German University in Cairo
Faculty of Media Engineering and Technology
Winter 2019

CSEN 1022 – Machine Learning

Assignment #3

(Due on December 10 at mid-night)

(This assignment can be done in teams of maximum 2 students – Please include a text files with your names and IDs in the submission)

Implement the K-means clustering algorithm to cluster the scanned images of the 10 digits (0 to 9) provided in the file "Assignment 3 Dataset.zip". The zip file contains one folder named "Images". The folder contains 240 images for each digit. Apply your algorithm to the data provided. 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 that has not been already chosen as a center

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 initialization strategy outlined above should be applied 30 different times. For each initialization, the code should then apply the K-means algorithm until it converges. Your code should then determine which of the 30 outputs is the best clustering result.

Deliverables:

- Your code.
- A text file explaining what method you used to evaluate the 30 clustering results to find the best one.
- A plot of the number of images clustered together for each digit in the best clustering result. The x-axis should show the digit number (0, 1, …, 9) while the y-axis should show the count. When the images of one digit are clustered in different clusters, use the count of the cluster that has the majority of images. Name the plot "Counts.jpg".

Important Notes:

- Convert the images to be binary instead of gray-scale. Use a threshold of 140 for binarization.
- Do not use Scikit learn or Scipy built-in functions for the K-means clustering. You have to implement your own version of all needed functions. You are allowed to use numpy functions.