07 March 2018
18:31

Tree-Based Methods

Decision trees can be applied to both regression and classification problems.

if we can come up with a set of splitting rules to segment or stratify the predictor space into simple region so that we can classify the observation for a class of outcome variable and we summarize these splitting rules in a form of a tree…….this approach of a statistical learning method is called “DECISION TREE”.

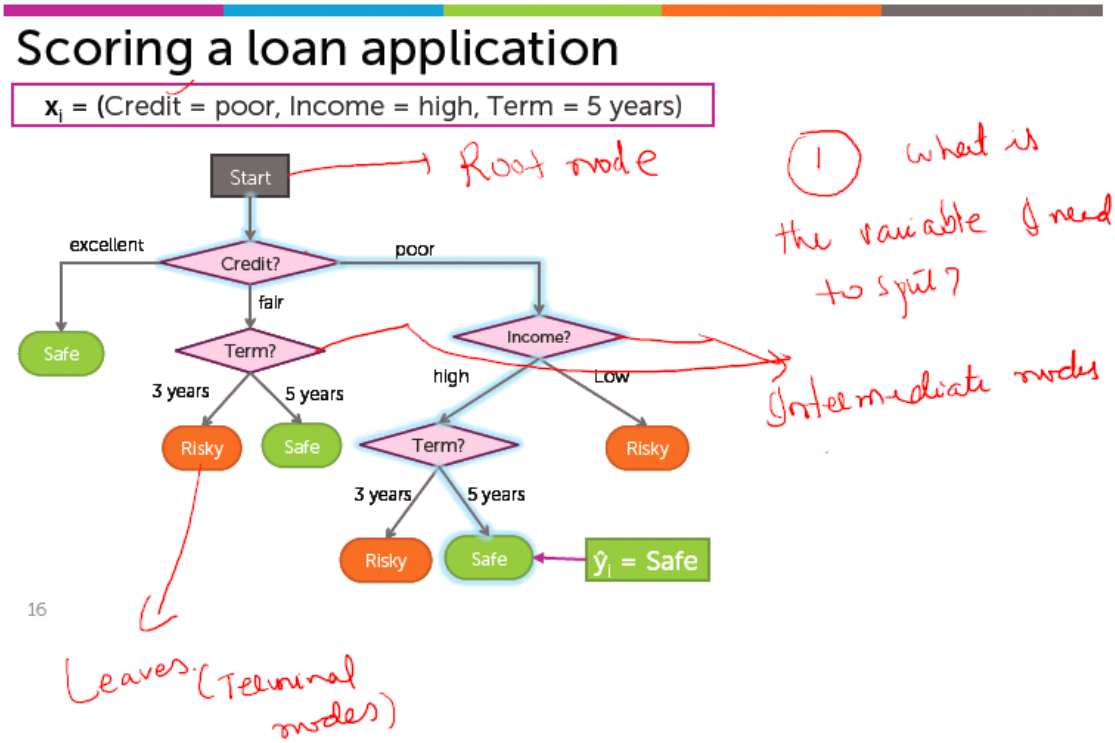
- can be applied for both regression and classification problems.
- good for non-linear data (relationship between outcome and predictors is non-linear)
- simple to implement and understand
- high interpretability
- not as accurate for prediction as other flexible methods.

but combining a large number of trees can often result in dramatic increase in prediction accuracy, at the expense of some loss in interpretability.

- Bagging, Random forests and Boosting are tree-based methods on this concept.

Ex: For particular person to give loan or not

Credit	Term	Income	y
excellent	3 yrs	high	safe
fair	5 yrs	low	risky
fair	3 yrs	high	safe
poor	5 yrs	high	risky
excellent	3 yrs	low	risky
fair	5 yrs	low	safe
poor	3 yrs	high	risky
poor	5 yrs	low	safe
fair	3 yrs	high	safe



- The Main variable what ever you choose which is going to split is called "ROOT NODE"
- The Next Level of Nodes is called "INTERMEDIATE NODES"
- The Nodes which are terminated at some point are "Leaves" or "Terminal Nodes"

STEP 1: What is the variable I need to split?

- Choose the variable such that your classification is Less.
- Suppose If we have a three variables i.e Credit, Term, Income.
- Split these variables till end and count the observation and which are having miss classification.
- Which ever is having Less value. Consider it has Root Node.

STEP 2: What Extent I need to split the tree?
The above process will be repeats until the number of observations will be closed at some terminal nodes.

SPLIT CRITERIA:

for Regression Tree:

- minimum RSS (Residual Sum of Squares).

for Classification Tree:

- Misclassification error (0 = perfect purity; 0.5 = no purity)
- Gini Index (0 = perfect purity; 0.5 = no purity)
- Information Gain (Cross-entropy) (0 = perfect purity; 1 = no purity)

Information gain uses log2, if loge then called the Deviance.
normally Gini-Index or information-gain is used to build trees as well as prune trees.
If prediction accuracy of the model is the goal then misclassification-error is used to prune the tree.