Tree Based Models

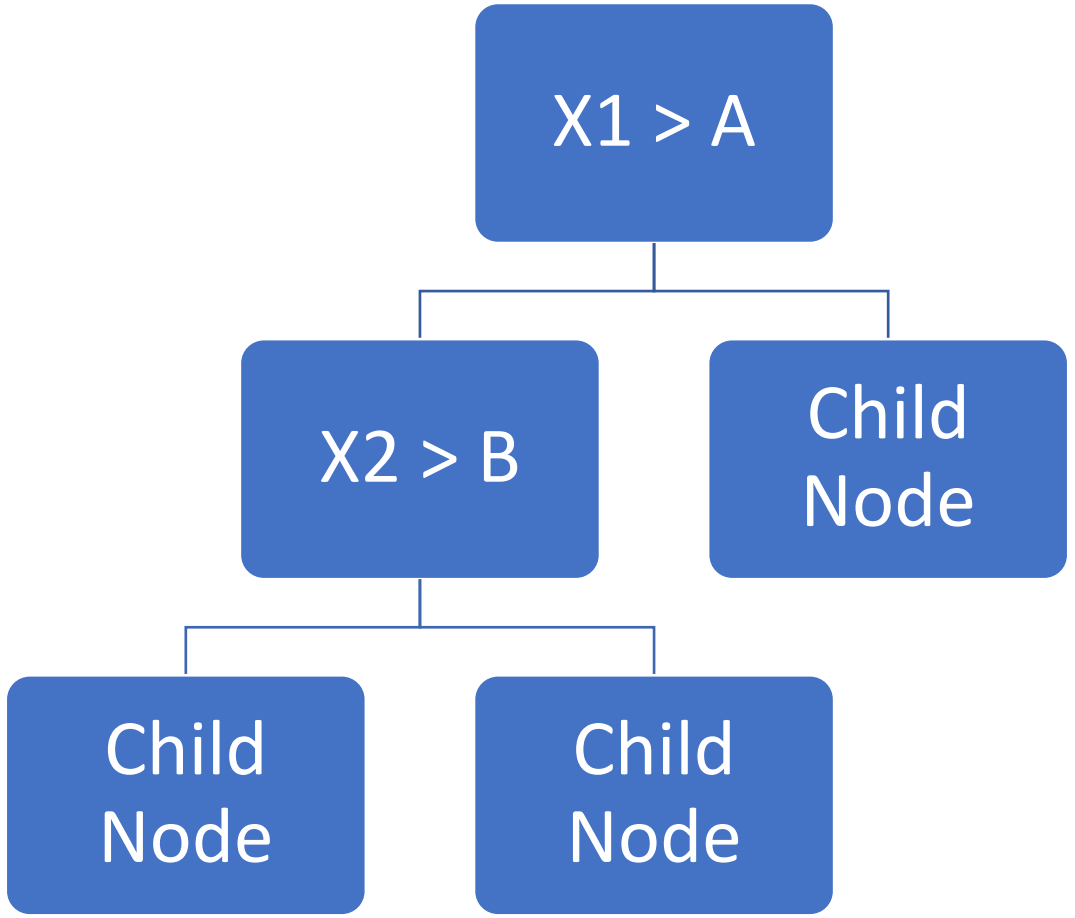
Bharat Bhushan Verma

What are Tree based Models

- Tree based models partition the predictor space into a number of simple regions.
- The challenge is to determine regional boundaries.
- We use the mean or mode of the region to predict a new observation.
- These splits determined by the threshold values of predictor variables can be summarized in form of decision rules and represented in form of trees. Hence called decision tree methods.
- They can be applied to both regression and classification problems.

Tree Terminologies

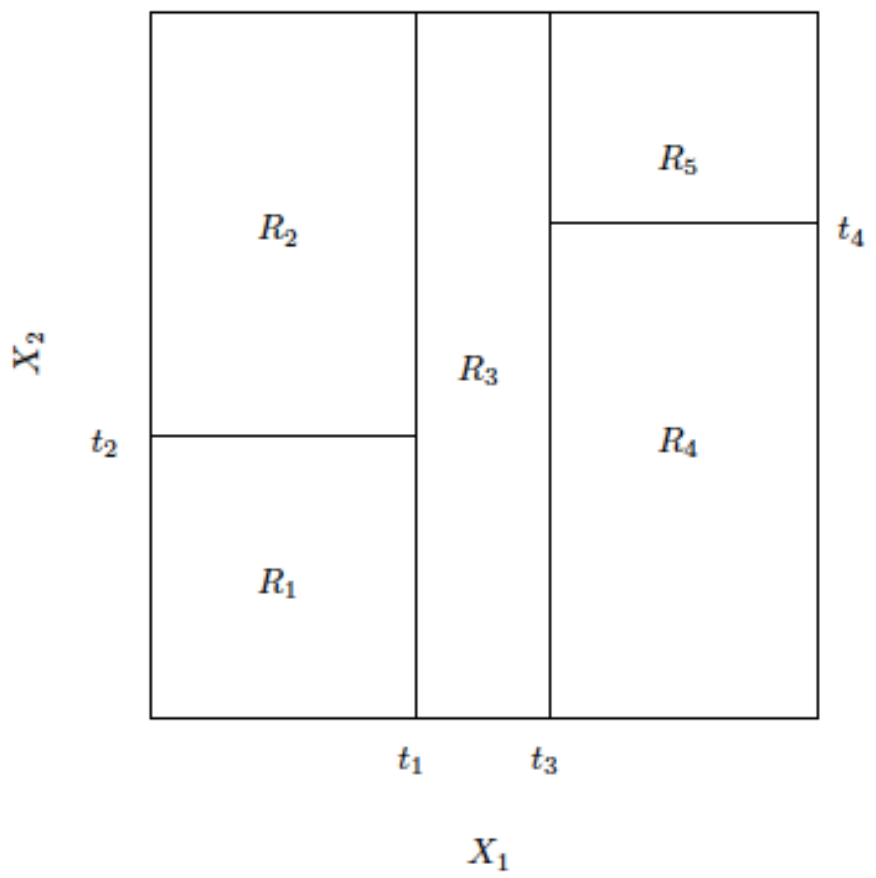
- Node on the top is called root or parent node.
- Node other than root node where the predictor space is split is called an internal node.
- Node which is not split further is called terminal node or leaves or child node.

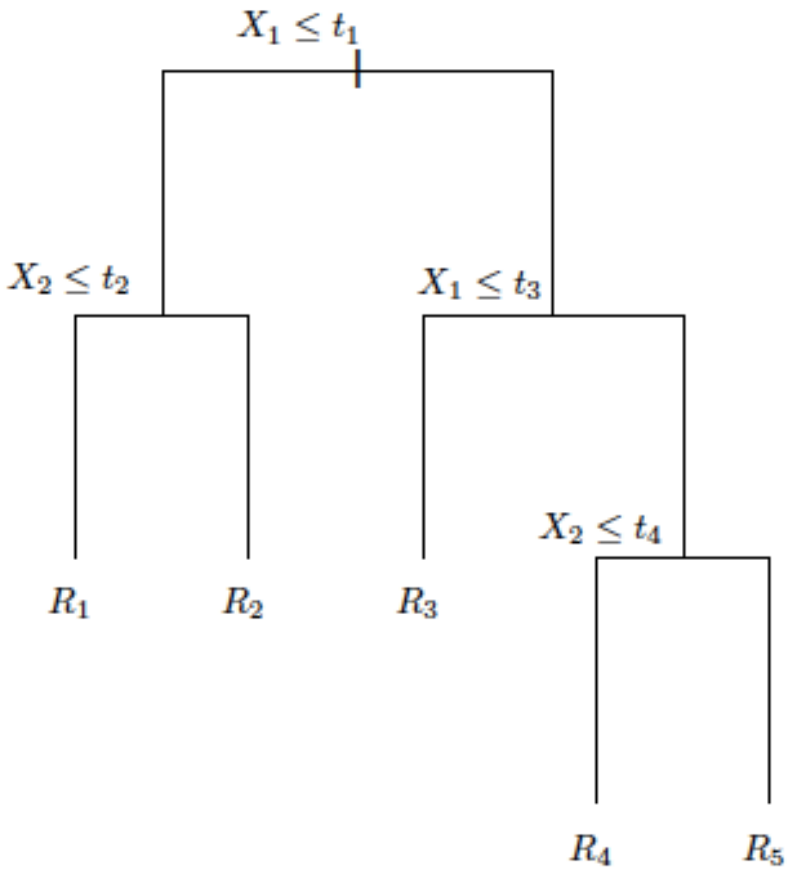


Recursive Binary Tree

- It involves two choices viz. Variables Xj and cut point A such that the predictor is split into {X|Xj < A} and {X|Xj >= A} which leads to greatest reduction in RSS.
- We consider all possible cut points for each predictor and choose the predictor resulting in lowest RSS.
- At every split we partition the one of the two spaces thus created.
- We stop when either a threshold child node size is reached or there are no more predictors to split.

Recursive Binary Tree





Tree Pruning

- Trees can quickly grow large and lead to overfit and hence poor predictions.
- A smaller tree with fewer splits can lead to lower variance and better interpretation with little bias.
- The strategy is to grow a large tree and then **prune** it back to subtree.
- Cost complexity pruning or weakest link pruning algorithm
  - We add tree complexity penalty alpha to SSR where alpha is a function of number of terminal node.
  - Alpha is a tuning parameter determined using cross validation.
  - So trade off between number of terminal nodes vs. number of splits.

Classification Trees

- In classification trees, it is not possible to calculate RSS as y is categorical.
- The prediction is made on the basis of most commonly occurring class.
- Instead of RSS, we use classification error rate = proportion of observations not belonging to most common class.
- Classification error rate is not sufficiently sensitive. So we use
  - Gini Index (Purity) is total variance across the classes
  - Cross Entropy is measure of underlying probability difference

Trees Vs Linear Models

Tree based models

- perform better when features and predicted variables have non-linear and complex relationships.
- have better interpretability and visualization.
- better mimic human decision making.
- Handle qualitative predictors without the need to create dummy variables.
- Have predictive lower accuracy in several feature response space.
- Are very sensitive to small changes in data.

Tree Vs Regression Models

- In a predictor space shown in first row, regression models work better.
- In a predictor space shown in last row, decision trees work better.