Machine Learning (CS60050): Assignment 1

Download data from
https://drive.google.com/open?id=1vJwvlt8Tp-mwbDj1V0ePOe0Ej1ndWXOx
In the first four columns of each row, you have four features corresponding to a house - area in square feet, number of floors, number of bedrooms and number of bathrooms. In the fifth column of each row, you will find the price of the house. The task is to predict the price of the houses from their corresponding features using linear regression.

For all the questions below, use the **first 80% of the rows for training and the remaining 20% for testing, i.e., measuring the performance of the learned model.** The performance of the learned model is measured using Root Mean Square Error (RMSE) which is defined as,

$$\text{RMSE}_{fo} = [\sum_{i=1}^{N} (z_{fi} - z_{oi})^2 / N]^{1/2}$$

where Z_{fi} is the predicted price and Z_{oi} is the actual price of the i-th house in the test set (size N). Note that the test set (last 20% of the rows) should NOT be used to train the model. Also the prices (last column) in the test set should not be visible to the model, rather these values should be used only for measuring the performance.

Part (a): implementing linear regression

- Use linear combination of the features.
- Minimize mean squared error cost function (as discussed in class).
- Use gradient descent to minimize the cost function (as discussed in class). Use learning rate of 0.05.
- Solve the problem with and without regularization. Show how the test RMSE varies with the weightage of the regularization terms (use same weightage for all features).

Part (b): experimenting with optimization algorithms

- Use linear combination of the features.
- Minimize mean squared error cost function (as discussed in class). Do not use any regularization.
- Solve the problem by minimizing the cost function using two optimization algorithms: (i) gradient descent with learning rate of 0.05, and (ii) iterative re-weighted least square method (described at http://www.cedar.buffalo.edu/~srihari/CSE574/Chap4/4.3.3-IRLS.pdf).
- Plot the test RMSE vs number of iterations for both the optimization algorithms. Which optimization algorithm would you prefer for this problem and why?

Part (c): experimenting with combinations of features

- Minimize mean squared error cost function. Do not use any regularization.
- Use gradient descent to minimize the cost function. Use learning rate of 0.05.
- Solve the problem using (i) linear, (ii) quadratic and (iii) cubic combinations of the features.

- Plot the test RMSE vs learning rate for each of the cases. Which one you would prefer for this problem and why?

Part (d): experimenting with cost functions

- Use linear combination of the features.
- Solve the problem by minimizing different cost functions: (i) mean absolute error, (ii) mean squared error, and (iii) mean cubed error. Do not use any regularization.
- Use gradient descent to minimize the cost function in each case.
- Plot the test RMSE vs learning rate for each of the cost functions. Which one would you prefer for this problem and why?

Submission instructions

For each of the four parts, 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, but you cannot use any standard library implementation of linear regression. Also you should not use any code available on the Web. We will use plagiarism detection tools over the submitted source codes. Submissions found to be plagiarised will be awarded zero.

Along with the source codes and the results, you should submit a report (pdf or Word) 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.{EXTENSION}.
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

Submission deadline: February 15, 2018

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

1. Soumya Sarkar (portkey1996 [AT] gmail [DOT] com)
2. Soumajit Pramanik (soumajit.pramanik [AT] gmail [DOT] com)