Assignment-4

Ginjala Lohith Reddy

June 3, 2023

1 What I Have Learnt

I have learnt about different methods of clustering like K-means, hierarchial. I have also learnt what and how to perform K-means clustering, Hierarchial clustering using **sklearn.cluster** library in python. I have learnt about Within-Cluster Sum of Squares(**wcss**), elbow method(which is used to find the appropriate k value for k-means clustering for a given dataset) and forming clusters. I have learnt how to perform **Agglomerative Clustering** and plotting different types of **Dendrograms** using **scipy** and **sklearn** libraries. I have learnt how to do image segmentation by converting image into appropriate data and performing k-means clustering on it and deriving the output.

2 K-Means Clustering

An unsupervised machine learning approach called **K-means clustering** is used to divide a dataset into a preset number of clusters. Iteratively assigning data points to clusters based on their proximity to cluster centroids seeks to minimise the within-cluster sum of squares(**wcss**). In unlabeled datasets, K-means clustering is frequently used to group similar data points and identify patterns.

I have used inbuilt **KMeans** function from sklearn library to perform K-Means clustering on the dataset. Then I have written function to compute wcss values from **kmeans.inertia_** function and after computing I have plotted the wcss values and we can see the plot in colab notebook file. We can see there is sharp bend near $k = 3$ and from k = 4 to k = 24 the difference in distance is also decreasing so hence the desired k is **3**. I have also plotted the points as clusters from the derived k value from elbow method(k = 3) with the centroids of clusters(shown as '**x**') as shown in the output in notebook.

3 Hierarchical Clustering

Unsupervised machine learning algorithm called "**Hierarchical Clustering**" groups data points into hierarchical cluster structures. It creates a structure resembling a tree called a **dendrogram** by repeatedly merging or dividing clusters based on how similar the data points are. Hierarchical clustering can be used for exploratory analysis to show the underlying structure of data because it doesn't require a predetermined number of clusters. It is helpful for identifying hierarchical patterns in the data and for visualising links since it allows for flexible interpretation of clusters at various granularity levels.

Dendrogram: A dendrogram is a graphical representation of hierarchical relationships among data points or clusters in hierarchical clustering. It is a tree-like structure that depicts how clusters or individual data points are merged or divided during the clustering process. Firstly, I have performed **Agglomerative Clustering** using the function **AgglomerativeClustering** imported from **sklearn.cluster** library. Then I have plotted **Dendrogram** using **dendrogram** function in **scipy.cluster.hierarchy** library from **linkage_matrix**. The dendrogram is shown in the notebook for given dataset array. I have then divided points into 3,4,5,6 and 7 clusters using **AgglomerativeClustering** function and stored the labels of points in dataset. I have then computed labels of points in dataset using **fcluster** function in **scipy** library from dendrogram. I have printed the both types of labels and we can see that there is an one-one mapping between the values of labels. There is change in cluster number but points belonging to cluster remains same.

Single Linkage Dendrogram:A single linkage dendrogram is a visual depiction of hierarchical clustering in which the clusters are linked based on their nearest neighbouring pair of data points. The smallest distance between any two data points from separate clusters is used to determine how clusters are merged. The generated dendrogram demonstrates the hierarchical relationships and the gradual fusion of groups based on their proximity.

Ward Linkage Dendrogram: The linkage between clusters in a Ward linkage dendrogram is based on minimising the increase in total within-cluster variance, which is a graphical depiction of hierarchical clustering. It demonstrates how clusters are combined to form evenly spaced, compact clusters with comparable sizes.

Complete Linkage Dendrogram: The linkage between clusters in a complete linkage dendrogram is based on the largest distance between any two data points from distinct clusters. This is a graphical depiction of hierarchical clustering. In order to create compact, spherical clusters, it demonstrates how clusters are combined based on the furthest pair of data points. Finally, I have plotted different types of dendrograms(single, ward, complete) by changing the method of linkage while computing the linkage_matrix for plotting of dendrogram. I have observed that dendrograms obtained by complete, ward linkage are similar but with single linkage there is a significant difference.

4 Image Segmentation

Image segmentation: Image segmentation is the task of partitioning an image into distinct regions or objects to facilitate analysis and understanding by separating foreground from background or identifying different objects within the image.
I have taken an image from my collection and used **OpenCV** library to convert image to RGB values tuples. After that I have changed the dimensions of my dataset to required shape I will scale down the values by **255** of dataset which is suitable to perform K-Means Clustering. The provided code uses various values of **k(1,3,5,7)** to conduct **K-means clustering** on an image. The code uses K-means clustering for each k value, labels each pixel with a cluster label, and then reconstructs the segmented image using the **cluster centres**. Following that, **Matplotlib** is used to display the segmented images. The code's goal is to investigate how changing **k** will affect the outcomes of picture segmentation. Different levels of image detail and segmentation can be achieved by varying the number of clusters. The titles of the subplots containing the split photos display the k-value used for clustering. This code offers a visual breakdown of the K-means clustering with various k values used for picture segmentation. It makes it possible to compare the outcomes and see how the number of clusters affects the clarity and level of information in the segmented images.