Description

Download the MNIST dataset from here. Pick its training and test sets. Each sample in the dataset is represented by a 784-dimensional vector. There are ten classes (the ten digits, `0’ to `9’), and each sample is associated with one class. In this assignment, we will use the raw binary feature vectors, assuming them to be “shingles”.

(1) Write a code to classify the test samples using the kNN algorithm and Jaccard similarity by varying the value of k in {1,2,3,4,5}. Report the classification accuracy and time required to classify all the test samples using one CPU core.
(2) Using any publicly available code for LSH, classify the test samples using the kNN algorithm. Vary the length of the signature vector in {40,60} and ‘s’ in {0.8, 0.9}. For each combination, run the experiment multiple times to calculate the average and standard deviation of classification accuracy and time required to classify all the test samples in all the set-ups using one CPU core.
(3) Compare (1) and (2) in terms of classification accuracy, time required, and peak RAM required.

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, analyses, observations, etc.
(3) Create a single zipped file name <RollNo_Assig2.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 be penalized. 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.