ML Labs

Delivering Enterprise Machine Leaning , AI Services

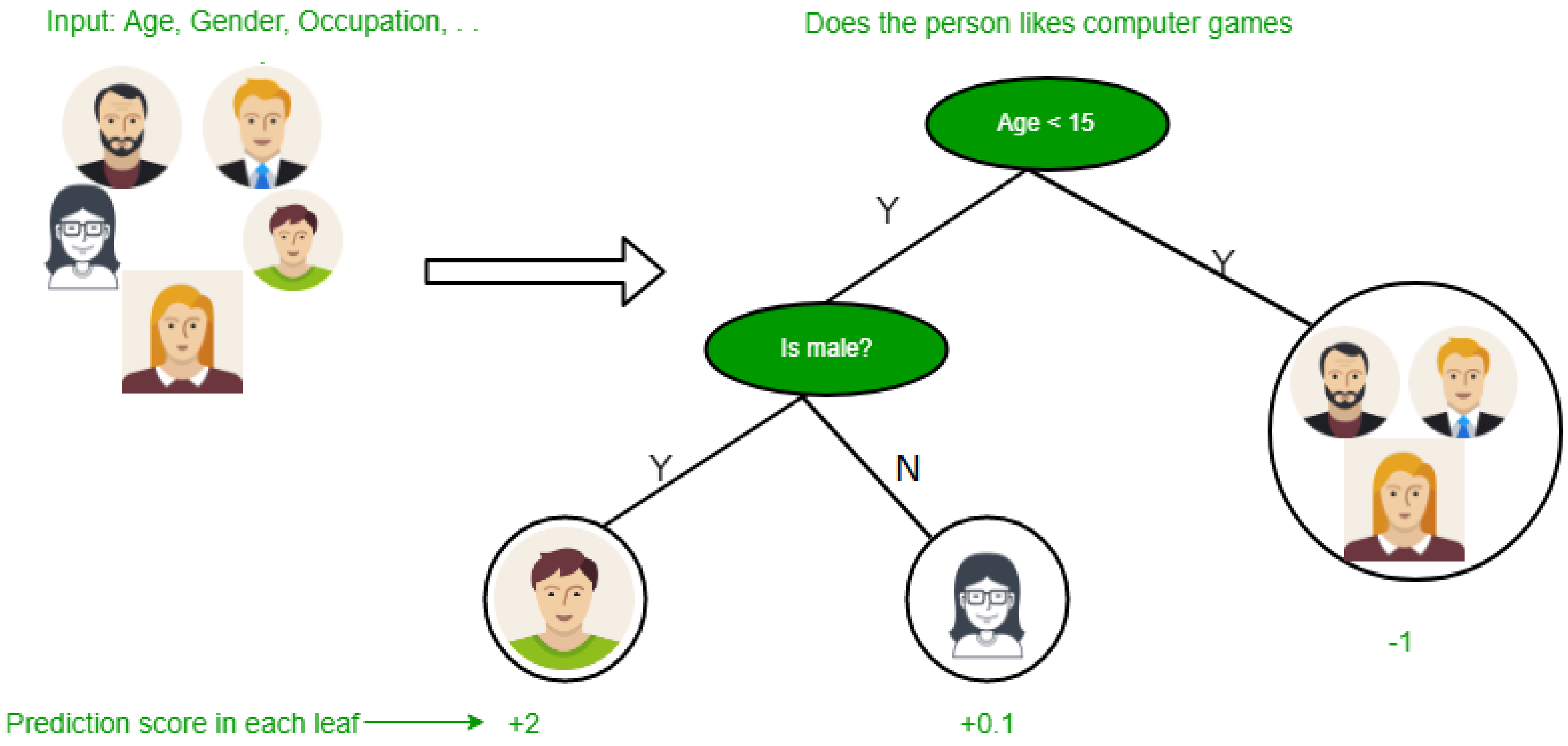
Decision Trees

Introduction :

Decision Trees are one of the supervised machine learning algorithms that are representation of Human Decision Making. Here we can deal with Categorical data unlike other ML Algorithms. There is no linear relationship between independent and target variable, so that they can be used to model high nonlinear data. They are very popular because the outputs can be easily understood by business people.

They form tree like models to make predictions, like how decisions are made in real world with series of questions. A decision tree splits the data into multiple splits and the further split is done by raising a question and the question is raised based on which attributes are the most important predictor and based on that the split occurs. They are easily **interpretable**, because we always can identify various factors to split the data .If a data is split into two are more partitions this is called **Multiway** decision tree.

Lets see an example how an Decision Tree look like:

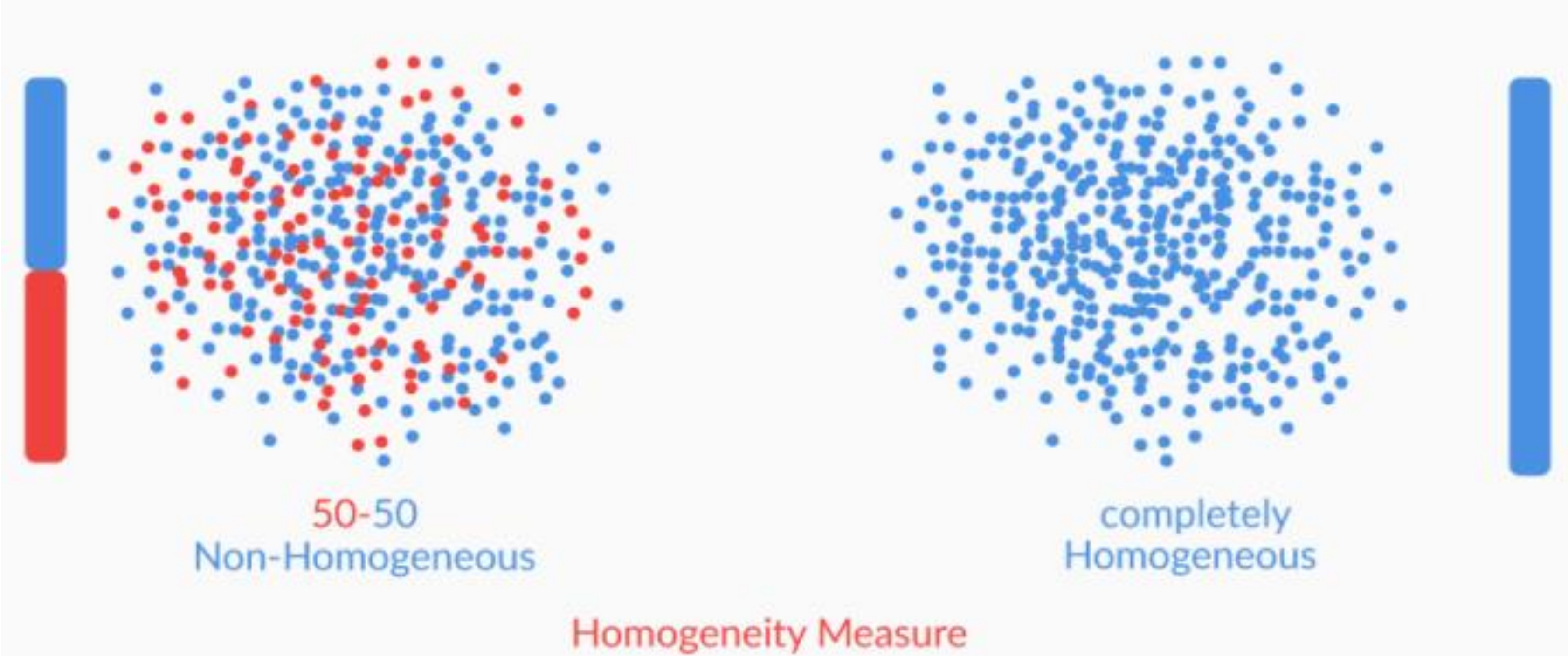


Here the first split he based on Age, we will discuss further how the split is made on Age

Algorithms of Decision Tree

Homogeneity Measures :

Suppose we have a dataset of 8 attributes, we can't randomly select a attribute to split the data. There is a selection criterion to choose the attributes, which is call Homogeneity Measure. We are going to pick the attribute which as more Homogeneity Measure. Lets see the fig below to understand better about Homogeneity .



Gini Index

The Homogeneity Measure can be calculate using the Gini index, if the Gini index of an attribute is equal to 1 then the Homogeneity of that attribute is high.

Formula for Gini Index : $Gini = \sum_{i=1}^{k} Pi^2$ **Key point: Gini = 1 Homogeneity is high if Gini = 0 less Homogeneity**

Lets see the example how to calculate Gini Index:

Lets first split the data set on Gender:

Gini Index(gender) = (fraction of total observations in male)*Gini index of male + (fraction of total observations in female)*Gini index of female.
=1/2((1/50)*2+ (49/50)*2) +1/2((3/5)*2+ (2/5)*2) = 0.74

Split on Age:
= 0.7((26/70)*2+ (44/70)*2) + 0.3 ((1/6)*2+ (5/6)*2) = 0.59

So the Gini Index for Gender is close to 1 so that we can split on Gender.

GENDER \ AGE	<50	>50
F	P - 10 N - 390	P - 0 N - 100
M	P - 250 N - 50	P - 50 N - 150

P = playing, N = Not playing

Entropy and Information Gained

Entropy quantifies the degree of disorder in the given data. Entropy ranges from 0 to 1 if Entropy =0 that means data has high Homogeneity

Formula for entropy : $\varepsilon[D] = -\sum_{i=1}^{k} p_i log_2 p_i,$

Where pi = Probability of finding label i
k = Number of different labels
ε[D] = Entropy of Dataset D

Information Gained Measures how much as the entropy decreases after splitting the data

Formula for Information Gained $Gain(D, A) = \varepsilon[D] - \varepsilon[D_A] = \varepsilon[D] - \sum_{i=1}^{k}((\frac{D_{A=i}}{D}) * \varepsilon[D_{A=i}]),$

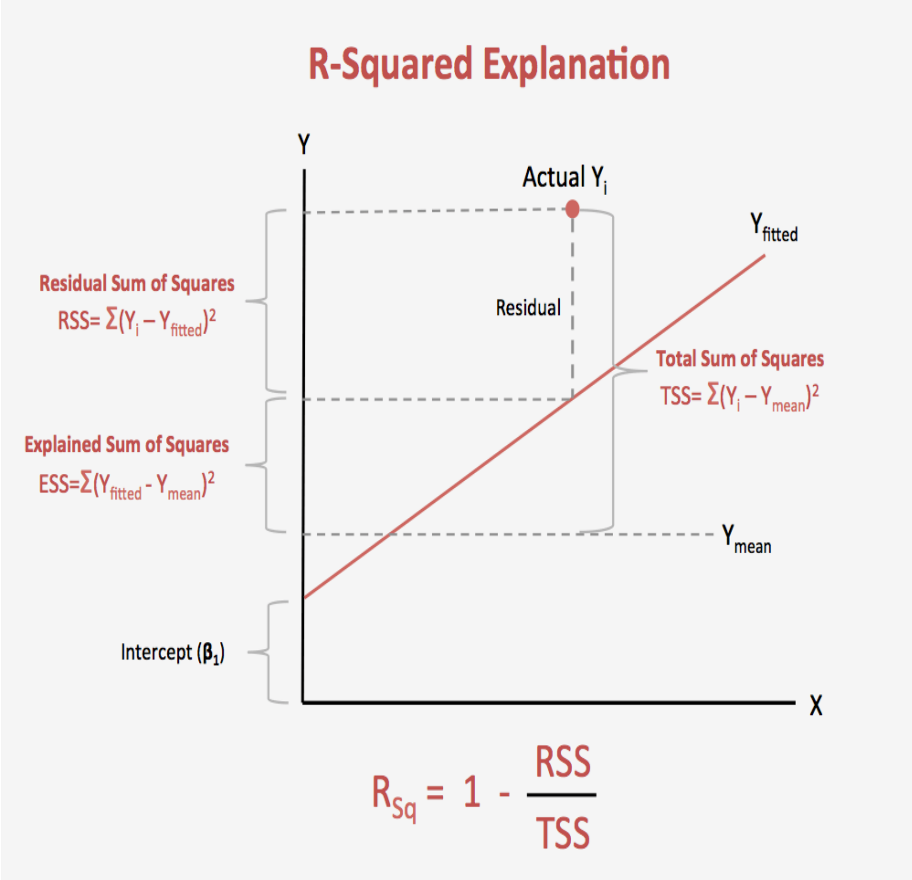
Where ε[D] = Entropy of parent set
ε[DA] = Entropy of Partitions Obtained after splitting
ε[DA=i] = Entropy of Partitions where value of attribute A for the data points is i
DA=i = Number of points where attribute A is i
K = Number of different labels

Splitting by R-squared

R-squared is calculated when there are continuous variables in the data. It is calculated in similar way how we work on a Regression model. To Split the date R-squared value should be greater than the original data

Formula for R-squared: 1- (RSS/TSS)

Where RSS = Residual Standard Error
TSS = Total sum of squares



Over fitting Control Techniques

Truncation :

This Technique will stop the tree while growing , So that it may not get over fit and not end up having leaves with few data points.

Pruning:

Let the Tree grow till the end, after that cut the branches of tree from deep. This is the most commonly used to avoid over fitting of data.

Advantages and Disadvantages of Decision Trees

Advantages

- Prediction made by Decision Trees are easily interpretable.
- Does not require normalisation, because it only compare values within data.
- They can seamlessly handle all kinds of data.

Disadvantages:

- Decision Tress tends to over fit the data, if it was allowed to grow.
- Decision Trees tend to very unstable.

ML Labs Pvt Ltd
Marathahalli ,3rd Floor Above Khazana Jewellery Bangalore 560066
91-7338339898
91-7829396922
Connect : Bharath@pylabs.com