Assignment 3

Machine Learning and Big Data for Economics and Finance

Consider the two variables in the dataset `Assign3.csv`. We are interested in predicting the second variable Y given the first variable X.

1. Fit a linear regression model to the data. Show the data scatter plot on the same figure with the values predicted by the linear model.
2. Fit a quadratic regression model to the data. Show the data scatter plot on the same figure with the values predicted by the quadratic model..
3. We are interested in constructing a step function learner as follows:

   First draw a random number U uniformly on the interval spanned by the minimum and maximum values of the inputs $(x_1, ..., x_n)$ and then use it to construct the following function whose purpose is to give the prediction of Y given $X = x$:

   $$f(x) = \alpha_1 I(U \leqslant x) + \alpha_2 I(U > x),$$

   where α_1 and α_2 are just unknown constants to be learned. It goes without saying that $I($`some statement`$)$ is the indicator function that equals 1 when the statement is true and 0 otherwise.

   a. Use two different methods to compute the estimate $\hat{f}(x) = \hat{\alpha}_1 I(U \leqslant x) + \hat{\alpha}_2 I(U > x)$. Is $\hat{f}$ a strong learner?

   b. Use one of the previous two methods to write an `R` function that takes as input x and the data $(x_1, ..., x_n, y_1, ..., y_n)$ and gives as output $\hat{f}(x)$.

      Make sure the function is capable of dealing with the case where x conatains more than one number.

   c. Using three different runs of the previous function, create three different plots where, on each, $\hat{f}$ is shown together with the scatter plot of the data.

4. Write an `R` function that applies boosting to the previous step function learner. That `R` function should take as inputs: the data, B the number of boosting iterations, λ the learning rate and an optional argument indicating the size of the test subsample in case a validation set approach is needed.

   As output the function should give: $\hat{f}_{\text{boost}}$ the boosted learner evaluated at the training data and the training mean squared error evaluated for each iteration $b = 1, ..., B$ of the boosting algorithm. Also, in case the size of the test subsample is greater than zero, the function should output: $\hat{f}_{\text{boost}}$ evaluated at the test sample and the test MSE evaluated for each iteration $b = 1, ..., B$.

   a. Use that function to plot $\hat{f}_{\text{boost}}$ on top of the data scatter plot for $\lambda = 0.01$ and for $B = 10000$. Show the same with different values of B.

   b. Plot the training MSE vs. the number of iterations.

   c. Was there overfitting when $B = 10000$?

Note: Even though the algorithm is described in detail in both the slides and textbook, for the sake of making the implementation easier, its special case pertaining to the questions in the assignment is presented here.

Boosting algorithm:

1. Inputs:
   - A sample of covariates (i.e. inputs) $x_1, ..., x_n$ and responses (i.e. outputs) $y_1, ..., y_n$.
   - A (weak) learner $\hat{f}$.
   - A learning rate $\lambda > 0$.
2. Initialize:
   - Set $\hat{f}_{\text{boost}}(x) \leftarrow 0$.
   - Compute the first learner $\hat{f}_0(x) = \hat{\alpha}_1 I(U \leqslant x) + \hat{\alpha}_2 I(U > x)$ on the original data.
   - Set $r_i \leftarrow y_i - \lambda \hat{f}_0(x_i)$ for $i = 1, ..., n$.
3. Do the following for $b = 1, ..., B$:
   a. Given $x_1, ..., x_n$ as covariates and $r_1, ..., r_n$ as responses, fit a learner $\hat{f}_b$ by first sampling U and then estimating $\hat{f}_b(x) = \hat{\alpha}_1 I(U \leqslant x) + \hat{\alpha}_2 I(U > x)$.
   b. Set $\hat{f}_{\text{boost}}(x) \leftarrow \hat{f}_{\text{boost}}(x) + \lambda \hat{f}_b(x)$.
   c. Set $r_i \leftarrow r_i - \lambda \hat{f}_b(x_i)$.
4. Output: $\hat{f}_{\text{boost}}(x)$.