Nile University
Communication and Information Technology Program
Fall 2018

CIT651 –Introduction to Machine Learning and Statistical Data Analysis

Machine Learning – Assignment #4

(Due on: January 30, 2019 at mid-night)

Implement the K-means clustering algorithm. Apply your algorithm to the data provided in the file "Data.txt" with K = 3. Each row in the file corresponds to one data point. One important aspect of K-means that changes the results significantly is the initialization. A good strategy for initializing cluster centers is as follows:

1- Pick one of the dataset points randomly as the center of the first cluster

2- For the next cluster, find the point with maximum distance to the center of the previous cluster

3- Choose this point as the center of the next cluster

4- Repeat steps 2 and 3 until you initialize the centers of all clusters

The K-means algorithm given in lectures can be then applied as explained in lectures. You should run the K-means algorithm with the initialization method above 100 times. The final output of the K-means clustering is the result that gives the minimum average distance between the points and the centers of their corresponding clusters.

Deliverables:

- Your code.
- A plot of the points provided in the dataset after clustering showing the three identified clusters. For this plot, use the best clustering result out of 100 repetitions you did. Name your plot "Kmeans.jpg".