CSC 535/635 – HW 3  
Due Date: Thursday, April 5, 2018

**Assignment description**: For this assignment, you will **implement the k-means clustering algorithm,** given in Dunham’s book on pages 140 and 141, and use it on the given dataset. The algorithm is also given on **slide 9** of the PowerPoint slides. In your implementation of k-means, you must choose an appropriate stopping criterion, as discussed in class. Please seed your random number generator with a constant value so that different runs of your code will provide the same results. This is done in many languages by using random.seed(n), where n is a constant integer.

Use the given two-dimensional synthetic dataset to test your code. The data contains a total of 500 objects from 3 clusters. There are 200 objects in the first cluster, 150 in the second cluster, and 150 objects in the third cluster. The third attribute in the dataset is for the class/cluster label, which is provided so that you can test the accuracy of your implementation. Use the Euclidean distance measure to calculate distances.

The output from your program **must** display the following:

* The initial k-means picked by the algorithm.
* The number of objects in each cluster.
* The number of objects misclassified in each cluster.
* The objects classified in each cluster, along with the original class label for each object. You may show the original or normalized objects here.
* The accuracy rate.

**Remarks:**

* Use the Euclidean distance measure to compute distances.
* You must **implement the k-means algorithm from scratch**:
  + You are not allowed to use any library implementation of the algorithm, which is provided in many languages.
  + You are not allowed to use any online sources. If you have questions about the algorithm, please feel free to ask either the Instructor or the TA.
* You may use the statistics module in Python or any built-in function to compute statistics such as means, modes, standard deviation …etc. or to help with your code.
* If you want to visualize the input data and the clustering, you may use **any** tool or visualization library for this. If you do so, please include the figures in your report under the experimental setup and results section in the template.

Sample output is as follows:

Initial k means are  
mean[0] is ((0.05841377449418182, 1.0301888224696443), 0)  
mean[1] is ((-1.810171639070208, -0.44470900915877676), 1)  
mean[2] is ((-0.9090375460986547, 1.050673514575594), 2)  
=====================  
Cluster 0  
Size of cluster 0 is 139  
Cluster label: 1  
Number of objects misclustered in this cluster is 8  
((0.2690684975264931, 0.05375183208601334), 2)  
((1.1272914432136873, -0.8680593126817495), 1)  
((2.859341388146024, 0.08106475489394709), 1)  
((0.8932306398444521, -1.305066077608689), 1)  
…  
((1.5056897419939497, 0.03326713998006349), 1)  
((0.5031293008957276, 0.4122339439401437), 2)  
=====================  
Cluster 1  
Size of cluster 1 is 200  
Cluster label: 0  
Number of objects misclustered in this cluster is 17  
((-0.6164615418871113, -0.10671158941059687), 0)  
((-1.3966642197845602, -0.2432762034502656), 0)  
((-0.5228372205394174, -0.6188288920593541), 0)  
…  
((-1.2913368582684048, -1.2675108087477802), 0)  
=====================  
Cluster 2  
Size of cluster 2 is 161  
Cluster label: 2  
Number of objects misclustered in this cluster is 24  
((0.12863201550495246, 2.1431904268929434), 2)  
((-0.47992607325505793, 0.914108900535926), 2)  
((0.05841377449418182, 1.0301888224696443), 2)  
((-0.04301257363248646, 0.28591167595345046), 1)  
((-0.323885537675568, 0.6887772873704727), 2)  
…  
((-0.0391115602429993, 2.6553077295417005), 2)

Accuracy rate is 90.20%

Class label is the majority class label for the objects in the current cluster.

Each normalized objects is formatted as ((normalized\_x\_value, normalized\_y\_value),   
 original\_class\_label)

**Submission instructions:**

* For this assignment you may either work alone or in groups of 2. Graduate and undergraduate students may work together in the same group. However, we expect more from the graduate students as defined in the next sub-section.
* You are allowed to use Python, IPython, R, Java or C++ for your implementation. Please provide sufficient instructions in your report about how to run your code.
* Upload your files including a copy of your report to your CSC 535-635 TRACE folder in a subfolder named **HW3.** Only one submission per group is needed. Please turn in a **stapled** hard copy of your report collection at the beginning of class on the due date.
* Name your program **hw3.extension**. Replace **extension** by the appropriate file extension based on your choice of programming language
* Name the input file containing the dataset: **synthetic\_2D.txt**

**Additional Requirements for Graduate Students:**

For Graduate Students, please select one or more of the following options to complete in addition to the standard requirements:

* Implement different techniques such as changing the stopping criteria, picking the initial k-means in a smart way, or changing other such variables and compare their influence on the performance of the algorithm.
* Use the K-Means Algorithm on the given (or real) dataset. Generate some figures/graphs and perform some visual analysis. (i.e. extract some meaning from the groups.)
* Improve K-Means, then compare your improved algorithm’s performance (speed, accuracy, etc.) against the given algorithm. Both algorithms must be implemented.

**What to turn in?**

Upload to TRACE copies of your source code and **a report (preferably) in MS word** describing the stopping criterion you selected for the algorithm, the accuracy of the algorithm, how you normalize the data (if you normalize), and any useful information about the assignment. Please make sure to follow the format of the template report provided on TRACE. If you do the extra credit part (see below), then add a separate section in the report to describe what you did. **Please make sure that your code is well organized and properly documented and commented.** Turn in hard copies of your report and code at the beginning of class on the due date. Please remember to include the name of your trace folder with your code. Please turn in only one report per group.

**Extra Credit [up to 25 points]** Apply your (maybe revised) implementation of the k-means algorithm to a real dataset. If you revise your original implementation for a real dataset, then create a separate file named: **hw3\_modified.extension.** Replace **extension** by the appropriate file extension based on your choice of programming language. Make sure to provide a link to the dataset and upload a copy of the dataset to your TRACE/HW3 folder. Also, make sure to mention in your report any preprocessing you do to the dataset. Please refer to homework 1 for possible places to look for real datasets.