Challenge - Advanced Regression Methods

November 2024

Description

We are given a statistical sample of size $n = 100$ with one dependent variable y and $p = 200$ independent variables $x^{(1)}, \ldots, x^{(p)}$. These data are provided in the file `data.txt` (training data).

The objective is to build the best predictive model for the dependent variable y using the independent variables $x^{(1)}, \ldots, x^{(p)}$. Finally, predictions must be made for the dataset provided in the file `Xtest.txt`, which contains 100×100 observations (100 rows, one for each prediction).

Evaluation Criterion

The performance of the predictive models will be evaluated using the standard criterion **Root Mean Square Error (RMSE)**:

$$\text{RMSE} = \sqrt{\text{MSE}} \quad \text{where} \quad \text{MSE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i - \hat{y}_i)^2$$

Here:

- $n_{\text{test}} = 100$: the number of predictions to be made,
- y_i: the observed values of y,
- $\hat{y}_i$: your predicted values of y.