Ch → 8

Tree Based Methods:

In this chapter we will learn tree based methods for regression & classification.
This involves segmenting a predictor space into a number of simple regions.

→ Since, the set of rules used to segment the predictor can be summarised in a tree
these types of approaches are known as decision tree methods.

Terminology for trees :

→ starting node is called root node.

→ Terminal nodes or leaves are nodes which don't split later.

→ Nodes between Root Node & Terminal Nodes are called internal nodes.

See illustrated example shown below:

Date :
Page :

Details of Tree building process:

→ A top-down greedy approach is used to perform reccursive binary splitting.

→ It is greedy because at each step of tree building process, the best split is chose (least RSS)

→ We first select a predictor "X_j" & the cutpoint "S" such that splitting leads to least RSS.

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

where j is the split parameter

→ We reccursively repeat the process until a stopping criterion is reached. ex: every split should have a minimum of 5 examples.

The procedure above might lead to Overfit: this is where pruning comes into picture.

Pruning a Tree:

A small tree with few splits might lead to lower variance & better interpretation.
We can do this by splitting using a high threshold to decrease RSS.
→ But this strategy is too short-sighted:
→ a seemingly worthless split early in the tree might be followed by a very good split, A better strategy would be to grow a large tree & prune it back inorder to select a tree.

→ Cost Complexity Pruning : also known as weakest link pruning is used to do this.

We consider a sequence of trees indexed by a non negative tuning parameter α.
For each value of α there corresponds a subtree $T \subset T_0$ such that

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

is as small as possible.

where, $|T|$ = # nodes in the tree
R_m = ~~Box~~ or the subset of predictor space.

α is a tuning parameter so, we will use cross validation.

Classification tree :-

The Only difference between a Regression & a classification tree is the target variable that is qualitative in nature

→ In classification tree we calculate the most commonly occuring class of training in the region to which it belongs.

Gini Index - Cross Entropy :

The Gini Index is defined as:

$$G = \sum_{K=1}^{K} \hat{p}_{mk} (1 - \hat{p}_{mk}),$$

a measure of total variance the K classes.
The more the purity, the lower the Gini Index.

An alternative to Gini Index is cross entropy, given by

$$D = - \sum_{K=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}.$$

(This is a part of Information Theory.)

This is called Information Gain

Bagging :-

Bootstrap aggregation or bagging, is a general purpose procedure for reducing the variance of a statistical learning procedure; we frequently used in context of decision trees.

& averaging a set of observations reduces variance (CLT)

This doesn't seem practical because we dont have multiple train sets.
Instead we can use bootstrap.
In this we create a bag of trainsets by repeated sampling from datasets.

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{*b}(x)$$

This is called bagging.

Random Forest:

→ Random forest provides an improvement over bagged trees by way of small tweaks that decorrelates the trees. This reduces the variance when we average the trees.

In Random Forest when building a decision tree, a random set of "m" predictors are chosen from "p" predictors & the split in trees are only allowed to use one of the "m" predictors.

"m" is typically chosen by p by the formula below

$$m = \sqrt{p}$$

Boosting trees:

Like bagging, boosting is a general approach that can be applied to many statistical learning methods for regression or classification.

We only discuss boosting for decision trees.

Boosting works in a similar way to bagging, except that trees are grown sequentially. Each tree is grown from previously grown tree.

Exact Explanation of all the algos will be done later in the notebook.