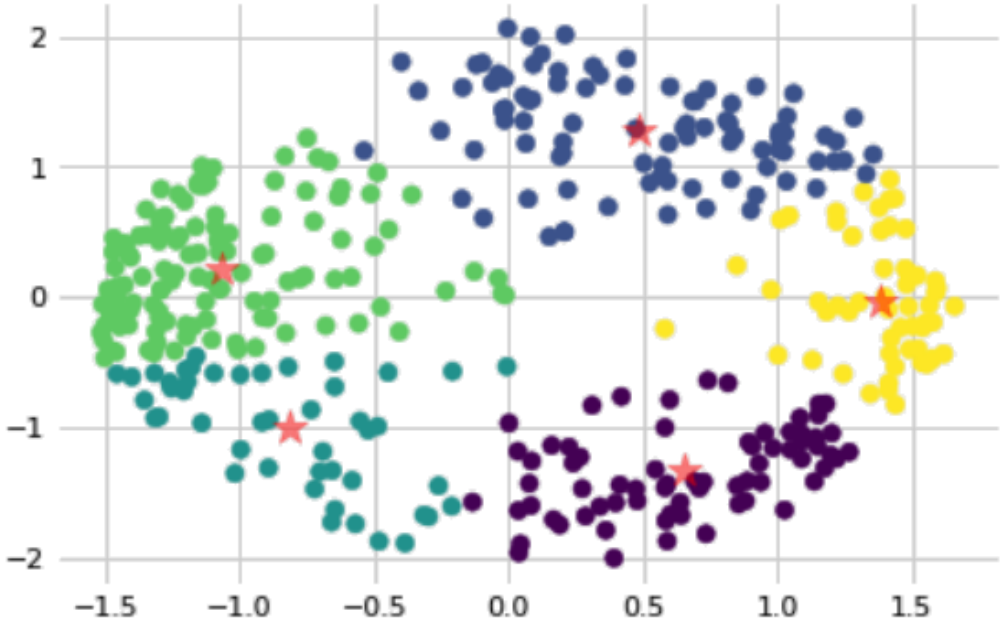
Project One

K-means and Spectral Clustering

Bashir Sadat

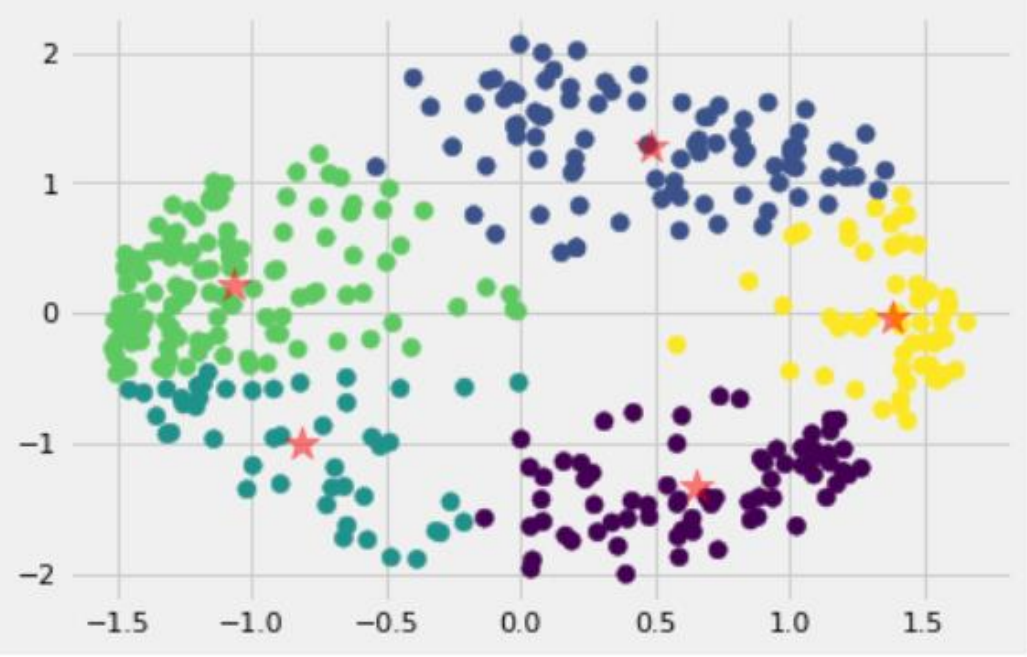
Lehigh University

Data-Mining Class
Spring 2020
Sas617@lehigh.edu



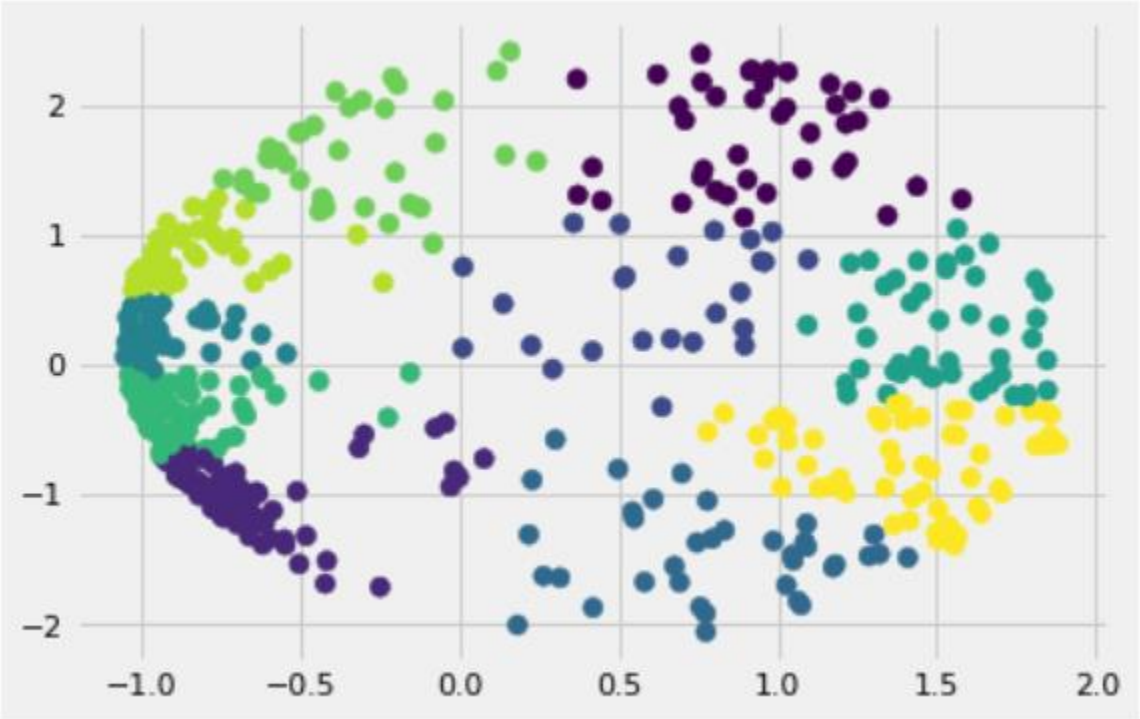
I followed the following steps to complete this project.

1. Clean the data: The data cleaner reduced the dimensionality of the data to two dimensions to be able to plot and see the graph. Later we apply the algorithms on all dimensions of the data. I also removed the ID and Ground Truth clusters columns too.
2. K-means Implementation: In this section, I implemented the K-means algorithm. I used different sources to complete this code. GeeksForGeeks' website was much helpful in this section.
3. Then, I called the K-means algorithm on the Cho dataset which is called X_std after being cleaned and reduced dimensions. In this first implementation, I used the reduced dataset which is just two columns. We can see in the bellow graph how the data is being clustered later we will find what is the number of clusters. I choose the current number of clusters based on the dataset ground truth column.



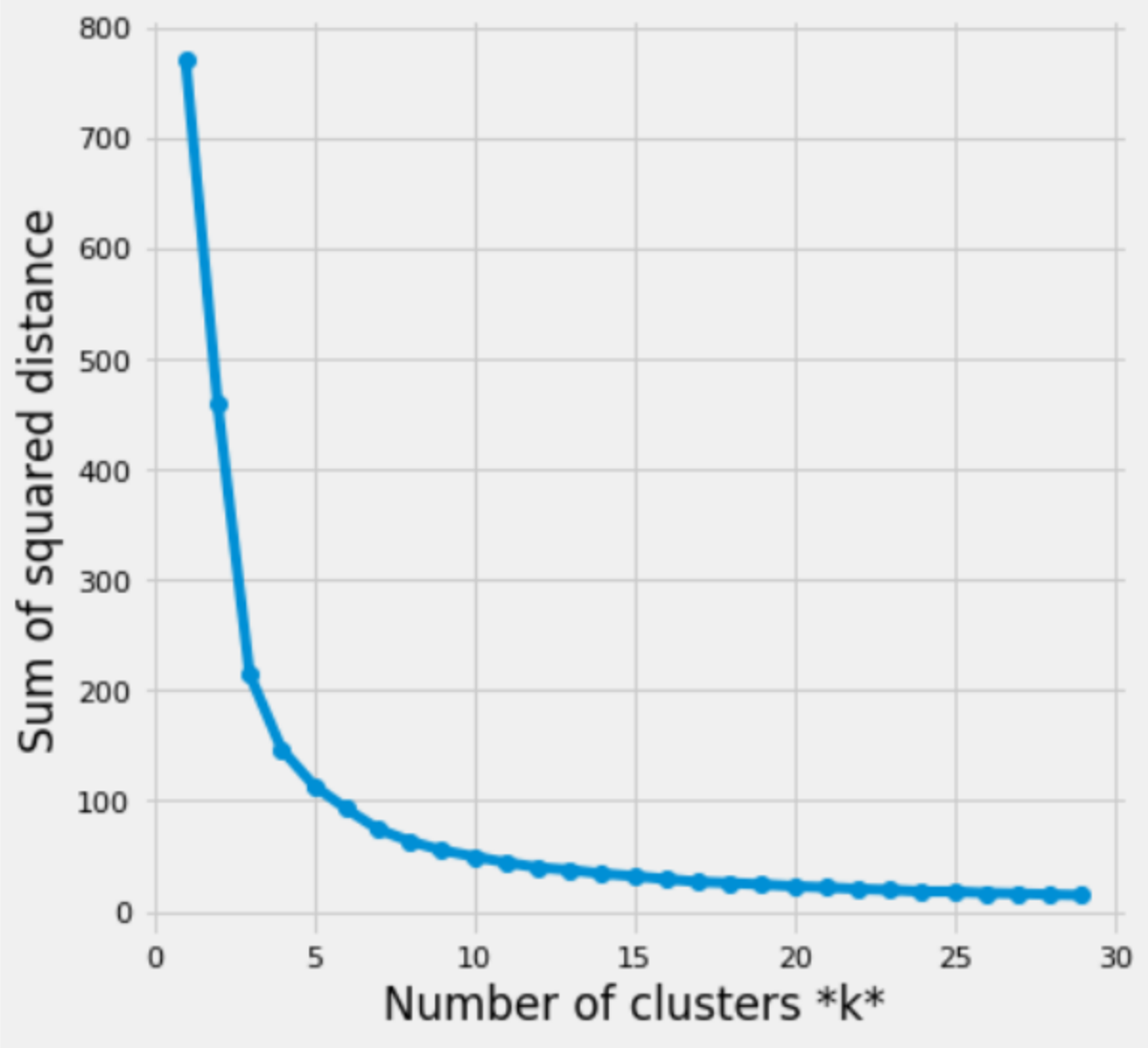
4. Now we call the K-means algorithm on the Iyer dataset which is called Z_std after being cleaned and reduced dimensions We can see in the bellow graph how the data is being

clustered later we will find the number of clusters. I choose the current number of clusters based on the dataset ground truth column.

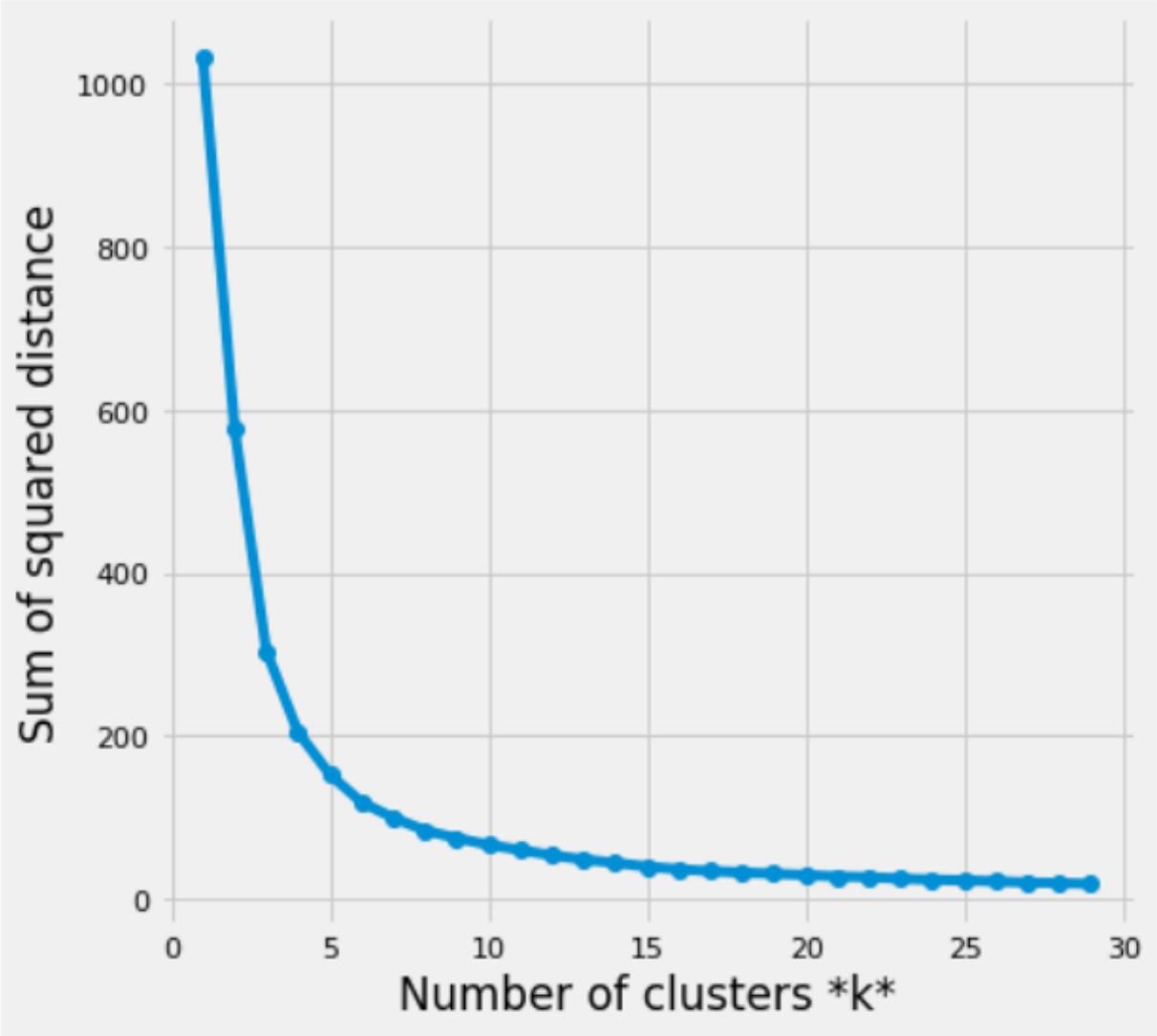


Evaluation:

5. Let's find out what is the right number of clusters to cluster this dataset. The graph below shows that 4 or 5 can be a good number of clusters for the Cho dataset.



6. We can say that any number between 5-10 can be a good number of clusters. while the ground truth clusters are 10, I don't recommend 10 clusters. somewhere around 7 could be a middle ground.



7. Now let's do the silhouette analysis to find out how good is our clustering?

   Clustering validation Silhouette plot: The silhouette coefficient (Si) measures how similar an object i is to the other objects in its cluster versus those in the neighbor cluster. Si values range from 1 to - 1: A value of Si close to 1 indicates that the object is well clustered. In other words, the object i is similar to the other objects in its group. A value of Si close to -1 indicates that the object is poorly clustered, and that assignment to some other cluster would probably improve the overall results. Please look at the codebook to

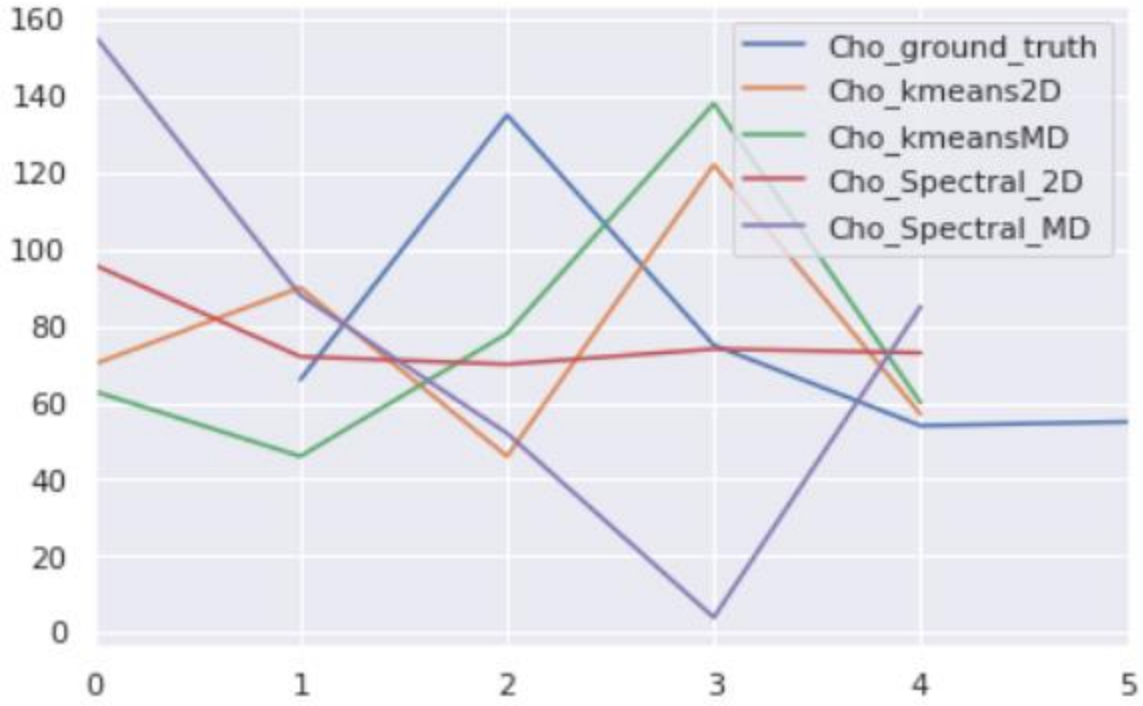
get a good grasp of the best number of clusters and different clusters and silhouette coefficients.

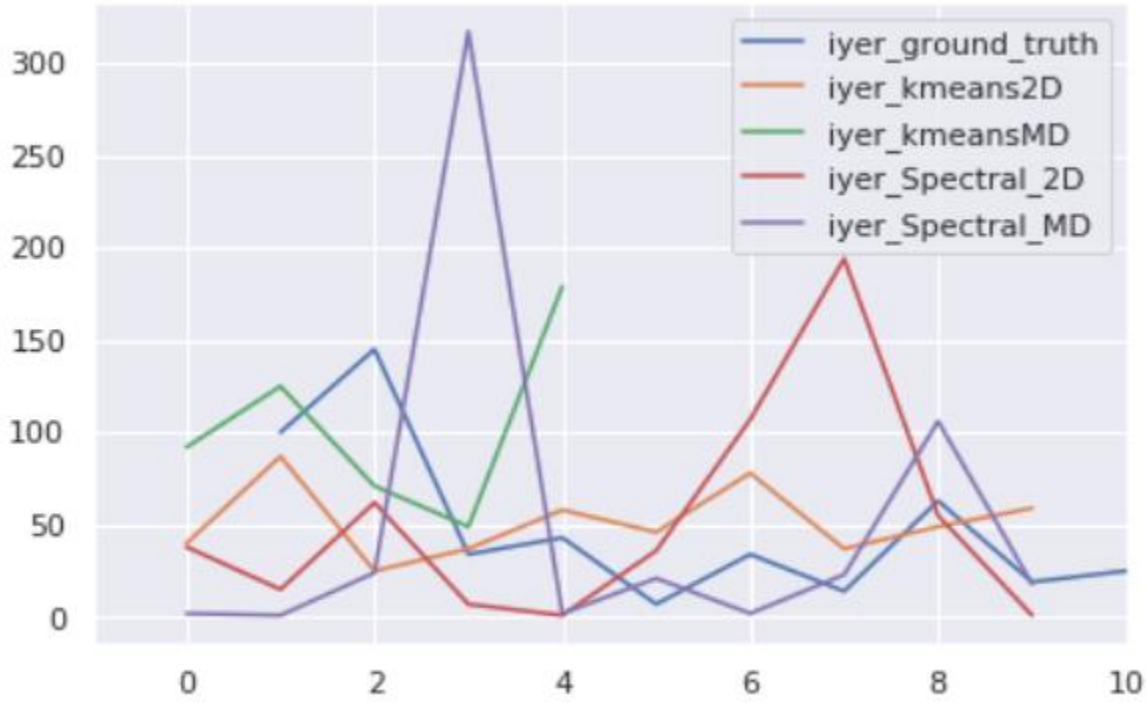
8. Spectral Clustering: most of the codes of this implementation were borrowed from the "Machine-Learning-From-The-Scratch" git hub repo.
9. Then I fed data in the function to implement spectral clustering
10. Then I fed the out of the Spectral Function to K-means
11. I did the same process on the second dataset
12. Then I applied both algorithms on all of the data not just reduced dimensions in the following way:
    a. Cleaned the data but doesn't reduce the dimensions of it.
    b. Implemented local implementation of k-means on full dimensions of both dataset

Here are the results of two algorithms on both datasets in two dimensions and full dimensions mode.

The results show that:

1. K-means is nearer to the ground truth on the Cho dataset, but multi-dimension had a closer performance.
2. But in the Iyer dataset Spectral clustering shows better results.


Cho_ground_truth
Cho_kmeans2D
Cho_kmeansMD
Cho_Spectral_2D
Cho_Spectral_MD
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5


iyer_ground_truth
iyer_kmeans2D
iyer_kmeansMD
iyer_Spectral_2D
iyer_Spectral_MD
300
250
200
150
100
50
0
0
2
4
6
8
10