

**CS5691: Pattern Recognition and Machine Learning**  
**Assignment 1**

**Course Instructor :** Arun Rajkumar.

**Release Date :** August - 23, 2022

**Submission Date: On or before 11:59 PM on September 11 ,2022**

**SCORING:** There are two questions in this assignment each with 4 sub questions. Each sub-question carries 12.5 points. The points will be decided based on the clarity and rigour of the report provided and the correctness of the code submitted.

**DATASETS** The data-set for both the questions is the same and is in a csv file titled Dataset.csv.

**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 POLICY** You are expected to submit your assignment on or before the deadline to avoid any penalty. Late submission incurs a penalty of points equal to 10 times the number of days your submission is late by. Any late submission post 3 days of the deadline would not be graded and will fetch 0 points.

## QUESTIONS

- (1) You are given a data-set with 1000 data points each in  $\mathbb{R}^2$ .
- 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?
  - Study the effect of running PCA without centering the data-set. What are your observations? Does Centering help?
  - Write a piece of code to implement the Kernel PCA algorithm on this dataset. Use the following kernels :
    - $\kappa(x, y) = (1 + x^T y)^d$  for  $d = \{2, 3\}$
    - $\kappa(x, y) = \exp \frac{-(x-y)^T(x-y)}{2\sigma^2}$  for  $\sigma = \{0.1, 0.2, \dots, 1\}$Plot the projection of each point in the dataset onto the top-2 components for each kernel. Use one plot for each kernel and in the case of (B), use a different plot for each value of  $\sigma$ .
  - 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  $\mathbb{R}^2$ .
- 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.
  - 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  $\mathbb{R}^2$ ).
  - 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.
  - 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  $\ell$  whenever

$$\ell = \arg \max_{j=1, \dots, k} v_i^j$$

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.