MACHINE LEARNING DAY 8 - DECISION TREE

CASE 1:-

=> Feature can be categorical } Classification Problem
outcome can be categorical

CASE 2:-

=> Feature can be continuous } Classification Problem.
outcome can be categorical

CASE 3:-

=> Feature can be continuous } Regression Problem
outcome can be continuous

★ Algorithms in Decision Tree :-

① ID3 - Iterative Dichotomiser 3
- It is used to solve classification Problem.
- It will work on categorical attribute.

② C4.5 - It is revision of ID3 algorithm, the way it divide the dataset is little bit different. It is also use for classification task.

③ CART :- Classification and Regression Tree :- It can be use for Regression and classification task.

④ CAID :- Chi-Square Automatic Interaction Detection

Example Dataset :-

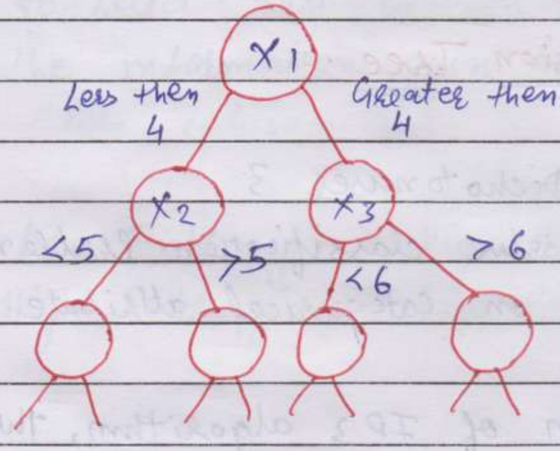
X_1	X_2	X_3	Y
1.1	7.5	2.5	A
2.2	8.8	5.5	B
3	9.2	6	A
3.6	5.1	6.7	A
5	5.4	7	B
5.8	2	8.9	B
8	1	9.1	A

Here Features → Continuous
Outcome → Categorical

How to select the node in this case?

The concept here is we have to create a threshold.

Suppose we take for X_1 Threshold = 4
X_2 Threshold = 5
X_3 Threshold = 6



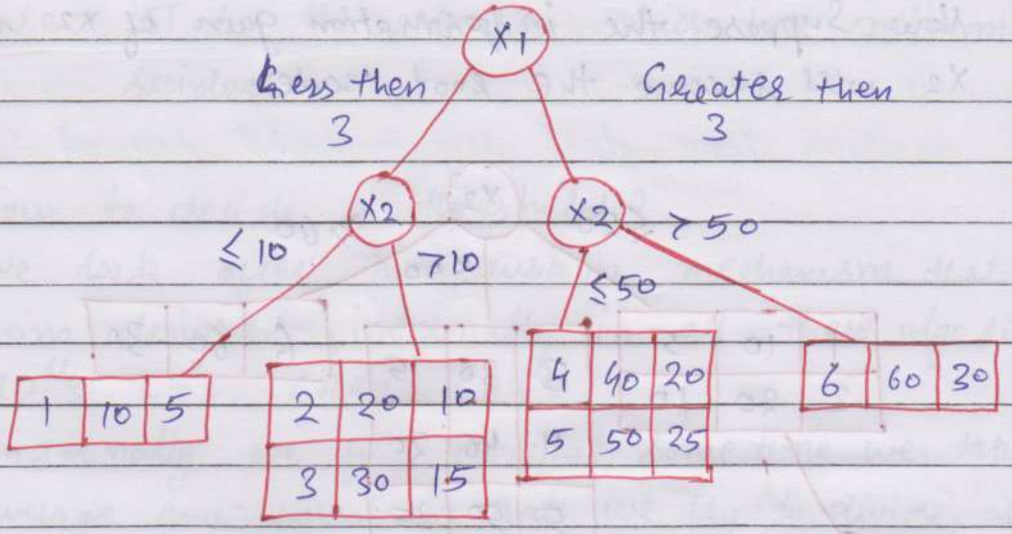
We will divide until we get the leaf node.

Another DATASET :-

X_1	X_2	Y
1	10	5
2	20	10
3	30	15
4	40	20
5	50	25
6	60	30

This is Regression Problem.

Suppose the information Gain X_1 is high. So X_1 will become root node.
For X_1 threshold = 3
X_2 threshold = 30



Now Calculate $Y - \hat{Y}$

Y : Actual value

$\hat{Y}$: Predicted value.

For

1	10	5

$Y = 5 \quad \hat{Y} = 5$

For

2	20	10
3	30	15

$Y = 10 \quad \hat{Y} = \left(\frac{10+15}{2}\right) = 12.5$

$Y = 15 \quad \hat{Y} = 12.5$

For

4	40	20
5	50	25

$Y = 20 \quad \hat{Y} = \left(\frac{20+25}{2}\right) = 22.5$

$Y = 20 \quad \hat{Y} = 22.5$

For

6	60	30

$Y = 30 \quad \hat{Y} = 30$

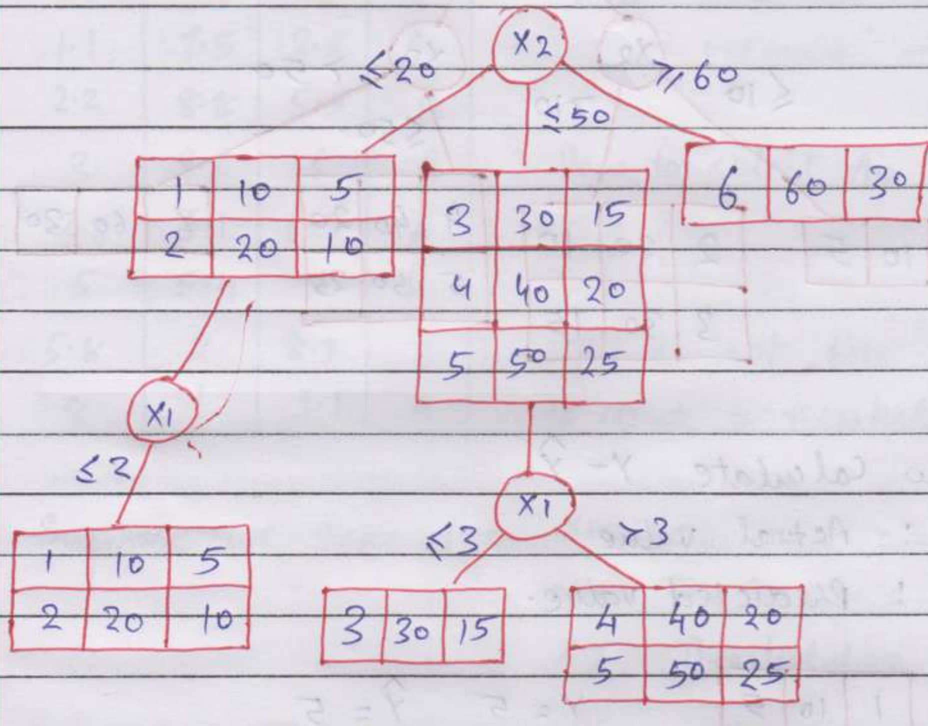
Total Residual :- $\sum (Y - \hat{Y})^2$

$\therefore (5-5)^2 + (10-12.5)^2 + (15-12.5)^2 + (20-22.5)^2 + (25-22.5)^2 + (30-30)^2$

$\therefore 0 + (2.5)^2 + (2.5)^2 + (2.5)^2 + (2.5)^2 + 0$

$\therefore$ Total Residual $= \underline{\underline{25}}$

Now Suppose the information gain of X_2 is high. And X_2 will become the root node.



Now test the tree $(Y - \hat{Y})^2$

For

1	10	5
2	20	10

$Y = 5 \quad \hat{Y} = \left(\frac{5+10}{2}\right) = 7.5$

$Y = 10 \quad \hat{Y} = 7.5$

For

3	30	15

$Y = 15 \quad \hat{Y} = 15$

For

4	40	20
5	50	25

$Y = 20 \quad \hat{Y} = \left(\frac{20+25}{2}\right) = 22.5$

$Y = 25 \quad \hat{Y} = \left(\frac{20+25}{2}\right) = 22.5$

$$\therefore \text{Total Residual} = (5-7.5)^2 + (10-7.5)^2 + (15-15)^2 + (20-22.5)^2 + (25-22.5)^2 + (30-30)^2$$

$$= (2.5)^2 + (2.5)^2 + 0 + (2.5)^2 + (2.5)^2 + 0$$

$$= 6.25 + 6.25 + 6.25 + 6.25$$

Total Residual $= 25$

Note :- Take the situation which has minimum residual.

Question How to decide a threshold?

Ans. We look after the custom mechanism that is being designed, internally in all these algorithms. ID 3, C4.5, CART, CAID.

Internally we find out that some time we take average and based on that we try to divide, or we make different different bins or different different blocks. This are some ways to create threshold.

★ **POST PRUNING** : This technique is used after the construction of desision tree.

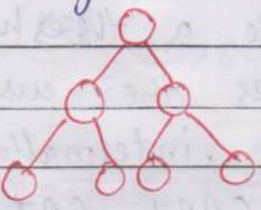
- This technique is used when decision tree will have very large depth and will show overfitting of model.
- It is also known as backward pruning.
- This technique is used when we have infinite grown decision tree.

★ **Pre - Pruning** :- This technique is used before construction of decision tree.

- Pre - Pruning can be done using Hyperparameter tuning.
- overcome the overfitting issue.
- Forward Pruning.
- In Pre - Pruning, we stop our tree to create insignificant branches before the construction of the tree.

Que How can we stop it ?

Ans. we can set criteria, that whenever you build decision tree don't try to build beyond 3 layer.

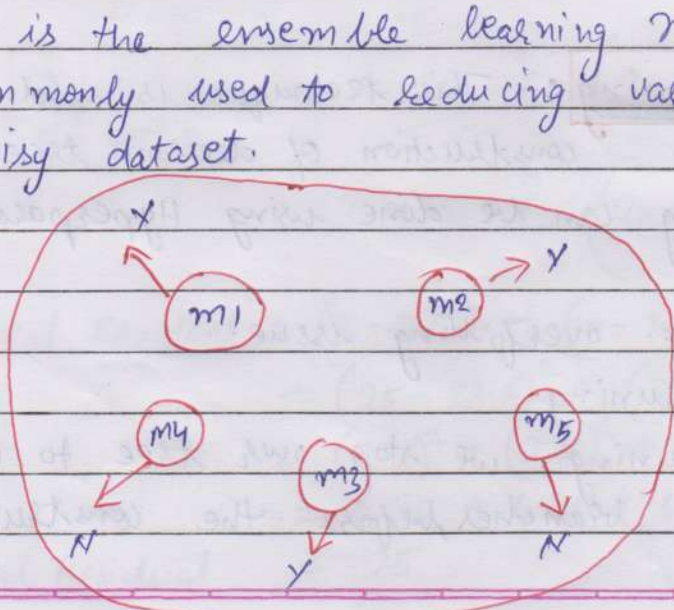


★ Ensemble Technique :-

- A technique where we involve multiple decision maker.
- Ensemble learning helps in improving machine learning results by combining several models.
- Where ever we are using multiple models to make a decision, so that kind of instance is called as Ensemble technique.
- we are not depend upon one single decision maker. So that we will be able to make a strong classifier or regressor.

① Bagging [Bootstrap Aggregation]

- It is the ensemble learning method that is commonly used to reducing variance within a noisy dataset.



We take a bag and put one one vote in it.

- In Bagging, we have a bag in which everyone will make a decision, and then the final decision will be average or may be majority

② **Boosting Technique :-**

- It is an ensemble modeling technique that builds a strong classifier from the number of weak classifiers. It is done by building a model by using weak model in series.

- Firstly a model is build from the training data. Then the second model is built which tries to correct the errors present in the first model.
- This procedure is continued and models are added until either the complete training dataset is predicted correctly or the maximum number of models are added.

★ Difference between Bagging and Boosting?

Ans - Bagging decreases variance, not bias and solves over-fitting issues in a model.
- Boosting decreases bias, not variance.

③ Stacking :- It is an ensemble method that enables the model to learn how to use combine predictional given by learner models with meta-models and prepare a final model with accurate prediction.

Bagging (Bootstrap Aggregation):- It is a way by which we divide a dataset into different different samples with an overlap. or with a replacement. In this a random sample of data in a training set is selected with replacement - meaning that the individual datapoints can be chosen more then once.

Posting:-

It is a way to divide a dataset into different different samples but there will be no overlap and there will be no replacement.

☆ Random Forest:-

As we know that forest is the combination of multiple trees. Random forest is the combination of decisions from many decision trees.

- It is one of the algorithm, which is a part of ensemble technique [Bagging] process. So inside bagging, we can use random forest directly.

- We can use any algorithm (SVM, KNN, Naive Biase, Decision Tree, Random Forest) in Bagging but the problem with Random Forest is that random Forest by default create multiple decision tree. internally. So, we will not be able to figure out how 1000 decision maker is created.

- Random Forest only take one algorithm i.e. decision tree.

★ Ada Boost :- [Adaptive Boosting]

Ada Boost is an ensemble learning method (also known as 'meta-learning') which was initialy created to increase the efficiency of linary classifiers.

Ada Boost ~~was~~ uses an iterative approach to learn from the mistakes of weak classifier and turn them into strong ones.

The steps to implement the Ada Boost algorithm using the decision trees are as follows.

Algorithm:-

Assume

Number of training samples — N

Number of Decision makers (models) = M

The possible decisions or class outputs are

$$Y = (-1, 1)$$

① Initialize the observation weights $w_i = \frac{1}{N}$

where $i = 1, 2, 3, \ldots N$ for all the samples.

② For $m = 1$ to M :

- fit a classifier $G_m(x)$ to the training data using weights w_i.
- Compute $err_m = \dfrac{\sum_{i=1}^{N} w_i \, I(y_i \neq G_m(x))}{\sum_{i=1}^{N} w_i}$

- compute $\alpha_m = \frac{1}{2} \log\left(\frac{(1 - err_m)}{err_m}\right)$.

This is the contribution of that tree to the final result.

- calculate the new weights using the formula:

$w_i \leftarrow w_i \cdot \exp\left[\alpha_m \cdot I(y_i \neq G_m(x))\right]$, where $i = 1, 2, 3 \cdots\cdots, N$

I means intersection of y where $y_i \neq G_m(x)$

- Normalize the new sample weights so that their sum is 1.
- Construct the next tree using the new weights.

3. At the end, compare the summation of results from all the trees and the final result is either the one with the highest sum (for regression) or it is the class which has the most weighted voted average (for classification).

$$\text{output } G_m(x) = \text{argmax}\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$$

(Regression)

$$\text{output } G_m(x) = \text{sign}\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$$

(Classification)

Note :-

We can consider $G_m(x)$ as $\hat{Y}$ which means Predicted value.