D

Decision tree → DT are non-parametric supervised learning method used for classification and regression. Goal is to create simple decision rules inferred from data features.
Eg- Customer will invest in Fixed Deposit (yes/no).

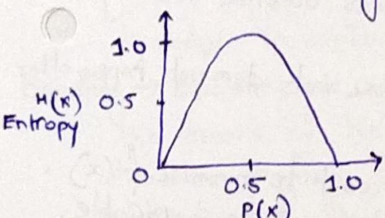
Assumptions → As it is a non parametric test, it does not assume anything.

Decision rules created → Whole training set considered as root, feature variable are to be categorical. If continuous, they are split to discretized. Creation of nodes are based on homogenity, and it will select the split which result in most homoganous sub-nodes.

Homogenity → Eg if a dataset contains only one label, then it is 100% homoganous/completely homogenous. More homogenity means most of the data points belong to same class and it will result in less error. Ultimate aim of DT is to increase homogenity.

DT to choose the split (homogenity) – i) Entropy ii) Information gain iii) Gini index iv) Chi-square.
The attribute with high value of information gain is placed at root.

Entropy — Measure of the randomness in the information being processed. Higher the entropy, harder it is to draw any conclusion from that information. Eg flipping a coin have highest entropy. Entropy is maximum when probability is 0.5, no chance to predict the outcome. Values lies between 0 and 1.



- Lower the value of entropy, higher is purity of the node.
- Entropy of homogenous node is 0 and a branch with entropy of zero is leaf node/terminal node. A branch with entropy more than zero need further splitting.

$$Entropy = -P_+ \log_2(P_+) - P_{(-)} \log_2 P_{(-)} \quad \text{or} \quad \sum_{i=1}^{n} -P \log P$$

Eg- suppose we have 3 yes and 2No, then entropy $= -\frac{3}{5}\log_2(\frac{3}{5}) - (\frac{2}{5})\log_2(\frac{2}{5}) = 0.78$

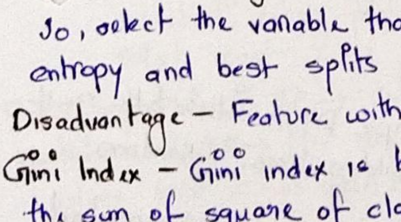
So we calculate entropy for all variables, variables which have lowest entropy selected for splitting. But when we select a node, it is splitted into many sub-node which will also have some entropy, so we need to do some kind of summation for all entropy (total entropy) for that we need information gain.

Information gain → Information Gain = Entropy (Before Split) – Entropy (After split)
So, select the variable that maximize the information gain, which in turns minimize the entropy and best splits the dataset into groups for effective classification.

Disadvantage – Feature with large number of values, generating larger decision trees.

Gini Index – Gini index is based on Gini impurity. Gini impurity is defined as 1 minus the sum of square of class probability in dataset. $Gini = 1 - \Sigma(P)^2 = 1 - [(P_+)^2 + (P_-)^2]$

Eg – For 3 yes and 3no, Gini Index $= 1 - [(P_+)^2 + (P_-)^2] = 1 - [(3/6)^2 + (3/6)^2] = 1 - [0.25 + 0.25] = 0.5$



Entropy $= -P_+ \log P_+ - P_- \log P_- = -\frac{3}{6}\log(\frac{3}{6}) - \frac{3}{6}\log(\frac{3}{6}) = 1$

So, Gini Index < Entropy. Because Gini Index range from 0 to 0.5.

Gini impurity are mostly used in ensemble technique like Random forest because it take less time than entropy as entropy contains log in formula.

So, Gini index ranges between 0 and 0.5. if dataset is pure, gini index = 0 or if two classes are equally distributed, gini index = 0.5. Feature with lowest Gini Index is used as next splitting features.

CHAID – Chi square Automatic Interaction Detector. It measure by sum of square of standardize differences between observed and expected frequencies of target variable.

$\chi^2 = \Sigma \frac{(O-E)^2}{E}$, χ^2 = Chi-sq obtained, O = Observed Score, E = Expected Score.

Higher the value of Chi-square, higher the statistical significance of difference between sub node and parent node.

Advantages of DT – Tree can be visualized, require little data preparation, (no normalisation etc), able to handle multiple output problem. Explaination of condition is easily explainable.

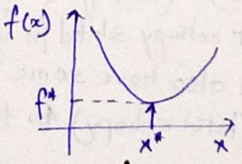
Issues faced in DT – Do not good at extrapolation, create biases tree if some class dominates, can be unstable because small variation in data might result to completly different tree being generated, overfit because tree become longer (solution is pruning), DT connot gurantee returning globally optima, instead they return local optima (solution ensemble learner, feature and sample with replacement).

Overfitting issue in DT → i) Pruning → Trim of the branches of tree, remove leaf node such that overall accuracy is not disturbed. ii) Random forest.

Interpolation and extrapolation – When we predict value that fall within range of data points taken it is called interpolation. When we predict value for points outside the range of data taken is called extrapolation.

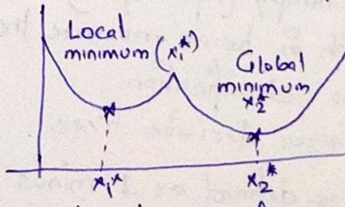
Disadvantage in intrpolation/extrapolation – Assume current trend to contine but does not happen often due to external factors, do not account undalying causes.

Local optima and Global optima – Objective function is $f(x)$ where we want to minimize $f(x)$. min $f(x)$ such that $x \in R$ where $f(x)$ = Objective function and x = Decision variable.



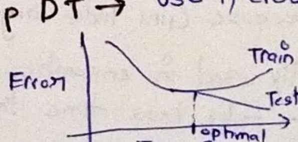
Suppose x^* is actual value of x at which function takes a minimum value and f^* is best value this function could possibly take.
So this function are called Convex f^n because we have only one minimum here.
So in this case minimum is both local and global minimum.

Non-convex f^n → Two points attains minimum. for x_1^*, function cannot take any better value from minimization. Same for x_2^*. But x_2^* is also global minimum if we take whole region.



Random forest → Eg of ensemble learning, in which we combine multiple DT to obtain better result.
why random – i) Random sampling of training dataset when building trees. ii) Random subsets of feature considered when splitting nodes.
Bagging is used to create an ensemble of trees where multiple training sets are generated with replacement.

When to stop DT → Use K cross validation and check optimal tree.



On error plot MSE (mean squ ar error) and on X axis plot no of trees.
So error will get same for train and test, and when division start, that point is optimal size of DT.

Ensemble technique → Techniques that create multiple models and combine them to produce improved results.

Ensemble techniques – Simple – i) Max Voting ii) Averaging iii) Weighted Averaging
Advance – i) Bagging ii) Boosting.

Max voting – Generally used for classification, where prediction of each model are considered as vote. The prediction which we get from majority of model are used as final prediction.
Eg – Suppose rate a movie for 5 colleage. 3 colg give 4 star and 2 colg give 5 star. So final rating will be 4.

Averaging – Average of prediction from all models. Eg (5+5+4+4+4)/5 = 4.4.

Weighted average – All models are assigned different weight defining the importance of each model prediction.
Weighted can be given on experience of colg – [(5*0.2) + (5*0.1) + (4*0.1) + (4*0.2) + (4*0.18)].

Out of Bag Sampling – OOB is random forest cross validation method. In this sampling, one third (1/3) of data is not used for training and can be used to evaluate its performance.
Very similar to leave-one-out-cross-validation but no additional computational burden.

Bagging → Bagging is combining the result of multiple models (for instance, all decision trees) to get a generalized result. Bagging (or Bootstrap aggregation) technique uses subsets (bags) to get fair idea of distribution (complete set). So Bootstrapping is a sampling technique in which we create observations from original dataset, with replacement.

Boosting → Boosting is sequential process, where each subsequent model attempts to correct the errors of the previous model. Thus boosting algorithm combine a number of weak learners to form strong learner. Individual model would not perform well on entire dataset, but they work well for some part of dataset. Thus each model actually boosts the performance of ensemble.

Bagging → Random forest   Boosting → AdaBoost, GBM, XGBoost etc.

Random Forest → Fits a number of decision tree, on various sub-samples of the dataset and uses averaging to improve predictive accuracy and control overfitting. Sub sample size is controlled with max_samples parameter if bootstrap= True (default), otherwise whole dataset is used to build each tree.

RF working → Select random samples from dataset provided, create a DT for each sample selected. Voting performed, classification use mode, regression use mean. And RF will select most voted prediction result as final prediction.

Feature importance in RF → Through Gini ~~im~~ impurity. Higher the value, more important is the feature.

Limitations in RF → For any data, RF has not seen before at best, it can predict avg of training values that it has seen before. RF cannot extrapolate the data. Eg what will be the population of India after 5 years.
So avoid feature ~~like~~ like Age, DOB, year made which are time dependent.

Solving the limitations → Use linear models ~~and~~ or avoid time feature variables for predictions.

CHAID – Chi-square automatic Interaction detector: Continuous predictors are split into categories with approximately equal no of observations. CHAID create all possible cross tabulations for each categorical variable until best outcome is acheived. and no splitting can be done. It is easy to visualize and use chi square test to check significance.

Limitations of CHAID – Since multiple split (buckets) it needs larger quantity of data to get desire result.