Assignment 1

Note:

This assignment should be done by each student individually. You can discuss it in general terms with other students; however, the files you hand in, e.g., the report and codes should be your own. If I find your reports/codes are the same as or similar to the other, both of you cannot get the scores for this assignment.

1. Problem Statement

In this assignment, you are required to investigate a supervised learning method for a regression problem. Suppose that there are n training cases, each of which has a vector of input x (d dimensions), and a real-valued output, y, you are required to get a linear regression model for y based on x. That is, y can be modeled as:

$$y = \beta_0 + \sum_{j=1}^{d} \beta_j x_j + noise \tag{1}$$

To get this model you are required to fit it by two different methods. The firs one is to fit the model by the following least squares:

$$\sum_{i=1}^{n} \left(y_i - (\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij}) \right)^2 \tag{2}$$

When $\beta_i (i = 0, 1, ..., d)$ are found, the residual for each training sample can be computed by the following equation:

$$r_i = y_i - (\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij}) \tag{3}$$

To predict the output y of a test sample with input x, y is computed by:

$$(\beta_0 + \sum_{j=1}^{d} \beta_j x_j) + \frac{1}{K} \sum_{i \in N(x)} r_i \tag{4}$$

where $N(x)$ is the neighborhood set of x with size K; that is to say $N(x)$ contains K nearest neighbors of x w.r.t Euclidean distance.

The second method to fit the model is to use the following penalized least squares:

$$\lambda \sum_{j=1}^{d} \beta_j^2 + \sum_{i=1}^{n} \left(y_i - (\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij}) \right)^2 \tag{5}$$

where λ is the parameter to balance the importance of two parts in Eq. 5.

Note that λ and K can be determined by n-fold cross validation (here, $n = 5$).

2. Submissions

Finally, you are required to submit the following files:

- A report which describes the details of your implementation, e.g., how to optimize the function, how to determine the values of λ and K, the effect of different values of λ and K, analyzing and comparing the test results of both methods. Moreover, you can also make other analysis; for example, what is the effect of standardizing the inputs (normalizing the input).

- The codes that are written with good style. I suggest to use Matlab; but if you like, you can also write them in other programming languages, e.g., Python, R, Java, C etc.

3. About the data

The file **trainingData-Ass1.txt** contains all the training samples, one sample one row. In each row, the first nine variables comprise the input x, and the last one is the output y. The file **testData-Ass1.txt** contains the test samples; its data format is the same as **trainingData-Ass1.txt**.

4. Deadline

All files should be submitted before May 4, 2016.