STA 621: Nonparametric Statistics
Breheny

Assignment 6
Due: December 13

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. Let n denote the number of observations.

   (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 in $O(3n)$ operations, 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 of the implementations 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. The course website contains a data set (`kyphosis.txt`) from a study of children undergoing a corrective spinal surgical procedure known as a laminectomy. Some patients develop a postoperative spinal deformity known as kyphosis (an over-curvature of the vertebra or "hunchback"; see `http://en.wikipedia.org/wiki/Kyphosis` for pictures if you are curious). The purpose of the study was to assess the incidence of kyphosis in patients undergoing laminectomies, as well as to determine whether certain types of patients or procedures were at increased risk.

   The explanatory predictors in the data set are:

   - `Age`: The age of the child (in months).
   - `Start`: The number of the first vertebra involved in the laminectomy procedure. Vertebrae are numbered from the top down, with 1 denoting the topmost thoracic vertebra. Vertebrae 1-12 are known as the *thoracic vertebrae*, while vertebrae 13-17 are known as the *lumbar vertebrae*.
   - `End`: The number of the last vertebra involved in the laminectomy procedure, according to the same numbering scheme.

   Note that, although there are only three explanatory variables, there are many ways in which this information can be represented. For example, the first patient in the data set has `Start=5` and `End=7`; a potentially important derived predictor would be the number of vertebrae involved in the procedure (3 for this patient). It may also be important whether, for example, the entire procedure was confined to the thoracic vertebrae, or whether the procedure involved both thoracic and lumbar vertebrae.

(a) Analyze this data using a generalized additive model. Write up your findings to include a "methods" section, in which you describe the details of the model and your model building process (did you use splines or local regression, how did you select the smoothing parameters, did you look at interactions, etc.) as well as a "results" section, in which you present your model's estimates. Use tables and/or figures as needed to represent the model in the results section.

(b) Analyze this data using a tree-based method. Again, include a methods section describing the methodology and a results section which presents your findings.

(c) Briefly (in one or two paragraphs), summarize your main conclusions from (a) and (b) in terms of answering the study's primary research questions: "What is the incidence of kyphosis in laminectomy patients?" and "Do certain types of patients or procedures present an increased risk of kyphosis?"