Description

Download the MNIST dataset from here. Pick its training test set which contains 60K samples, each represented by a 784-dimensional vector. There are ten classes (the ten digits, `0' to `9'), and each sample is associated with one class. Preprocess the features using appropriate technique(s), and store them in a file.

(1) Implement the BFR and CURE clustering algorithms on this data assuming that you can store `K_1' samples in the main memory at a time. One may use any existing libraries for performing the initial k-means clustering in case of BFR, and agglomerative clustering in case of CURE. After this, every step needs to be implemented from scratch.

(2) For the BFR algorithm, ensure that the number of clusters obtained at the end of the training process is 10.

(3) For the CURE algorithm, keep the number of clusters as 10.

(4) After clustering, calculate the percentage of samples from each class and convert it into probability values. Using these, calculate the entropy of each cluster. Also calculate the total entropy of all the clusters by summing the entropy of individual clusters.

(5) Re-run the two algorithms five times assuming K_1 = {100,200,500}, and report the above result.

Deliverables

(1) A folder containing your codes and a detailed readme file. You may use any programming language.

(2) A report (PDF) describing the experimental details, results, observations, etc.

(3) Create a single zipped file name <RollNo_Assig1.zip> containing the above two and upload.

General instructions

(1) Do not paste your codes in the report.

(2) Cite all the resources in the report.

(3) If anything is missing or not clear from the above description, you may make appropriate assumptions and clearly mention them in the report.

(4) A submission which does follow any of the guidelines will be awarded a penalty.

(5) Any submission received after the deadline will not be evaluated. The time recorded in google-classroom will be considered.

(6) Plagiarism of any kind will result in a zero in this assignment, and an additional penalty in the total score in the course.