Homework 2

March 23, 2023

1 Dataset

The Forest Fires dataset (forestfires.csv) has 13 columns where the first to the 12 -th columns are attributes (Please see forestfires.txt for an introduction to these attributes). The last column is the burned area of the forest fire. Given a sample $X \in \mathbf{R}^{12}$, you need to find a function f to predict the burned area of forest fires $y \in \mathbf{R}$. The MSE loss is given as

$$\text{Loss}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - f\left(X_i\right)\right)^2,$$

where X_i denotes the i-th training sample and y_i denotes the burned area of the forest fire. N denotes the number of training samples.

Specifically, the dataset (forestfires.csv) has 517 samples. **We have already divided the provided data into a training set and a test set for you in the python script,** with the first 450 samples designated as the training set and the remaining 67 samples as the test set.

More information about the dataset can be referred to as follows: `https://archive.ics.uci.edu/ml/datasets/Forest+Fires`.

2 Tasks

Solve a simple regression problem using Mean Squared Error (MSE) loss with different penalty or regularization terms. Here are five methods you need to implement:

(1) Find a linear model to solve the regression problem. In other words, the linear model is $f\left(X_i\right) = \beta^\top X_i$, where $\beta \in \mathbf{R}^{12}$. To this end, you need to compute the MSE loss and use gradient descent to find a weight matrix β.

(2) Using Ridge Regression to solve the regression problem. The ridge regression you need to implement is on page 3 of lecture 3 .

(3) Using RBF kernel regression to solve the regression problem. The kernel regression you need to implement is on page 5 of lecture 3 , and the RBF kernel is on page 12 of lecture 2 .

(4) Using Spline Regression to solve the regression problem. The spline regression you need to implement is on page 6 of lecture 3.

(5) Using Lasso Regression to solve the regression problem. The lasso regression you need to implement is on page 8 of lecture 3.

Please provide the following results for five methods:

$$Error_{\text{test}} = \sum_{i=1}^{m}\left(y_i - f\left(X_i\right)\right)^2,$$

where m denotes the number of the testing samples on the test set using four methods.

1. Try different penalty coefficients λ in methods (2)-(4), and give the prediction errors of at least two penalty coefficients λ for each method in methods (2)-(4).
2. Please visualize the weight vector β for each method. Specifically, given a weight vector $\beta = [\beta_1, \beta_2, \ldots, \beta_{12}]^\top \in \mathbf{R}^{12}$, you first need to sort $\{\beta_1, \beta_2, \ldots, \beta_{12}\}$ according to their values from the largest to the smallest. Then, you should visualize the sorted weight vector.
3. Please provide some analysis and discussion about your experimental results.

3 Requirements

To clarify, students must write their code and cannot use code from previous sources. Any plagiarized code will make you fail in this class. Specifically, students are not allowed to use any external packages or libraries except for basic algebraic operations and visualizations.

1. The report and the source code should be submitted separately, with the report in PDF or DOC format, and the code packaged into a compressed file.
2. The dataset and corresponding loading interface have been packaged in the "hw2" compressed file. Students can implement their regression models based on the interface.
3. A README document is needed. Please package it together with the source code. Please refer to the comments for more information.