Machine Learning (CS60050)

Assignment 1: Linear Regression
Due date: February 2, 2020, 23:59 IST

You are given two files train.csv and test.csv containing the **training data** and **testing data** respectively. You can download the files from here
Each file contains two columns -- a feature and a label.

1. Understanding the data and simple curve fitting **[5 + 25 = 30 marks]**

a. Plot a feature vs label graph for both the training data and the test data.

b. Write a code to fit a curve that minimizes *squared error cost function* using gradient descent (with learning rate 0.05), as discussed in class, on the training set while the model takes the following form $y = W^T \Phi_n(x)$, where $W \in R^{n+1}$, and $\Phi_n(x) = [1, x, x^2, x^3 \ldots, x^n]$. Squared error is defined as $J(W) = \frac{1}{2m} \sum_{i=1}^{m} \left(W^T \Phi_n(x) - y \right)^2$.

In your experiment, vary n from 1 to 9. In other words, fit 9 different curves (polynomials of degree 1, 2, …, 9) to the training data, and hence estimate the parameters. Use the estimated W to predict labels on test data and measure squared error on the test set, name it as test error.

2. Visualization of the fitted curves **[10 + 10 = 20 marks]**

a. Draw separate plots of all 9 different curves that you have fit for the training dataset in 1b.

b. Report squared error on both train and test data for each value of *n* in the form of a plot where along x-axis, vary *n* from 1 to 9 and along y-axis, plot both training error and test error. Explain which value of *n* is suitable for the dataset that you have, and why.

3. Regularization **[15 + 15 = 30 marks]**

Perform the following regularizations on the curves for which you obtain the minimum and maximum training error from above.

a. Perform Lasso regression on the cost function as follows, vary $\lambda = 0.25, 0.5, 0.75, 1$:

$$J(W) = \frac{1}{2m} \sum_{i=1}^{m} \left(W^T \Phi_n(x) - y \right)^2 + \lambda W$$

b. Perform Ridge regression on the cost function as follows, vary $\lambda = 0.25, 0.5, 0.75, 1$:

$$J(W) = \frac{1}{2m} \sum_{i=1}^{m} \left(W^T \Phi_n(x) - y \right)^2 + \lambda W^2$$

Plot both training and test error for the two types of regularization (a) and (b). What differences do you notice between the two kinds of regression? Which one would you prefer for this problem and why?

Submission instructions

For each part, you should submit the source code and all result files. Write a separate source code file for each part. **You should include a README file describing how to execute each of your codes**, so that the evaluators can test your code.

You can use C / C++ / Java / Python for writing the codes; no other programming language is allowed. You cannot use any library/module meant for Machine Learning or Deep Learning. You can use libraries for other purposes, such as formatting and pre-processing of data, but NOT for the ML part. Also you should not use any code available on the Web. Submissions found to be plagiarised or having used ML libraries will be awarded zero marks for all the students concerned.

All source codes, data and result files, and the final report **must be uploaded via the course Moodle page, as a single compressed file (.tar.gz or .zip)**. The compressed file should be named as: **{ROLL_NUMBER}_ML_A1.zip or {ROLL_NUMBER}_ML_A1.tar.gz**
Example: If your roll number is 16CS60R00, then your submission file should be named as 16CS60R00_ML_A1.tar.gz or 16CS60R00_ML_A1.zip

*****Note that the evaluators can deduct marks if the deliverables are not found in the way that has been asked for the assignment.**

Submission deadline: February 2, 2020, 23:59 IST [hard deadline]

For any questions about the assignment, contact the following TAs:

1. Paheli Bhattacharya (paheli.cse.iitkgp @ gmail . com)
2. Soham Poddar (sohampoddar26 @ gmail . com)