Decision Tree Classifier (DTC)

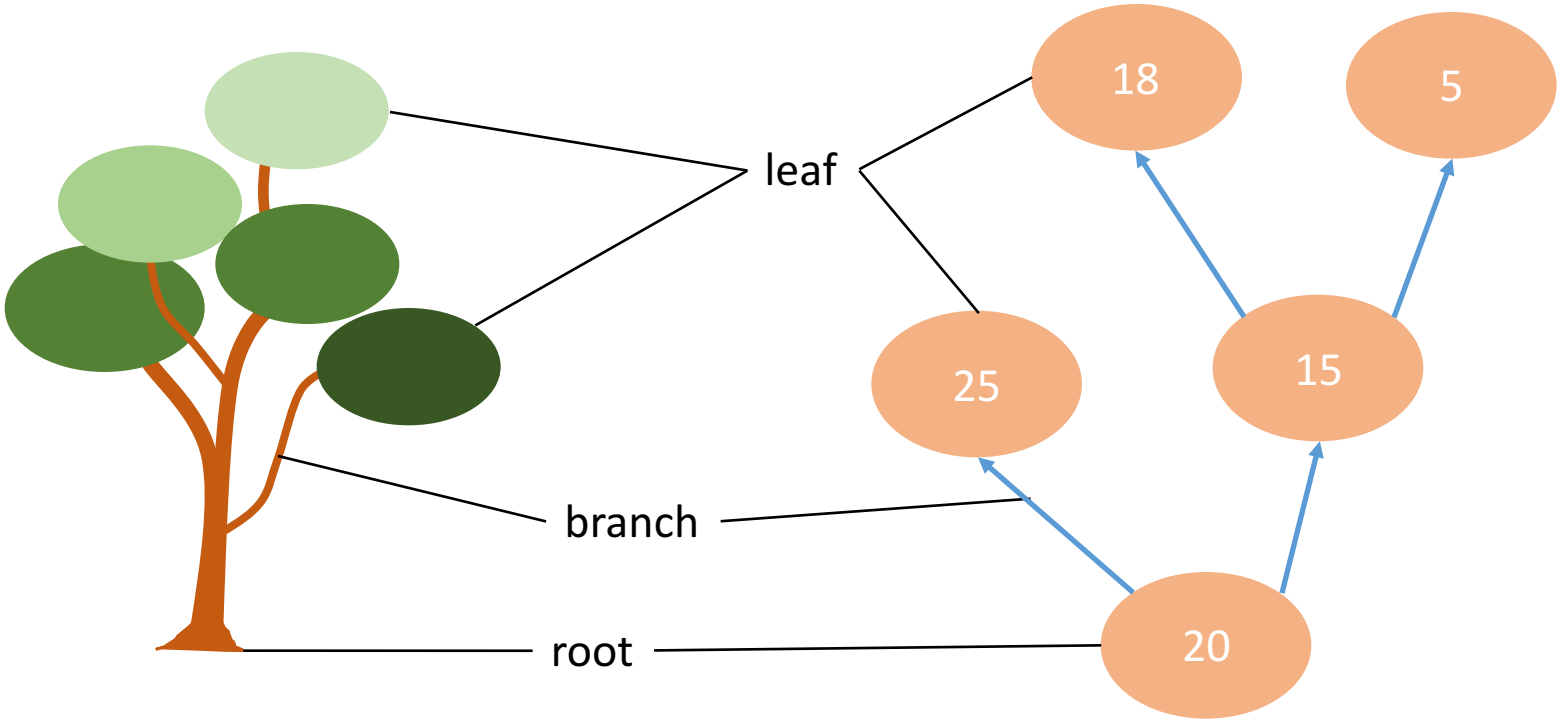
Khundrakpam Veeshel Singh

NIELIT Imphal

Outlines

- What is tree?
- Decision Tree Classifier
- Algorithm
- Measure of selection the best split
- Types
- Pros and Cons

What is Tree?



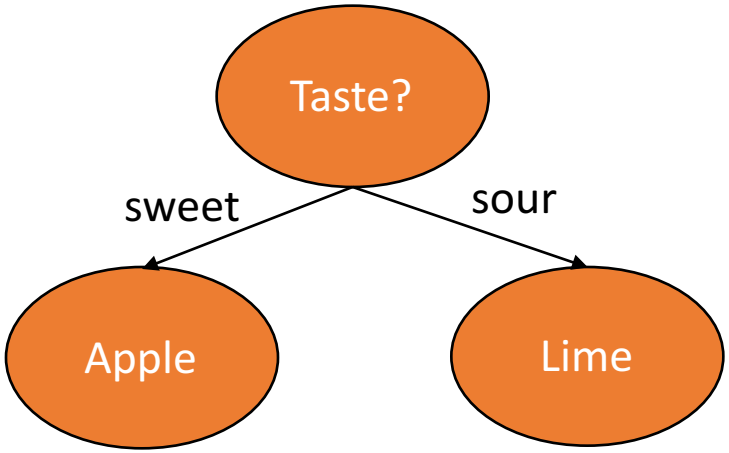
Decision Tree Classifier

- Decision Tree is a Supervised Machine Learning Algorithm that uses a set of rules to make decisions, similarly to how humans make decisions.
- The intuition behind Decision Trees is that you use the dataset features to create yes/no questions and continually split the dataset until you isolate all data points belonging to each class.

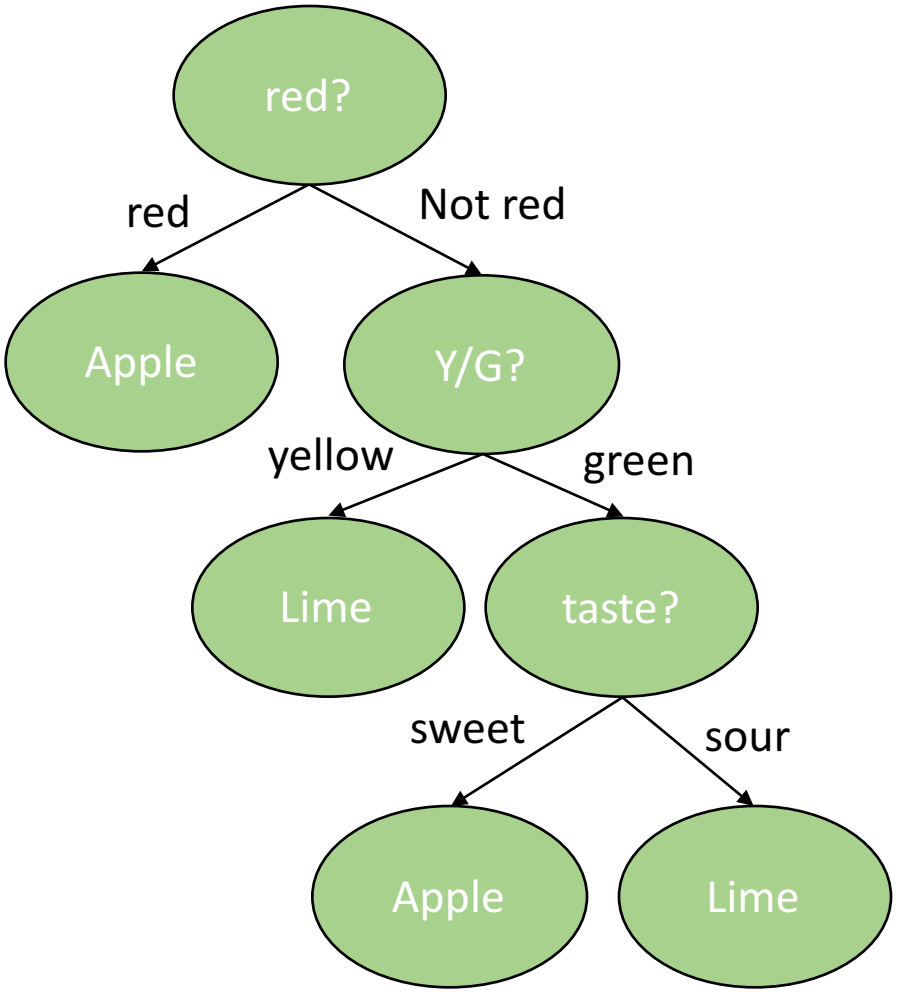
Algorithm

- ALL decision tree are based on Hunt's Algorithm
- Steps
  - If all the data points in the node belong to same class c then it is leaf node labeled as c
  - If the data points belong to more then one class then an attribute test condition is selected to partition the data points into smaller nodes A child node is created for each outcome of the test condition and the data points are distributed to the children based on the outcome. The algorithm is then applied recursively to each child node

Sl No	Taste	Colour	Class
1	Sweet	Green	Apple
2	Sour	Yellow	Lime
3	Sour	Green	Lime
4	Sweet	Red	Apple



DTC 1

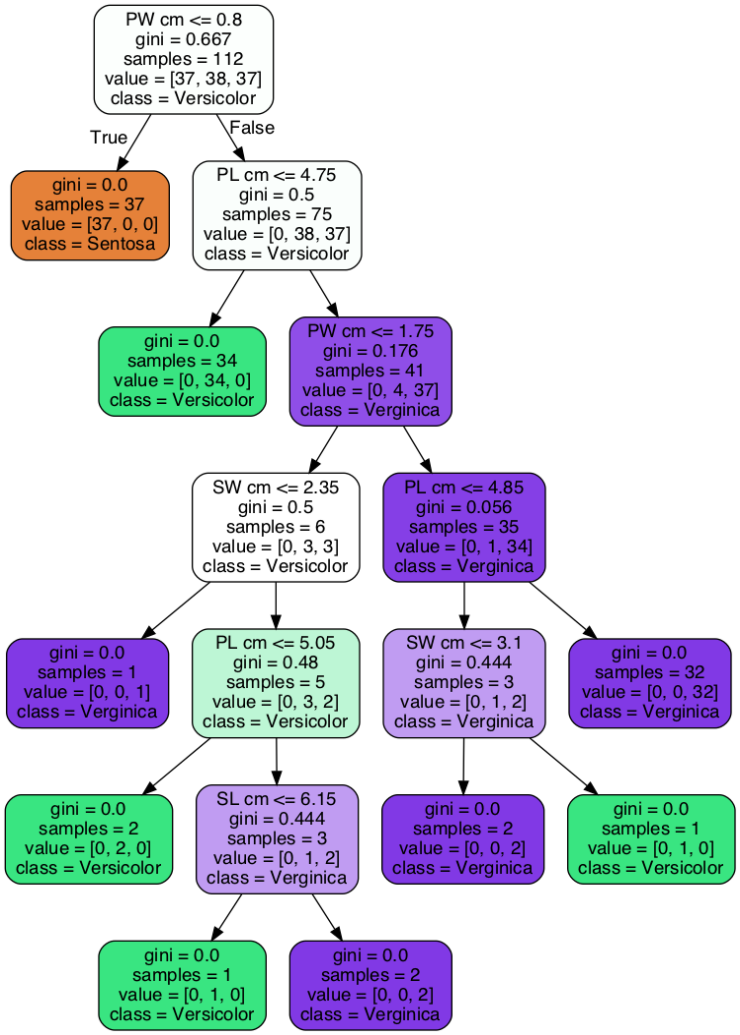


DTC 2

Measures for Selecting the Best Split

- Entropy(t): $-\sum_{i=0}^{c-1} p(i|t)\log_2 p(i|t)$
- Gini(t): $1-\sum_{i=0}^{c-1}[p(i|t)]^2$
- Classification error(t):$1-\max_i[p(i|t)]$
- Gain Ratio: $\frac{\Delta_{info}}{Split\ Info}$ or $\frac{I(parent)-\sum_{j=1}^{k}\frac{N(v_j)}{N}Iv_j}{-\sum_{i=1}^{S}P(v_i)\lim_2 P(v_i)}$

Example from the Iris flower classification



Types

- ID3
- C4.5
- CART

Pros & Cons

- Pros
  - Inexpensive to construct.
  - Extremely fast at classifying unknown records.
  - Easy to interpret for small-sized trees
  - Accuracy comparable to other classification techniques for many simple data sets.
  - Excludes unimportant features.
- Cons
  - Easy to overfit.
  - Decision Boundary restricted to being parallel to attribute axes.
  - Decision tree models are often biased toward splits on features having a large number of levels.
  - Small changes in the training data can result in large changes to decision logic.
  - Large trees can be difficult to interpret and the decisions they make may seem counter intuitive.