file

August 22, 2024

Decision Tree

A decision tree is a flowchart-like tree structure where an internal node represents a feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome.

The time complexity of decision trees is a function of the number of records and attributes in the given data. The decision tree is a distribution-free or non-parametric method which does not depend upon probability distribution assumptions. Decision trees can handle high-dimensional data with good accuracy.

How Does the Decision Tree Algorithm Work?

The basic idea behind any decision tree algorithm is as follows:

Select the best attribute using Attribute Selection Measures (ASM) to split the records.

Make that attribute a decision node and breaks the dataset into smaller subsets.

Start tree building by repeating this process recursively for each child until one of the conditions will match:

All the tuples belong to the same attribute value.

There are no more remaining attributes.

There are no more instances.

[ ]:

```python
import math
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import tensorflow as ts, keras
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LinearRegression, LogisticRegression
from sklearn.preprocessing import OneHotEncoder,OrdinalEncoder, StandardScaler
```

```
2024-08-22 12:23:15.199750: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:485] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
```

```
already been registered
2024-08-22 12:23:15.464054: E
external/local_xla/xla/stream_executor/cuda/cuda_dnn.cc:8454] Unable to register
cuDNN factory: Attempting to register factory for plugin cuDNN when one has
already been registered
2024-08-22 12:23:15.538784: E
external/local_xla/xla/stream_executor/cuda/cuda_blas.cc:1452] Unable to
register cuBLAS factory: Attempting to register factory for plugin cuBLAS when
one has already been registered
2024-08-22 12:23:19.131417: W
tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not
find TensorRT
```

```
[ ]: df = pd.read_csv('salaries.csv')
```

```
[ ]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16 entries, 0 to 15
Data columns (total 4 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   company             16 non-null     object
 1   job                 16 non-null     object
 2   degree              16 non-null     object
 3   salary_more_then_100k  16 non-null     int64
dtypes: int64(1), object(3)
memory usage: 640.0+ bytes
```

```
[ ]: df.head(n=10)
```

```
[ ]:      company                  job     degree  salary_more_then_100k
     0     google      sales executive  bachelors                      0
     1     google      sales executive    masters                      0
     2     google     business manager  bachelors                      1
     3     google     business manager    masters                      1
     4     google  computer programmer  bachelors                      0
     5     google  computer programmer    masters                      1
     6  abc pharma      sales executive    masters                      0
     7  abc pharma  computer programmer  bachelors                      0
     8  abc pharma     business manager  bachelors                      0
     9  abc pharma     business manager    masters                      1
```

What is Gini Impurity?

Gini Impurity is a measurement used to build Decision Trees to determine how the features of a dataset should split nodes to form the tree. More precisely, the Gini Impurity of a dataset is a number between 0-0.5, which indicates the likelihood of new, random data being misclassified if it were given a random class label according to the class distribution in the dataset.

What is Entropy?

Entropy is an information theory metric that measures the impurity or uncertainty in a group of observations. It determines how a decision tree chooses to split data. The image below gives a better description of the purity of a set.

What is Information Gain?

Claude Shannon invented the concept of entropy, which measures the impurity of the input set. In physics and mathematics, entropy is referred to as the randomness or the impurity in a system. In information theory, it refers to the impurity in a group of examples. Information gain is the decrease in entropy. Information gain computes the difference between entropy before the split and average entropy after the split of the dataset based on given attribute values. ID3 (Iterative Dichotomiser) decision tree algorithm uses information gain.