BST 764: Applied Statistical Modeling
Breheny

Assignment 4: Classification and regression trees
Due: Tuesday, November 22

Mathematical concepts and derivations

1. Consider the problem of solving for a regression tree split in a single dimension. Suppose x and y are both continuous, and all of their values are unique.

   (a) How many different split values $\{s_j\}$ must be evaluated in order to consider all possible splits of the form $x \leq s_j$?

   (b) For each of the split values in part (a), let

   $$\begin{aligned} u_j &= \sum_{i:x_i \leq s_j} y_i \\ v_j &= \sum_{i:x_i \leq s_j} y_i^2 \\ y_+ &= \sum_i y_i \\ y_+^2 &= \sum_i y_i^2. \end{aligned}$$

   Note that once you have obtained $\{u_j\}$ and $\{v_j\}$, calculating y_+ and y_+^2 is trivial. Derive RSS_j in terms of these four quantities, where

   $$\text{RSS}_j = \sum_{i:x_i \leq s_j} (y_i - \hat{c}_1)^2 + \sum_{i:x_i > s_j} (y_i - \hat{c}_2)^2.$$

   Your final answer should be a simple expression of u_j, v_j, y_+, and y_+^2 with no summations or other derived quantities (like $\hat{c}$) in it. (Note that if the $\{x_i\}$ values have been sorted, calculating the entire list of $\{u_j\}$, $\{v_j\}$, y_+, and y_+^2 can be done with the same computational burden as finding the variance of y).

   (c) Linear regression (provided that the design matrix is of full rank), has the nice property that if you consider RSS as a function of β, any local minimum is the one unique global minimum. Do regression trees have this property? In other words, if you were to plot RSS_j versus s_j, are you guaranteed to have exactly one local minimum? If "yes", prove it[1]; if "no", give a counterexample.

[1] For the proof, you may consider the simpler special case where $\{x_i\} = 1, 2, 3, 4$

Simulation

2. Conduct a simulation study comparing linear regression to regression trees. Generate data according to the following setup: For $i = 1, 2, \ldots, 100$, Let x_i follow a uniform distribution and let $y_i = x_i + \epsilon_i$, where ϵ_i follows a standard normal distribution. You may use either (or both) tree-based algorithms we discussed in class (`rpart` or `party`).

   To evaluate the two modeling approaches, generate test data sets with 1,000 observations from the same mechanism as above. For a criterion, use the mean squared prediction error minus the irreducible error (*i.e.*, the variance of y given x). This quantity is called the *model error*. Comment on which approach performs better and give an explanation for why it performs better.

3. Repeat problem 2 with the following data-generating mechanism: Let x_{1i}, x_{2i}, and x_{3i} follow independent random Bernoulli distributions with $p = 0.5$, and let $y_i = x_{1i}x_{2i} + x_{2i}x_{3i} + \epsilon_i$. Again, comment on the model error, and if your results differ from those of problem, comment on the reasons why.

4. Repeat problem 3, only compare the two tree-based approaches (`rpart` and `party`), and use the following data-generating mechanism: Let x_{1i} and x_{2i} follow independent random Bernoulli distributions with $p = 0.5$, and let $y_i = x_{1i}(1 - x_{2i}) + (1 - x_{1i})x_{2i} + \epsilon_i$. In words, y has a higher expected value if x_1 happens or x_2 happens, but not if they both happen. Again, comment on the model error and explain why the approaches performed as they did.

Application

5. Revisit our WHO data from earlier in the semester concerning the prediction of pneumonia based on clinical signs. Analyze the data using a tree-based method and comment on your results. In particular, comment on any statistical decisions you made during the analysis (*e.g.*, using `party` or `rpart`, treating the outcome as categorical or continuous, etc.), as well as on the medical/scientific interpretation of your model/algorithm.

   You may face the dilemma that the optimal model from a statistical perspective is too large to be easily interpreted. If you do, comment on whether it is possible to simplify the model, and if so, at what cost to its predictive accuracy?