In The name of God

University Tehran

Engineering Facility

Electrical and Computer Engineering

Pattern Recognition

HW # 7

Amin Fadaeinedjad

810195442

Spring 98

List

Abstract

In this homework we are going to see different methods of clustering like agglomerative hierarchical clustering, sequential clustering, K-mean clustering the Separation index which is a number that shows us the goodness of a clustering and at the last part we used the K-mean algorithm on an image to see that we can use clustering to reduce the size of the image so in this case we are going to memory.

Question 1

In this problem we used the predefined function called $AgglomerativeClustering$ this function will use Agglomerative hierarchical clustering to cluster the data that we have and we have two data sets of train and test and for clustering it doesn't matter this two data sets because are only going to cluster the data.

This algorithm is like this at first we have the data point and we look at the data set that we have we choose the two data points that are so close to each other and we do this until all the data point are clustered and in that case we are in the end and in the end all the data points are clustered together.

The accuracy is: 74.15%

And the confusion matrix is Figure 1.

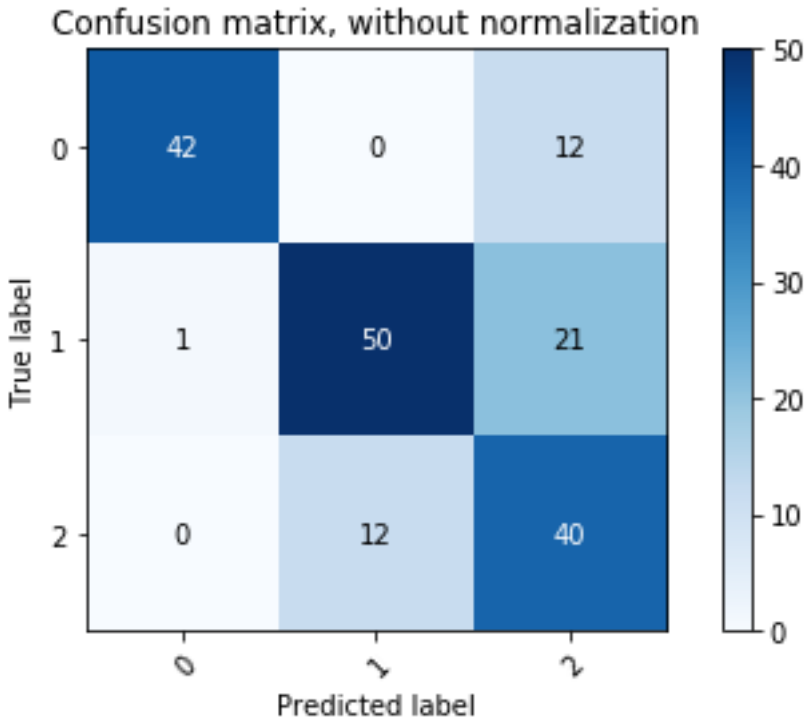

Figure 1

We can see the result of the clustering and the accuracy was 74.15% and that is quit a good number.

The average distance of the center of the clusters and the data sets are:

label 1 average distance 141.11171

label 2 average distance 61.09111

label 3 average distance 88.465515

and the distance of the center of the center of the data set is

distance 1 and 2: 747.186

distance 3 and 2: 258.52728

distance 1 and 3: 488.7512

And we can see from the confusion matrix that the clustering was quite good and the mean distance of the canter and the data of that label is much smaller of the distance of the centers of the labels.

Question 2

In this part we are going to use sequential algorithm to cluster the data that we have and we have this results that are mentioned below.

In this algorithm will start like this at first we start will a data point and make one cluster and after that we compare the next data points with the previous one and if it has a threshold that is acceptable we put it in the cluster if not be ignore that data and make another cluster and we keep on doing it until all the data points are in a cluster.

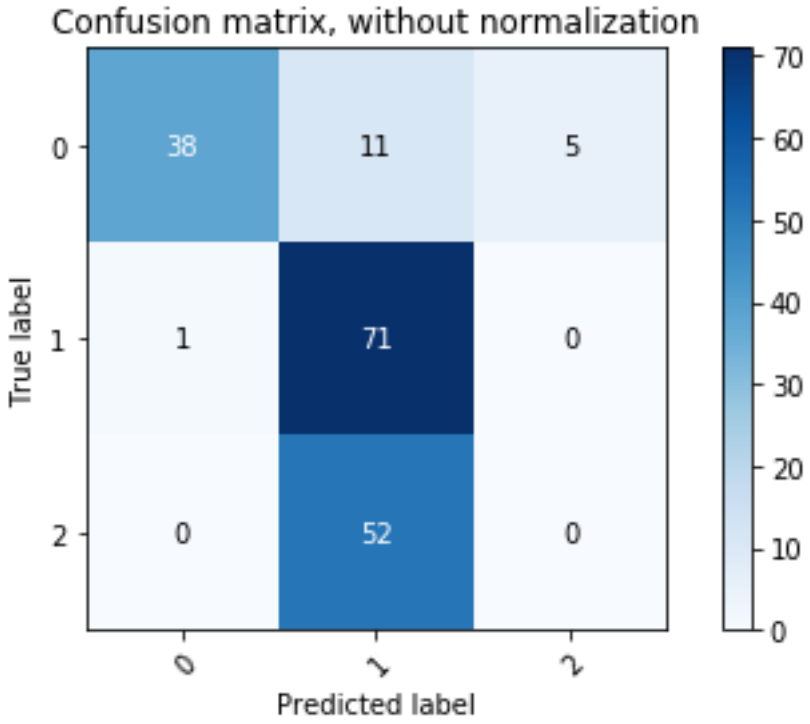

Figure 2

The accuracy of this clustering is 61.23%

The average distance of the center of the cluster and the center of that cluster is:

label 1 average distance 111.44512

label 2 average distance 130.17384

label 3 average distance 11.015416

And we can see the distance of the center of clusters

distance 1 and 2 560.556

distance 3 and 2 941.54504

distance 1 and 3 381.04678

And we can see from the confusion matrix that the clustering was quite good and the mean distance of the canter and the data of that label is much smaller of the distance of the centers of the labels.

Question 3

In this question we are going to use the $K-mean$ algorithm to determine the clusters of the data points that we have at the first part we used the script code that was written by myself and for the next part we used.

a) For the first part we have to set a random value for the center of the clusters and what I have done was that I declare 3 variables which one was the average and the another was the maximum and the last one was minimum of all the data points that we have.

$$\mu_1 = \min(all_{data})\ \ \mu_3 = \min(all_{data})\ \ \mu_2 = avg(all_{data})$$

And after that we start the algorithm of this question and choose the nearest cluster for each data point and after clustering all the data points we should calculate the new center of the following clusters and do all this algorithm again and again until we reach a point that no data point from one cluster changes in next step and then the following clusters are the final ones and we can see the result from out code.

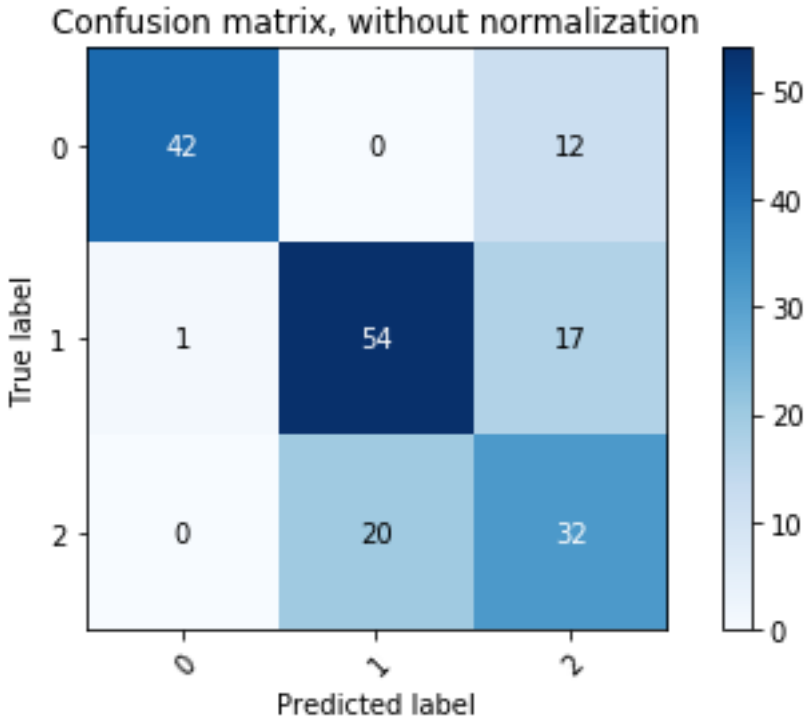

Figure 3

The accuracy of the clustering is 71.90% which is a god accuracy

The average distance of the data point and the centers are:

label 1 average distance 726.40216

label 2 average distance 263.30646

label 3 average distance 463.49734

And the distance of the center of labels are:

distance 1 and 2 263.07245

distance 3 and 2 463.2886

distance 1 and 3 726.2675

Time computing: 0.09375

b) In the second part we are using predefined function in python library and we get some results, the following are the result for the predefined functions.

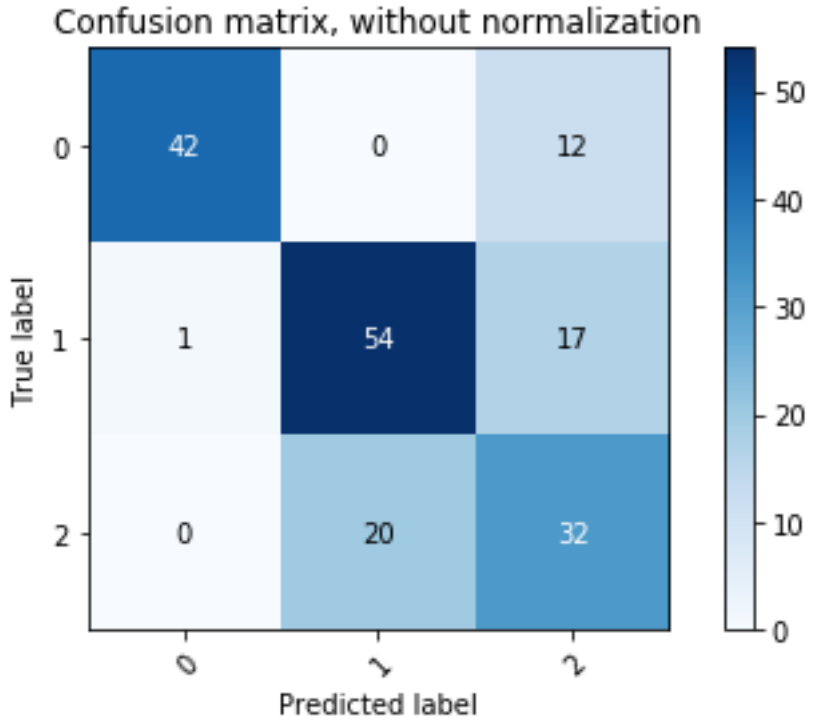

Figure 4

The accuracy of the clustering is 71.90% which is good accuracy.

The average distance of the data point and the centers are:

label 1 average distance 141.11171

label 2 average distance 68.4817

label 3 average distance 81.927315

And the distance of the center of the clusters to each other is:

distance 1 and 2 726.2675

distance 3 and 2 263.07245

distance 1 and 3 463.2886

Time computing: 0.078125

c) We can now compare the result of the two clustering the good thing is that the accuracy of the two clustering are the same which means that we have done a good thing in the code and the confusion matrix are the exact same thing which is also a good thing, and we need to compare the other parameters with each other to see the difference and similarities.

One of the differences is the average distance of the data and the center of those data which the predefined function has done it better than me and also the distance of the center of the clusters are quite closer in the predefined function.

And also the speed of the clustering n the predefined function are faster than our code.

So after comparing the two clustering we can see we have done well in it and have a good accuracy, as good as predefined ones but the other condition like the mean distance the predefined have done it better.

Question 4

In this part we are going to use the Separation Index to show us the performance and the accuracy of the clustering and the duration that takes to do it.

Method	Separation Index	Accuracy	Computing Time
Agglomerative	6.436	74.15%	0.3125
Sequential	6.62	61.23%	1.0625
K-means	2.97	71.91%	0.3281

We can parameters of the clustering and the performance of the three different methods of the clustering and we can see that the agglomerative method is better than k-mean because it has better accuracy and a better separation index but the sequential method has the biggest separation index which means it has better performance in separating the clusters.

$$SI = \min\{\underbrace{\min}_{i!=j}\{\frac{d(Si\,, S_j)}{\max d(S_l, S_l)}\}\}$$

Question 5

In this part we are going read the image and after that we going cluster the pixels as out data point and the three dimension RGB as our features and what we have done we used the method of K-mean algorithm to cluster the data of all the pixels in the image that we have and we have done this with different number of clusters and it is obvious that the bigger the K is the bitter image will be.

The reason that this method will help to reduce the image size is that we have no need to save all the $256 * 256 * 256$ possible data we can save the index of the cluster and we use the average of the cluster as the cluster point and so we only have to put one number with is between 1 to K to show the meaning of the image.

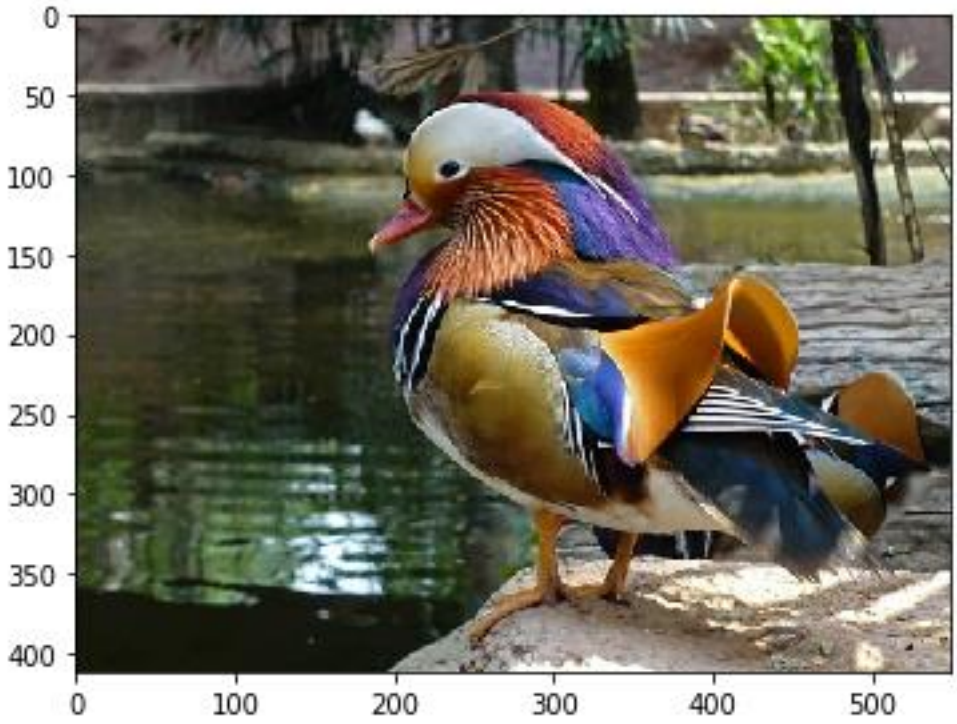

Figure 5 Main image

As we see Figure 5 is the main image that we have and every pixel will need a number of bits to represent the image and it has $256 * 256 * 256$ case that I might be.

K=13:

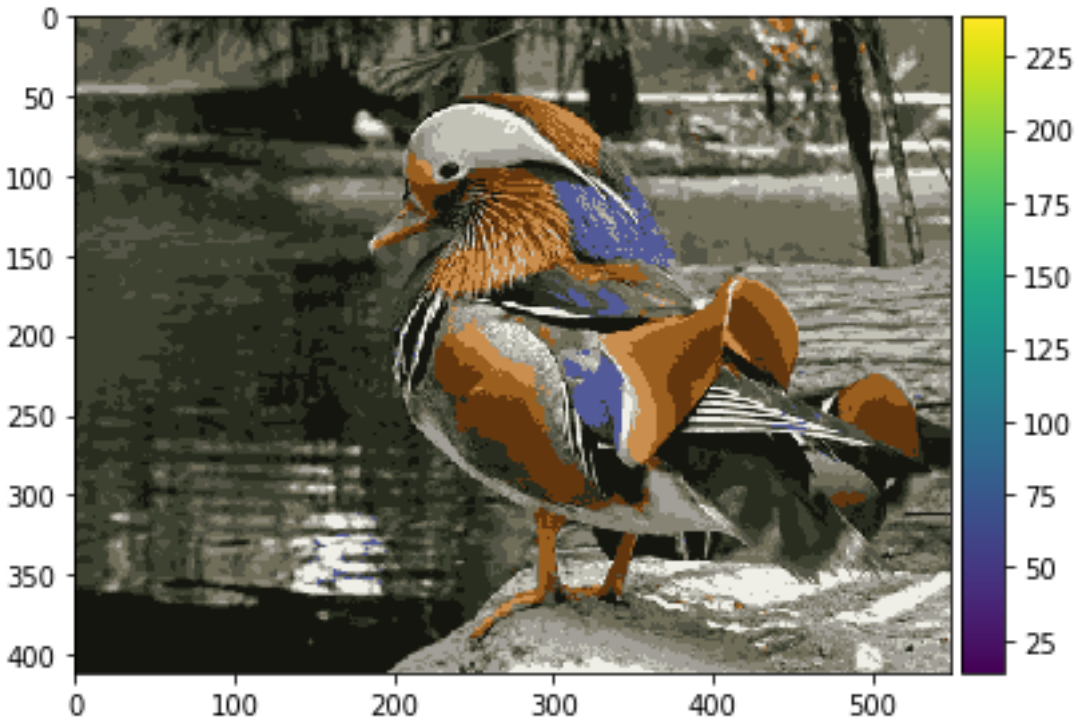

Figure 6 K=13

There are 13 different centers if the clusters that we have.

[array([61.25606461, 66.29746607, 49.21760538]),
array([161.71240992, 159.92213265, 151.54965189]),
array([239.05595409, 239.43285509, 232.06685796]),
array([20.76986652, 21.85120199, 15.24790922]),
array([202.32234273, 142.79392625, 79.69132321]),
array([135.43942389, 132.5970065 , 118.12386332]),
array([84.7340673 , 85.76269215, 70.58317408]),
array([100.02670378, 54.30776369, 14.23414785]),
array([82.11104895, 90.78097902, 153.7351049]),
array([40.77816437, 45.33965599, 31.6994304]),
array([195.21102554, 194.69928625, 182.43895567]),
array([154.79114763, 94.23800689, 30.39663398]),
array([112.03812825, 109.39145003, 88.02780666])]

K=20:

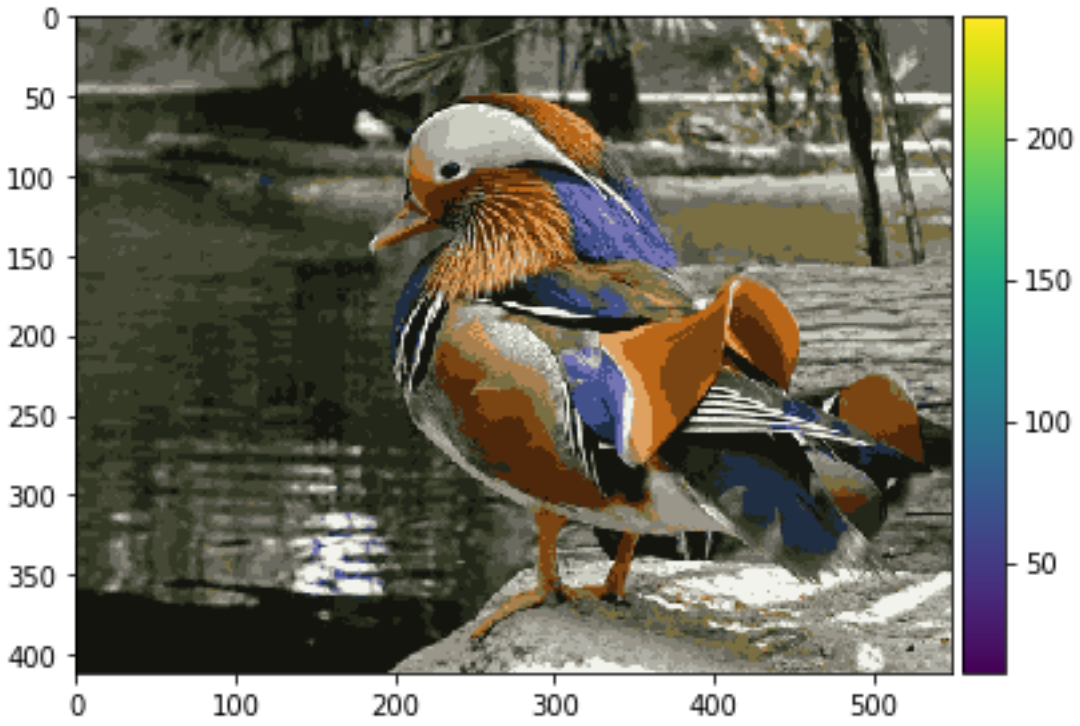

Figure 7 K=20

[array([154.85419451, 151.79888641, 138.8047513]),
array([35.023157 , 39.85506692, 26.66879116]),
array([106.15695039, 106.97777364, 97.11100764]),
array([242.75077667, 243.23023818, 236.28702106]),
array([31.96771175, 46.02620496, 67.25924193]),
array([51.68410406, 57.16324797, 37.85952988]),
array([175.94133532, 175.83406276, 166.9070677]),
array([168.76367443, 128.37189217, 79.00026171]),
array([123.82892275, 69.75126971, 19.38104785]),
array([86.95191945, 87.51219512, 73.7674043]),
array([116.26395409, 115.10693792, 180.00156495]),
array([205.81162123, 206.44648494, 194.66025825]),
array([121.11312349, 111.13598063, 66.13230024]),
array([79.72537449, 40.82009381, 11.92162203]),

array([19.01832298, 20.1380823 , 14.13396739]),

array([186.35843281, 102.49433731, 24.22742577]),

array([68.42126081, 72.17843745, 53.06638395]),

array([66.39300938, 81.13682864, 140.00809889]),

array([130.31090108, 128.64567936, 114.66131325]),

array([223.37249782, 161.8537859 , 96.92167102])]

And this are the center of the 20 clusters that we have.

K=6:

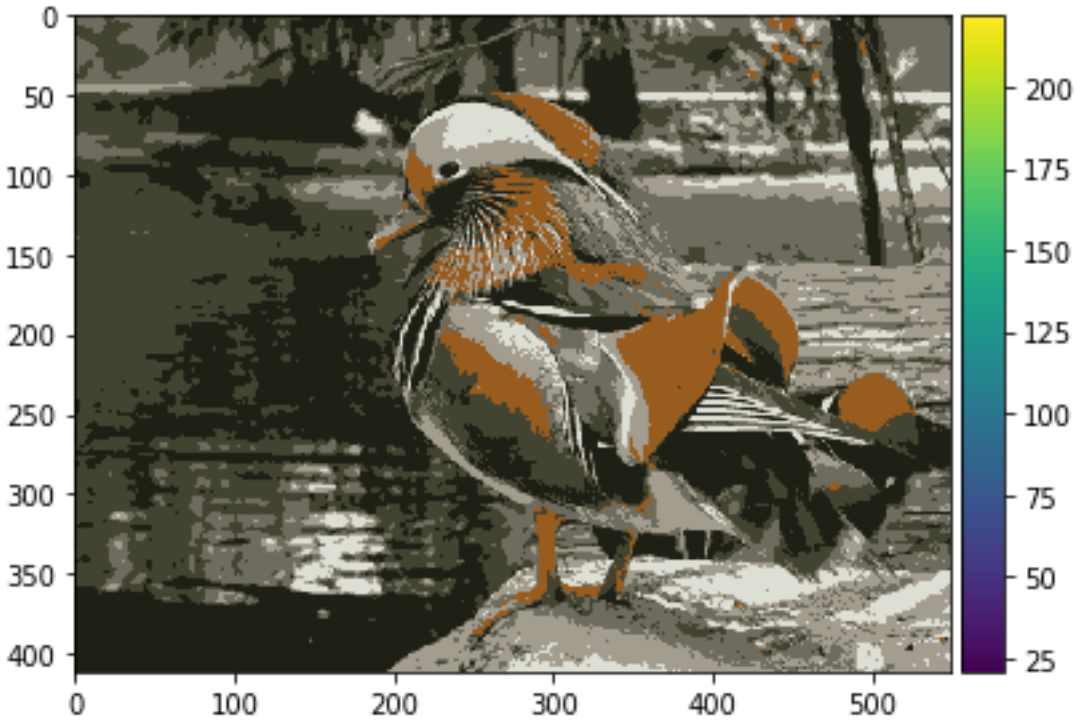

Figure 8 K=6

[array([30.03864669, 31.68955522, 21.89970506]),

array([152.32080407, 92.57450306, 31.99686146]),

array([109.12053957, 108.09678575, 94.86847929]),

array([163.26702989, 157.0500934 , 143.86562889]),

array([66.54332998, 67.13354663, 48.8633839]),

array([221.77122984, 222.32821218, 212.87716885])]

K=3:

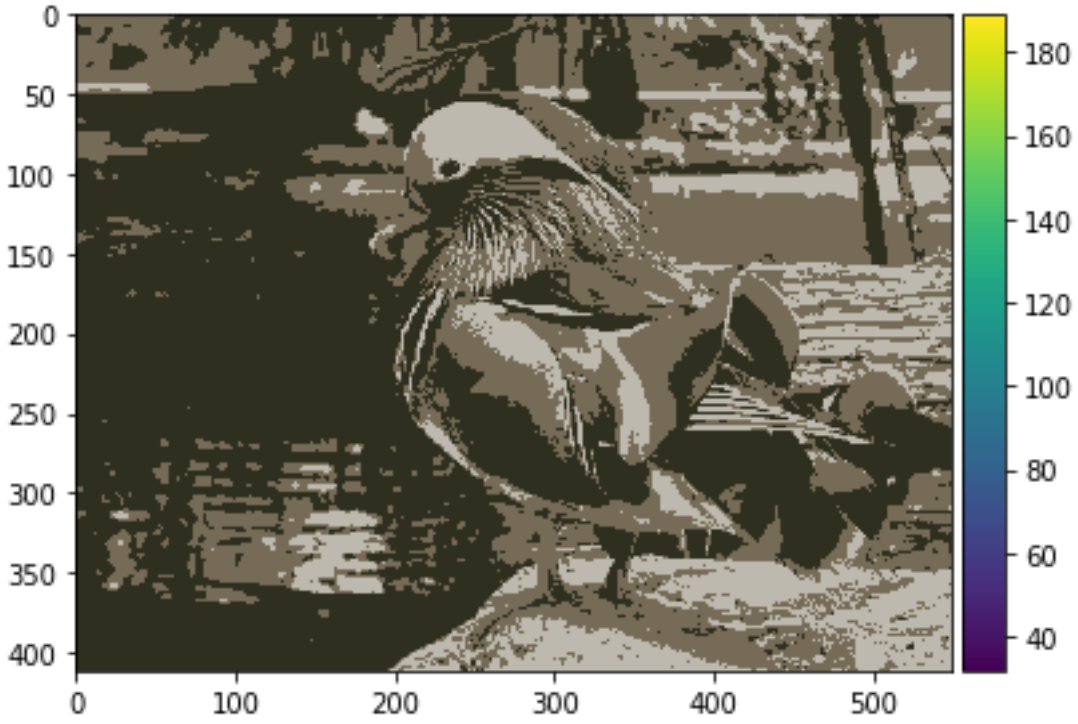

Figure 9 K=3

[array([117.25025051, 107.25437425, 86.29545747]),

array([47.03579091, 47.02289129, 32.42934021]),

array([189.03072517, 185.49371594, 174.08286064])]

As we see the result of the clustering the K near amount of 15 has a good result but we must know that the bigger the K is the more time it takes to compute the new image so there is a kind of a tradeoff between time and the quality of the image that we have.

The smaller the K is the faster we can make the new image.

Process

In this Homework are going to use different methods to cluster the data points that we have and for this we learned different methods to do this and also different ways to compare the performance of the clustering, and we can us clustering to reduce the size of data that we need and use less memory.

Reference

Automatic speech recognition a deep learning approach

stackoverflow.com

youtube.com

https://scholar.google.com/scholar?q=Forward+Algorithm+using+hidden+markov+model&hl=fa&as_sdt=0&as_vis=1&oi=scholart

https://eu.udacity.com/course/introduction-to-computer-vision--ud810

https://scikit- https://scikit-learn.org/stable/

https://sebastianraschka.com/Articles/2014_kernel_density_est.html

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AffinityPropagation.html