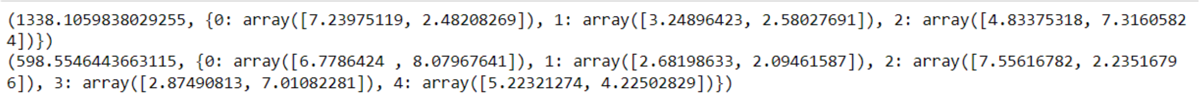
CSE-575 Project-2 Report

In this project I were given a sample which I tried two different strategies for K-means clustering algorithm on. The dataset was made up of points which had x and y values for each row. In order to implement the K-Means clustering algorithm, I have first defined a k-means formula which would be used for the algorithms. In K-means algorithm we first pick centroids with our picking of choice from the data and then try to minimize our own objective function changing centroids as means for each iteration of the algorithm.

In my case the first strategy was to randomly pick k different centroid points from the given data. After picking the centroids I have calculated minimum Euclidean distances of each data point to centroids and assigned them to their clusters. After this I have found means of each cluster and then updated centroids until convergence for the objective function of squared Euclidean distance. I have tried two different methods to find convergence for the objective function: first was to get the function until objective function does not decrease and the other one was to do epochs until convergence. First method worked on the data, but in the end, I used epochs to see full convergence for myself. For the first strategy the results that I get are below:



```
(1338.1059838029255, {0: array([7.23975119, 2.48208269]), 1: array([3.24896423, 2.58027691]), 2: array([4.83375318, 7.3160582
4])})
(598.5546443663115, {0: array([6.7786424 , 8.07967641]), 1: array([2.68198633, 2.09461587]), 2: array([7.55616782, 2.2351679
6]), 3: array([2.87490813, 7.01082281]), 4: array([5.22321274, 4.22502829])})
```

Figure 1 – Strategy 1 Objective Loss and Centroid Values Dictionary

I have also plotted the values of k and objective function in the graph below:

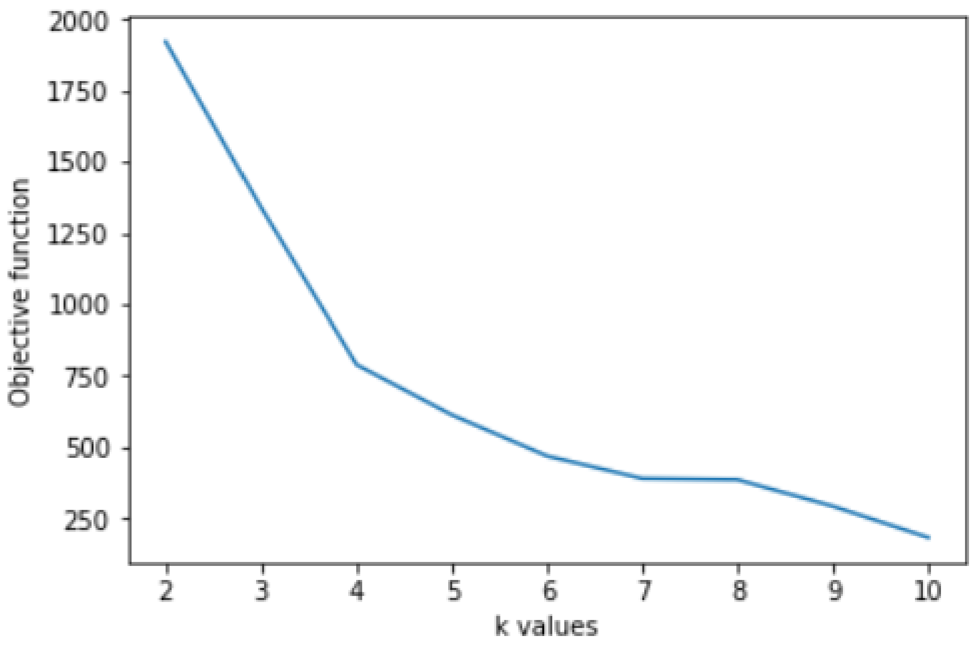
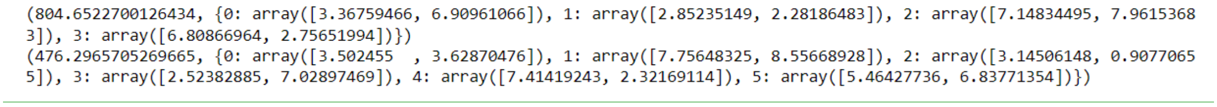


Figure 2-Strategy 1 Objective Function vs K Values

Overall higher k value yields to a lower objective function value in strategy 1.

In the second strategy rather than picking random centroids in the beginning I have selected the data points that are farthest away from the previous centroids. This was the most challenging part of the project as for each ith initial centroid point, I had to find the max distanced data point for (i-1) points. After this initialization the algorithm was the same for strategy 1. In this one, the results I got are below:



```
(804.6522700126434, {0: array([3.36759466, 6.90961066]), 1: array([2.85235149, 2.28186483]), 2: array([7.14834495, 7.9615368
3]), 3: array([6.80866964, 2.75651994])})
(476.2965705269665, {0: array([3.502455  , 3.62870476]), 1: array([7.75648325, 8.55668928]), 2: array([3.14506148, 0.9077065
5]), 3: array([2.52382885, 7.02897469]), 4: array([7.41419243, 2.32169114]), 5: array([5.46427736, 6.83771354])})
```

Figure 3 - Strategy 2 Objective Loss and Centroid Values Dictionary

And the resulting graph was:

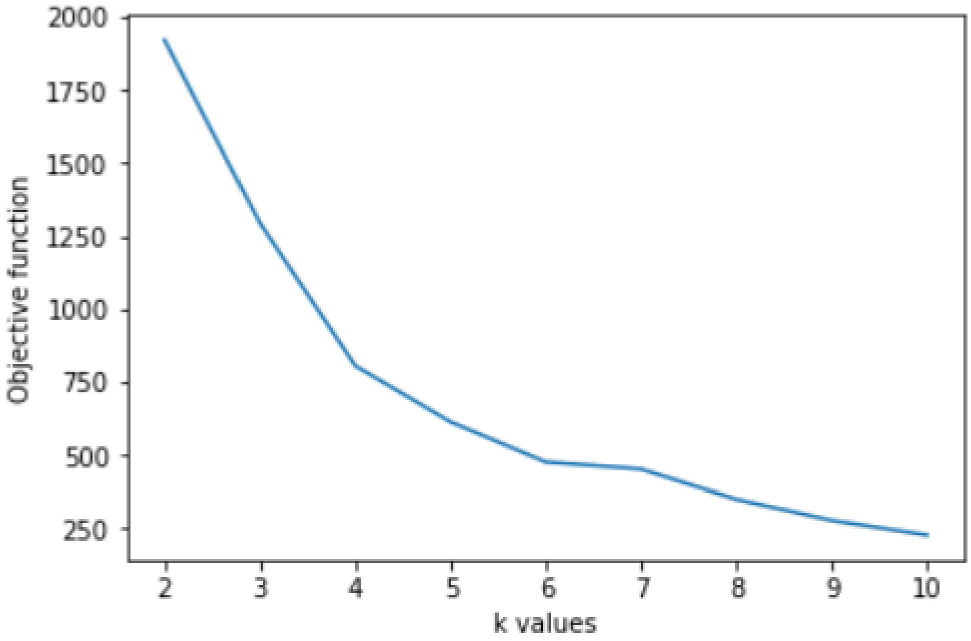


Figure 4 Strategy 2 Objective Function vs K Values

Overall, I have found that higher k value yields a smaller value for objective function. I have expected this result, but I think that this can lead to overfitting as k increases so seeing convergence in objective function vs k values graph should indicate optimal k value. I have also seen that both strategies had similar results when it comes to objective function according to corresponding k values.

Barlas YARDIMCI

1223149330