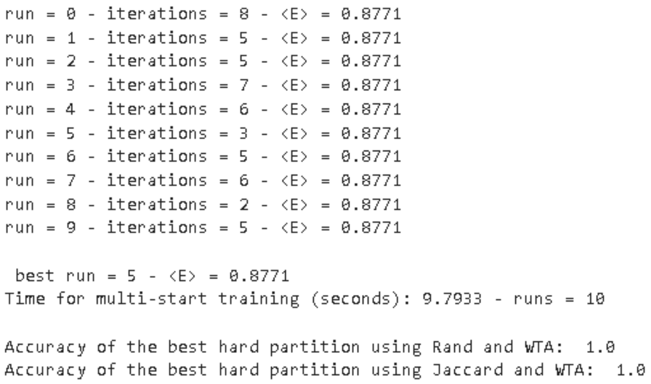


```
run = 0 - iterations = 3 - <E> = 0.2994
run = 1 - iterations = 7 - <E> = 5.2399
run = 2 - iterations = 2 - <E> = 0.2994
run = 3 - iterations = 2 - <E> = 0.2994
run = 4 - iterations = 6 - <E> = 5.2396
run = 5 - iterations = 7 - <E> = 5.2396
run = 6 - iterations = 3 - <E> = 0.2994
run = 7 - iterations = 3 - <E> = 0.2994
run = 8 - iterations = 3 - <E> = 0.2994
run = 9 - iterations = 2 - <E> = 0.2994

 best run = 0 - <E> = 0.2994
Time for multi-start training (seconds): 1.4762 - runs = 10

Accuracy of the best hard partition using Rand and WTA:  1.0
Accuracy of the best hard partition using Jaccard and WTA:  1.0
```

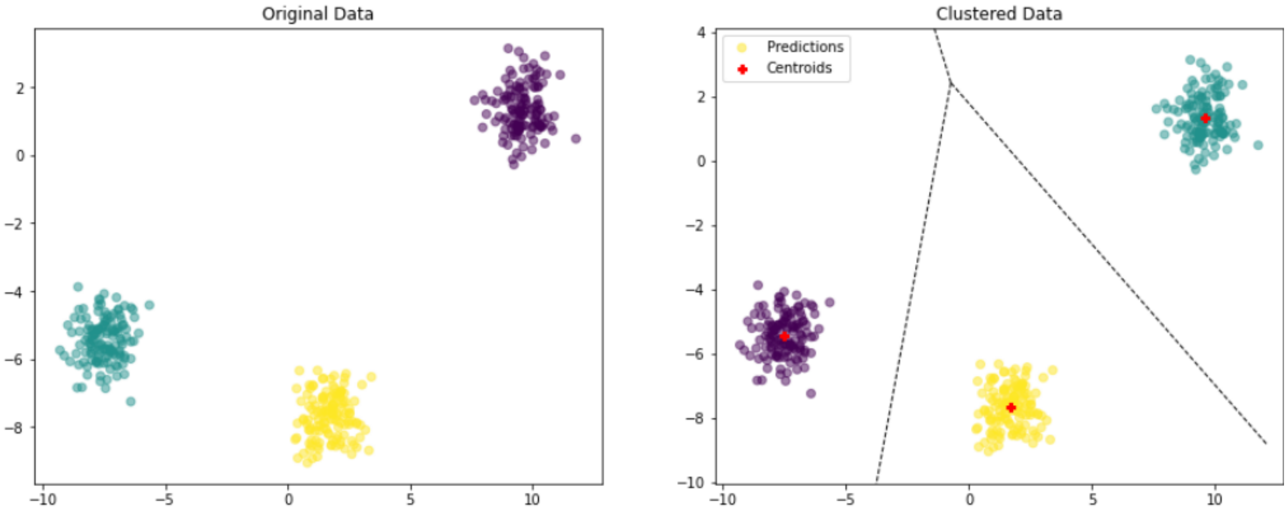


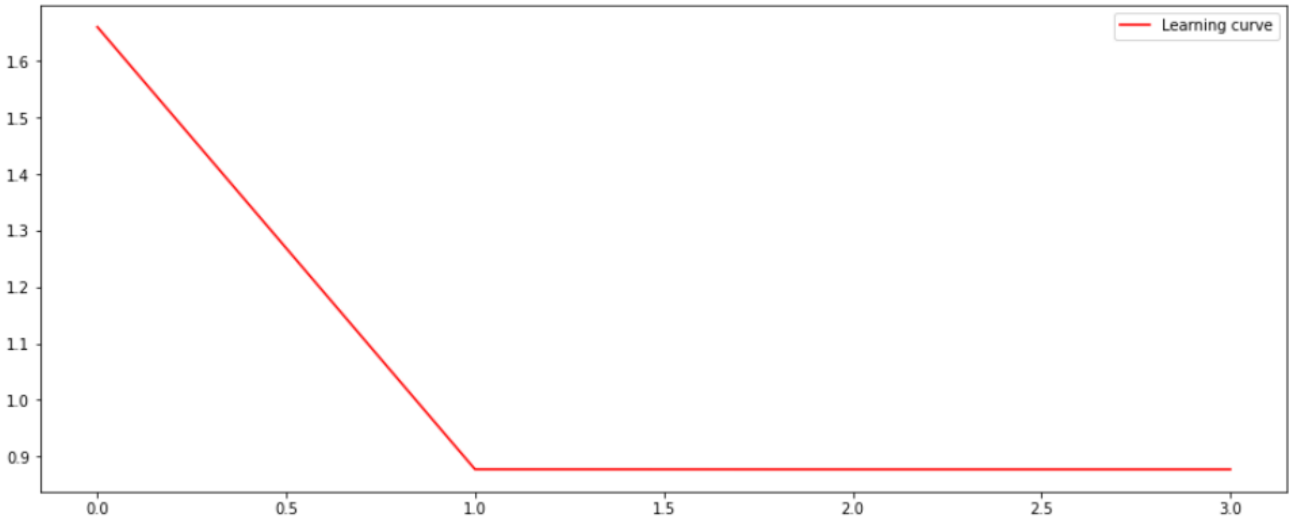


- C-Means:

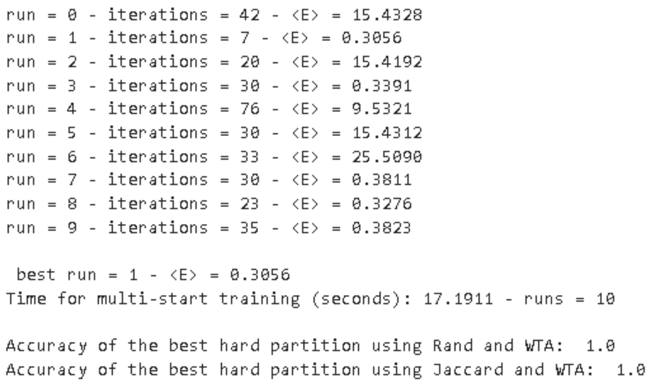


```
run = 0 - iterations = 8 - <E> = 0.8771
run = 1 - iterations = 5 - <E> = 0.8771
run = 2 - iterations = 5 - <E> = 0.8771
run = 3 - iterations = 7 - <E> = 0.8771
run = 4 - iterations = 6 - <E> = 0.8771
run = 5 - iterations = 3 - <E> = 0.8771
run = 6 - iterations = 5 - <E> = 0.8771
run = 7 - iterations = 6 - <E> = 0.8771
run = 8 - iterations = 2 - <E> = 0.8771
run = 9 - iterations = 5 - <E> = 0.8771

 best run = 5 - <E> = 0.8771
Time for multi-start training (seconds): 9.7933 - runs = 10

Accuracy of the best hard partition using Rand and WTA:  1.0
Accuracy of the best hard partition using Jaccard and WTA:  1.0
```

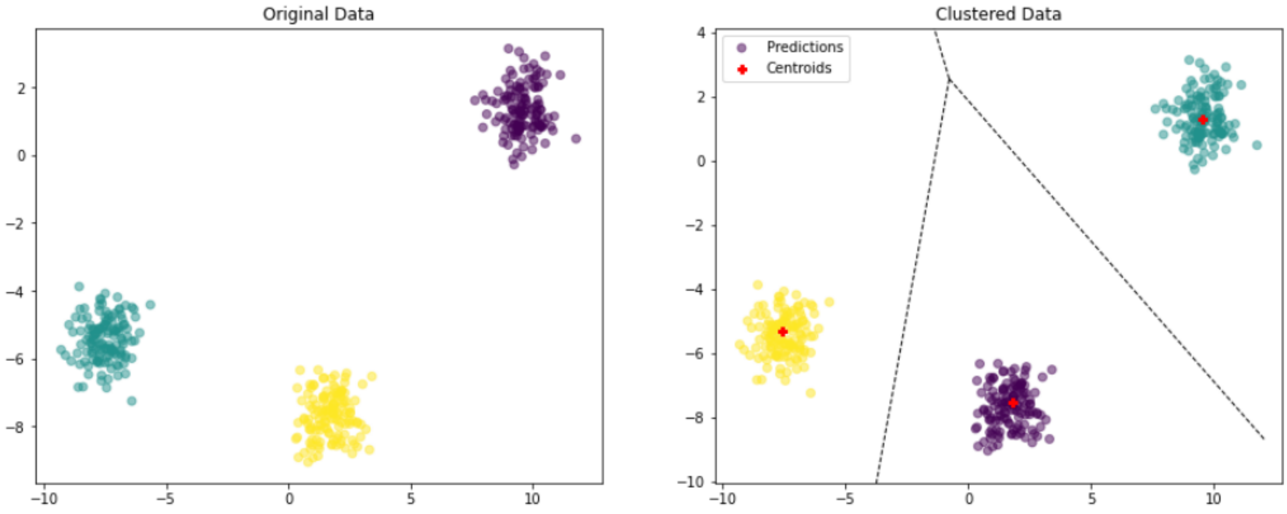


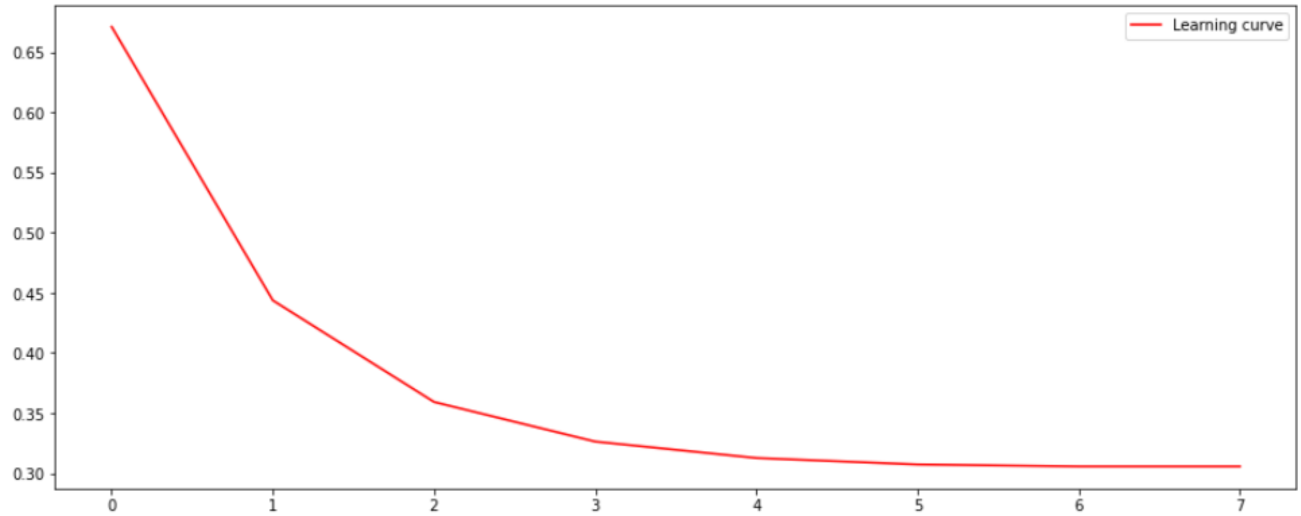


- GPC-Means:



```
run = 0 - iterations = 42 - <E> = 15.4328
run = 1 - iterations = 7 - <E> = 0.3056
run = 2 - iterations = 20 - <E> = 15.4192
run = 3 - iterations = 30 - <E> = 0.3391
run = 4 - iterations = 76 - <E> = 9.5321
run = 5 - iterations = 30 - <E> = 15.4312
run = 6 - iterations = 33 - <E> = 25.5090
run = 7 - iterations = 30 - <E> = 0.3811
run = 8 - iterations = 23 - <E> = 0.3276
run = 9 - iterations = 35 - <E> = 0.3823

 best run = 1 - <E> = 0.3056
Time for multi-start training (seconds): 17.1911 - runs = 10

Accuracy of the best hard partition using Rand and WTA:  1.0
Accuracy of the best hard partition using Jaccard and WTA:  1.0
```

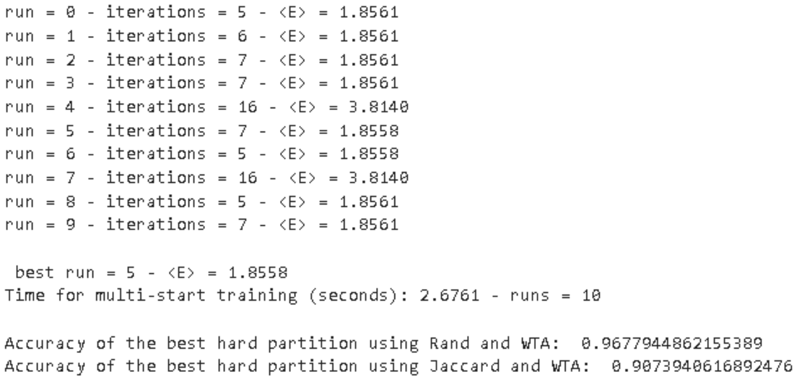




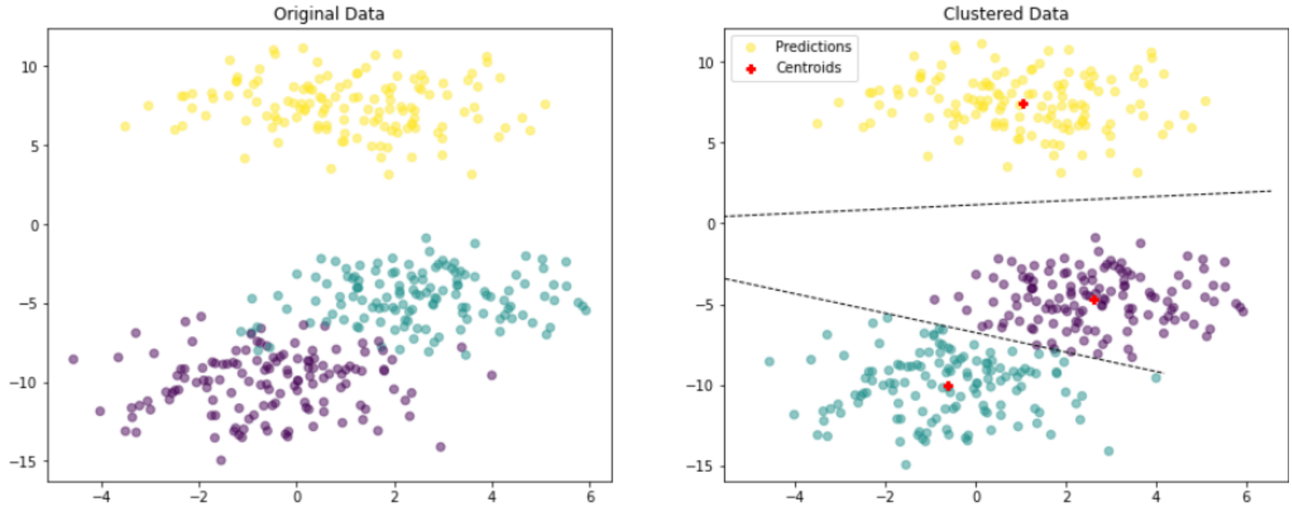
Data are perfectly clustered in all three cases, but the best result is obtained with K-Means (best run=0,2994). The accuracy is equal to 1 using both RAND and Jaccard indeces. The evolution of the learning curve of K-Means and C-Means is the same, while the learning curve of GPC-Means is a little bit more fluid.

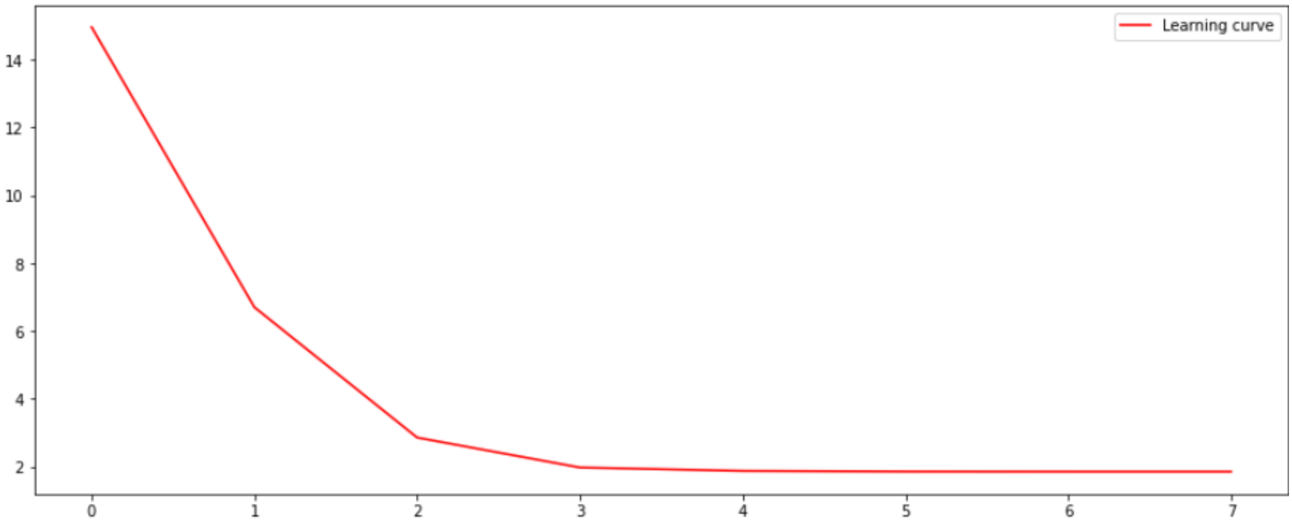
Second dataset (400 points, 2 features, 3 clusters and standard deviation=1.70):

- K-Means:



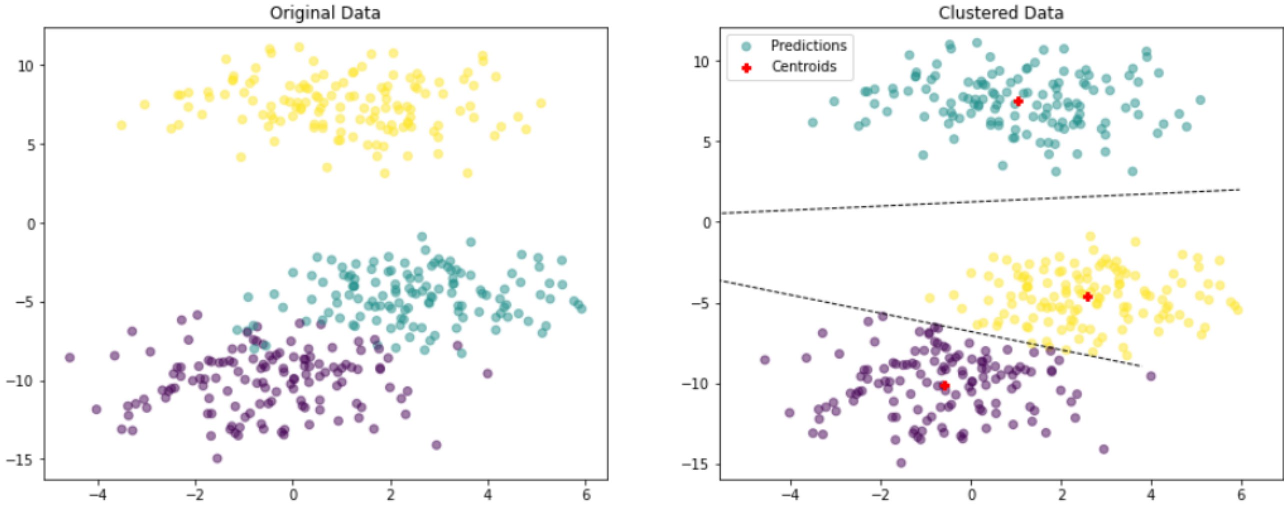
```
run = 0 - iterations = 5 - <E> = 1.8561
run = 1 - iterations = 6 - <E> = 1.8561
run = 2 - iterations = 7 - <E> = 1.8561
run = 3 - iterations = 7 - <E> = 1.8561
run = 4 - iterations = 16 - <E> = 3.8140
run = 5 - iterations = 7 - <E> = 1.8558
run = 6 - iterations = 5 - <E> = 1.8558
run = 7 - iterations = 16 - <E> = 3.8140
run = 8 - iterations = 5 - <E> = 1.8561
run = 9 - iterations = 7 - <E> = 1.8561

 best run = 5 - <E> = 1.8558
Time for multi-start training (seconds): 2.6761 - runs = 10

Accuracy of the best hard partition using Rand and WTA:  0.9677944862155389
Accuracy of the best hard partition using Jaccard and WTA:  0.9073940616892476
```



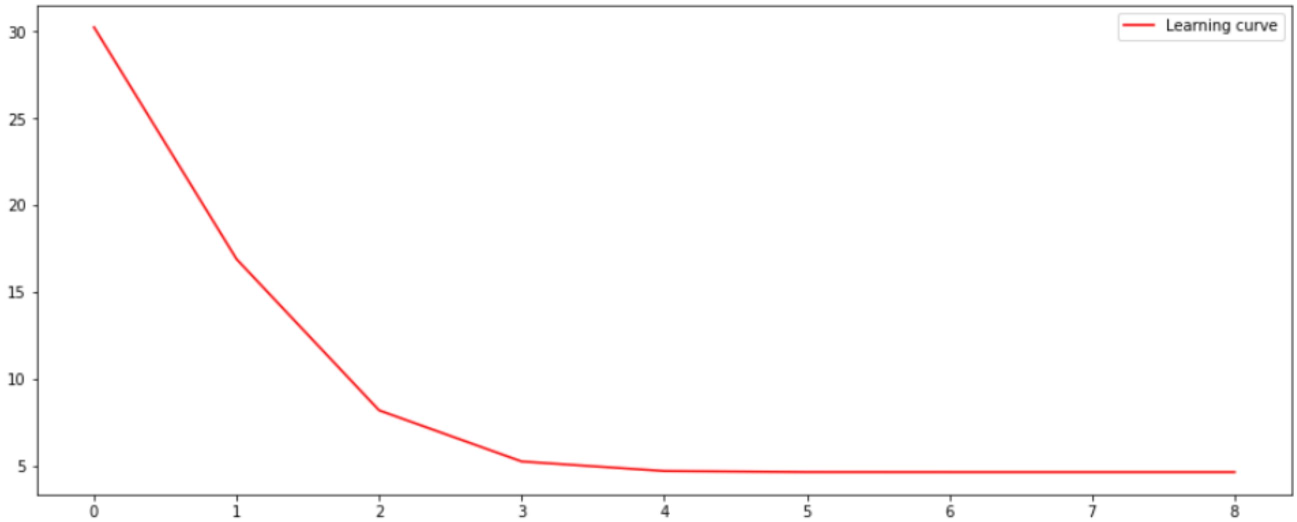


- C-Means:

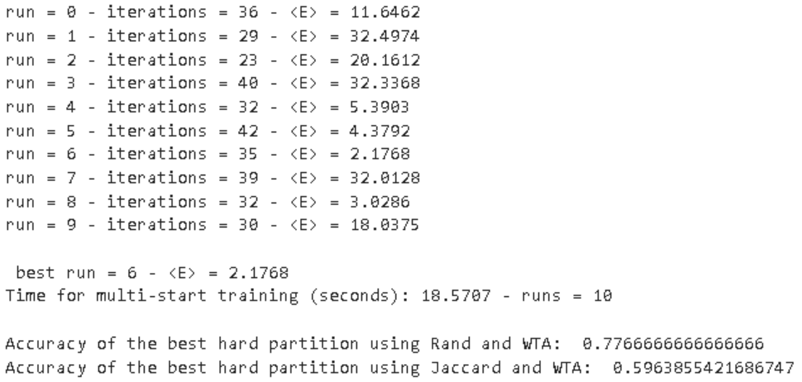
```
run = 0 - iterations = 8 - <E> = 4.6151
run = 1 - iterations = 8 - <E> = 4.6151
run = 2 - iterations = 10 - <E> = 4.6151
run = 3 - iterations = 9 - <E> = 4.6151
run = 4 - iterations = 15 - <E> = 4.6151
run = 5 - iterations = 8 - <E> = 4.6151
run = 6 - iterations = 6 - <E> = 4.6151
run = 7 - iterations = 7 - <E> = 4.6151
run = 8 - iterations = 8 - <E> = 4.6151
run = 9 - iterations = 10 - <E> = 4.6151

 best run = 5 - <E> = 4.6151
Time for multi-start training (seconds): 15.3513 - runs = 10

Accuracy of the best hard partition using Rand and WTA:  0.9677944862155389
Accuracy of the best hard partition using Jaccard and WTA:  0.9073940616892476
```


Original Data

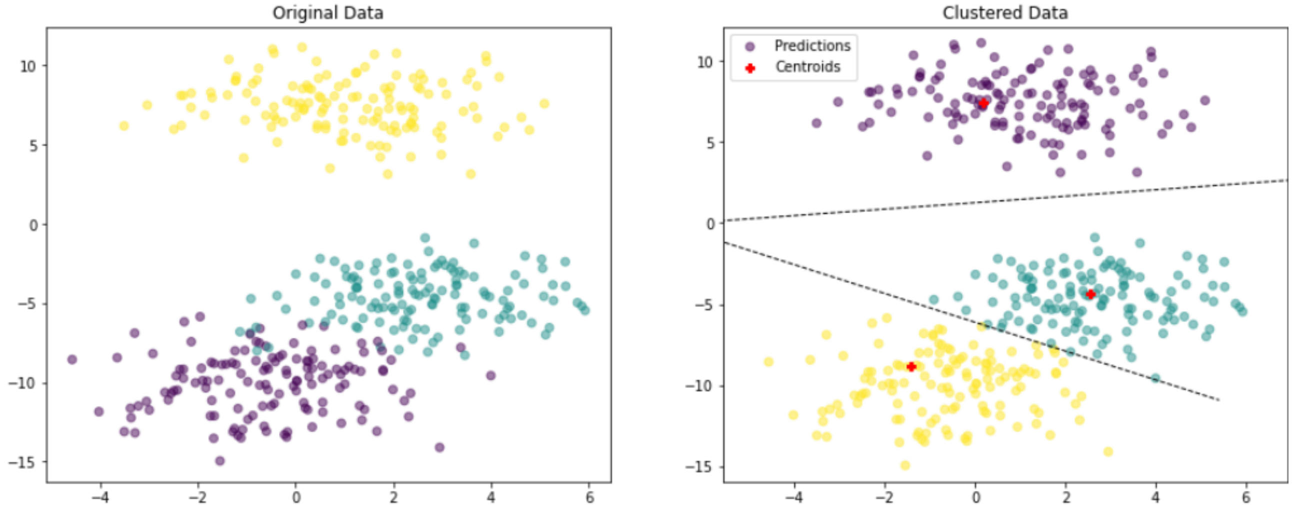
Clustered Data

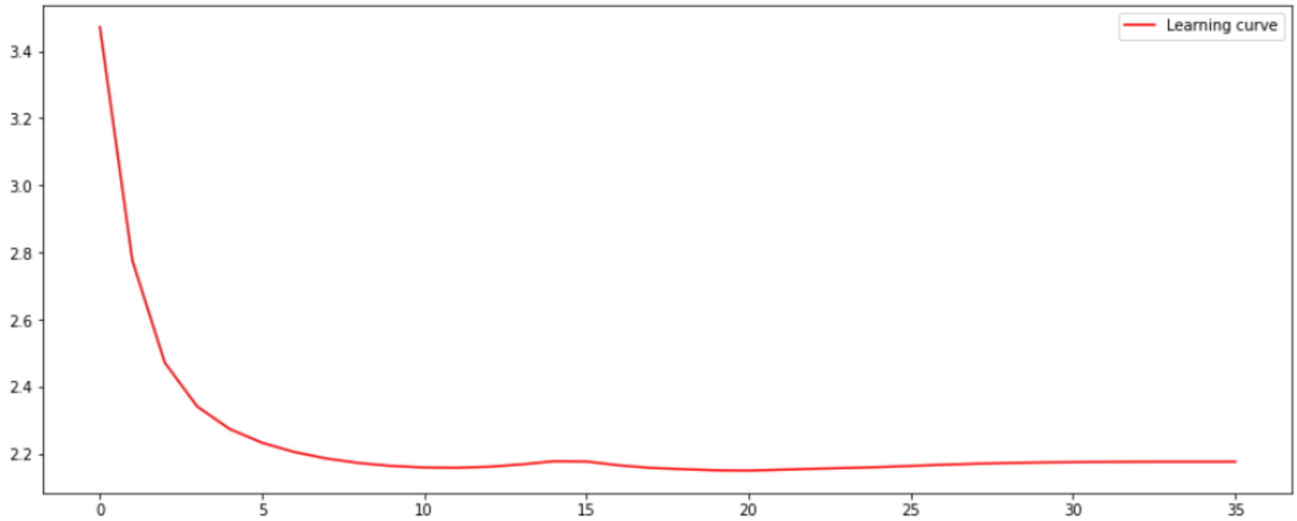


- GPC-Means:



```
run = 0 - iterations = 36 - <E> = 11.6462
run = 1 - iterations = 29 - <E> = 32.4974
run = 2 - iterations = 23 - <E> = 20.1612
run = 3 - iterations = 40 - <E> = 32.3368
run = 4 - iterations = 32 - <E> = 5.3903
run = 5 - iterations = 42 - <E> = 4.3792
run = 6 - iterations = 35 - <E> = 2.1768
run = 7 - iterations = 39 - <E> = 32.0128
run = 8 - iterations = 32 - <E> = 3.0286
run = 9 - iterations = 30 - <E> = 18.0375

 best run = 6 - <E> = 2.1768
Time for multi-start training (seconds): 18.5707 - runs = 10

Accuracy of the best hard partition using Rand and WTA:  0.7766666666666666
Accuracy of the best hard partition using Jaccard and WTA:  0.5963855421686747
```

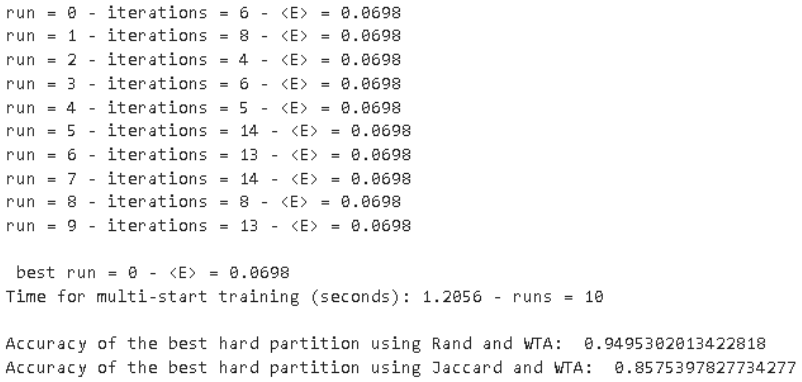




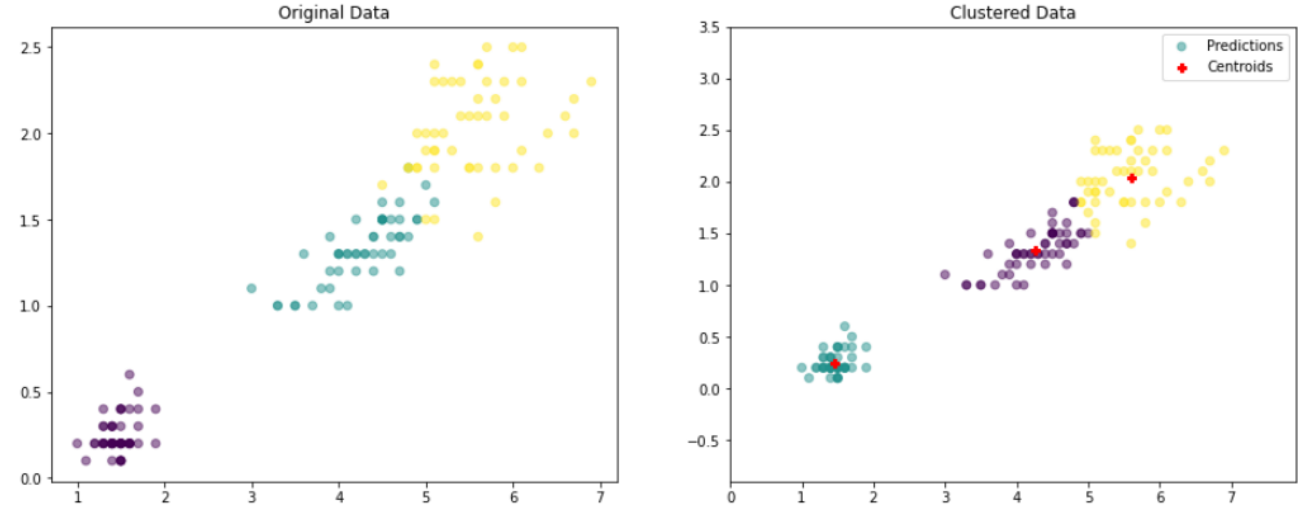
Data are perfectly clustered in all three cases, but the best result is obtained again with K-Means (best run=1,8558). The accuracy is equal to 0.96 using RAND index, 0.90 with Jaccard index. The evolution of the learning curve of K-Means and C-Means is almost the same, while the learning curve of GPC-Means is a little bit more fluid and it has a faster descent.

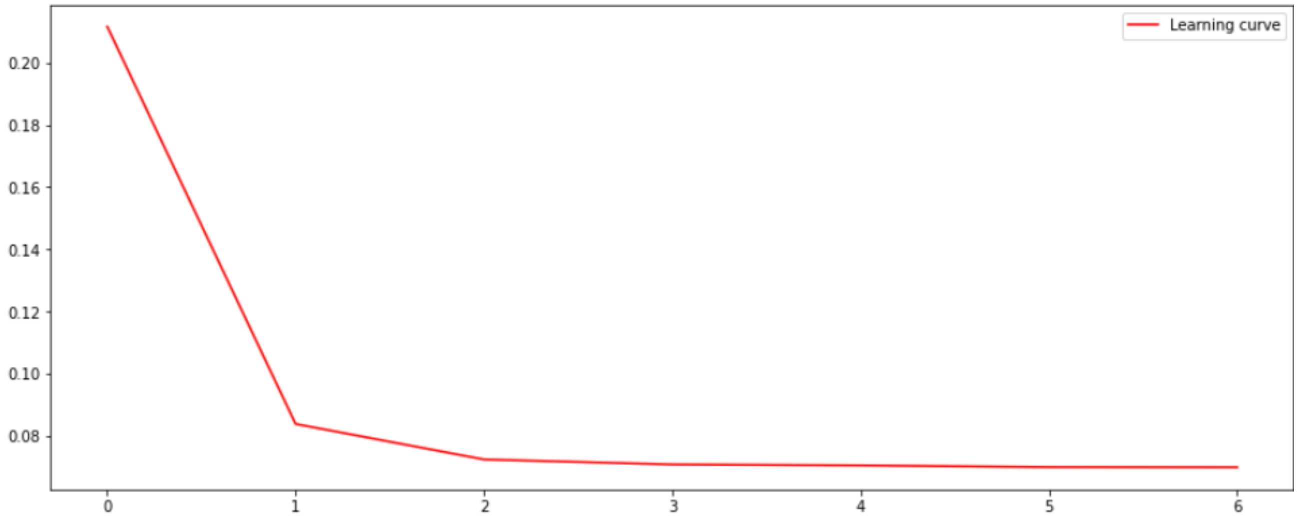
Iris dataset:

- K-Means:

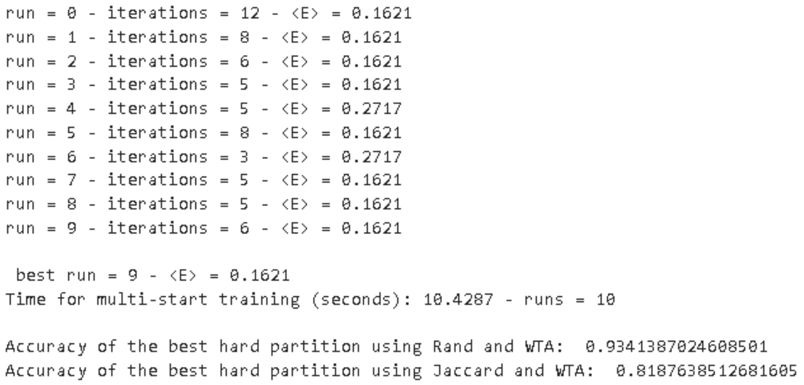


```
run = 0 - iterations = 6 - <E> = 0.0698
run = 1 - iterations = 8 - <E> = 0.0698
run = 2 - iterations = 4 - <E> = 0.0698
run = 3 - iterations = 6 - <E> = 0.0698
run = 4 - iterations = 5 - <E> = 0.0698
run = 5 - iterations = 14 - <E> = 0.0698
run = 6 - iterations = 13 - <E> = 0.0698
run = 7 - iterations = 14 - <E> = 0.0698
run = 8 - iterations = 8 - <E> = 0.0698
run = 9 - iterations = 13 - <E> = 0.0698

 best run = 0 - <E> = 0.0698
Time for multi-start training (seconds): 1.2056 - runs = 10

Accuracy of the best hard partition using Rand and WTA:  0.9495302013422818
Accuracy of the best hard partition using Jaccard and WTA:  0.8575397827734277
```

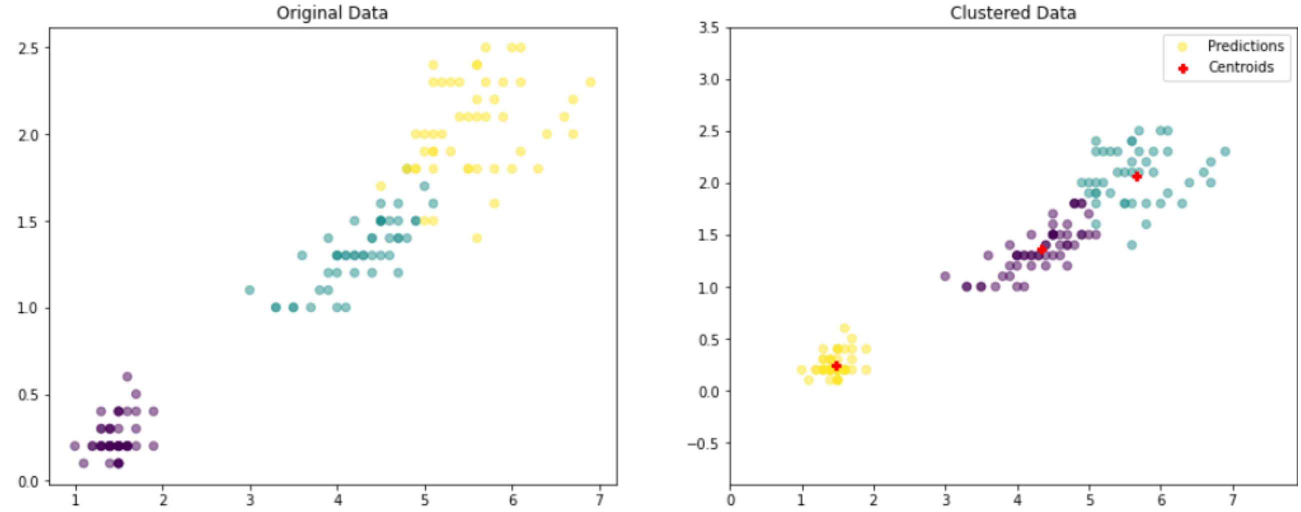




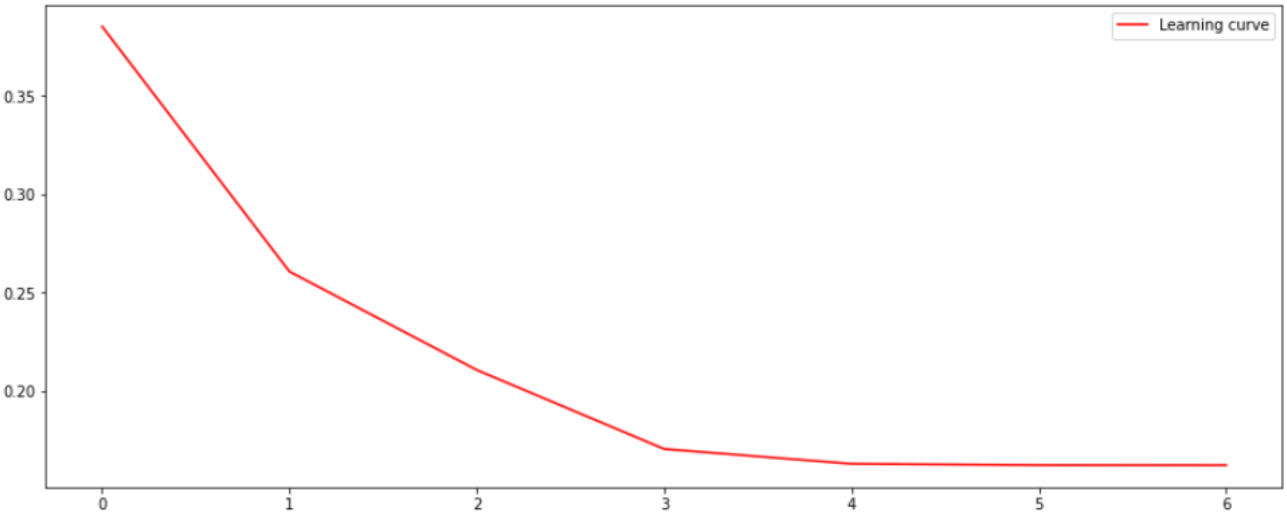
- C-Means:



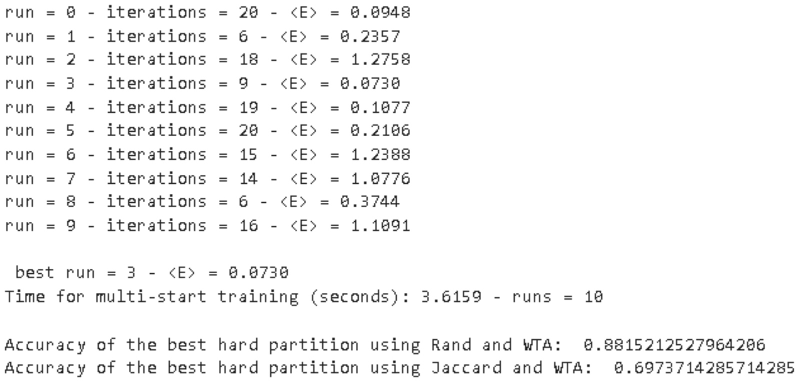
```
run = 0 - iterations = 12 - <E> = 0.1621
run = 1 - iterations = 8 - <E> = 0.1621
run = 2 - iterations = 6 - <E> = 0.1621
run = 3 - iterations = 5 - <E> = 0.1621
run = 4 - iterations = 5 - <E> = 0.2717
run = 5 - iterations = 8 - <E> = 0.1621
run = 6 - iterations = 3 - <E> = 0.2717
run = 7 - iterations = 5 - <E> = 0.1621
run = 8 - iterations = 5 - <E> = 0.1621
run = 9 - iterations = 6 - <E> = 0.1621

 best run = 9 - <E> = 0.1621
Time for multi-start training (seconds): 10.4287 - runs = 10

Accuracy of the best hard partition using Rand and WTA:  0.9341387024608501
Accuracy of the best hard partition using Jaccard and WTA:  0.8187638512681605
```


Original Data

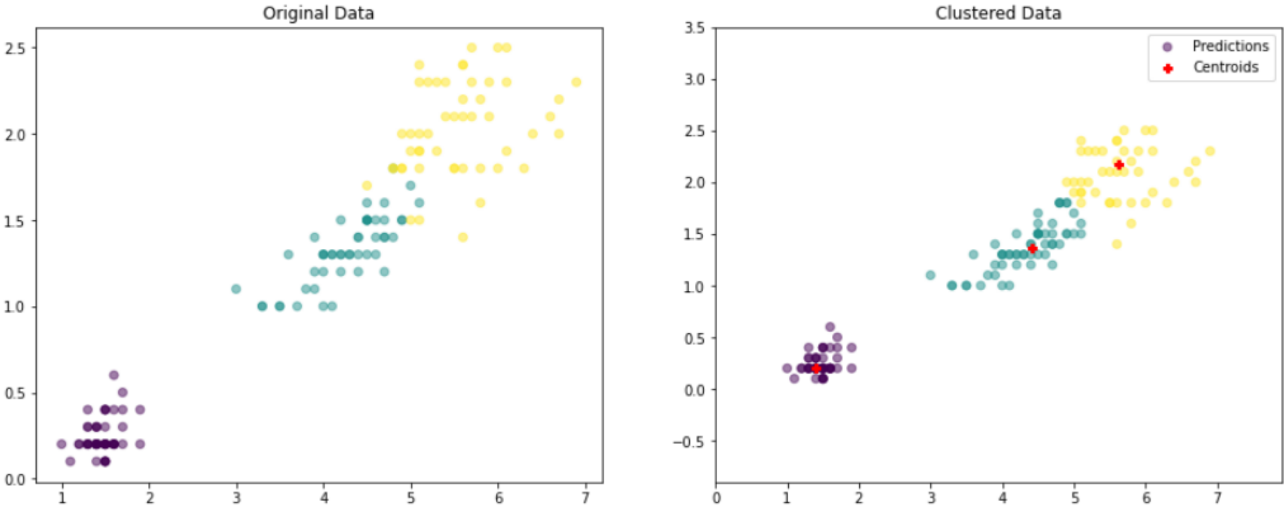
Clustered Data

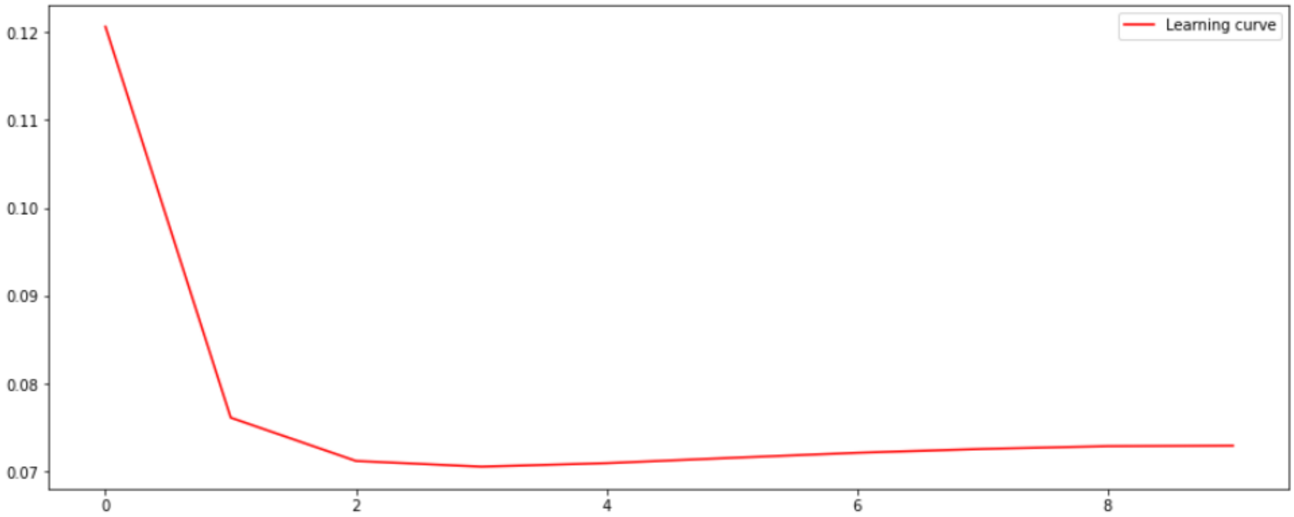


- GPC-Means:



```
run = 0 - iterations = 20 - <E> = 0.0948
run = 1 - iterations = 6 - <E> = 0.2357
run = 2 - iterations = 18 - <E> = 1.2758
run = 3 - iterations = 9 - <E> = 0.0730
run = 4 - iterations = 19 - <E> = 0.1077
run = 5 - iterations = 20 - <E> = 0.2106
run = 6 - iterations = 15 - <E> = 1.2388
run = 7 - iterations = 14 - <E> = 1.0776
run = 8 - iterations = 6 - <E> = 0.3744
run = 9 - iterations = 16 - <E> = 1.1091

 best run = 3 - <E> = 0.0730
Time for multi-start training (seconds): 3.6159 - runs = 10

Accuracy of the best hard partition using Rand and WTA:  0.8815212527964206
Accuracy of the best hard partition using Jaccard and WTA:  0.6973714285714285
```





The best result is obtained with K-Means (best run=0,0698). The accuracy is equal to 0.94 using RAND index, while using Jaccard index it's equal to 0.85. The evolution of the learning curve of K-Means and GPC-Means shows a fast descent, while the learning curve of C-Means is more fluid.

In all three datasets, the K-Means is the algorithm which shows the best results in terms of run and the highest accuracy, while the C-Means is the algorithm which shows the worst results in terms of run and the GPC-Means in terms of accuracy.