UBIT	CVIP Project - 3	UB-mail
anand6	Task2	anand6@buffalo.edu

K-means clustering algorithm:

Unsupervised machine learning algorithm where we start clustering data based on k clusters as we don't have true label to validate that so we have to know the dataset thoroughly to start with optimal k. It's hard clustering algorithm where one data point belongs to only one cluster by enhancing K-means with probabilistic approach where we can predict the probability of the data point belonging to the specific cluster there by achieving the soft clustering.

Advantages:

- Easy to implement
- Scalable to larger datasets
- Guarantees convergence after specific number of iterations

Disadvantages:

- Time consuming for convergence if initial centroids are away from optimal
- Choosing k manually and to find optimal k we have to run for different k and choose best k
- Prone to outliers

Algorithm:

Step1: Manually assign k
Step2: Initialize k centroids randomly from the given dataset
Step3: Iterate till max iterations or convergence:
find distance from each data point to centroids
and assign centroid to data point with minimum distance

Recalculate the new centroids with mean of data points
belonging to the cluster.

Note: If k is large then convergence will take longer as data points keep assigning to different clusters on every iteration till it converges. There are many ways to improve efficiency of k-means to choose optimal initial centroids one of the methods is "Naive Sharding Centroid Initialization".

Here for color quantization : Instead of parallel processing and optimal initialization of centroids, maximum number of iterations is used as stopper for k-means.

Maximum number of iterations – 50

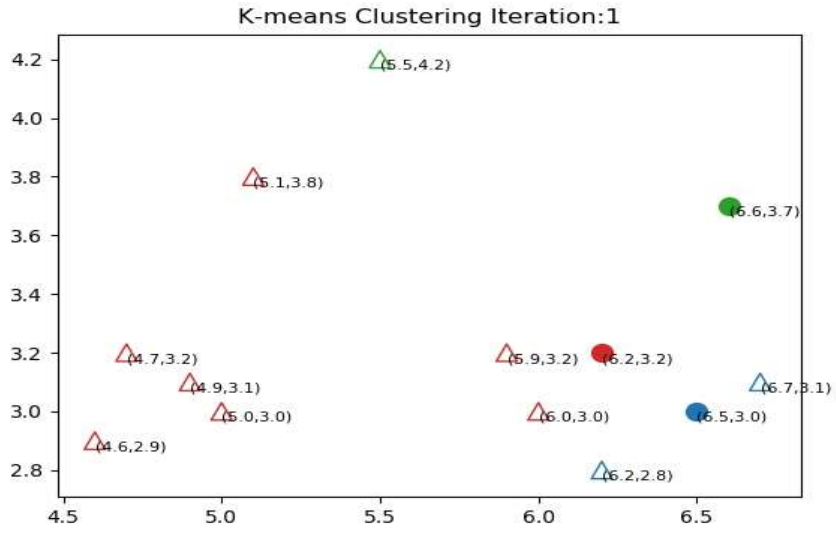
K-means on given dataset:

First Iteration:

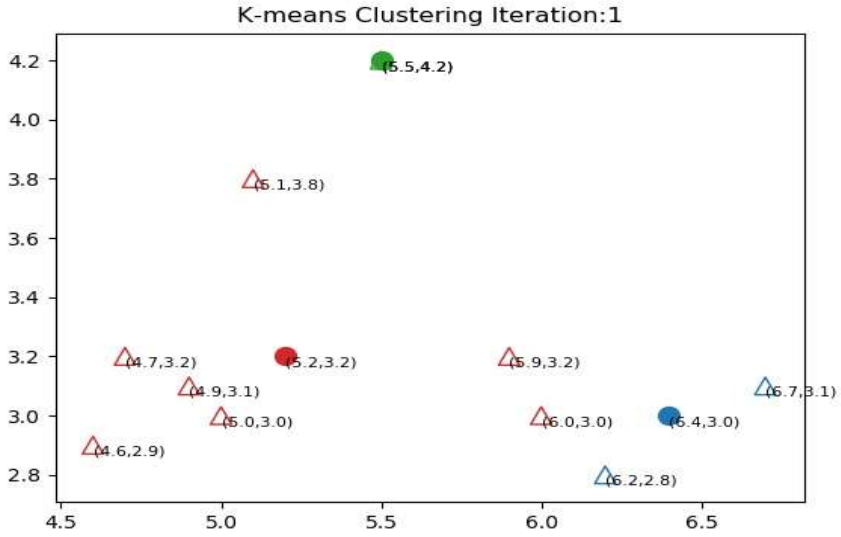
Classification Vector :

```
Classification Vector for iteration  1 : [0 0 2 0 1 0 0 2 0 0]
```

task2_iter1_a.jpg – Shows how data points move to new clusters based on distance



task2_iter1_b.jpg – Shows how new centroids move towards the new clusters.

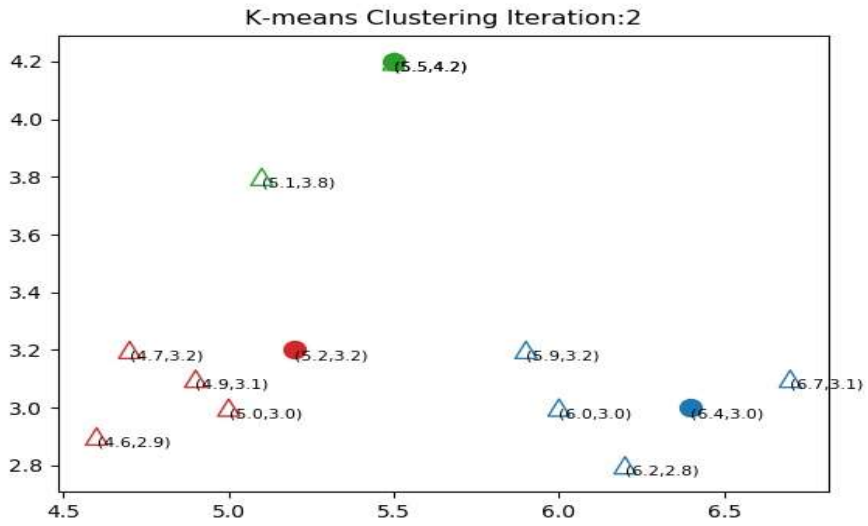


Second Iteration:

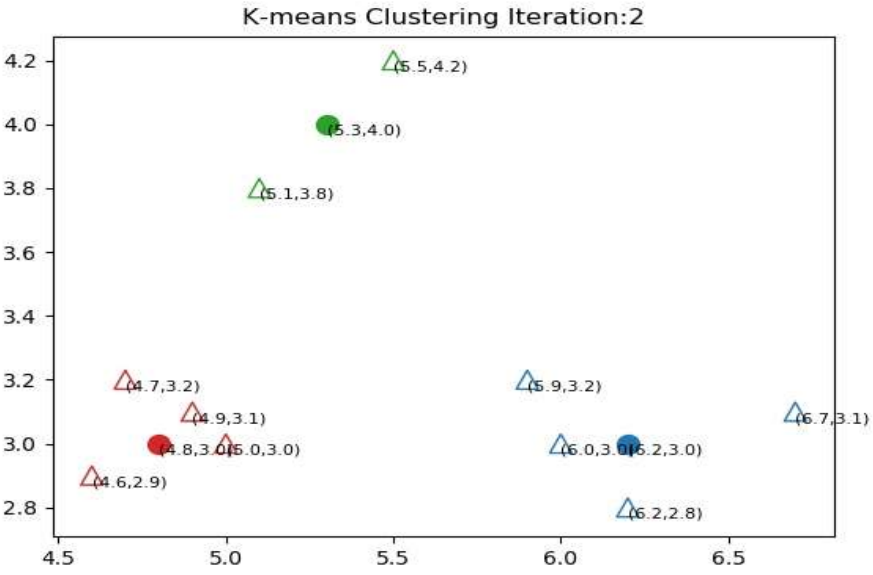
Classification Vector :

```
Classification Vector for iteration  2 : [2 0 2 0 1 0 0 2 1 2]
```

task2_iter2_a.jpg – Shows how data points move to new clusters based on distance



task2_iter2_b.jpg – Shows how new centroids move towards the new clusters.



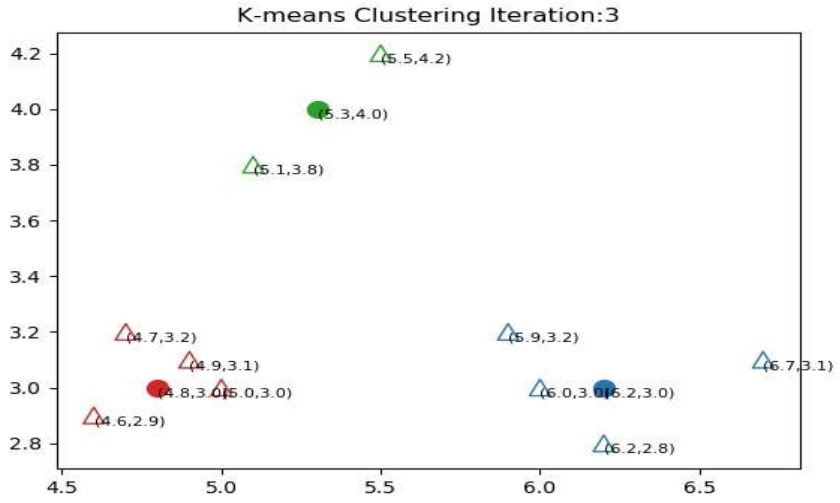
For given dataset and initial centroids it converges in 3 iterations:

Third Iteration:

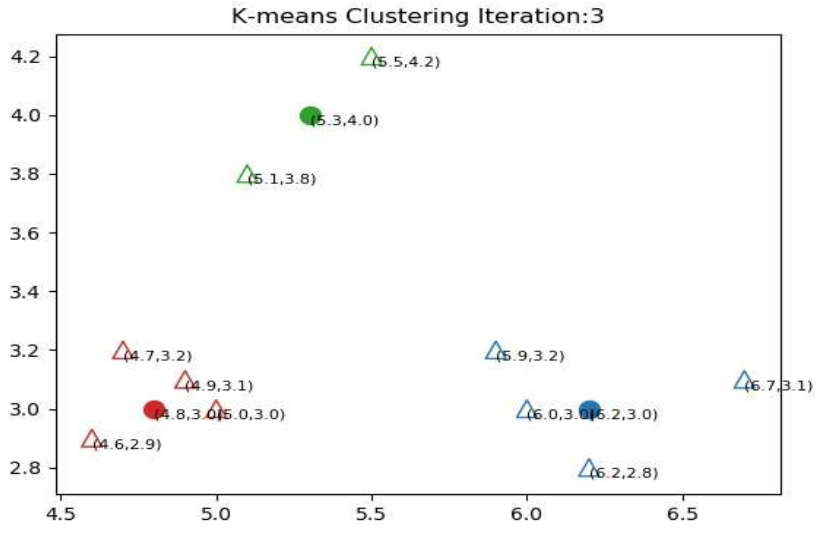
Classification Vector :

```
Classification Vector for iteration  3 : [2 0 2 0 1 0 0 2 1 2]
```

task2_iter3_a.jpg – Shows how data points remain in the same cluster



task2_iter3_b.jpg – Shows how new centroids don't change remain same- It converged.



Color Quantization:

Process of reducing the number of distinct colors used in the image and the new image should be visually similar to the original image. Its used in image compression. Here we use k-means algorithm to perform color quantization of the given image.

We have to find the k colors to represent the image and value of k varies from 3,5,10,20. Use those k-colors to create image which will be compressed version of the image.

For given image we have – **`347633`** unique colors representing the image(RGB encoding 3 dimensional). We have to reduce number of colors to represent image with just k colors and to id entify these k colors we will be employing k means algorithm.

Procedure:

1. *Initialize k centroids chosen randomly from the 347633 colors and call k-means algorithm*
2. *Initial centroids form cluster of colors and then new centroids are created based on the colors in the clusters till it converges*
3. *As number of colors are more it takes much time when value of k is large to converge so we will be using the max number iterations as 50 to stop iterations if convergence takes more time.*
4. *We use these centroids of clusters as the color replacement of all member colors in the cluster and then create image just with these new centroid colors.*

Results:

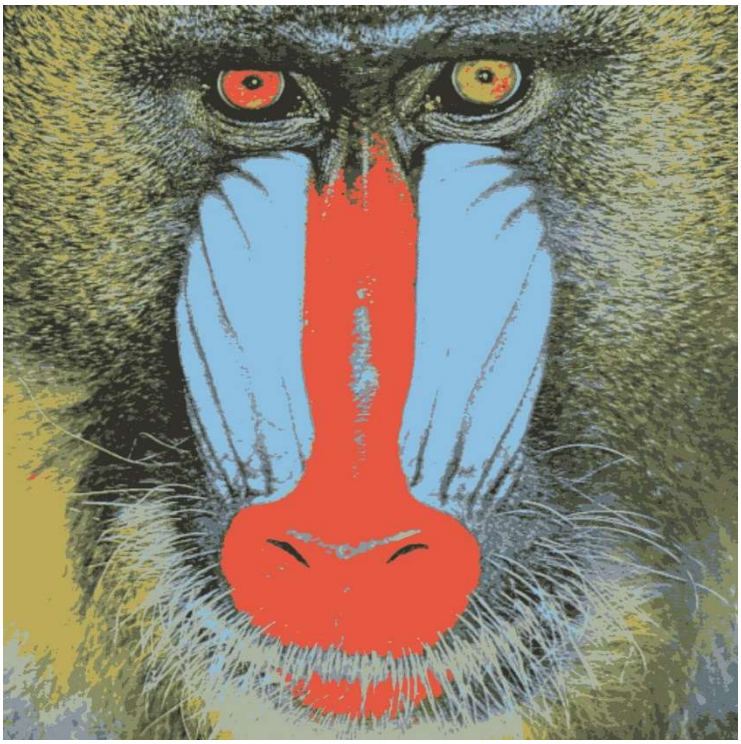
```
Number of unique colors used to represent current image: 347633

Number of colors:  3
Quantized color values:
[[ 97 123 136]
 [173 121 132]
 [145 177 169]]
```

Number of colors: 5

Quantized color values:
```
[[ 73 125 134]
 [169 121 154]
 [124 121 137]
 [186 122 112]
 [144 181 169]]
```

Number of colors: 10

Quantized color values:
```
[[ 87 123 136]
 [157 126 122]
 [147 119 150]
 [184 122 137]
 [179 123 172]
 [115 121 145]
 [192 120 106]
 [124 124 122]
 [144 182 169]
 [ 52 128 132]]
```



Number of colors: 20

Quantized color values:

```
[[ 65 126 135]
 [151 124 103]
 [127 120 133]
 [163 117 154]
 [155 136 177]
 [108 120 144]
 [187 119 102]
 [108 126 116]
 [160 158 136]
 [ 39 130 128]
 [190 116 151]
 [116 159 154]
 [200 123 117]
 [134 120 155]
 [147 179 159]
 [187 125 179]
 [143 188 180]
 [176 124 130]
 [150 119 133]
 [ 87 122 137]]
```