Decision trees

October 2022

Decision tree

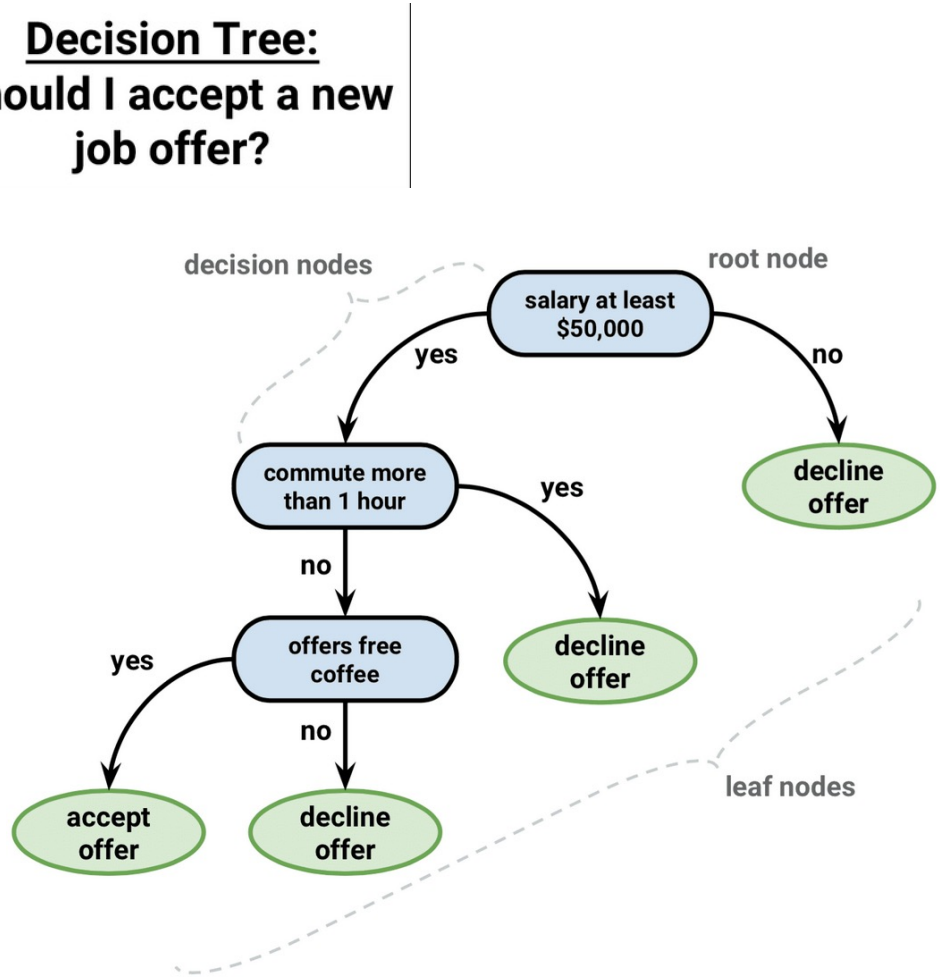
Decision Tree is a non-parametric supervised learning method used for classification and regression.

- supervised Machine Learning method
- approximate a sine curve with a set of if-then-else decision rules
- deeper the tree, the more complex the decision rules

Decision tree

Classification or regression model in the form of a tree structure.

Idea: break down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.

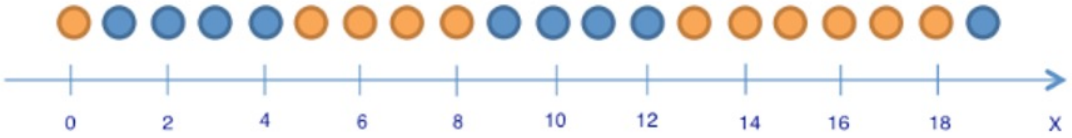


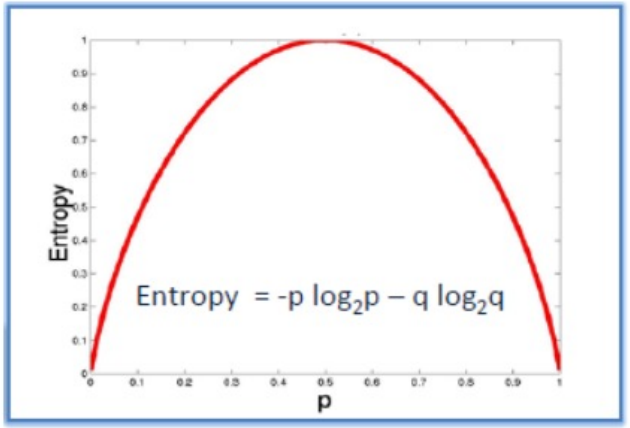
Entropy

Entropy is a measure of chaos. The more homogeneous the set is, the lower its combinatorial entropy, and Vice versa.

Formula: $$S = -\sum p_i \cdot \log_2 p_i$$

$$S_o = -\left(\frac{9}{20}\right)*\mathrm{Ln}\left(\frac{9}{20}\right) - \left(\frac{11}{20}\right)*\mathrm{Ln}\left(\frac{11}{20}\right) \approx 0{,}69$$



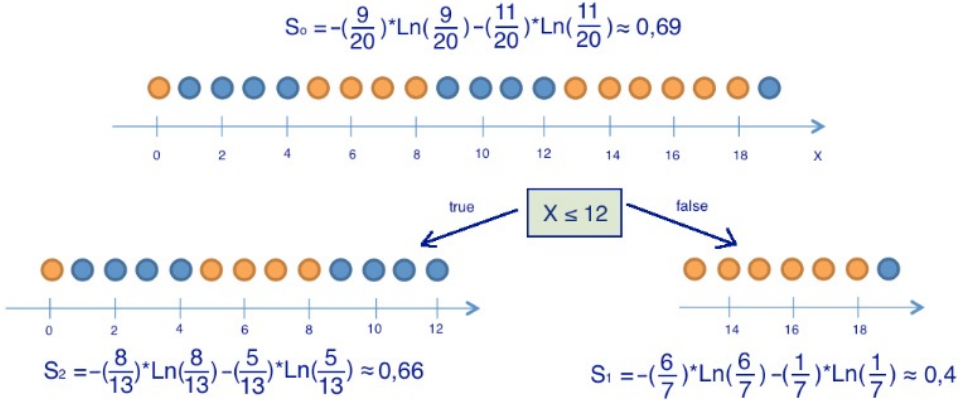


Entropy = $-0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$

Information gain

The information gain is based on the decrease in entropy after a dataset is split on an attribute.

Idea: choose the attributes and find the predicates in such a way that after splitting the entropy decreases.



Algorithm

If the stopping criteria is not satisfied

Step 1 Calculate entropy of the target.

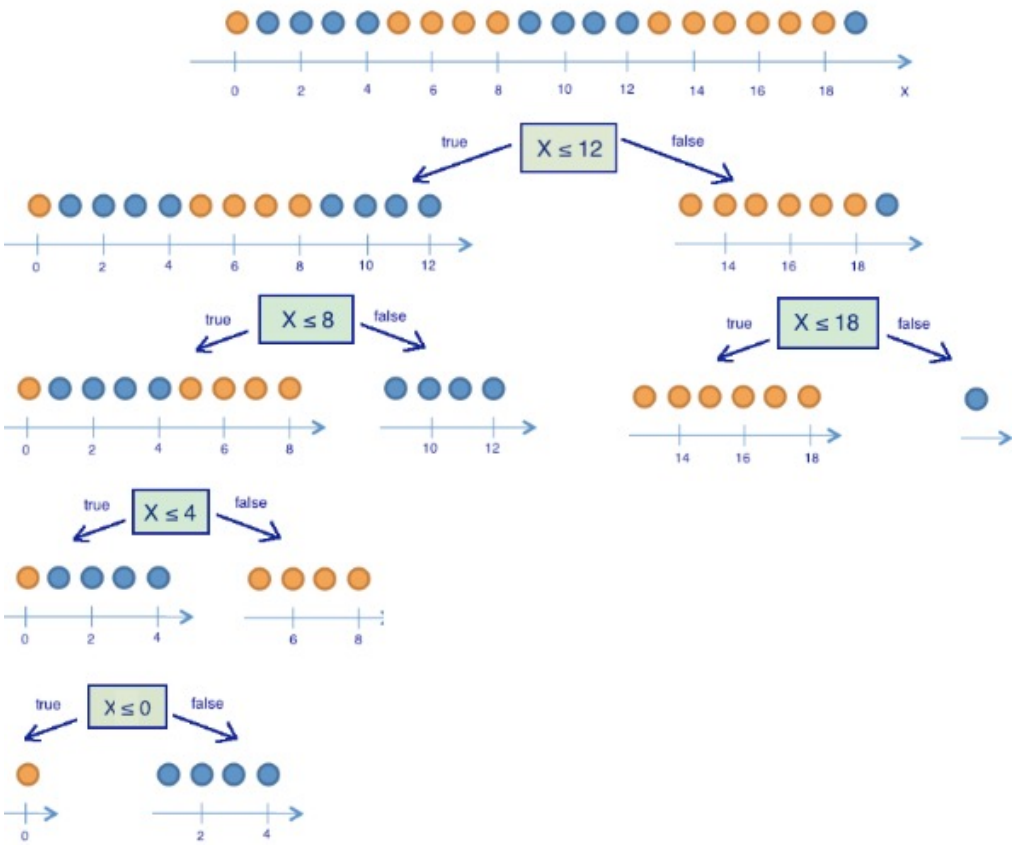
Step 2 For every predicate and for every possible split:

- split the data and calculate entropy for each branch
- add proportionality to get the entropy of each split
- subtract the resulting entropy from the entropy before the split

(information gain)

Step 3 Choose the predicate with the largest information gain as the decision node and divide the dataset

Step 4 Repeat the process on every branch



Advantages and disadvantages:

Advantages:

- simple to understand and interpret
- help determine worst, best and expected values for different scenarios.
- a white box model
- can be combined with other decision techniques

Disadvantages:

- unstable (small change can lead to a large change in the structure)
- prone to overfitting
- relatively inaccurate
- calculations can get very complex
- Data driven approach!

Decision tree VS Linear regression

- Decision trees supports non linearity, where linear regression supports only linear solutions.
- When there are large number of features with less datasets (with low noise), linear regressions may outperform Decision trees/random forests. In general cases, Decision trees will be having better average accuracy.
- For categorical independent variables, decision trees are better than linear regression.
- Decision trees handles colinearity better than Linear regression

Conclusion

- A supervised learning method that can be used for solving regression and classification problems.
- Builds non-linear models.
- Idea of learning decision rules inferred from prior data (training data).
- Decision tress often mimic the human level thinking so it’s simple to understand the data and make some good interpretations.