DSCI 440W Implementation Assignment 2 (100 points)
Due 11:59 pm March 5, 2023

General instructions.

This is an individual assignment. You will only need to submit one copy of the source code and report. Please, type your report in LaTeX.
2. Your source code and report must be submitted through the Moodle site.
3. Be sure to answer all the questions in your report. You will be graded based on your code as well as the report. So please write your report in a clear and concise manner. Clearly label your figures, legends, and tables.

Linear regression, regularization

In this assignment you will use the Boston Housing dataset from the CMU StatLib Library that concerns the housing prices in Boston suburbs. The data set contains 13 attributes describing each area (e.g., crime rate, accessibility to major highways) and the target variable is the median value of housing (in thousands) for that area. The description of the data is in the file desc.txt associated with this assignment on Moodle. The goal is to predict the median value of housing of an area based on 13 attributes. For your convenience the data has been divided into two datasets: (1) a training dataset housing_train.txt you should use in the learning phase, and (2) a testing dataset housing_ttest.txt to be used for testing. Your task is to implement the linear regression learning algorithm presented in class and explore the effect of the regularization. In particular:

1. Given the training data, load the data into the corresponding X and Y matrices, where X stores the features and Y stores the desired outputs. The rows of X and Y correspond to the examples and the columns of X correspond to the features. Add an extra column of ones to X (Make this extra column to be the first column).

2. Compute the optimal weight vector w^*. You can use the code that you wrote for IA1. Report the learned weight vector.

3. Apply the learned weight vector to the training and testing data and compute the sum of squared error (SSE) on training and testing data sets. Report the SSE values.

4. Consider a variant of linear regression, where we minimize the following objective function:

$$SSE = \frac{1}{2}\left(\sum_{i=1}^{N}(y^i - \mathbf{w}^T\mathbf{x}^i)^2 + \lambda\|\mathbf{w}\|_2^2\right), \tag{1}$$

where the first term is the regular SSE and the second term is called a regularization term and computes the squared Euclidean norm ($L2$-norm) of the weight vector $\mathbf{w}$. Compute the optimal $\mathbf{w}$ using the corresponding formula with different values of λ (choose the range for λ and explore it). Evaluate each of the learned $\mathbf{w}$ by computing the SSE on both training and testing data. Plot the training and testing SSE values as a function of λ. What behavior do you observe? Explain.

5. Compare the different $\mathbf{w}$'s that you got in part 4. As the λ value gets bigger, what impact do you observe it has on the weight values? Plot $\|\mathbf{w}\|_2$ (the Euclidian norm of $\mathbf{w}$) as a function of λ. What behavior do you observe? Explain. You may use the objective function in (1) to explain the behavior that you observe.

Your report should have the following structure:

(a) You full name and assignment number.

(b) Introduction (Briefly state the problem you are solving).

(c) Optimal w^* from part 2.

(d) Train and test SSE from part 3.

(e) Plot of train SSE vs λ, plot of test SSE vs λ, discussion for part 4.

(f) Plot of $\|\mathbf{w}\|_2$ vs λ, discussion for part 5.