Assignment 10 - Clustering

June 29, 2019

0.1 Assignment 10

0.1.1 Problem

Given a dataset, look into the scikit-learn.org Python library and find the function implementing the k-means algorithm. Then cluster the dataset into two clusters and represent the result in a plot.

0.1.2 Resolution

We can preprocess the data and use this function in order to solve the above problem with a function in sklearn package called cluster.KMeans that allow to run the K-means algorithm.

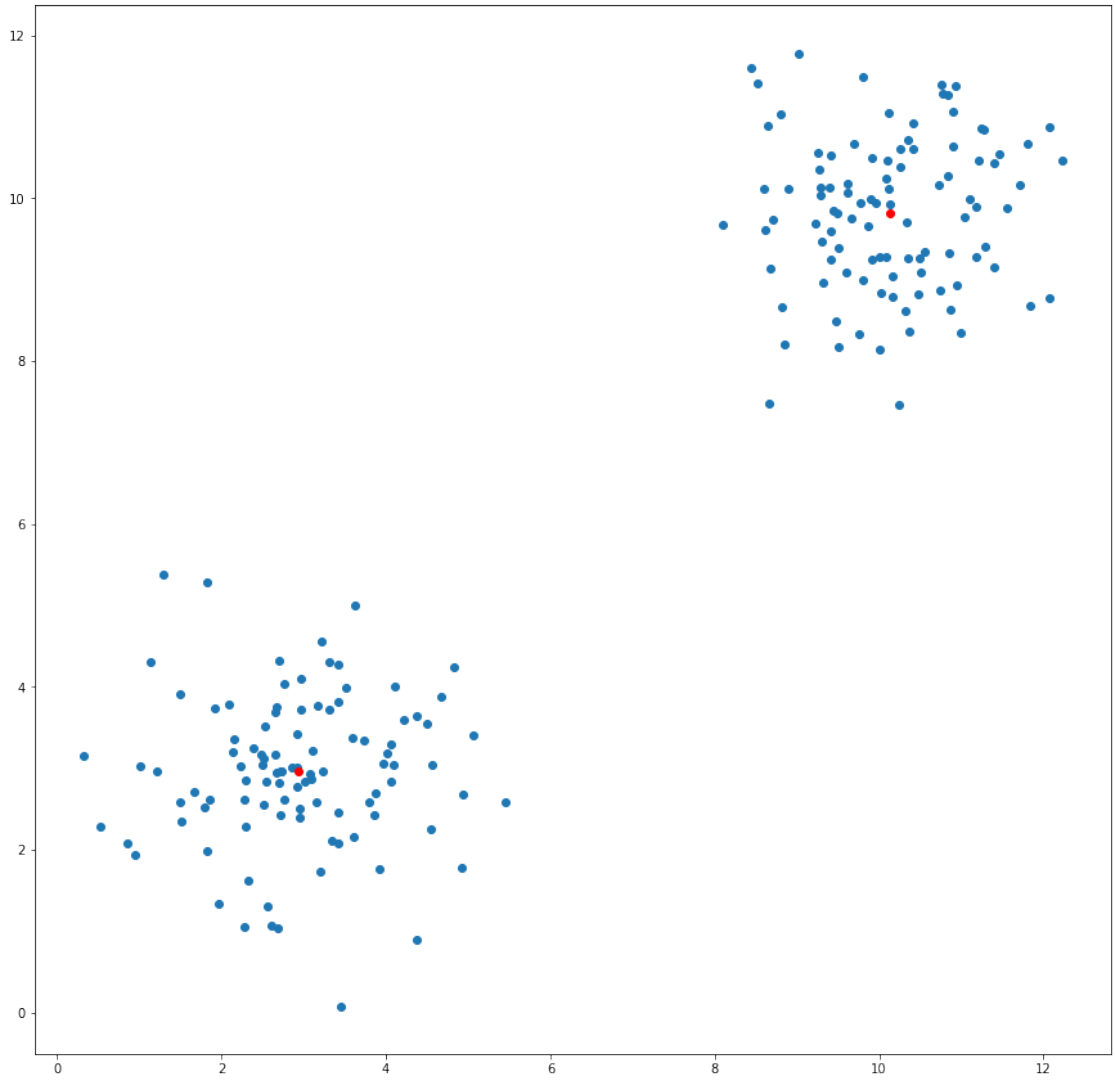
```
In [1]: from sklearn.cluster import KMeans
        import numpy as np
        import matplotlib.pyplot as plt
```

```
In [2]: file = open("clusterData.txt", "r")
        list = file.readlines()
        list = list[0].split()
        X = []
        Y = []
        for i in range(0, len(list), 2):
            X.append(float(list[i]))
            Y.append(float(list[i + 1]))
        arr = np.array([X, Y])
        arr = np.transpose(arr)
        kmeans = KMeans(n_clusters=2).fit(arr)
        print("Cluster centers: ", kmeans.cluster_centers_)
```

```
Cluster centers:  [[ 2.92833957  2.96072918]
 [10.12649163  9.8176424 ]]
```

```
In [3]: cent = np.transpose(kmeans.cluster_centers_)
        plt.figure(figsize=(15, 15))
        plt.scatter(X, Y)
        plt.scatter(cent[0], cent[1], c='r')

        plt.show()
```


12
10
8
6
4
2
0
0
2
4
6
8
10
12