Foundations of Machine Learning
Assignment 3
Course Instructor : Arun Rajkumar.
Release Date : Oct 23, 2024
Submission Date: On or before 11:59 PM on November 17, 2024

SCORING: There are two questions in this assignment each with 4 sub questions. Each question carries 5 points towards your final grade. The points will be decided based on the clarity and rigour of the report provided and the correctness of the code submitted.

WHAT SHOULD YOU SUBMIT? You should submit a zip file titled 'Solutions_rollnumber.zip' where rollnumber is your institute roll number. Your assignment will NOT be graded if it does not contain all of the following:

- A text file titled 'Details.txt' with your name and roll number.
- A PDF file which includes explanations regarding each of the solution as required in the question. Title this file as 'Report.pdf'
- Clearly named source code for all the programs that you write for the assignment .

CODE LIBRARY: You are expected to code all algorithms from scratch. You cannot use standard inbuilt libraries for **algorithms**. You are allowed to use libraries for plotting, for Eigenvector computations, etc. You can use either Python or Matlab or C.

GUIDELINES: Keep the below points in mind before submission.

- Plagiarism of any kind is unacceptable. These include copying text or code from any online sources. These will lead to disciplinary actions according to institute guidelines.
- Any graph that you plot is unacceptable for grading unless it labels the x-axis and y-axis clearly.
- Don't be vague in your explanations. The clearer your answer is, the more chance it will be scored higher.

LATE SUBMISSION You are expected to submit your assignment on or before the deadline. Late submission after the deadline would not be graded and will fetch 0 points.

QUESTIONS

(1) Download the MNIST dataset from https://huggingface.co/datasets/mnist. Use a random set of 1000 images (100 from each class 0-9) as your dataset.

i. Write a piece of code to run the PCA algorithm on this data-set. Visualize the images of the principal components that you obtain. How much of the variance in the data-set is explained by each of the principal components?

ii. Reconstruct the dataset using different dimensional representations. How do these look like? If you had to pick a dimension d that can be used for a downstream task where you need to classify the digits correctly, what would you pick and why?

(2) You are given a data-set with 1000 data points each in $\mathbb{R}^2$ (cm_dataset_2.csv).

i. Write a piece of code to implement the Llyod"s algorithm for the K-means problem with $k = 2$. Try 5 different random initialization and plot the error function w.r.t iterations in each case. In each case, plot the clusters obtained in different colors.

ii. For each $K = \{2, 3, 4, 5\}$, Fix an arbitrary initialization and obtain cluster centers according to K-means algorithm using the fixed initialization. For each value of K, plot the Voronoi regions associated to each cluster center. (You can assume the minimum and maximum value in the data-set to be the range for each component of $\mathbb{R}^2$).

iii. Is the Llyod's algorithm a *good* way to cluster this dataset? If yes, justify your answer. If not, give your thoughts on what other procedure would you recommend to cluster this dataset?