Fondations of Machine Learning
Assignment 1
Course Instructor : Arun Rajkumar.
Release Date : Sep 2, 2024
Submission Date: On or before 11:59:59 PM on Sep 22, 2024. Strictly no extension will be given.

SCORING: There is *a* question in this assignment with 5 parts. The contribution of points scored in this assignment towards your final grades will be 10 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-sets are in the corresponding google drive folder shared in Moodle.

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** taught in class. You are free to use inbuilt libraries for plots, for performing matrix operations, etc. 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 will not be accepted under any circumstances.

QUESTIONS

(1) You are given a data-set in the file FMLA1Q1Data_train.csv with 10000 points in $(\mathbb{R}^2, \mathbb{R})$ (Each row corresponds to a datapoint where the first 2 components are features and the last component is the associated y value).

i. Write a piece of code to obtain the least squares solution $\mathbf{w}_{ML}$ to the regression problem using the analytical solution.

ii. Code the gradient descent algorithm with suitable step size to solve the least squares algorithms and plot $\|\mathbf{w}^t - \mathbf{w}_{ML}\|_2$ as a function of t. What do you observe?

iii. Code the stochastic gradient descent algorithm using batch size of 100 and plot $\|\mathbf{w}^t - \mathbf{w}_{ML}\|_2$ as a function of t. What are your observations?

iv. Code the gradient descent algorithm for ridge regression. 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 $\mathbf{w}_R$. Compare the test error (for the test data in the file FMLA1Q1Data_test.csv) of $\mathbf{w}_R$ with $\mathbf{w}_{ML}$. Which is better and why?

v. Assume that you would like to perform kernel regression on this dataset. Which Kernel would you choose and why? Code the Kernel regression algorithm and predict for the test data. Argue why/why not the kernel you have chosen is a better kernel than the standard least squares regression.