Foundations of machine learning
Assignment 1

Course Instructor : Arun Rajkumar.

Release Date : Sep-02, 2025

Submission Date: On or before 11:59 PM on Sep-21,2025

SCORING: There are 2 questions in this assignment. Each question carries 5 points. The points will be decided based on the clarity of the report provided and the code submitted.

DATASETS Each question has an associated data-set indexed by the question number (Eg: Dataset1 corresponds to Question 1).

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

1. A PDF file which includes explanations regarding each of the solution as required in the question. Title this file as 'Report.pdf'
2. Source code for all the programs that you write for the assignment clearly named.

CODE LIBRARY: You are expected to code all algorithms from scratch. You cannot use standard inbuilt libraries for **computations**. The only allowed library are those that compute the Eigenvectors and Eigenvalues of matrices. If your code calls any other library function for computation, it will fetch 0 points. You are free to use inbuilt libraries for plots. You can code using 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 POLICY You are expected to submit your assignment on or before the deadline to avoid any penalty. Late submission fetches 0 points. Moodle is usually busy close to assignment deadlines - it is up to the student to make sure a version of the assignment is submitted well in advance of the deadline so that last minute server crashes/other issues do not affect the submission.

QUESTIONS

(1) You are given a dataset with 1000 data points each in $\mathbb{R}^2$.

(a) Write a piece of code to run the PCA algorithm on this data-set. How much of the variance in the data-set is explained by each of the principal components?

(b) Write a piece of code to implement the Kernel PCA algorithm on this dataset. Explore various kernels discussed in class. For each Kernel, plot the projection of each point in the dataset onto the top-2 principal components. Use one plot for each kernel - In case of RBF kernel, use a different plot for each value of σ that you use.

(c) Which Kernel do you think is best suited for this dataset and why?

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

(a) Write a piece of code to run the algorithm studied in class for the K-means problem with $k = 4$. 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.

(b) Fix a random initialization. For $K = \{2, 3, 4, 5\}$, 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 $\mathbb{R}^2$).

(c) Run the spectral clustering algorithm (spectral relaxation of K-means using Kernel- PCA) $k = 4$. Choose an appropriate kernel for this data-set and plot the clusters obtained in different colors. Explain your choice of kernel based on the output you obtain.

(d) Instead of using the method suggested by spectral clustering to map eigenvectors to cluster assignments, use the following method: Assign data point i to cluster ℓ whenever

$$\ell = \arg \max_{j=1,\ldots,k} v_{ji}$$

where $v_j \in \mathbb{R}^n$ is the eigenvector of the Kernel matrix associated with the j-th largest eigenvalue. How does this mapping perform for this dataset? Explain your insights.