Rohit Ravishankar
rr9105@rit.edu

# Home Work 1

## CSCI - 720 Big Data Analytics

Collaborators: None

1.

a. Set aside the durations of less then half a second.

a. What are we doing when we do this? Why are we doing this? What insights are there into this process?

**Ans.** We are trying to clean the data by separating out the stop duration where the stop duration is lesser than 0.5 seconds. We are doing this because a stop duration of less than 0.5 seconds would imply that the car is in motion or there is missing data for the time when the data was recorded. In such a situation, the stop duration data is redundant. Cleaning the data/removing the noise in the data would help us better visualize the trends in the data.

b. Implement Otsu's method to separate the data into two clusters with the minimum average variance

**Ans.** In the first split, the data was split into 2 clusters of 270 points (titled **'largest_cluster'** on the code) and 90 points (titled **'smallest_cluster'** on the code) each.

c. Run Otsu's method again on the largest of the two clusters.

**Ans.** The data from the previous solution(**'largest_cluster'**) was split into 2 clusters of 180 points(**'cluster_1'**) and 90 points(**'cluster_2'**) each.

d. Repeat the clustering process on the largest cluster until you have four clusters of data. Sort the clusters from lowest average duration to largest avg. duration.

__Ans.__ From the previous result the 180 points(**'cluster_1'**) was split into 2 clusters of 120 points(**'cluster_3'**) and 60 points(**'cluster_4'**).

i. The number of data points in that cluster

__Ans.__ So, finally, we end up with **4 clusters**, viz., 90 points(**'smallest_cluster'**), 90 points (**'cluster_2'**), 120 points(**'cluster_3'**) and 60 points(**'cluster_4'**)

ii. The average value of the durations in the cluster

__Ans.__

cluster_2: 1052.5222222222221 s
cluster_3: 16.346666666666664 s
cluster_4: 94.72 s
smallest_cluster: 3296.5366666666673 s

iii. The standard deviation of the data in that cluster

__Ans.__

cluster_2: 111.96726582927685
cluster_3: 14.852171184338298
cluster_4: 25.755658278004336
smallest_cluster: 681.614621395404

iv. The minimum data value of the data in that cluster

__Ans.__

cluster_2: 902.1 s
cluster_3: 2.6 s
cluster_4: 56.7 s
smallest_cluster: 2406.7

v. The maximum data value of the data in that cluster

__Ans.__

cluster_2: 1339.4 s
cluster_3: 55.2 s
cluster_4: 159.7 s
smallest_cluster: 5435.8 s

vi. The threshold that Otsu's method found as the maximum value for this cluster.

__Ans.__

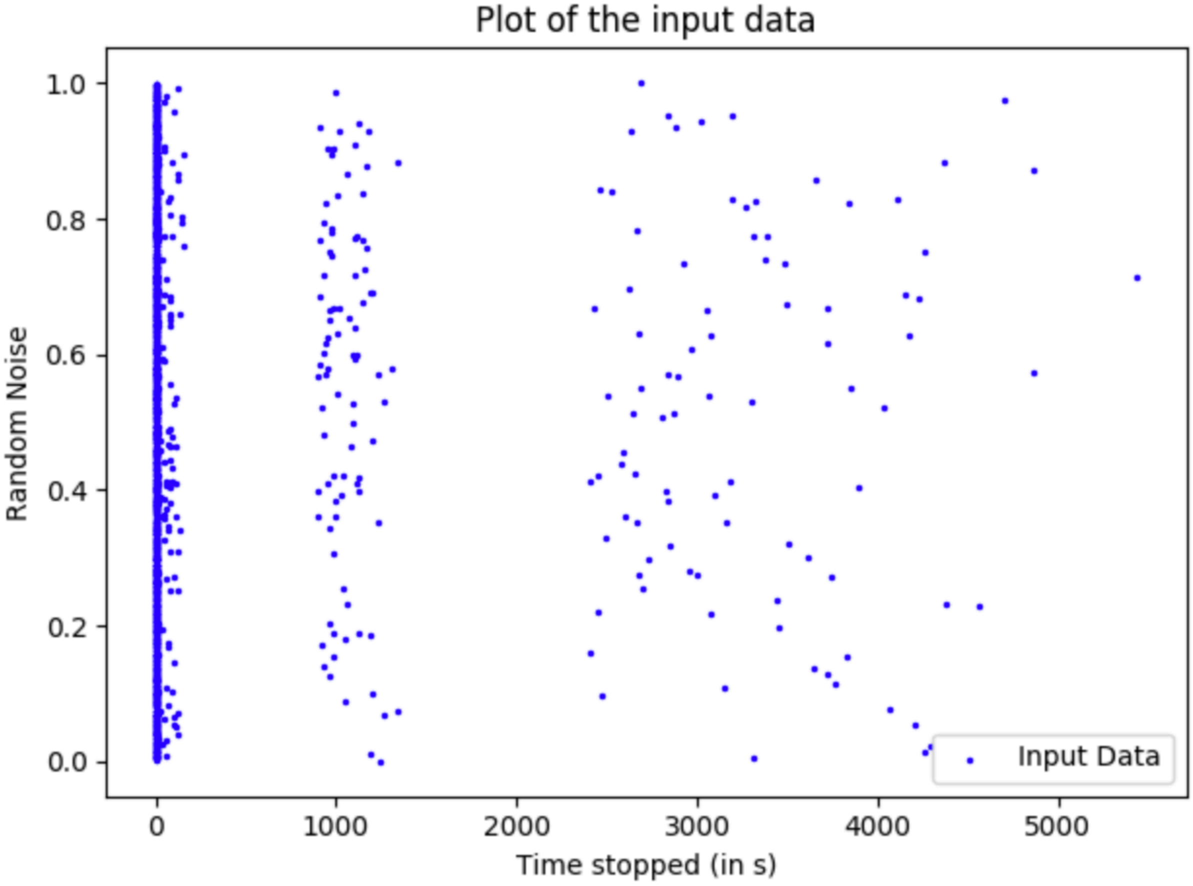
largest_cluster and smallest_cluster: 1339.4 s
cluster_1 and cluster_2: 159.7 s
cluster_3 and cluster_4: 55.2 s

e. Considering the clusters that resulted, does this make sense to you? Why or why not?

**Ans.** Yes, the results make sense to me. If you observe the condition for Otsu's method, i.e., that all points less than or equal to the threshold are in 1 cluster, for example, we would notice that the threshold value for largest_cluster and smallest_cluster is the largest value of cluster_2 which is subsequently calculated by splitting largest_cluster. Since, we see that this lines up for all cases we can say that the results make sense and hold good.

2.

a. Plot the input data, or a sub-set of it, against random noise



b. Describe the plot you rendered.

What does this data visualization show?

What did you learn about the data?

The plot I rendered is a scatter plot which shows the stop duration for all records across random noise. From the plotted graph, we observe a distribution of the stop times into 3 distinct clusters, i.e., around 0s, around 1000s and a third cluster for values in excess of 2500s.

c. Perform Parzen estimation, and try to get a smooth plot of the data.

3.

Conclusion and Discussion.

Write a conclusion and discussion about what you learned in this homework.

**Ans.** From this homework, at a high level, we learn how to implement clustering techniques on a given data set.

In Otsu's method for 1-D clustering we learn how to iteratively cluster points by choosing a threshold value and use brute-force to find the best mixed variance across the given set of points. With Parzen window density estimation technique we implement the the clustering technique where we move a window and find all the points within that window and based on that plot the density estimate for that values