CS4100/CS5100 Assignment 2

This assignment must be submitted by Friday 27 November 2020, 10:00. Feedback will be provided by 5 January 2021.

Learning outcomes assessed

The learning outcomes assessed are:

- develop, validate, evaluate, and use effectively machine learning models
- apply methods and techniques such as decision trees and ensemble algorithms
- extract value and insight from data
- implement machine-learning algorithms in R

Instructions

In order to submit, copy all your submission files into a directory, e.g., DA2, and create a compressed file, e.g., DA2.zip. Upload DA2.zip to Moodle through one of the following links:

Coursework assignment 2 CS4100
Coursework assignment 2 CS5100

You should submit files with the following names:

- `BDS.R` should contain your source code for boosted decision stumps (these can be split into files such as `BDS1.R` for task 1, `BDS2.R` for task 2, and `BDS3.R` for task 3); please avoid submitting unnecessary files such as old versions, back-up copies made by the editor, etc.;
- `report.pdf` should contain the numerical results and (optionally) discussion.

The files you submit cannot be overwritten by anyone else, and they cannot be read by any other student. You can, however, overwrite your submission as often as you like, by resubmitting, though only the last version submitted will be kept. Submissions after the deadline will be accepted but they will be automatically recorded as late and are subject to College Regulations on late submissions.

The deadline for submission is **Friday, 27 November 2020, 10:00**. An extension can only be given by the department office (but not by the lecturer).

Note: All the work you submit should be solely your own work. Coursework submissions are routinely checked for this.

Tasks

You are to implement decision stumps (DS) and boosted decision stumps (BDS) for regression. As explained in Chapter 6, decision stumps are decision trees with one split. Decision trees are covered in Chapter 1 and Lab Worksheet 2. Boosting is covered in Chapter 6 and Lab Worksheet 5. (See also Chapter 8 of [1].) For further details, see below.

Your programs should be written in R. *You are not allowed to use any existing implementations of decision trees or boosting* in R, or any other language, and should code DS and BDS from first principles.

You should apply your DS and BDS programs to the `Boston` data set to predict `medv` given `lstat` and `rm`. (In other words, `medv` is the label and `lstat` and `rm` are the attributes). There is no need to normalize the attributes, of course.

Split the data set randomly into two equal parts, which will serve as the training set and the test set. Use your birthday (in the format MMDD) as the seed for the pseudorandom number generator. The same training and test sets should be used throughout this assignment.

Answer the following questions:

1. Train your DS implementation on the training set. Find the MSE on the test set. Include it in your report.

2. Train your BDS implementation on the training set for learning rate $\eta = 0.01$ and $B = 1000$ trees. Find the MSE on the test set. Include it in your report.

3. Plot the test MSE for a fixed value of η as a function of $B \in [1, B_0]$ (the number of trees) for as large B_0 as possible. Do you observe overfitting? Include the plot and answer in your report.

Feel free to include in your report anything else that you find interesting.

Decision stumps

Decision trees in general are described on slides 39–53 of Chapter 1. Decision stumps are a special case corresponding to stopping after the first split. The description below is a streamlined (for this special case and for our data set) version of the general description.

The DS algorithm A decision stump is specified by its attribute (`lstat` or `rm`) and the threshold s. (Consider, e.g., $s = 1.8, 1.9, \ldots, 37.9$ in the case of `lstat` and $s = 3.6, 3.7, \ldots, 8.7$ in the case of `rm`.) The training RSS of a decision stump (`lstat`, s) is

$$\sum_{i:\texttt{lstat}_i < s} (y_i - \hat{y}_<)^2 + \sum_{i:\texttt{lstat}_i \geq s} (y_i - \hat{y}_\geq)^2$$

where both sums are over the training observations, $\hat{y}_<$ is the mean label y_i for the training observations satisfying $\texttt{lstat}_i < s$ and $\hat{y}_\geq$ is the mean label y_i for the training observations satisfying $\texttt{lstat}_i \geq s$. The training RSS of a decision stump $(\texttt{rm}, s)$ is defined similarly. Find the decision stump with the smallest training RSS. This will be your prediction rule.

Suppose the decision stump with the smallest training RSS is $(\texttt{rm}, s)$ (i.e., this is your prediction rule). The test RSS of this decision stump is

$$\sum_{j:\texttt{rm}_j<s} (y_j - \hat{y}_<)^2 + \sum_{j:\texttt{rm}_j\geq s} (y_j - \hat{y}_\geq)^2$$

where both sums are over the test observations, $\hat{y}_<$ is the mean label y_i for the training observations satisfying $\texttt{rm}_i < s$, and $\hat{y}_\geq$ is the mean label y_i for the training observations satisfying $\texttt{rm}_i \geq s$. The test MSE (to be given in your report) is the test RSS divided by the size m of the test set.

Remark. Using both RSS and MSE is somewhat superfluous, but both measures are standard. Remember that the test MSE is the test RSS divided by the size m of the test set, and the training MSE is the training RSS divided by the size n of the training set.

Boosted decision stumps

Boosting regression trees is described on slide 23 of Chapter 6. The description below is a more detailed version of that description.

The BDS algorithm

1. Set $\hat{f}(x) := 0$ and $r_i := y_i$ for all $i = 1, \ldots, n$.
2. For $b = 1, 2, \ldots, B$, repeat:
   (a) fit a decision stump $\hat{f}^b$ to the training data (x_i, r_i), $i = 1, \ldots, n$; in other words, $\hat{f}^b$ is the decision stump with the smallest training MSE
   (b) remember the decision stump $\hat{f}^b$ for future use
   (c) update $\hat{f}$ by adding in a shrunken version of the new stump: $\hat{f}(x) := \hat{f}(x) + \eta\hat{f}^b(x)$ (but we do not really need this!)
   (d) update the residuals: $r_i := r_i - \eta\hat{f}^b(x_i)$, $i = 1, \ldots, n$

The prediction rule is

$$\hat{f}(x) := \sum_{b=1}^{B} \eta\hat{f}^b(x).$$

To compute the test MSE (to be given in your report) of this prediction rule use the formula

$$\frac{1}{m}\sum_j \left(y_j - \sum_{b=1}^{B} \eta\hat{f}^b(x_j)\right)^2,$$

where the sum is over the test set and m is the size of the test set.

Marking criteria

To be awarded full marks you need both to submit correct code and to obtain correct results on the given data set. Even if your results are not correct, marks will be awarded for correct or partially correct code (up to a maximum of 75%). Correctly implementing decision stumps (Task 1) will give you at least 50%.

References

[1] Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*, Springer, New York, 2013.