German University in Cairo
Faculty of Media Engineering and Technology
Winter 2021

CSEN 1022 – Machine Learning

Assignment #3

(Due on January 3 at mid-night)
(This assignment can be done in teams of maximum 2 students)

Implement the K-means clustering algorithm to cluster the images provided in the dataset available on https://classroom.github.com/a/pCY-PkJx. These images are part of the machine learning benchmark CIFAR-10 dataset for three types (airplane, bird and truck). You need to implement the following steps:

1. Apply your K-means algorithm with K = 3 to the data provided in the train folder. One important aspect of K-means that changes the results significantly is the initialization. In your implementation, you should run K-means 10 different times starting with a different random initialization each time.
2. Use the Davis Bouldin index to choose the best outcome out of the 10 outcomes you obtained in Step 1.
3. For each class of images (airplane, bird and truck), identify the cluster to which the majority of images belong and, hence the corresponding center in this case. For instance, if the **1st 5K images** were clustered as (500 in Cluster 1, 1000 in Cluster 2,**3500 in Cluster 3**), this means that the majority of images belonging to the **airplane class** were clustered into **Cluster 3**. Thus, Cluster 3 should be considered the cluster representing the airplane class and the center of cluster 3 should be used in Step 4.
4. Use the cluster centers identified in Step 3 to classify the images in the test folder to one of the three types (airplane, bird or truck) based on the shortest Euclidean distance to the centers.

Deliverables:

- Your code.
- A plot of the **maximum number of images** clustered together for each type in the best clustering result of the train folder. The x-axis should show the type (airplane, bird, truck) while the y-axis should show the count.
- The confusion matrix obtained when classifying the test data using this method.

Important Notes:

- Save the notebook results before you submit it on GitHub. DO NOT CLEAR THE OUTPUT.
- Make sure the data type of all images is ***np.int64*** in order to avoid logical errors when calculating the Euclidean distance.
- **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.**