Tree model

Contents

Section 1: Decision Tree

Decision tree is a supervised machine learning model used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. Tree based models empower predictive models with high accuracy, stability and ease of interpretation.

1.1 Basic background of decision tree model

1.1.1 Two types of decision trees

Types of decision tree is based on the type of target variable we use.

a. **Classification Trees**: A tree with categorical target variable is called classification or categorical decition tree. Such as True/False, Yes/No, education levels.

b. **Regression Trees**: A tree with continuous target variable(typically real numbers) is called regression tree. Such as income, area.

1.1.2 Important Terminology related to Decision Trees

1. **Root Node**: The topmost decision node in a tree which corresponds to the best predictor called root node.

2. **Splitting**: The process of dividing a node into two or more sub-nodes.

3. **Decision Node**: When a sub-node splits into further sub-nodes, then it is called decision node.

4. **Leaf/Terminal Node**: Nodes that do not split anymore is called leaf or terminal node.

5. **Brance/Sub-Tree**: A sub section of entire tree is called a branch or sub-tree.

6. **Pruning**: Removing sub-nodes of a decision node called pruning, it's the opposite process of splitting.

7. **Parent and Child Node**: A node which is divided into sub-nodes is called parent node of sub-nodes where as sub-nodes are the children of parent node.

1.1.3 Advantages of tree model

1. Easy to understand and interpret. Decision tree output is very easy to understand even for people without analytical background. No requirment of statistical knowledge to read and interpret them. Its graphical representation is very intuitive and users can easily relate their hypothesis.

2. Useful in data exploration and variable selection. Decision tree is one of the fastest way to identify most significant variables and relation between two or more variables. With the help of decision tree, we can create new variables/features that has better power to predict target variable.

3. Less data cleaning required. It required less data cleaning or preprocessing procedure compared to some other modeling method.

4. Tree models have good robustness and stability. t is not influenced by outliers and missing values to a fair degree.

5. Data type is not a constraint: It can handle both numerical and categorical variables.

6. Non Parametric Method: Decision trees have no assumptions about the space distribution and the classifier structure.

1.1.4 Disadvantages of tree model

1. Over fitting. Over fitting is one of the most practical difficulty for decision tree models. This problem gets solved by setting constraints on model parameters and pruning (discussed in detailed below).

2. Not fit for continuous variables: While working with continuous numerical variables, decision tree looses information when it categorizes variables in different categories.

1.2 Tree spliting procedure

Tree models follow a *top-down greedy* approach known as recursive binary splitting. We call it as 'top-down' because it begins from the top of tree when all the observations are available in a single region and successively splits the predictor space into two new branches. We call it as 'greedy' because the algorithm looks for best variable available in the current stage and only cares the current split.Future splits will not be considered even if it lead to a better tree.
Generally, we want to increase the purity of a node. When the tree model is analysing an attribute to partition the data at the node, we want that each partition is as homogeneous as possible. This means we would like to see most of the instances in each partition belonging to as few classes as possible and each partition should be as large as possible. There are several ways or measurements we can generate to calculate the purity:

1. Entropy

2. Gini Index

3. Reduction in Reduction
   We will discuss more detail about each measurement and its relative model in the following part.

The splitting process will continue until all leaves reach to maximum purity, or a user defined stopping criteria is reached. For example, we set the minimum number of a nude to 50, then once a node has less than 50 observations, spliting will stop on this node. Such splitting process results in a fully grown tree until the stopping criteria is reached. However, the fully grown tree is likely to overfit data, leading to poor accuracy. Pruning is one of the technique used tackle overfitting. We'll learn more about it in a later section.

Section 2: Entropy, ID3 & C4.5 algorithm

2.1 Entropy

Entropy, or **Information Entropy**, is the average rate at which information is produced by a stochastic source of data. The measure of information entropy associated with each possible data value is the negative logarithm of the probability mass function for the value. Generally, entropy refers to disorder or uncertainty. A data set with smaller entropy has greater certainty. Shannon defined the entropy H(Greek capital letter eta) of a discrete random variable X with possible values $\{x_1, ..., x_n\}$ and probability distribution function $P(X)$ is:

$P(X = x_i) = p_i,\ i = 1, 2, ..., n$

The Entropy can be written as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) log_2 p(x_i)$$

When the log base equals to 2, then the corresponding units of entropy are the bits. Nats for base on e, and bans for base on 10.In the case of $P(x_i) = 0$ for some i, the value of the corresponding summand $0 log_b(0) = 0$, which is consistent with the limit:

$$\lim_{p \to 0+} p log(p) = 0$$

2.2 ID3 Algorithm

ID3 is one of the most common decision tree algorithm. It was first introduced in 1986 by John Ross Quinlan and it is acronym of **Iterative Dichotomiser**. It calculates the entropy and information gains of each atrribute to look for the most dominant attribute and put it on the tree as decision node. Thereafter, entropy and gain scores would be calculated again among the other attributes and next most dominant attribute will be picked as decision node. We use following dataset about decisions making on playing tennis ourside or not as an example to illuminate the calculating process of ID3.

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Information gain is calculated as:

$Gain(S, A) = Entropy(S) - \sum[p(S|A) Entropy(S|A)]$

We calcualte the entropy of the target variable first, here is the decision column. Decision has two labels,

among 14 obervations, there are 9 yes and 5 no. So the entropy of decision would be:

$Entropy(Decision) = -p(Yes) * log_2 p(Yes) - p(No) * log_2 p(No)$

$Entropy(Decision) = -9/14 * log_2(9/14) - 5/14 * log_2(5/14) = 0.940$

Gain from attribute wind

2.2 ID3 Algorithm

2.3 C4.5 Algorithm

Section 3: Gini index & CART algorithm

3.1 Gini index

3.2 CART Algorithm