Machine Learning (CS60050)

Assignment 1: Linear Regression
Due date: **15th February 2019**

---

1. Synthetic data generation and simple curve fitting [10 + 5 + 25 = 40 marks]

a. Generate a synthetic dataset as follows. The input values $\{x_i\}$ are generated uniformly in range $[0, 1]$, and the corresponding target values $\{y_i\}$ are obtained by first computing the corresponding values of the function $sin(2\Pi x)$, and then adding a random noise with a Gaussian distribution having standard deviation 0.3. Generate **10 such instances of** (x_i, y_i). [You can use any standard module to generate random numbers as per a gaussian / normal distribution, e.g., numpy.random.normal for python.]

b. Split the dataset into two sets randomly: (i) Training Set (80%) (ii) Test Set (20%).

c. 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 following form $y = W^T \Phi_n(x)$, $W \in R^{n+1}$, $\Phi_n(x) = [1, x, x^2, x^3 \ldots, x^n]$. Squared error is defined as $J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(W^T \Phi_n(x) - y \right)^2$. In your experiment, vary n from 1 to 9. In other words, fit 9 different curves to the training data, and hence estimate the parameters. Use the estimated W to measure squared error on the test set, and name it as test error on test data.

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

a. Draw separate plots of the synthetic data points generated in 1 (a), and all 9 different curves that you have fit for the given dataset in 1 (c).

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 train error and test error. Explain which value of n is suitable for the synthetic dataset that you have generated and why.

3. Experimenting with larger training set [10 marks]

Repeat the above experiment with three other datasets having size 100, 1000 and 10,000 instances (each dataset generated similarly as described in Part 1a).
Draw the learning curve of how train and test error varies with increase in size of datasets (for 10, 100, 1000 and 10000 instances).

4. Experimenting with cost functions [20 + 10 = 30 marks]

a. Solve the problem by minimizing different cost functions (Do not use any regularization, Use gradient descent to minimize the cost function in each case) :

i. Mean absolute error i.e. $J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left| W^T \Phi_n(x) - y \right|$

ii. Fourth power error i.e. $J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(W^T\Phi_n(x) - y\right)^4$

b. Plot the test RMSE vs learning rate for each of the cost functions. Vary the learning rates as 0.025, 0.05, 0.1, 0.2 and 0.5. 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. You can use libraries for other purposes, such as generation and formatting 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.

Along with the source codes and the results, you should submit a report (pdf) including the following:

- Final learned values of the model parameters for each of the parts.
- The plots as described above. You can use any standard plotting tool / library to generate the plots. The data files and the scripts (if any) used to generate the plots should be included in your submission.
- The choices, as described above, with proper justifications.

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 15, 2019, 23:59 IST [hard deadline]

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

1. Abhisek Dash (assignmentad @ gmail . com)
2. Paheli Bhattacharya (paheli.cse.iitkgp @ gmail . com)