Machine Learning - Final Exam

*K-Mean Algorithm to form Clusters on the dataset

Muhammad Shahzeb Kayani
Dept. of computer engineering
Air University
Islamabad,Pakistan
171209@students.au.edu.pk

***Abstract*—We are faced with an emerging disaster situation, where an earthquake has wreak havoc in the ICT region of our capital. We are receiving reports that there are people with critical, severe and minor injuries. To address the situation we are required to formulate three teams of doctors and paramedics who are going to address the requirements of the injured people.**

***Index Terms*—Machine, Learning, K-Means, Clustering, unsupervised learning**

I. Introduction

We are given the number of years of experience and number of cases handled of each individual. Based on the two features, we are required to deploy k-means clustering algorithm and come up with the formulation of teams, who can start working on the relief efforts.

II. Dataset Generation

I started with generating the dataset for my model using the data given in the problem statement. I used the given means for the three different PDFs and their respective dataset.

After generating the dataset, I added gaussian noise in it by generation 300 training sets with zero means and a standard deviation of 1.Then, added this noise in my present dataset.

After doing all this, I wrote my dataset into a csv file using the csv package in order to be able to use the dataset in my model.The final csv has 300 rows and 2 columns.

III. K-means algorithm deployment

The next part was to deploy K-Means algorithm to the dataset I had created using the means and standard deviation. I started that with reading the dataset csv file with the help of read csv module of pandas library.Before proceeding, i checked the shape of the data by using the shape module of pandas library. Then to convert the data into an array, i read the columns individually and used the array module of numpy library.

Now I had my dataset in an array named ”X”. I started with initializing a variable ”k” which contained the number of cluster the algorithm has to form in the dataset. then I initialized the cluster centeroids by choosing random values between 0.1 and the maximum values in our dataset. this initialization decides how our solution will be formed. Then i visualized our training sets along with the initial cluster centers. The initial clusters centers are marked with ”+” in magenta color.

Then I created a function to calculate the Euclidean Distance which takes two arguments ”a” and ”b” and finds their distance. Then i created two arrays , one to store the previous cluster centeroid which has the same size as our cluster centroid variable and will help us to find the error in the future and one to store the assigned cluster labels to each training set. Along with it, i created a variable ”t” to store the number of iterations in which our algorithm converges.

Then I started my main while loop which will run until our error becomes zero. Then I started a for loop which will run from one to the length of the dataset. I then calculated the deistance between the centroids and the training sets and store them in the variable distance and assigned the training sets to one of the three available clusters. Then I calculated mean of the clusters and declared them as the new cluster centroids and began from the start of the while loop again. The result is shown below:

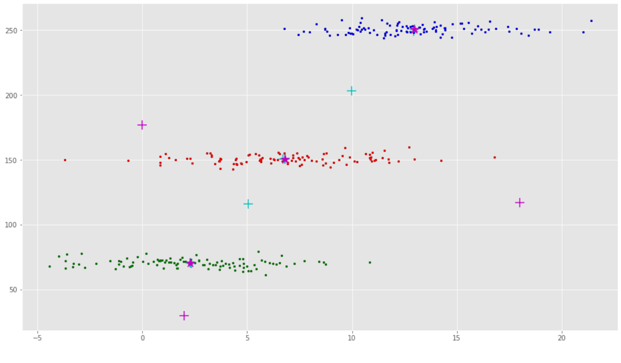

Fig. 1. Final Figure

Execution #	Initial Centroids	Number of iteration
1	[[13. 70.] [19. 182.] [13. 74.]]	10
2	[[14. 31.] [0. 239.] [18. 111.]]	4
3	[[7. 121.] [15. 163.] [4. 154.]]	3
4	[[9. 236.] [5. 73.] [8. 146.]]	2
5	[[1. 125.] [1. 126.] [12. 167.]]	3
6	[[1. 37.] [15. 87.] [18. 167.]]	8
7	[[3. 208.] [14. 119.] [3. 113.]]	2
8	[[15. 226.] [6. 226.] [5. 96.]]	6
9	[[16. 175.] [17. 104.] [2. 145.]]	3
10	[[10. 2.] [1. 1.] [3. 170.]]	13
11	[[16. 119.] [13. 126.] [20. 38.]]	12
12	[[10. 241.] [2. 258.] [6. 106.]]	5
13	[[9. 254.] [0. 28.] [19. 102.]]	7
14	[[18. 170.] [14. 42.] [2. 107.]]	11
15	[[13. 87.] [7. 93.] [13. 68.]]	12
16	[[3. 130.] [16. 48.] [17. 107.]]	14
17	[[5. 28.] [10. 109.] [16. 127.]]	16

This tables the number of iterations it took the algorithm to coverge to the solution every time I run the algorithm with a random initial cluster centroid values.

IV. K-means algorithm analysis

A. Which step was computationally most extensive?

In the algorithm, finding the new controids by taking the average value of the clusters elements and assigning each element to its closest cluster was computationally most extensive.

B. How many features which were generated from the same Gaussian in step (i) of 'Dataset Generation', got clustered together?

In the best case scenario, all the elements generated from the same Gaussian got clustered together.But this was not the only case, the initial cluster centroids also clustered features generated from two different Gaussians into one cluster.

C. Why some features of the same Gaussian did not get clustered together?

The initial value of the clustroid centers decides in which direction the solution will converge. Some features of the same Gaussian did not get clustered together because the cluster centroid moved in such a direction that the features fromsame Gaussian got allocated to different clusters which were closer to each feature.

D. Based on the given scenario, what is a reasonable clustering result in terms of the application?

Based on the given scenario, a reasonable clustering result will be that all the features from the same Gaussian get clustered together because visually the data from different Gaussians have a lot of distance between them.

E. Does our algorithm give a reasonable enough result?

For some initial values of intial cluster centroids, our algorithm gave a reasonable enough result while for others,it drove past the reasonable solution towards an unconventional solution.

F. What numbers can we put in the table to show that the performance of our algorithm is correct?

The relation between the iterations to converge to a solution and the initial error can tell us about the performance of our algorithm.

V. Conclusion

Our algorithm did a goo enough job but it can be improved by using the latest means of initializing a cluster centroid. There are a lot of extended K-Means algorithm which initialize the cluster centroids in different ways like K-means ++ or naive sharding methods. In our algorithm, we used random data points as our initial cluster centroids.