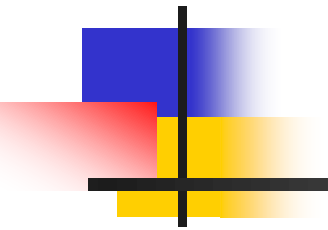

Pros and Cons of Trees and Multiple Trees

Dr. Goutam Chakraborty

Outline

- Pros and Cons of decision tree models
- From a single tree to multiple trees
  - Gradient Boosting
  - Random Forest

Decision Trees Pros and Cons

- Rules are transparent and easy to understand for non-technical people
- No need for imputing missing values
- No need for transformation
- No need for variable selection
- No need for assuming linear relationships between X's and Y

- One of the problems of a single decision tree is that any small change in the data can easily change the size and the shape of a tree.
  - There is an inherent tendency to overfit the data and it's difficult to determine the appropriate size.
- Trees tend to favor X's with many potential split points
- Trees don't work well when classes are separable via linear equations

Multiple Tress: Boosting

- Boosting is a form of *ensemble model*, where predictions from a set of multiple decision trees are combined into a single prediction.
  - *Boosting uses varying probabilities in selecting an observation* to be included in the sample.
  - All observations that had poor prediction performance, as indicated by a validation of the original decision tree, have a *greater probability of being selected* for the boosted sample

Main Features of Gradient Boosting in SAS EM

- Data set re-sampled several times
- The results is a weighted average of re-sampled data
- A *series* of base learner models are created
  - In Gradient Boosting, the base learner is a decision tree
- The *series* of base learners combined together forms a single predictive model (gradient boosting)

Multiple Trees: Random Forests

- A random forest is an average of multiple decision trees.
  - In each node, a branch search is performed *on a random set of inputs*, instead of on the full set of inputs.
  - The training data *is a random sample* of the original data set. A portion of the random sample is *set aside as a test sample*. Multiple decision trees are grown independently (in parallel).
- At each node of the developed decision tree, *a subset of inputs is selected at random out of the total number of inputs* that are available. The branch that is used is the one that produces the best split on this subset of inputs.
- Random forest approach could handle hundreds and thousands of input variables with no degeneration in accuracy

Gradient Boosting vs. Random Forest

- Trees in a **forest** are formed from a series of independent samples.
- Training data for an individual tree in a **boosting** model depends on the predictions of the trees already trained.
- Trees in a boosting model are generally small; trees in a forest are generally large.
- Which works better in a data set?

Demo of Gradient Boosting and Random Forest

- Continue to use the current diagram
  - Add a HP Random Forest node (HPDM tab) to the data partition node
  - Add a HP Gradient Boosting node (Model tab) to the data partition node
- Run both of the new nodes using default options
- Explore results
- Self study: look at SAS EM help guide for these two nodes