STAT 2270: Data Mining
University of Pittsburgh
Fall 2020

Homework 4

1. Are regression trees linear smoothers? If they are, provide a proof; if not, explain why not. How about random forests?

2. Show that boosting with stumps (regression trees with only one split) produces a estimate for the regression function that is an additive model.

3. Read the paper "Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers" by Wyner et al. that was discussed in class. (You can find that paper here: `http://www.jmlr.org/papers/volume18/15-240/15-240.pdf`).

   (a) Figure 5 in Section 3.3 was one figure shown in class demonstrating the performance of different classifiers on what the authors call the "pure noise" model. Write down this model formally to make clear why the authors describe it this way. Provide some explanation.

   (b) Rewrite the model in (a) so that this model form produces exactly the same data, but now would appear to contain signal. Again, provide a few lines of explanation.

   (c) The authors argue that in this context, the interpolating nature of boosting and random forests is what is causing them to perform so well. Provide a brief summary of this argument.

   (d) In this particular context, that argument is completely valid. It does not, however, generalize to "real data" settings or really explain why these methods work so well across different datasets. Explain.

   (e) On a related note to the previous question, the authors are quite harsh in their critique of the classical statistical setup, arguing that in practice, most modern datasets contain and extremely high amount of signal. Explain why this represents a naive view of the standard regression setup where we assume $Y = f(\boldsymbol{X}) + \epsilon$.

4. Suppose that we have a dataset with many predictors and a binary response; $y \in \{0, 1\}$, for example.

   (a) Explain why you might still prefer to build a regression tree as opposed to a classification tree. What advantage does a regression tree have, even for (eventually) doing classification?

   (b) Give an example where a regression tree is used to perform classification and outperforms a traditional classification tree, as measured by misclassification error:

$$\sum_{i=1}^{n} \mathcal{I}\{\hat{y}_i \neq y_i\}.$$

That is, find a dataset where this is the case. The data can be real or simulated but should contain at least a few hundred observations on at least a handful of predictors. Are there any properties of the dataset that make it apparent that the regression tree approach might be preferable?

5. In class we discussed the fact that the out-of-bag measures have a tendency to artificially inflate the perceived importance of correlated predictors. Here we'll walk through a simple example demonstrating this.

   (a) Create a training dataset with 1000 observations on 10 predictors where the response is given by $Y = \sum_{i=1}^{10} X_i + \epsilon$. You can do this however you like, but $X_3, ..., X_10$ should be independent of each other as well as X_1 and X_2, X_1 and X_2 should be strongly correlated (let's say around 0.9), and ϵ should have mean 0 and relatively small variance. Generate an additional (independent) test dataset with 100 observations.

   (b) Using the `randomForest` function in `R`, build a random forest with the training data and be sure to set `importance = TRUE`. Look at the random forest importance measures and plot the `%IncMSE` scores for each predictor. You should see that X_1 and X_2 appear more important than the others.

   (c) Now let's measure the importance differently. Construct a total of 10 additional random forests. For the i^{th} forest, randomly permute the values of the i^{th} predictor variable before constructing it and record the difference in MSEs on the test set:

   $$\text{MSEperm}_i = \text{MSE}_{-i} - \text{MSE}$$

   where MSE_{-i} is the MSE from the random forest with the i^{th} predictor permuted, and MSE denotes the MSE from the forest built with the original data and evaluated on the test set. Compare these importance measures to those from part (b).

   (d) Repeat part (c) but instead of permuting the i^{th} predictor on the i^{th} iteration, simply remove it from the data frame. Compare these new importance measures to the two calculated previously. Would you prefer the importance measures calculated here, or those calculated in part (c)? Why?

6. *Demonstrating Double Descent:* Daniella Witten recently gave an excellent demonstration of the double descent phenomenon using regression splines (see the link). These models produce a spline fit via ordinary least squares on a transformed feature space. You can get more information about regression splines in the ISLR textbook. Repeat her analysis here. In particular, produce the "double-descent" curve in this context and see what happens when we use something like a ridge loss instead.