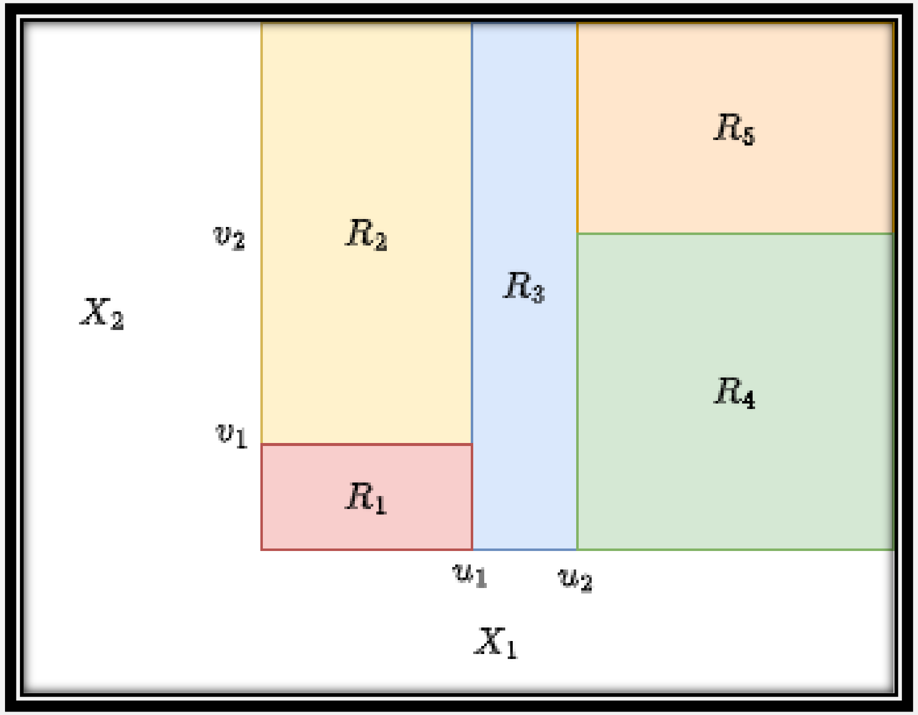
Decision Tree Learning

Concept of Decision Tree, Recursive induction of Decision Tree, Picking the Split variable, Entropy, Information Gain, Searching of Simple Trees, Computational Complexity, Pruning

Concept of Decision Tree

- Tree based model is simple, powerful and easy to explain.
- Tree based model partition the input space into a set of regions and then fit a simple model to each region.
- Input space is partitioned by the lines that are parallel to the coordinate axes
- Each partition can be modeled with different constants. Partitioning line can be described as $x_1 = c$.



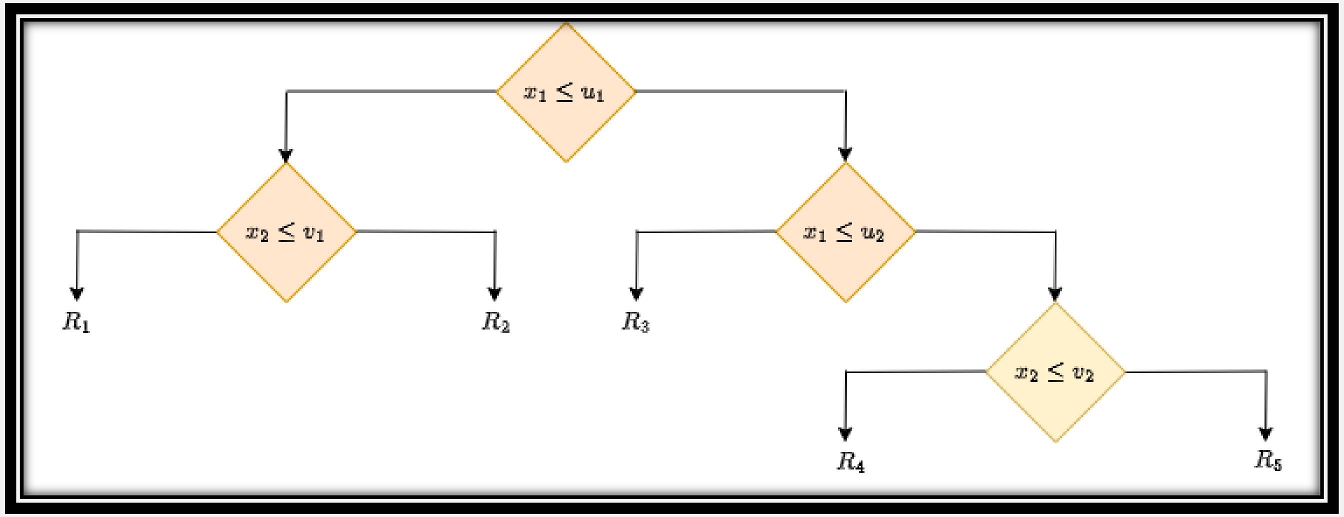
Recursive Induction of Binary Tree

- Determining split variable and split point is not easy task.
- Recursive binary partitioning can solve the problem.
- Split the space into two regions, and model the response by the mean of Y in each region.
- Chose the variable and split point to achieve the best fit.
- One or both regions are split into two more regions and these process is continued until some stopping criteria is applied.

Recursive Induction of Binary Tree

- The corresponding regression model predicts Y with a constant C_m in region R_m, that is

$$\hat{f}(x) = \sum_{i=1}^{5} C_m \mathrm{I}\{(x_1, x_2) \in R\}$$



Regression Tree

- For the dataset (X, Y) of N observation having p variable input $(x_1, x_2, \dots, x_p)$. The algorithm needs to decide splitting variable and splitting point automatically and also topology or structure of the tree we should have.
- For M regions $R_1, R_2, \dots, R_M$, a constant C_m is fit for the response variable.

$$f(x) = \sum_{m=1}^{M} C_m I(x \in C_m)$$

- Objective is to minimize the sum squared error $\sum\left(y_i - f(x_i)\right)^2$, then the best constant $\widehat{C_m}$ can be taken as the average of y_i in region R_m.

Picking the Split Variable

- Finding the best partition i.e. split variable and split point at each step is computationally infeasible.
- Greedy approach is used.
- For the split variable x_j and split point s, the regions defined as
- $R_1 = \{X|x_j \leq s\}$ and $R_2 = \{X|x_j > s\}$
- Algorithm has to find out x_j and s that solves
- $\min_{j,s}[\min_{c1} \sum_{x_i \in R_1\{j,s\}}(y_i - C_1)^2 \ - \min_{c2} \sum_{x_i \in R_2\{j,s\}}(y_i - C_2)^2 \ ]$

Picking the Split Variable

- For any choice of j and s, the inner minimization is solved by
- $\hat{C}_1 = avg\ \{y_i | x_i \in R_1(j, s)\}$
- $\hat{C}_2 = avg\ \{y_i | x_i \in R_2(j, s)\}$
- How large should we grow the tree?
- Very large tree may overfit the data and very small tree may miss important structure of the data.

Pruning

- Tree size is hyperparameter which requires tuning.
- Early Stopping: Stop splitting when the value of sum squared error is not decreasing significantly.
- Pruning: Grow a large tree. Stop when minimum size is reached then prune the tree.

Reduced Error Pruning

- Use training data for building the tree and validation dataset for pruning the tree.
- After building the tree, we replaced an internal node with it leaves then compare the performance of the new tree with the original tree.
- If the performance of the new tree does not change then keep the new tree, otherwise keep the original tree.

Classification Tree

- Let $p\hat{}_{mk}$ is the proportion of class k is in the region R_m with total $N_m$observation. Then $\hat{p}_{mk}$ can be defined as
$$\hat{p}_{mk} = \frac{1}{|N_m|} \sum_{x_i \in R_m} I(y_i = k)$$
- Then the observations of region R_m can be classified to the class
$$k(m) = \arg\max_{\mathrm{k}} \hat{p}_{mk}$$

The majority class in node m

Measures of Impurity

- Misclassification Error: $\frac{1}{|N_m|}\sum_{x_i \in R_m} I(y_i \neq k) = 1 - \hat{p}_{mk}$
- Gini Index: $\sum_{k \neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$
- Cross Entropy: $-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$
- If $k = 2$, i.e. two class classification, p is the probability that observations belong to 2nd class
- Cross Entropy: $-p \log p - (1 - p) \log(1 - p)$
- Gini Index: $2p(1 - p)$

Measures of Impurity

- The figure depicts that all the measures are similar.