Assignment3 - Clustering

Ji

Dec 2022

Contents

1 Verification of Correctness

This section verifies the correctness of my implementation of three clustering algorithms: K-Means, Gaussian Mixture, and Spectral Clustering.

1.1 K-Means

In the experiment of testing K-Means, the dataset has six points: $[[1,2],[1.5,1.8],[5,8],[8,8],[1,0.6],[9,11]]$. By setting $k=2$, the three points of $[[1,2],[1.5,1.8],[1,0.6]]$ should be assigned into one cluster, and the remaining three points are expected to be assigned into the other. The centroid of the first cluster should be $[1.167,1.467]$, and that of the second cluster should be $[7.333,9.0]$.

The results of the experiment conform to the expectation. As shown in Figure 1, the three blue points are $[[1,2],[1.5,1.8],[1,0.6]]$, and the three red points represent $[[5,8],[8,8],[9,11]]$, which have been assigned to two clusters respectively. More clustering results by K-Means are shown in Section 2.

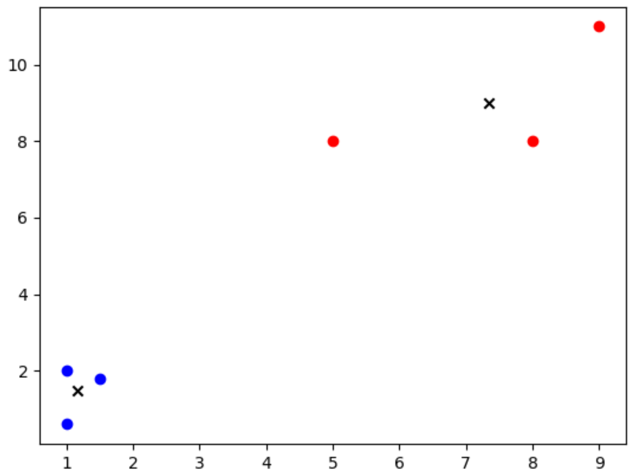
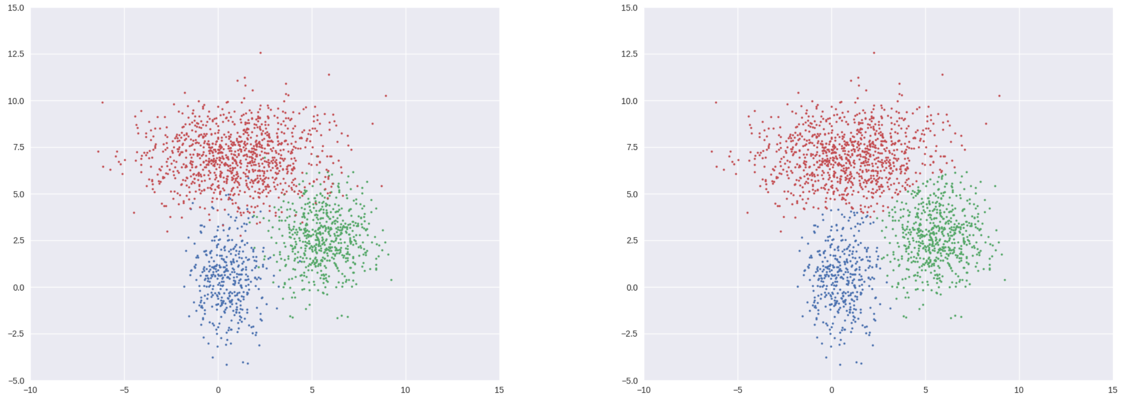


Figure 1: Visualization of the clustering results by K-Means.

1.2 GMM

The experiment for verifying the implementation of GMM is more complicated than that described in Section 1.1. The test dataset has 1000 2D data points generated by three different multivariate normal distributions, which is visualized in Figure 2 (a).

By comparing the ground truth labels with the clustering results by the gaussian mixture model, 96.2% of the points are assigned to their correct cluster. Furthermore, the three clusters shown in Figure 2 (b) look almost the same as those of the ground truth. Therefore, the implementation of GMM is correct.



(a) ground truth

(b) clustering results

Figure 2: Visualization of the ground truth and clustering results of the test dataset.

1.3 Spectral Clustering

When verifying the implementation of the spectral clustering algorithm, I used the so-called moon dataset including 150 2D data points. Since K-Means and GMM both fail in the moon dataset, it is a good test dataset for verifying the implementation of the spectral clustering algorithm.

Firstly, I also compared the ground truth labels with the clustering results by the spectral clustering model. The accuracy is as high as 99.3%. In regard to the visualization of the clustering results, the model can split the dataset into two clusters correctly, as shown in Figure 3.

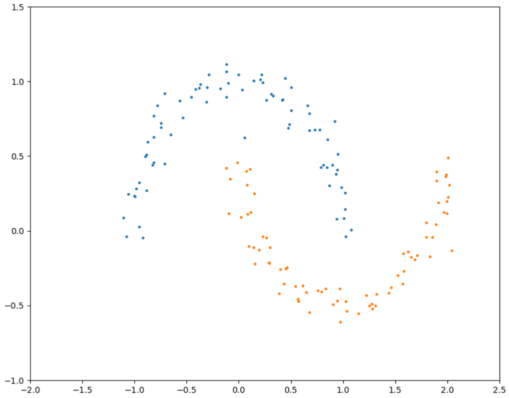


Figure 3: Visualization of the clustering results by Spectral Clustering.

2 Comparison to Sklearn

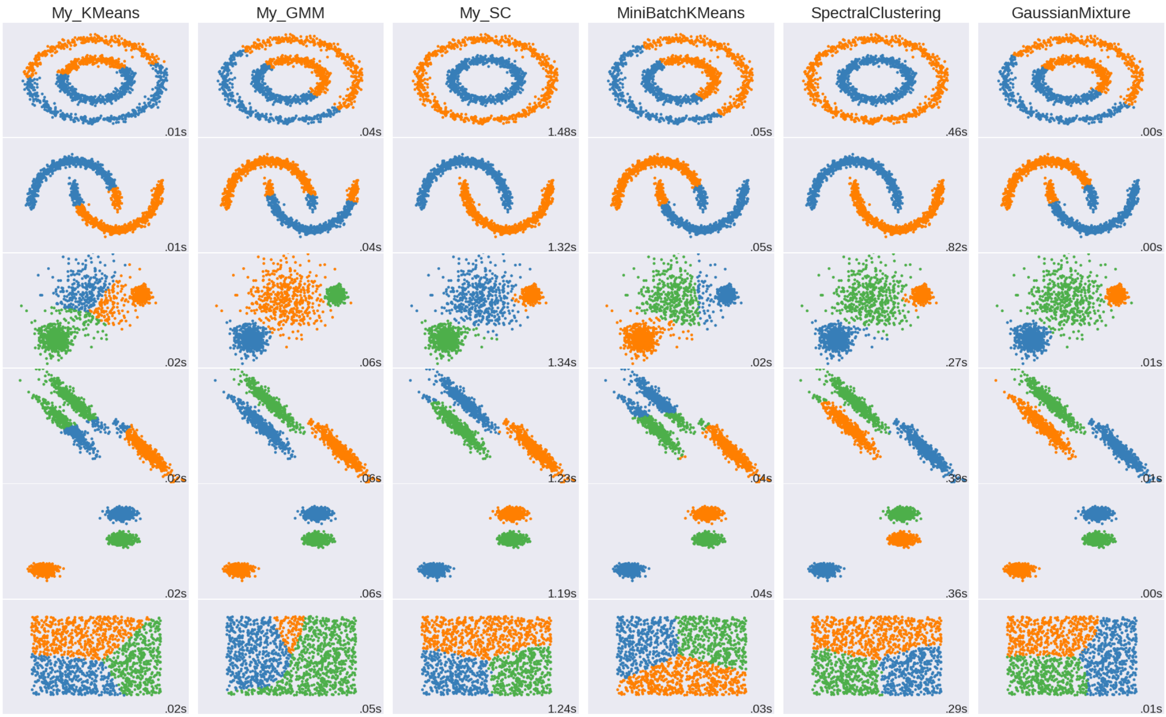


Figure 4: Comparison of three algorithms to their sklearn implementation in six datasets.

As shown in Figure 4, My_KMeans fails row 1, 2, 3, and 4, as MiniBatchKMeans. The k-means algorithm can only succeed at row 5.

In addition to My_KMeans, My_GMM can succeed at row 3 and 4, as GaussianMixture. However, neither My_GMM nor GaussianMixture can solve the problems of clustering row 1 and 2.

As SpectralClustering, My_SC can successfully cluster row 1 and 2. Notably, the spectral clustering algorithm cannot obtain the perfect clustering result at row 4 as the gaussian mixture algorithm.