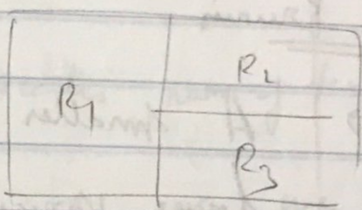
8.5 x 6.3

Decision Trees 10.35 x 7.75

→ Decision trees can be applied to both regression & classification problems.



(i) Regression Trees:-

Step 1:- Divide predictor space - i.e set of possible values for $X_1, X_2, .. X_p$ - into J distinct and non-overlapping regions, $R_1, R_2 .. R_J$

Step 2:- for every observation that falls in to R_j, we make same prediction (mean of responses in that region)

→ How to construct regions $R_1, R_2, .. R_J$?

- **Goal:-** find $R_1, .. R_J$ regions that minimize the RSS given by

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

- Top-down, Greedy approach :- (Recursive binary Splitting)
  because at each step of the tree-building process, the best split is made at the particular step, rather than looking ahead & picking a split that will lead to a better tree in future step

⇒ Stopping Criterion =

process of minimizing RSS continues till a stopping criterion (Ex: no region contains $\not>$ 5 observation)

Tree Pruning

→ A smaller tree with fewer splits ~~eg~~ may lead to lower variance & better interpretation at cost of a little bias.

⇒ ~~Alternative~~ possible alternative:- build the tree only so long as the decrease in RSS due to each split exceeds some (high) threshold..

→ But this process is short-sighted since a seemingly worthless split early on in the tree might be followed by a very good split, i.e a split that leads to a large reduction in RSS later on.

→ Better strategy is to grow a very large tree T_0, and then prune it back in order to obtain a subtree.

Goal: Subtree with lowest test error rate.

Method: <u>Cost Complexity pruning</u> or (<u>weakest link pruning</u>)

↳ Algorithm 8.1 (ISLR 7th)

⇒ Cost function now is

$$\sum_{m=1}^{|T|} \sum_{x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

$|T| \Rightarrow$ indicates num. of terminal nodes.

⇒ Note:- For a higher value of α, we get a tree which is sub tree of large tree ($\alpha = 0$).

→ use k-fold CV to find 'α' that gives best test Error.

(FYI: Each value of 'α' corresponds to one sustree)

(2) Classification Trees

→ Assign most commonly occuring class of training observations in that region.

→ Classification Error $= 1 - \max_k (\hat{p}_{mk})$

$\hat{p}_{mk}$ - proportion of training observations in the m^{th} region that are from k^{th} class.

Not sensitive to tree-growing, hence we use other alternative Gini index, Entropy etc

$= \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$ → measure of total variance across classes.

If Relationship b/w features & Response

is well approximated by a linear model
↓
linear Regression

Non-linear Relation
↓
Trees perform better than linear Model.

Yet to → Popular Tree-based methods
(a) → CART (Classification And Regression Trees)
- uses Gini Index
(b) → ID3 (Iterative Dichotomiser 3)
- uses Entropy & Information Gain
(c) → C4.5

Yet to → Interview questions on Decision Tree