Programming Assignment - 1
Linear Regression
EE5180
Release Date : 28 February, 2023
Submission Date: On or before 11:59:59 PM on March 9, 2023

Notes:

1. Please use moodle discussion threads to post your doubts and check it before posting if the same question has been asked earlier.
2. **What to Submit? :** Submit a single zip file in the moodle named as PA1_Rollno.zip containing reports named "Rollno_Report.pdf" and folders containing corresponding codes.
3. Read the problem fully to understand the whole procedure.
4. **NO inbuilt functions** are allowed in this task. Code everything yourself.
5. Any plagiarism/cheating will be dealt very, very strictly. You may end up with U-grade. All your reports and codes will be matched through Turnitin with each other and all previous years submissions.
6. You should thoroughly know what you are doing, and it will be asked in your viva.
7. Late submissions will be evaluated for reduced marks, and for each day after the deadline, we will reduce the weightage by 10%.

Questions

1 Question

1. There are two datasets for polynomial regression:
   - 1-dimensional data data1.txt
   - 2-dimensional data data2.txt

   Hint: Use "$df = pd.read\_csv(r'path', delim\_whitespace = True, header = None)$" to read the files into a dataframe.

2. For Dimension k, there are k+1 columns in the data. First k columns are the features, and the last is the outcome.

3. Use 70% of the data for training, 20% for validation and 10% of data for testing.

TASK REQUIREMENTS : For both datasets, do the following tasks. NO inbuilt functions are allowed in this task.

1. Try different orders of the polynomials (such as 0th, 1st, 2nd , upto 10th order) and show the predicted curves for the various polynomial orders.

2. Plot the validation loss as a function of the polynomial order. Which polynomial order gives the best result.

3. Perform ridge regression. Cross-validate for various choices of lambda and plot the error in the validation set as a function of lambda.

2 Question

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

1. Obtain the least squares solution W_{ML} to the regression problem using the analytical solution.

2. Code the gradient descent algorithm with suitable step size to solve the least squares algorithms and plot $||\mathrm{W}_t$ - $W_{ML}\ ||^2$ as a function of t. What do you observe?

3. Code the gradient descent algorithm for ridge regression. Cross-validate for various choices of lambda and plot the error in the validation set as a function of lambda. For the best chosen lambda, obtain w_R. Compare the test error (for the test data in the File test.csv) of W_R with W_{ML}. Which is better and why?