CS5691: Pattern recognition and machine learning
Programming Assignment 2

Course Instructor : Arun Rajkumar.

Release Date : March-15, 2023

Submission Date: On or before 11:59 PM on April 2,2023

SCORING: There are 2 questions in this assignment. Each question carries 5 points. The points will be decided based on the report provided, code submitted and a final oral examination that covers all the assignments together.

DATASETS There are two csv files provided along with this assignment.

WHAT SHOULD YOU SUBMIT? You should submit a zip file titled 'Solutions_rollnumber1.zip'. Your assignment will NOT be graded if it does not contain all of the following:

- A PDF file which includes explanations regarding each of the solution as required in the question. Title this file as 'Report.pdf'
- 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, Eigenvalues, inverses, pseudo inverses of matrices if needed. 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 equal to the number of days your submission is late by.Late submission beyond 3 days will not be considered. The deadline is strict.

QUESTIONS

(1) You are given a data-set with 10000 points in $(\mathbb{R}^{100}, \mathbb{R})$ (Each row corresponds to a datapoint where the first 100 components are features and the last component is the associated y value).

i. Obtain the least squares solution $\mathbf{w}_{ML}$ to the regression problem using the closed form expression.

ii. Code the gradient descent algorithm with suitable step size to solve the least squares algorithms and plot $\|\mathbf{w}^t - \mathbf{w}_{ML}\|$ as a function of t. What do you observe?

iii. Code the stochastic gradient descent algorithm using batch size of 100 and plot $\|\mathbf{w}^t - \mathbf{w}_{ML}\|$ as a function of t. What are your observations?

(2) Consider the same data-set as in Question (1). You are additionally given a data-set with 500 points for testing which you cannot use during train/cross-validation.

i. Code the gradient descent algorithm for ridge regression.

ii. Cross-validate for various choices of λ and plot the error in the validation set as a function of λ. For the best λ chosen, obtain $\mathbf{w}_R$. Also obtain $\mathbf{w}_{ML}$ for the training data. Compare the test error of $\mathbf{w}_R$ with $\mathbf{w}_{ML}$. Which is better and why?

iii. Code the co-ordinate descent algorithm to obtain a LASSO solution. Cross validate for the same train-validation splits as in part (ii). Plot the error in the validation set as a function of λ and obtain the best λ. Compare the test error of $\mathbf{w}_{LASSO}$ with $\mathbf{w}_R$. Which is better and why?