Decision Tree for classification

The decision tree Algorithm belongs to the family of supervised machine learning algorithms. It can be used for both a classification problem as well as for regression problem. The concept behind the decision tree is that it helps to select appropriate features for splitting the tree into subparts and the algorithm used behind the splitting is ID3(Iterative Dichotomies-3).

The goal of this algorithm is to create a model that predicts the value of a target variable, for which the decision tree uses the tree representation to solve the problem in which the **leaf node** corresponds to a **class label** and **attributes are represented on the internal node of the tree**.

In decision tree we have to keep in mind two things:

1. Entropy.
2. Information gain.

Decision Tree

Entropy:

Entropy helps us to build an appropriate decision tree for selecting the best splitter. Entropy can be defined as a measure of the purity of the sub split. Entropy always lies between 0 to 1. The algorithm calculates the entropy of each feature after every split and as the splitting continues on, it selects the best feature and starts splitting according to it.

The formula of Entropy:

$$Entropy(S) = \sum_{i=1}^{n} -P_i log_2 P_i$$

P = Probability

Decision Tree

Gain:

The internal working of Gini impurity is also somewhat similar to the working of entropy in the Decision Tree. In the Decision Tree algorithm, both are used for building the tree by splitting as per the appropriate features but there is quite a difference in the computation of both the methods. Gini Impurity of features after splitting can be calculated by using this formula.

The formula of Entropy:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \times Entropy(S_v)$$

Note: If entropy is E_A before partition and entropy is E_B after partition then our target is to increase the difference between E_A and E_B as much as possible. Means $E_A \gg E_B$

Note: We will always try to minimize the Entropy and maximize the Gain.

Decision Tree

Let’s try to understand how to make a decision tree using an Example:

1	Gender	Car	Travel cost	Income	Transport
2	male	0	cheap	low	bus
3	male	1	cheap	medium	bus
4	female	1	cheap	medium	train
5	female	0	cheap	low	bus
6	male	1	cheap	medium	bus
7	male	0	standard	medium	train
8	female	1	standard	medium	train
9	female	1	expensive	high	car
10	male	2	expensive	medium	car
11	female	2	expensive	high	car
12					

Decision Tree

Iteration:1 and Root Selection

At first we have to find the root node of the decision tree. In our dataset four features and one target features. So, we have to find for gain for all four features. Which will give us the maximum gain we will choose that features as a root node.

Note: Before partition we have to find the entropy for target feature. In our target feature total number of data is ten and number of possible value is three that is bus, train, car. We will have to calculate entropy for each possible value.

Note: Gender column possible value is two that is male, female. Car column possible value is three that is 0, 1, 2. Travel cost column possible value is three that is cheap, standard, expensive. Income column possible value is three that is low, medium, high.

We will have to calculate each possible for each features based on the target column possible value.

Decision Tree

Iteration-1

1	Gender	Car	Travel cost	Income	Transport
2	male	0	cheap	low	bus
3	male	1	cheap	medium	bus
4	female	1	cheap	medium	train
5	female	0	cheap	low	bus
6	male	1	cheap	medium	bus
7	male	0	standard	medium	train
8	female	1	standard	medium	train
9	female	1	expensive	high	car
10	male	2	expensive	medium	car
11	female	2	expensive	high	car
12					

Entropy before partition: Transport

$$E(s) = -(\frac{4}{10}log_2\frac{4}{10} + \frac{3}{10}log_2\frac{3}{10} + \frac{3}{10}log_2\frac{3}{10})$$

= -(- 0.528 – 0.521 – 0.521)

= 1.57

Total data = 10
For 10 data possible
Bus=4, train=3, car=3

Now calculate the entropy for each features:

$$E(\text{gender}\{male\}) = -(\frac{3}{5}log_2\frac{3}{5} + \frac{1}{5}log_2\frac{1}{5} + \frac{1}{5}log_2\frac{1}{5})$$

= -(- 0.442 – 0.464 – 0.464)

= 1.37

Total male = 5
For 5 male possible
Bus=3, train=1, car=1

$$E(\text{gender}\{female\}) = -(\frac{1}{5}log_2\frac{1}{5} + \frac{2}{5}log_2\frac{2}{5} + \frac{2}{5}log_2\frac{2}{5})$$

= -(- 0.464 – 0.528 – 0.528)

= 1.52

Total female = 5
For 5 female possible
Bus=1, train=2, car=2

Now, Information gain: (Gender)

Total data = 10
For 10 data Male=5, Female=5

$$Gain = 1.57 - ((\frac{5}{10} \times 1.37) + (\frac{5}{10} \times 1.52))$$

= 0.125

Decision Tree

Iteration-1

1	Gender	Car	Travel cost	Income	Transport
2	male	0	cheap	low	bus
3	male	1	cheap	medium	bus
4	female	1	cheap	medium	train
5	female	0	cheap	low	bus
6	male	1	cheap	medium	bus
7	male	0	standard	medium	train
8	female	1	standard	medium	train
9	female	1	expensive	high	car
10	male	2	expensive	medium	car
11	female	2	expensive	high	car
12					

Now calculate the entropy for each features:

$$\mathrm{E}(\mathrm{car}\{0\}) = -(\tfrac{2}{3}\, log_2 \tfrac{2}{3} + \tfrac{1}{3}\, log_2 \tfrac{1}{3})$$

= -(- 0.387 – 0.528)

= .915

$$\mathrm{E}(\mathrm{car}\{1\}) = -(\tfrac{2}{5}\, log_2 \tfrac{2}{5} + \tfrac{2}{5}\, log_2 \tfrac{2}{5} + \tfrac{1}{5}\, log_2 \tfrac{1}{5})$$

= -(- 0.528 – 0.528 – 0.464)

= 1.52

$$\mathrm{E}(\mathrm{car}\{2\}) = -(\tfrac{2}{2}\, log_2 \tfrac{2}{2})$$

= 0

Now, Information gain: (Car)

$$Gain = 1.57 - ((\tfrac{3}{10} \times 0.915) + (\tfrac{5}{10} \times 1.52) + (\tfrac{2}{10} \times 0))$$

= 1.57 – (0.274 + 0.76 + 0)

= .537

Decision Tree

Iteration-1

1	Gender	Car	Travel cost	Income	Transport
2	male	0	cheap	low	bus
3	male	1	cheap	medium	bus
4	female	1	cheap	medium	train
5	female	0	cheap	low	bus
6	male	1	cheap	medium	bus
7	male	0	standard	medium	train
8	female	1	standard	medium	train
9	female	1	expensive	high	car
10	male	2	expensive	medium	car
11	female	2	expensive	high	car
12					

Now calculate the entropy for each features:

$$\mathrm{E}(\mathrm{cost}\{cheap\}) = -(\frac{4}{5} log_2 \frac{4}{5} + \frac{1}{5} log_2 \frac{1}{5})$$

= -(- 0.257 - 0.464)

= .721

$$\mathrm{E}(\mathrm{cost}\{standard\}) = -(\frac{2}{2} log_2 \frac{2}{2})$$

= 0

$$\mathrm{E}(\mathrm{cost}\{expensive\}) = -(\frac{3}{3} log_2 \frac{3}{3})$$

= 0

Now, Information gain: (Travel cost)

$$Gain = 1.57 - ((\frac{5}{10} \times 0.721) + 0 + 0)$$

= 1.57 – 0.35

= 1.21

Decision Tree

Iteration-1

1	Gender	Car	Travel cost	Income	Transport
2	male	0	cheap	low	bus
3	male	1	cheap	medium	bus
4	female	1	cheap	medium	train
5	female	0	cheap	low	bus
6	male	1	cheap	medium	bus
7	male	0	standard	medium	train
8	female	1	standard	medium	train
9	female	1	expensive	high	car
10	male	2	expensive	medium	car
11	female	2	expensive	high	car
12					

Now calculate the entropy for each features:

$$\mathrm{E}(\text{income}\{low\}) = -(\frac{2}{2} log_2 \frac{2}{2})$$

= 0

$$\mathrm{E}(\text{income}\{medium\}) = -(\frac{2}{6} log_2 \frac{2}{6} + \frac{3}{6} log_2 \frac{3}{6} + \frac{1}{6} log_2 \frac{1}{6})$$

= -(- 0.528 – 0.5 – 0.430)

= 1.459

$$\mathrm{E}(\text{income}\{high\}) = -(\frac{2}{2} log_2 \frac{2}{2})$$

= 0

Now, Information gain: (Income)

$$Gain = 1.57 - ((\frac{2}{10} \times 0) + (\frac{6}{10} \times 1.459) + (\frac{2}{10} \times 0))$$

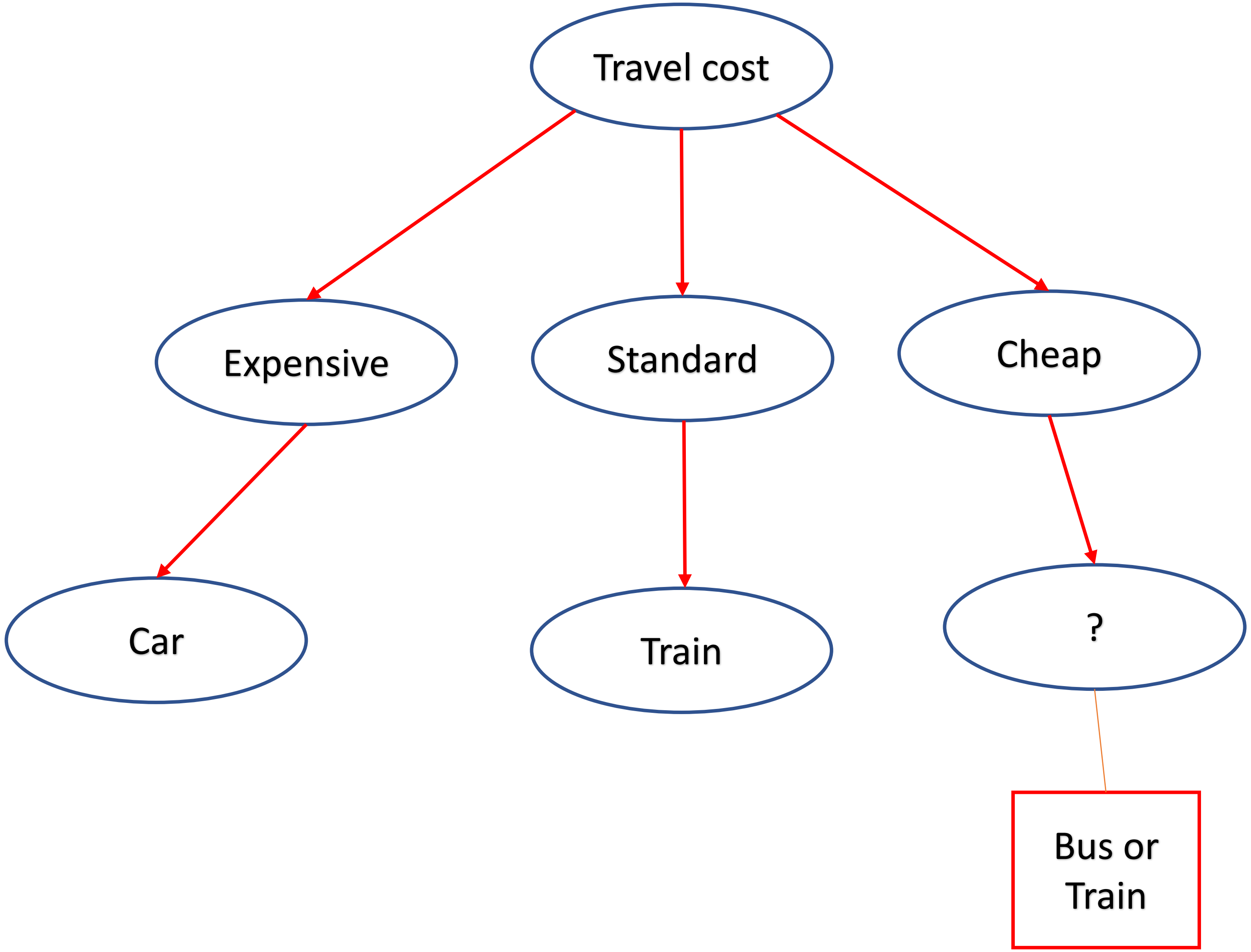
= 1.57 – 0.875

= .695

Decision Tree

Attributes	**Gain**
Gender	0.125
Car	0.537
Travel cost	**1.21**
Income	0.695

	Gender	Car	Travel cost	Income	Transport
1	Gender	Car	Travel cost	Income	Transport
2	male	0	cheap	low	bus
3	male	1	cheap	medium	bus
4	female	1	cheap	medium	train
5	female	0	cheap	low	bus
6	male	1	cheap	medium	bus
7	male	0	standard	medium	train
8	female	1	standard	medium	train
9	female	1	expensive	high	car
10	male	2	expensive	medium	car
11	female	2	expensive	high	car
12					



Decision Tree

Iteration-2

In second iteration we will work for only cheap so we can omit the Travel cost column.

Gender	Car	Income	Transport
male	0	low	bus
male	1	medium	bus
female	1	medium	train
female	0	low	bus
male	1	medium	bus

Entropy before partition: Transport

$$\mathrm{E(s)} = -\left(\frac{4}{5} log_2 \frac{4}{5} + \frac{1}{5} log_2 \frac{1}{5}\right)$$

= -(- 0.257 – 0.464)

= 0.721

Now calculate the entropy for each features:

$$\mathrm{E(gender}\{male\}) = -\left(\frac{3}{3} log_2 \frac{3}{3}\right)$$

= 0

$$\mathrm{E(gender}\{female\}) = -\left(\frac{1}{2} log_2 \frac{1}{2} + \frac{1}{2} log_2 \frac{1}{2}\right)$$

= 1

Now, Information gain: (Gender)

$$Gain = 0.721 - \left(\left(\frac{2}{5} \times 1\right) + 0\right)$$

= 0.321

Again same procedure:

Decision Tree

Iteration-2

In second iteration we will work for only cheap so we can omit the Travel cost column.

Gender	Car	Income	Transport
male	0	low	bus
male	1	medium	bus
female	1	medium	train
female	0	low	bus
male	1	medium	bus

Again same procedure:

Now calculate the entropy for each features:

$$E(\text{car}\{0\}) = -(\tfrac{2}{2}\, log_2 \tfrac{2}{2})$$

$= 0$

$$E(\text{car}\{1\}) = -(\tfrac{2}{3}\, log_2 \tfrac{2}{3} + \tfrac{1}{3}\, log_2 \tfrac{1}{3})$$

$= -(-0.389 - 0.528)$

$= 0.917$

Now, Information gain: (Car)

$$Gain = 0.721 - ((0) + (\tfrac{3}{5} \times 0.917))$$

$= 0.721 - 0.550$

$= 0.170$

Decision Tree

Iteration-2

In second iteration we will work for only cheap so we can omit the Travel cost column.

Gender	Car	Income	Transport
male	0	low	bus
male	1	medium	bus
female	1	medium	train
female	0	low	bus
male	1	medium	bus

Again same procedure:

Now calculate the entropy for each features:

$$E(\text{income}\{low\}) = -(\frac{2}{2} log_2 \frac{2}{2})$$

= 0

$$E(\text{income}\{medium\}) = -(\frac{2}{3} log_2 \frac{2}{3} + \frac{1}{3} log_2 \frac{1}{3})$$

= - (- 0.389 – 0.528)

= 0.917

Now, Information gain: (Income)

$$Gain = 0.721 - ((0) + (\frac{3}{5} \times 0.917))$$

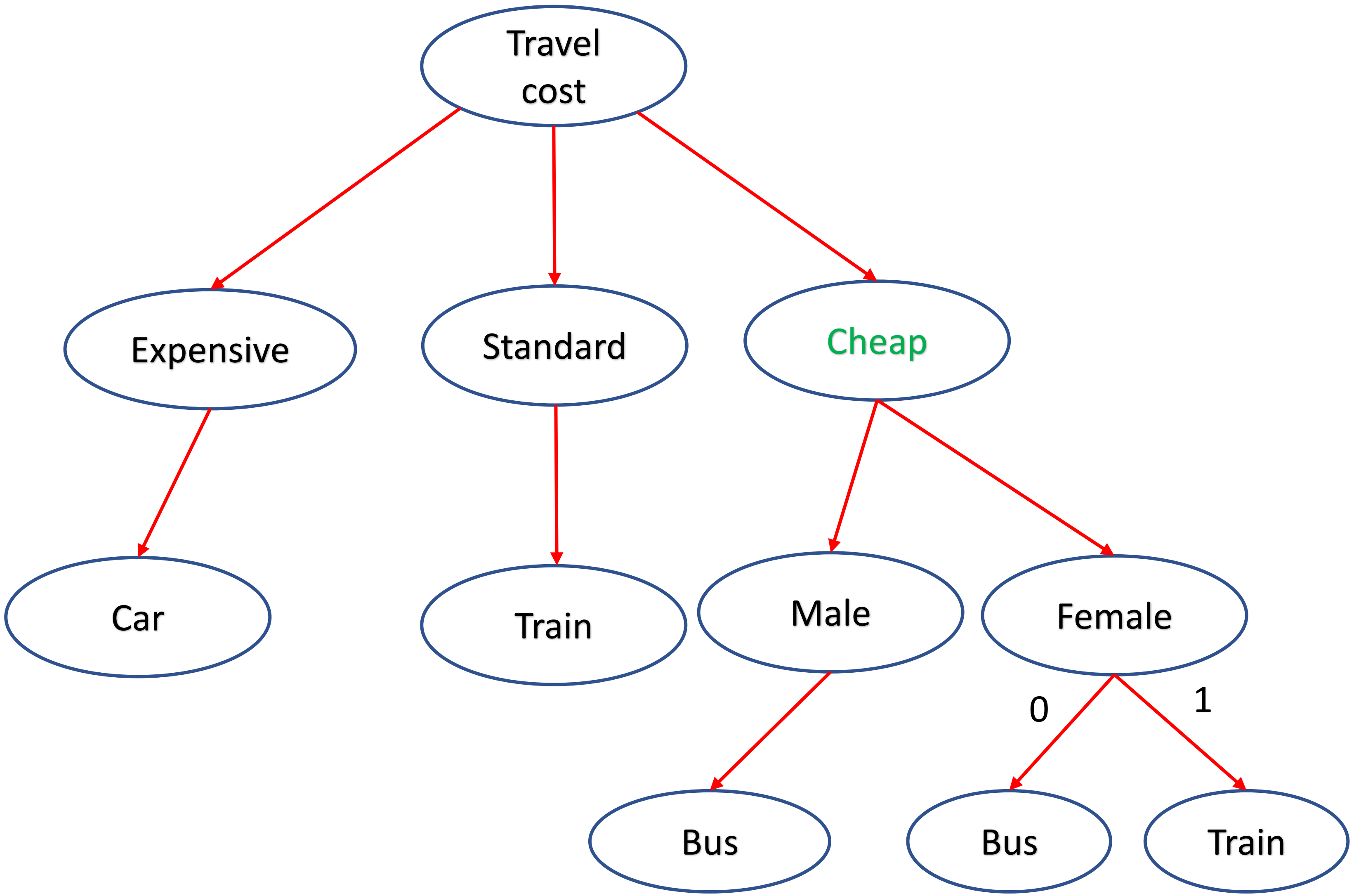
= 0.721 – 0.550

= 0.170

Decision Tree

Attributes	**Gain**
Gender	**0.322**
Car	0.170
Income	0.170

Gender	Car	Income	Transport
male	0	low	bus
male	1	medium	bus
female	1	medium	train
female	0	low	bus
male	1	medium	bus



Full Calculation for one(before I calculate shortcut as you can seen):

$$E(s) = -\left(\frac{4}{10} log_2 \frac{4}{10}\right)$$

$$= -\left(\frac{4}{10} \times \frac{log_2 \frac{4}{10}}{log_2}\right)$$

$$= -\left(\frac{4}{10} \times \frac{\log(4)-\log(10)}{\log(2)}\right)$$

$$= -(0.4 \times -1.322)$$

$$= -(-0.528)$$

$$= 1.57$$

Decision Tree for Regression

For decision tree regression the same formula use but here is slight different. In decision regressor threshold value consider and compare with threshold and also calculate the entropy and gain then make the decision tree. The decision criteria is different for classification and regression trees. **Decision trees regression normally use mean squared error (MSE) to decide to split a node in two or more sub-nodes.**

Decision Tree for Regression

Suppose we are doing a binary tree the algorithm first will pick a value, and split the data into two subset. For each subset, it will **calculate the MSE separately**. The tree chooses the value with results in smallest MSE value.

Let's examine how is Splitting Decided for Decision Trees Regressor in more details. The first step to create a tree is to create the first binary decision. How are you going to do it?

- We need to pick a variable and the value to split on such that the two groups are as different from each other as possible.
- For each variable, for each possible value of the possible value of that variable see whether it is better.
- How to determine if it is better? Take weighted average of two new nodes (**mse*num_samples**)

To sum up, we now have:

- A single number that represents how good a split is which is the weighted average of the mean squared errors of the two groups that create.
- A way to find the best split which is to **try every variable and to try every possible value of that variable** and see which variable and which value gives us a split with the best score.

This is the entirety of creating a decision tree regressor and will stop when some stopping condition (defined by hyperparameters) is met:

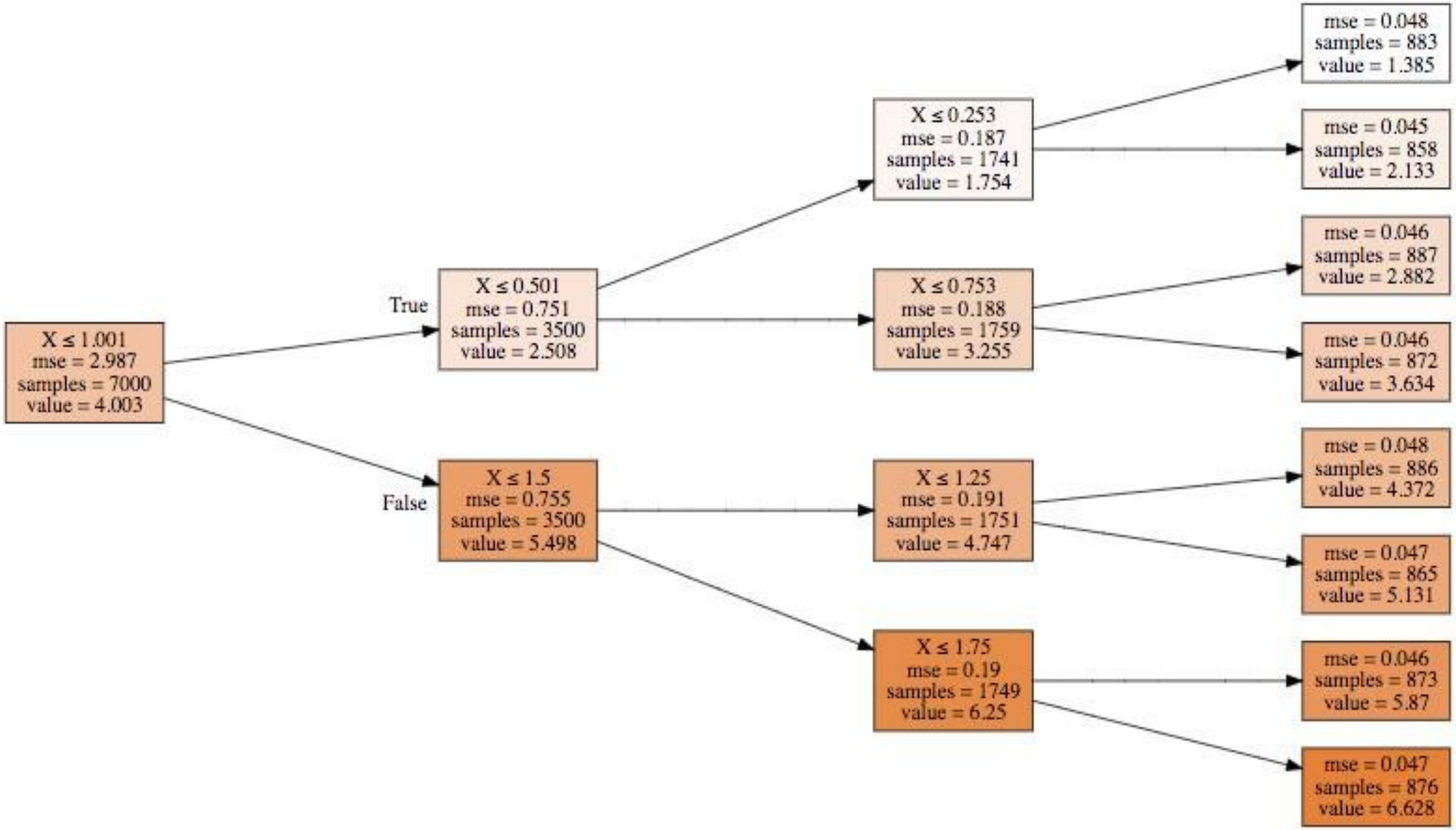
- When you hit a limit that was requested (for example: max_depth)
- When your leaf nodes only have one thing in them (no further split is possible, MSE for the train will be zero but will overfit for any other set -not a useful model)

Decision Tree for Regression

How it makes predictions?

- Given a data point you run it through the entirely tree asking True/False questions up until it reaches a leaf node. The final prediction is the average of the value of the dependent variable in that leaf node.

Decision Tree for Regression



Decision Tree for Regression

As you can see we're taking a subset of the data, and deciding the best manner to split the subset further. Our initial subset was the entire data set, and we split it according to the rule X<=1.001. Then, for each subset, we performed additional splitting until we were able to correctly predict the target variable while respecting the constraint of max_depth=3.

Scikit-learn use CART algorithm to create decision tree. In CART algorithm only produce binary tree means non-leaf nodes always two children.

ID3 algorithm also use for making decision tree. In id3 algorithm can create multiple child node or leaf node.