CS5691: Pattern recognition and machine learning
Assignment 1
Course Instructor : Arun Rajkumar.
Release Date : Oct-06, 2020
Submission Date: On or before 5 PM on October 24,2020

SCORING: There are 4 questions in this assignment. The first three questions are mandatory and carry 5 points each. The last question is a *play* question and is not mandatory. It carries 5 points. The contribution of points scored in this assignment towards your final grades will be calculated as

$$\min\{\text{your score}, 15\}$$

For instance if you scored 12 in the first 3 questions and 4 in the last question, your total score would be $\min\{12+4, 15\} = 15$. If you scored 14 in the first 3 questions and did not attempt the last question, your score would be $\min\{14+0, 15\} = 14$.

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 Question 3 is in a csv file titled Dataset3.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 **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 incurs a penalty in points equal to the number of days your submission is late by. Any late submission post October 30 would not be graded and will fetch 0 points.

QUESTIONS

(1) The column space of a matrix $A \in \mathbb{R}^{d \times d}$ is the set of all vectors that can be obtained as linear combinations of the column vectors of A. The row space of A is the column space of A^T

- Is the column space of a matrix $A \in \mathbb{R}^{d \times d}$ same as the columnn space of AA^T? If yes, prove. If not, give a counter example.
- Is row space of A same as column space of A? If so, prove. If not, argue why.

(2) Let $K1$,$K2$ be two arbitrary kernel functions mapping vectors from $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. For each of the cases below, show if it is a valid Kernel or if it is not, argue why.

- $K3(x, y) = K1(x, y) + K2(x, y) + 7.5$
- $K4(x, y) = 5 * K1(x, y) - 3 * K2(x, y)$
- $K5(x, y) = K1(x, y) * K2(x, y)$
- $K6(x, y) = (x^T y + 1)^3$

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

i. 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?

ii. Study the effect of running PCA without centering the data-set. What are your observations? Does Centering help?

iii. Write a piece of code to implement the Kernel PCA algorithm on this dataset. Use the following kernels :

A. $\kappa(x, y) = (1 + x^T y)^d$ for $d = \{2, 3\}$

B. $\kappa(x, y) = \exp \frac{-(x-y)^T (x-y)}{2\sigma^2}$ for $\sigma = \{0.1, 0.2, \ldots, 1\}$

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 σ.

iv. Which Kernel do you think is best suited for this dataset and why?

(4) **Play Question: Non-Mandatory** Create your own image dataset as follows. Use your phone to capture 25 photos of some type of vessel (category 1) in your kitchen and 25 photos of chairs/tables in your home (category 2). (If these are not available, pick any two sets of distinct items. You can take multiple pictures of the same item with slightly different angles). Pick only 20 images in each category and convert them to black and white. Make sure they are of the same size. Let the pixel intensity be your features. Run standard PCA and project all the 40 images onto the Eigenspace spanned by the top $K\%$ ($K = 10\%, 20\%, 30\%, 50\%, 75\%, 100\%$ of Eigenvectors of the Covariance matrix. For each of the remaining 10 images and for each value of K, project them onto the top $K\%$ of Eigenvectors, compute the average distance of their projections with the projections of the 20 images in category 1 and the 20 images in category 2. What do you observe from this experiment? Draw insights.