Lecture: Bagging, Boosting and Random Forest

Dr. Goutam Chakraborty

SAS® Professor of Marketing Analytics

Director of MS in Business Analytics and Data Science* (http://analytics.okstate.edu/mban/)
Director of Graduate Certificate in Business Data Mining (http://analytics.okstate.edu/certificate/grad-data-mining/)
Director of Graduate Certificate in Marketing Analytics (http://analytics.okstate.edu/certificate/grad-marketing-analytics/)

1

1

Outline

- An Overview of Multiple Decision Trees
  - Cross validation
  - Bootstrapping
  - Bagging
  - Boosting
  - Random Forest

2

2

Characteristics of a Single Decision Tree

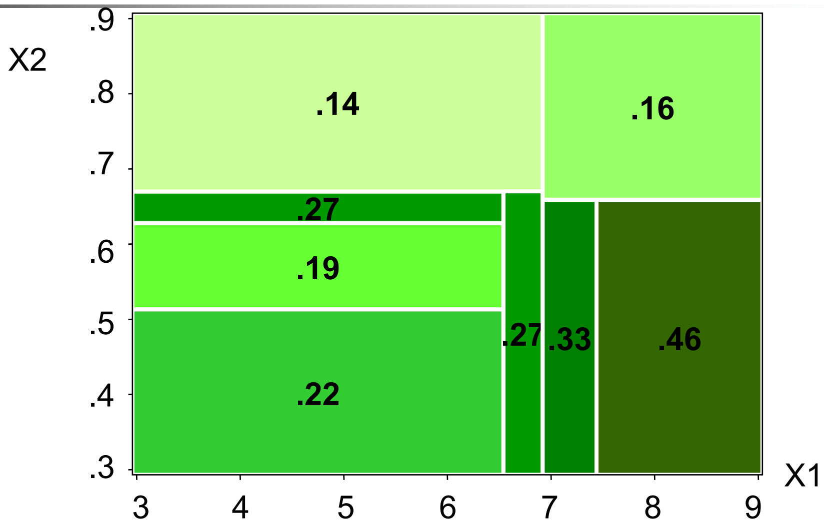
- Works reasonably well and very easy to understand
- But, one of the main problems of a single decision tree is that any small change in the data can easily change the size and the shape of a tree.
  - There is an inherent tendency to overfit the data and it's difficult to determine the appropriate size.
  - Using training-validation sample to trim a large tree to a small tree works ok but can be improved via multiple trees approach

Leaves in A Tree = Boolean Rules

If X1 $\in$ {*values*} and X2 $\in$ {*values*}, then $\hat{Y}$=*value*.

Leaf	X1	X2	Predicted Y
1	<6.5	<.51	.22
2	<6.5	[.51, .63)	.19
3	<6.5	[.63, .67)	.27
4	[6.5, 6.9)	<.67	.27
5	<6.9	≥.67	.14
6	[6.9, 7.4)	<.66	.33
7	≥7.4	<.66	.46
8	≥6.9	≥.66	.16

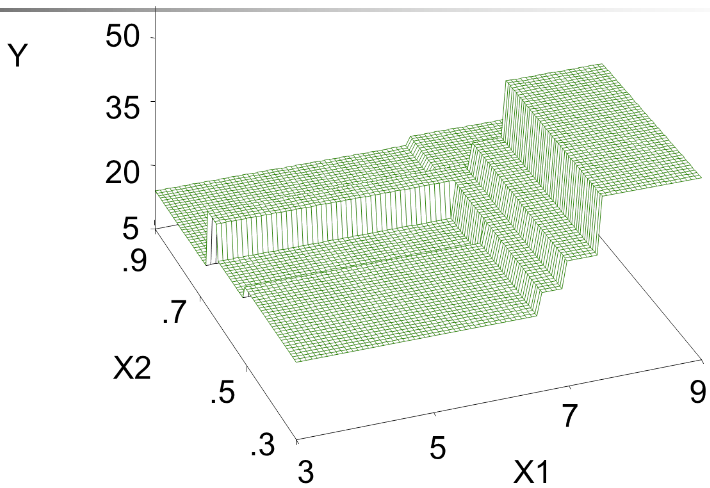
Partitioned Input Space



5

5

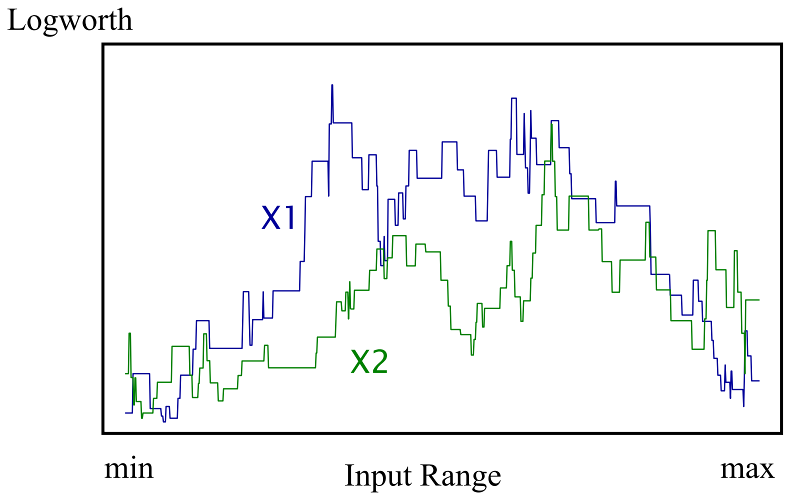
Multivariate Step Function



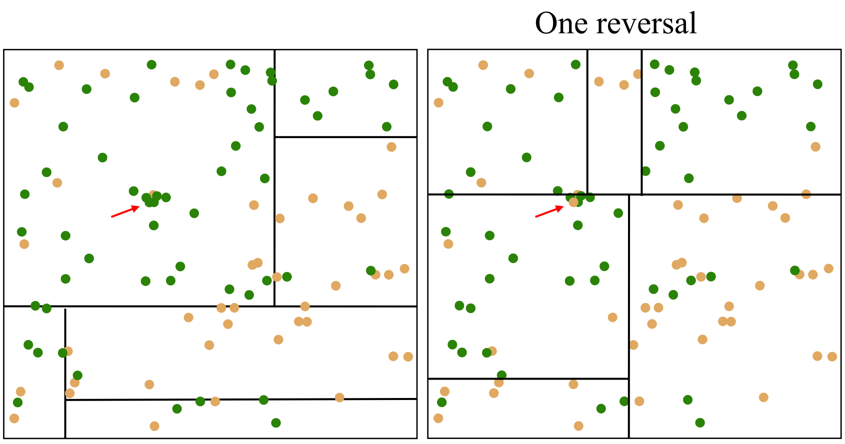
6

6

Competitor Splits



Instability



Accuracy = 81%   Accuracy = 80%

Major Multiple Decision Tree Methods

- Cross-validation
  - V-fold cross validation
- Bootstrap based methods
  - Bagging and boosting
  - Gradient boosting
  - Random Forests

9

9

K-Fold Cross-Validation

- The parent data set is partitioned into groups called folds.
- Typically, 10 folds are used; this is called 10-fold cross-validation.
  - Nine of the partitions are used as a new cross-validation training data set.
  - The 10% of the data that was held back is used as an independent test sample for the test decision tree.
- A different set of nine partitions is again collected into a cross-validation training data set.
  - The partition held back this time is different from the partition held back for the first test decision tree.
  - A second test decision tree is built and its classification error rate is computed.
- This process is repeated 10 times, building 10 separate test decision trees.

10

10

K-Fold Cross-Validation (Contd.)

- Once the 10 test decision trees have been built, their classification error rates (which is a function of *decision tree size*) are averaged.
  - This averaged error rate for a decision tree size is known as the cross-validation cost.
- The cross-validation cost for *each size* of the test decision tree is computed.
  - The decision tree size that produces the minimum cross-validation cost is found.
  - The parent decision tree is pruned to the number of nodes matching the size that produces the minimum cross-validation cost.

11

11

Bootstrapping

- Bootstrapping consists of constructing many subsamples, 50 to 2000, from an original data set.
  - Each subsample is a random sample *with replacements* from the full sample.
    - So, the same observation may be in multiple subsamples.
  - Each of these subsamples are used to train and test a model.
  - Collecting and displaying the pooled information from all the models will indicate how well the model (and the predictors) will perform in new data sets.
  - May be used with any predictive modeling tools (not just Trees)

12

12

Bagging

Bagging stands for bootstrap aggregation and refers to the creation of a pooled estimate of the target .

- Successive samples from the original data set are taken and the decision tree is trained in this sample.
- Typically, a *random sample with replacement* is taken.
- The non-sample observations can be used as validation data. These are called OOB (Out of Bag) observations.

- Bagging often improves accuracy of the predictions by helping to smooth out predictions (but, there is a cost of loss of interpretability)
- For continuous targets (regression tree), the predictions are *averaged*.
- For classification (categorical) targets, the predictions may be based on plurality voting.
  - An alternative strategy is to average the probabilities of the various categories occurring in the bootstrap samples, and to base the predicted class on these averaged posterior probabilities

13

13

Boosting

Boosting is a form of *ensemble model*, where predictions from a set of models are combined into a single prediction.

- Boosting operates much like bagging; however, *boosting uses varying probabilities in selecting an observation* to be included in the sample.
- In bagging, each observation is **equally likely** to be selected each time a new sample is created.
  - Therefore, no matter how many rules are developed, each decision tree that is produced from a boosting iteration has no dependence on any previous decision tree.
- The goal of boosting is to increase the probability of selecting an observation that performs well when predicting the target.
  - 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

14

14

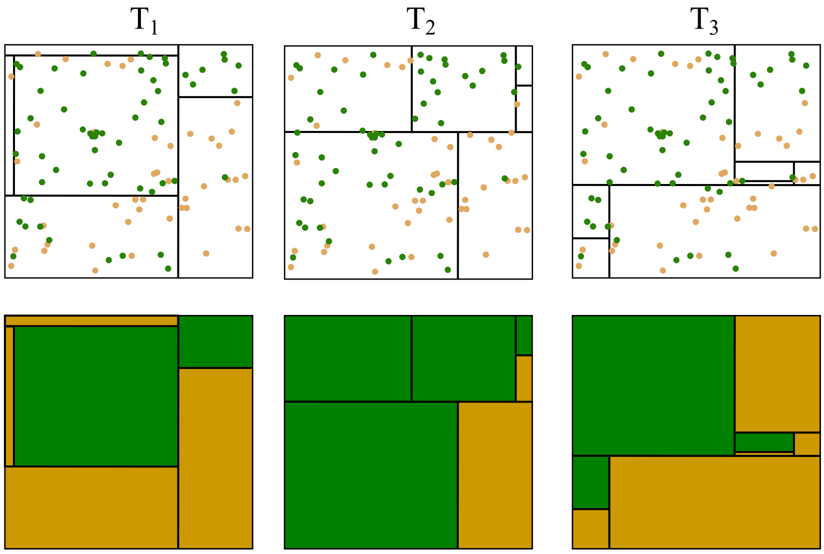
Bagging Vs. Boosting

- Bagging builds the decision trees in parallel and they vote on the prediction; boosting builds a series of decision trees and the prediction receives incremental improvement by each decision tree in the series.
- Bagging produces good results, but only if a single decision tree is reasonably effective to start with.
- Boosting has been shown to produce lower error rates than bagging in many situations.

15

15

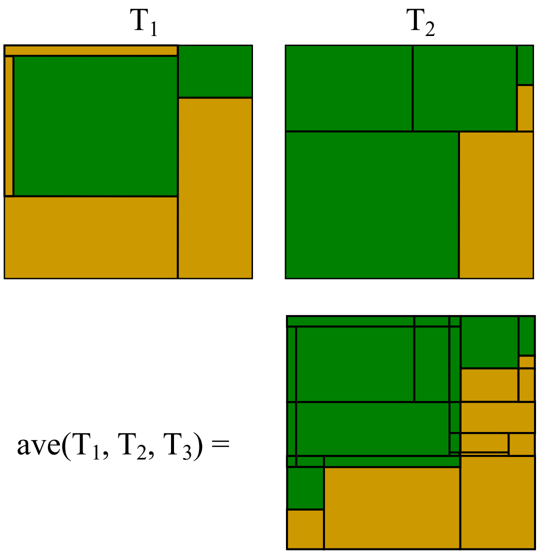
Perturb



16 …

16

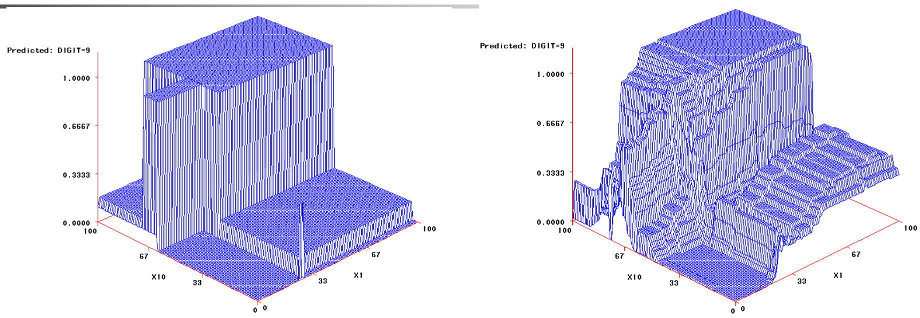
Combine



17 …

17

Single versus Bagged Trees



18

18

Random Forests

- A random forest is an average of 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. Like in bagging, decision trees are grown independently (in parallel).
- The randomness makes the variable selection less greedy (i.e., less likely to overfit)
- Each decision tree in the random forest is grown in a bootstrap sample of the training data set.
- 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

19

19

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 machine depends on the predictions of the trees already trained.
  - The data for training successive trees changes in two ways:
  - The target is the residual of the original target from the current prediction,
  - The training data for one tree is a sample without replacement of the available data
- Trees in a boosting machine are generally small; trees in a forest are generally large.

20

20

Summary Points

- Trees automatically handle missing values and variable reduction. Therefore, the input data requires less preparation.
- Forests tend to give better prediction than any specific tree, and often outperform other classes of models.
- Forests are **challenging to interpret**, but they can be considered an “ideal” model for other models to be compared against.

21

21