Tracy Michaels

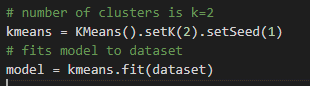
Big Data Programming – Assignment 6

After initial set up, the program loads the data from the kmeans\_input.txt file into a variable called dataset

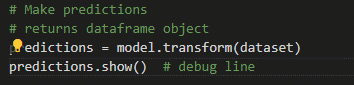


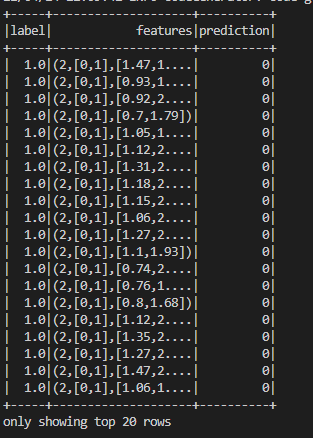
Next I created a variable called kmeans that is initialized with the number of clusters set to k=2

Then calls the fit() method on passing the dataset variable as a parameter this returns a model of type Transformer of a fitted model to the passed data set

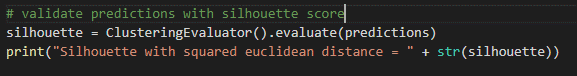


Next I used the transform() method on the model while passing the dataset as a parameter, this returns a DataFrame object with the predictions of which cluster each pair of data points belong to in a third column



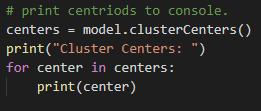


To verify the results from the prediction I calculated the silhouette score. This is a range between 1 and -1 where a value close to 1 means that the points in a cluster are close to other points in the same cluster and far from the points of the other cluster(s)



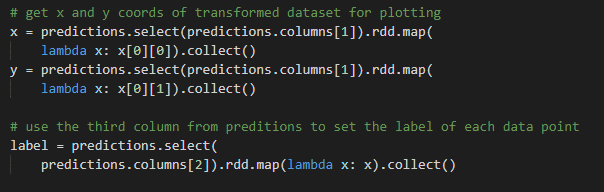


Then I calculated the centroids of the clusters and printed the result to the terminal





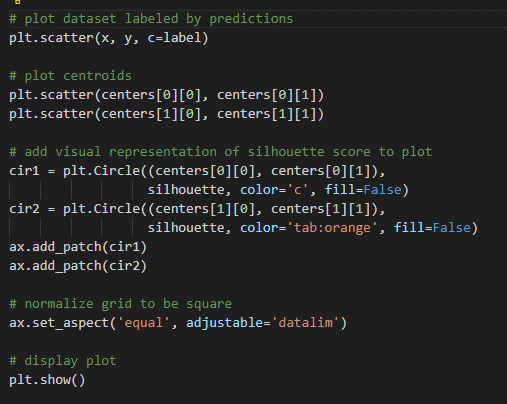
Next I collected the x and y coordinates for plotting purposes as well as a label for each cluster as determined by the prediction



Next I plotted each point color coded by cluster

I also added the centroids to the plot, as well as a visual representation for the silhouette score as a circle centered around each centroid

And normalized the grid to a square and doesn’t distort the visual



Final result of plotting: