ASSIGNMENT-6

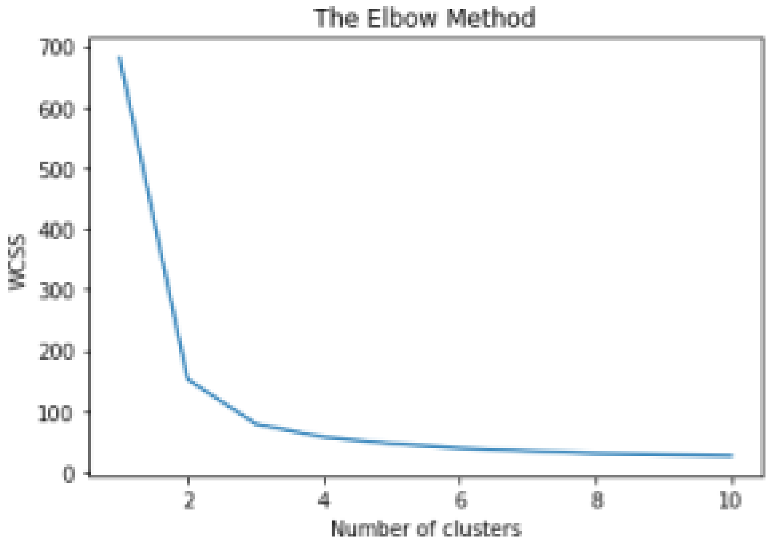
Arsh Pratap

2018IMT-021

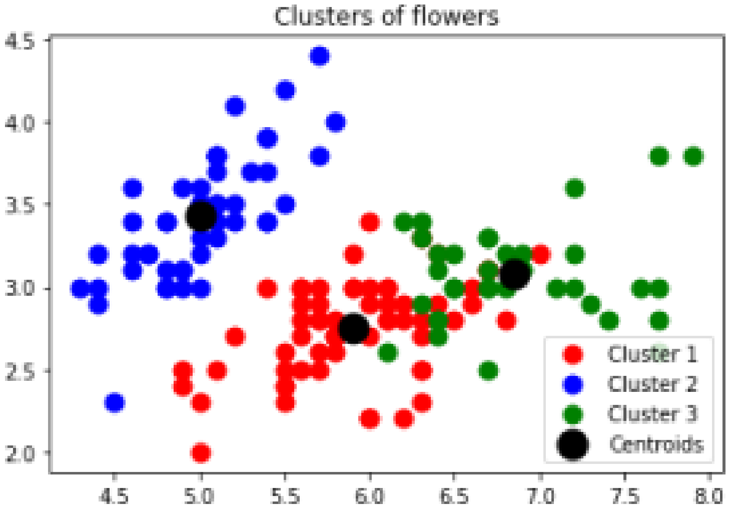
Note - The code for the assignment can be found here

1. **Considering the IRIS dataset discussed in the previous assignment, apply the EM algorithm to cluster the data (without considering the output labels) Use the same dataset for clustering using the K-means algorithm. Compare the results of these two algorithms.**

   We are actually adjusting the number of clusters (K) in the Elbow approach from 1 to 10. We calculate WCSS for each value of K. (Within-Cluster Sum of Square). WCSS is the sum of squared distances between each point and the centroid in a cluster. When we plot the WCSS with the K value, the plot looks like an Elbow. As the number of clusters increases, the WCSS value will decrease. When K = 1, the WCSS value is the highest. When we examine the graph, we can see that it will shift rapidly at a point, forming an elbow shape. The graph begins to travel practically parallel to the X-axis at this point.



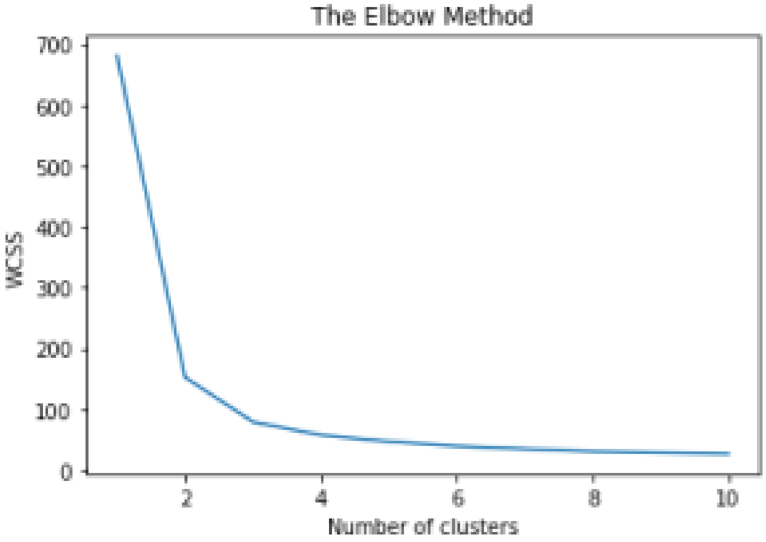
Since the elbow lies at approximately 3 on the x-axis. We can conclude the no of clusters is 3.

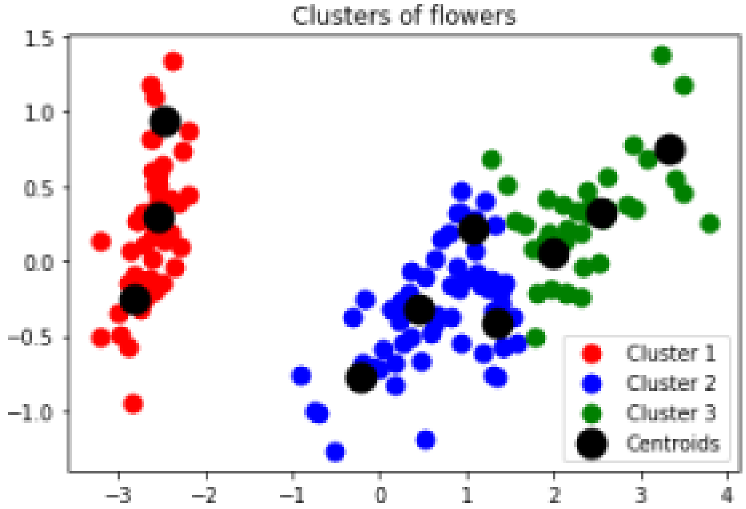


An insight we can get from the scatterplot is the model's accuracy in determining Cluster 2 is comparatively more to Cluster 1 and Cluster 3.

2. Apply PCA algorithm to obtain the first two principal components and perform the clustering using both algorithms on the resultant data. Compare the results of these

two algorithms.





An insight we can get from the scatterplot is the model's accuracy in determining Cluster 1 is comparatively more to Cluster 2 and Cluster 3.

Accuracy of K-means and EM models

1. The accuracy of the K-Mean model is: 0.27
2. The accuracy of the EM model is: 0.31666666666664

Accuracy of K-means and EM models on applying PCA

1. The accuracy of the K-Mean model with PCA is: 0.86666666666667
2. The accuracy of the EM model is: 0.94

Inference:

In both raw data and PCA data, it can be demonstrated that the EM approach behaves and performs better than the K-means model (dimensionally reduced data). The EM Algorithm is a promising alternative to standard k-means clustering in semi-supervised learning. To provide reliable solutions, it finds multivariate Gaussian distributions for each cluster.