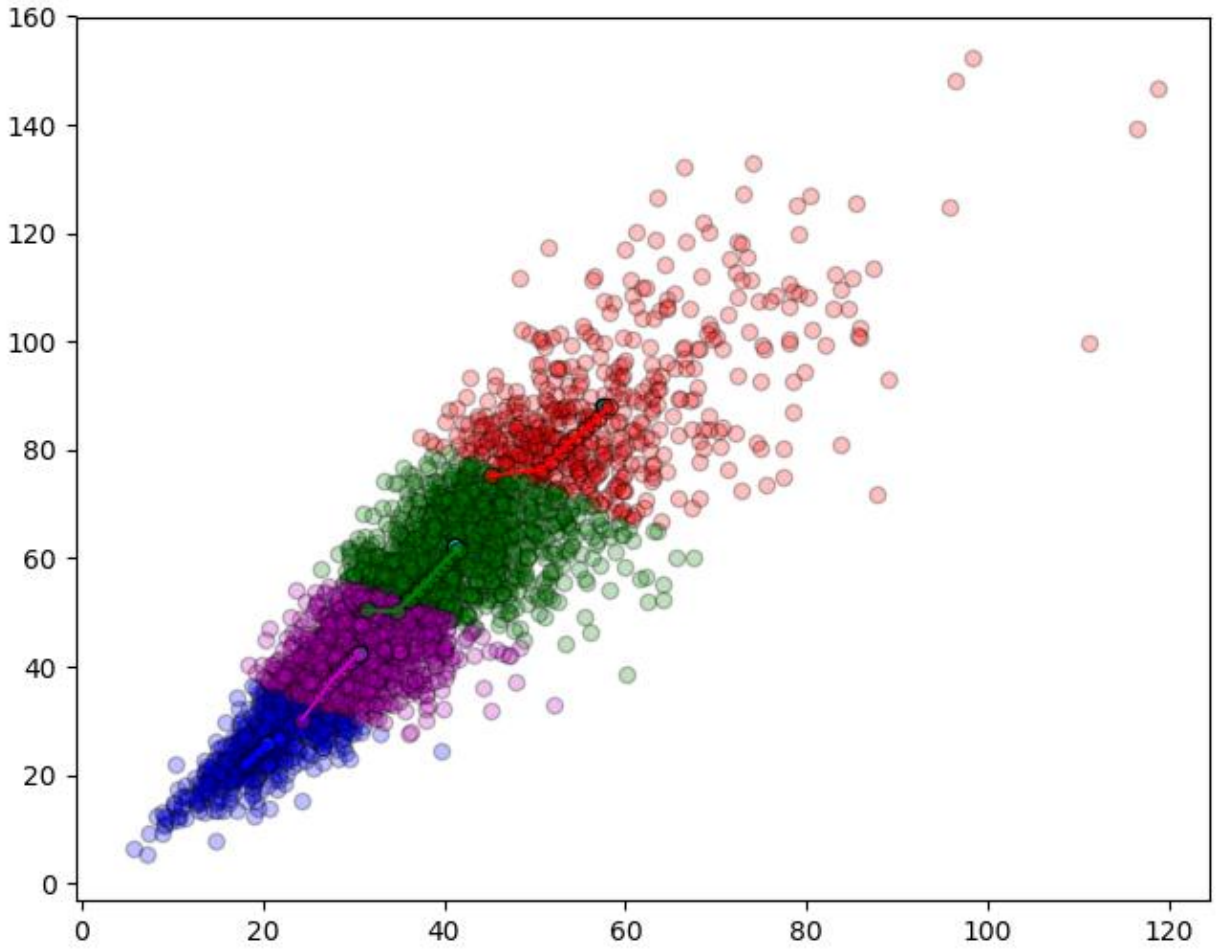
K-Means Clustering

1 Find the Cluster

All the raw program output can be found in the *Resut-KMeans.txt* file.

1.1 Plot

Here is the plot for **Training Set** and the movements of the cluster centroids:



1.2 Closest Training Samples

From the program output, we can know the centroid of each cluster and the samples are the closest one to their cluster center:

- For centroid [58.170153, 88.071845] (red)
  - closedLeftEyes\closed_eye_0427.jpg_face_1_L.jpg 
- For centroid [41.573942, 61.726335] (green)
  - openRightEyes\Bill_Simon_0001_R.jpg 

- For centroid [20.570045, 26.003195] (blue)
  - openRightEyes\Raymond_Arthurs_0001_R.jpg 
- For centroid [30.627886, 42.633633] (magenta)
  - openRightEyes\Zinedine_Zidane_0001_R.jpg 

For more images, please check the *Resut-KMeans.txt* file, for every centroid I output the 10 most closest samples.

2 Clustering

For the initialize of the cluster's centroids, just random choose 4 samples in the data set as the centroids.

In here, I think only the blue centroid is almost converge to the optimal because it starts moving like the cycle. For other centroids, even so the steps are smaller and smaller, but they are still not converging yet.

2.1 Converge

With each iteration, we will get a new centroid for each cluster that have a less cost. So, it will become better and better to converge to the optimal.

And also the Loss function we use is convex, so we can say that the clustering will converge.

2.1.1.1 Mathematically Prove

First, there are at most kN ways to partition N data points into k clusters. This is a large but finite number.

For each iteration of the algorithm, we produce a new clustering based only on the old clustering. And it has only 2 ways:

- if the old clustering is the same as the new, then the next clustering will again be the same.
- If the new clustering is different from the old one, then the newer one has a lower cost

Since the algorithm iterates a function whose domain is a finite set, the iteration must eventually enter a cycle. The cycle cannot have length greater than 1, because otherwise, there will have some clustering which has a lower cost than itself which is impossible.

Hence the cycle must have length exactly 1. That's also means K-Means Clustering converges in a finite number of iterations.

3 Closest Training Samples

The closest samples to the cluster centers I think still make sense, even though it not the original cluster we have: left or right, open or close. Because of the featurization logic we use only generate 2 features for each sample, and this let the clustering logic only check these 2 average Y-Gradient to try it best to produce the clustering.

If we want to do the clustering to split the eye status, such as: left or right, open or close, then we need to give the model more features to compute the cluster.

4 Another Plot

Because there has a bug in the featurization logic, the following image shows the plot for **Training Set** and the movements of the cluster centroids without update the **Featurize** function in *Assignment5Support.py*.