CSE 4553

Machine Learning

Lecture 3: Decision Tree

Winter 2024

Dr. Hasan Mahmud | hasan@iut-dhaka.edu

Contents

- Introduction
- What is Decision tree?
- Machine Learning problem types
- Machine Learning applications

Introduction

- Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks.

- The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

- It is one of the most widely used and practical methods for supervised learning.

- The decision rules are generally in form of if-then-else statements. The deeper the tree, the more complex the rules and fitter the model.

- Decision trees classify the examples by sorting them down the tree from the root to some leaf node, with the leaf node providing the classification to the example.

Decision tree example

- Each node in the tree acts as a test case for some attribute, and each edge descending from that node corresponds to one of the possible answers to the test case. Leaf nodes are the decisions or class labels.

- This process is recursive in nature and is repeated for every subtree rooted at the new nodes.

Day	Weather	Temperature	Humidity	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	Weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	Rainy	Mild	High	Strong	No
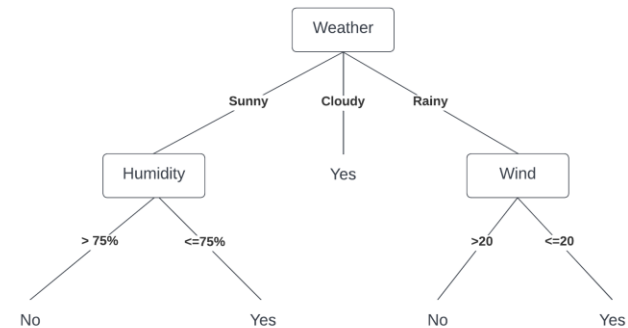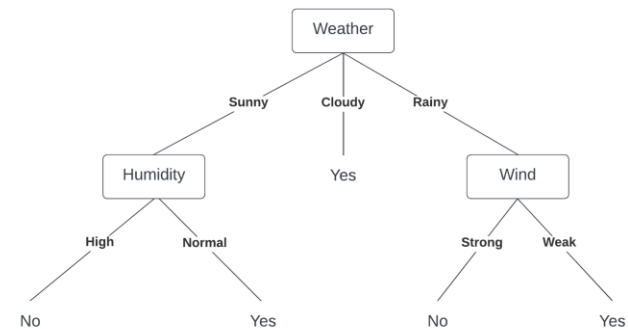
Decision tree example

Height (cm)	Weight (kg)	Gender
175	85	Male
170	70	Female
185	83	Male
160	55	Female
165	60	Female

$H > 180$
Yes — *Male*
No — $W > 80$
Yes — *Male*
No — *Female*

$H > 180$
Yes — *Male*

185	83	Male

No — $W > 80$

175	85	Male
170	70	Female
160	55	Female
165	60	Female

Yes — *Male*

175	85	Male

No — *Female*

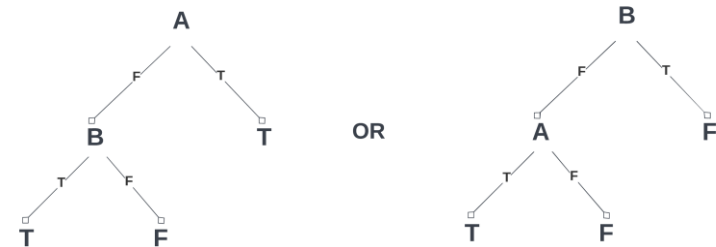| 170	70	Female
160	55	Female
165	60	Female
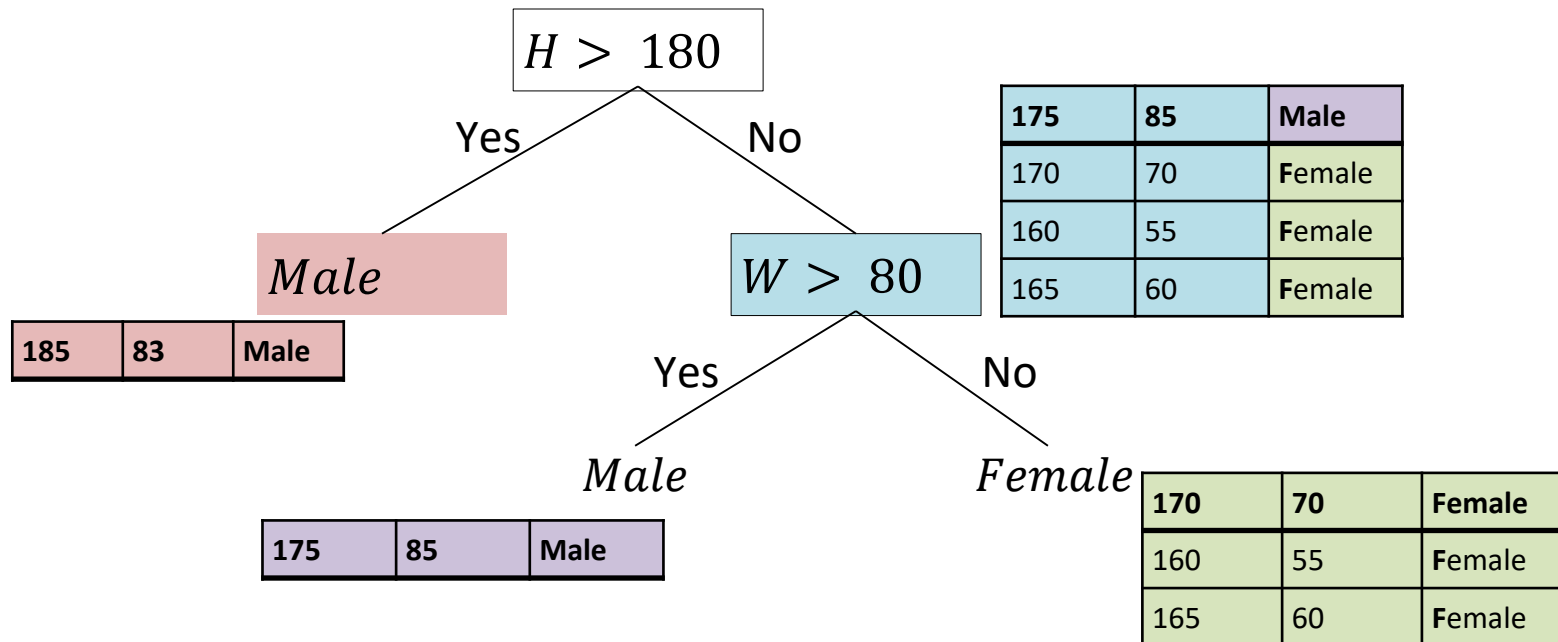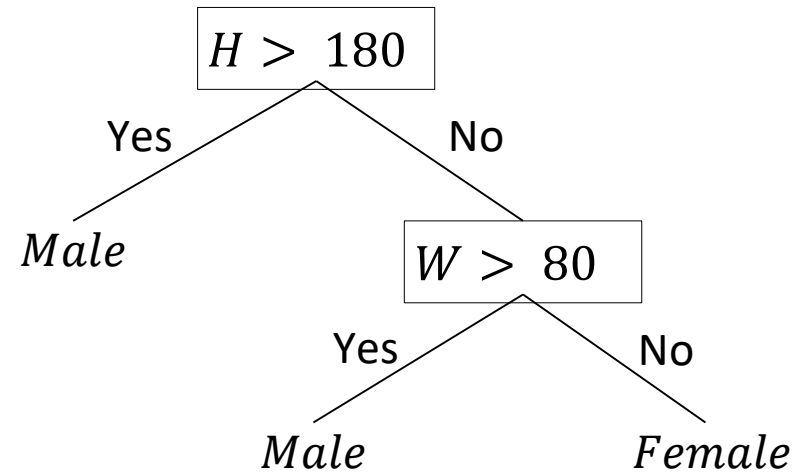
Expressiveness of decision tree

- The space of decision tree (i.e. the hypothesis space) is very much expressive.

- Same function may represent different decision tree because of attribute selection

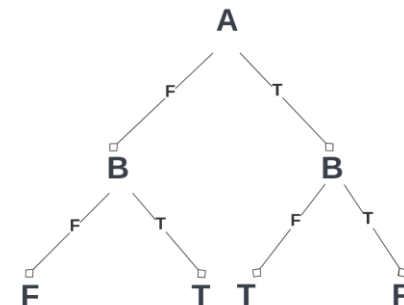- So, a clever way is needed to search the best tree among them.

A	B	A AND B
F	F	F
F	T	F
T	F	F
T	T	T
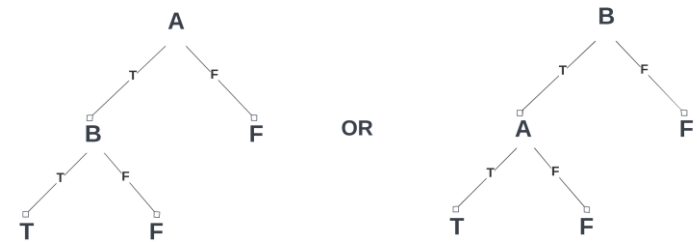
A	B	A OR B
F	F	F
F	T	T
T	F	T
T	T	T

| A	B	A XOR B
F	F	F
F	T	T
T	F	T
T	T	F

Measures of decision tree construction

- Information theory
- Information gain
- Entropy
- Equations

Information Theory: Entropy, Information Gain

- Entropy:
  - Defined as the randomness or measuring the disorder of the information being processed in Machine Learning.
  - Entropy is the machine learning metric that measures the unpredictability or impurity in the system.
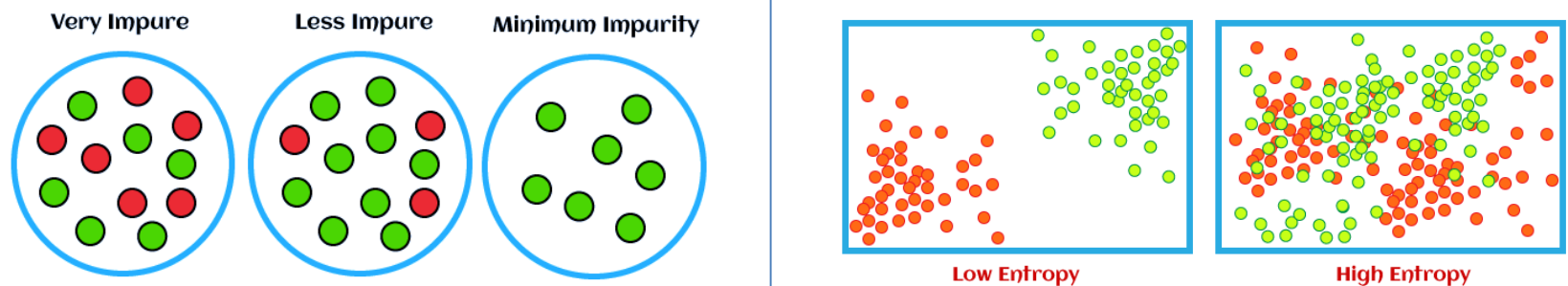
  - In information theory, every piece of information has a specific value associated with it which is used to draw a conclusion.
  - Easy to draw a conclusion from a piece of information means less entropy. Entropy higher means, it is difficult to draw a conclusion from the information

Equation of Entropy

- Entropy can defined as the expected value of the information, i.e.

$$Entropy, E = -\sum_{i=1}^{n} p_i \log_2 p_i$$

p_i = probability of randomly picking an element of class i (i.e. the proportion of the dataset made up of class i), and n = Total number of classes

- Suppose we have a dataset having three colors of fruits as red, green, and blue.
- Suppose we have 2 red, 2 green, and 4 blue observations throughout the dataset.

- $E = -\left(\frac{2}{8}\log_2\frac{2}{8} + \frac{2}{8}\log_2\frac{2}{8} + \frac{4}{8}\log_2\frac{4}{8}\right)$ 🔴🔴 🟢🟢 🔵🔵🔵🔵

$$= -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{2}\log_2\frac{1}{2}\right)$$

$$= -(-1.5) = 1.5$$

- What is the entropy of a dataset containing only one color?
- $E = -1\,log_2 1 = 0$ 🔵🔵🔵🔵

Equation example

- If all the samples in a set belongs to same class then the Entropy is 0 and node consisting that subset is called a pure node.

- If the samples are equally distributed then the Entropy is 1

- When entropy becomes 0, then the dataset has no impurity. Datasets with 0 impurities are not useful for learning. Further, if the entropy is 1, then this kind of dataset is good for learning.
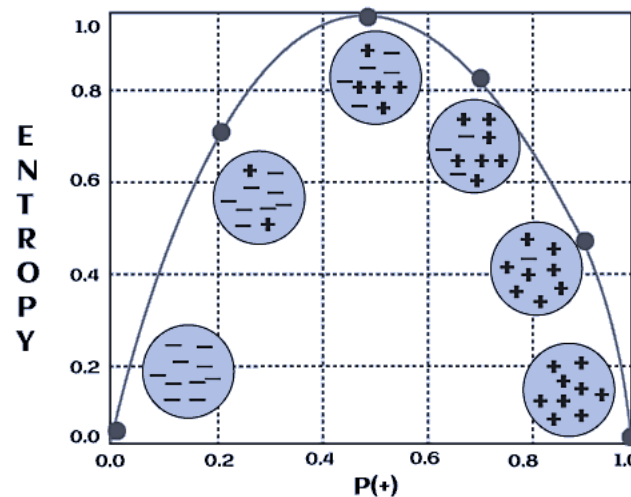
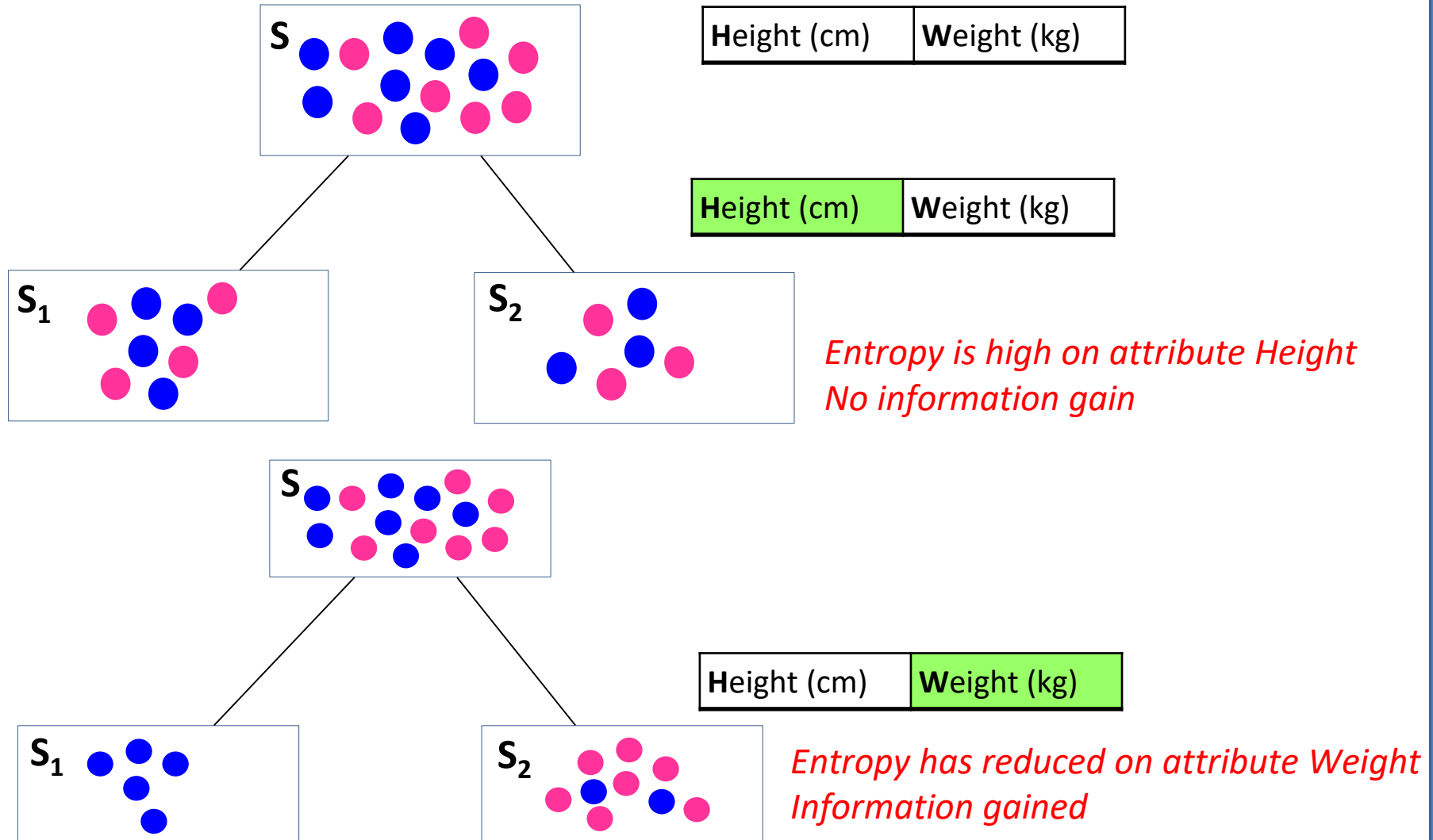Two-class problem
– Class 1
+ Class 2

Equation of Information Gain

- It quantifies the quality of a split. It measures the effectiveness of an attribute in choosing training data.

- Calculated for a split by subtracting the weighted entropies of each branch from the original entropy. It measures the reduction in Entropy.

- Information gain is used to select the best splitting attribute while constructing a decision tree.

- Information Gain (S, A) of an attribute/feature A, related to a set of training examples, S, is defined as:

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

$i.e.\ Information\ Gain = Entropy\ (ParentNode) - Avg.Entropy\ (ChildNode)$

Equation example

S

Height (cm)	**W**eight (kg)

S₁

S₂

Height (cm)	**W**eight (kg)

Entropy is high on attribute Height
No information gain

S

S₁

S₂

Height (cm)	**W**eight (kg)

Entropy has reduced on attribute Weight
Information gained

Equation example

S

S₁

S₂

Height (cm)	**W**eight (kg)

Entropy has reduced on attribute weight
Information gained
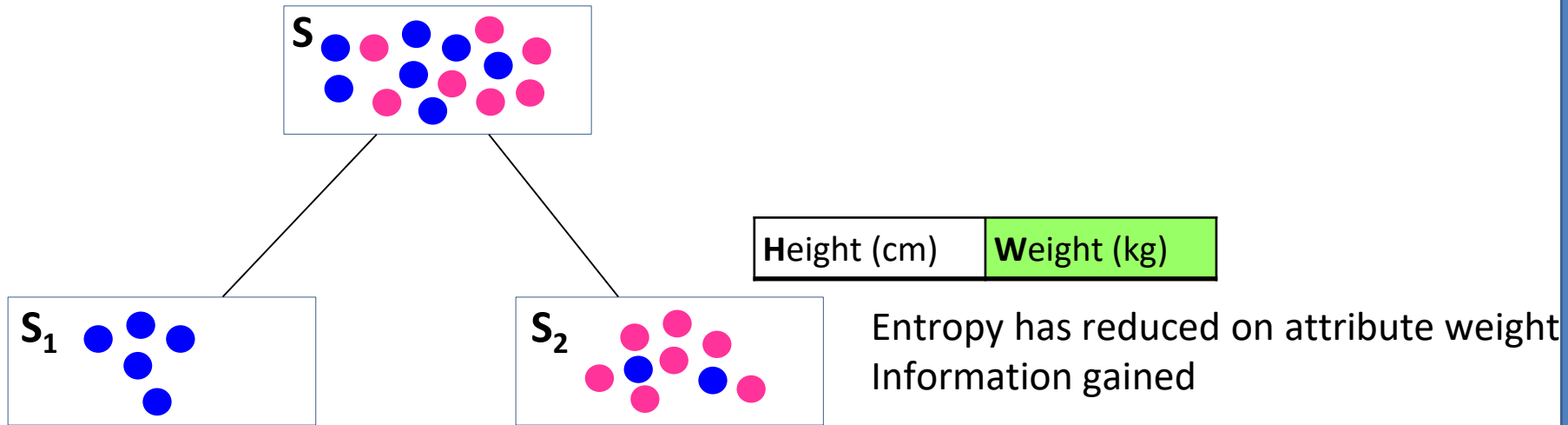
$$Gain\ (S, A) = Entropy\ (S) - \sum_{v \in A} \frac{|S_v|}{|s|} \times Entropy\ (S_v)$$

$$Gain\ (S, Weight) = Entropy\ (S) - \frac{5}{14} Entropy\ (S_1) - \frac{7}{14} Entropy\ (S_1)$$

The feature that gives maximum information gain will be selected as the best splitting feature

Construction of Decision Tree using Entropy and Information Gain

SN	Color	Size	Shape	Edible
1	Yellow	Small	Round	No (–)
2	Yellow	Large	Irregular	No (–)
3	Green	Large	Round	No (–)
4	Yellow	Large	Round	No (–)
5	Green	Small	Round	Yes (+)
6	Green	Small	Irregular	Yes (+)
7	Yellow	Small	Irregular	Yes (+)

Information Gain Calculation

- Formula for Information Gain,

SN	Color	Size	Shape	Edible
1	Yellow	Small	Round	No (−)
2	Yellow	Large	Irregular	No (−)
3	Green	Large	Round	No (−)
4	Yellow	Large	Round	No (−)
5	Green	Small	Round	Yes (+)
6	Green	Small	Irregular	Yes (+)
7	Yellow	Small	Irregular	Yes (+)

Entropy before splitting,

$$E(S) = -\frac{3}{7}\log(\frac{3}{7}) - \frac{4}{7}\log(\frac{4}{7})$$
$$= 0.985$$

COLOR

Color
Yellow   Green
1+, 3-   2+, 1-
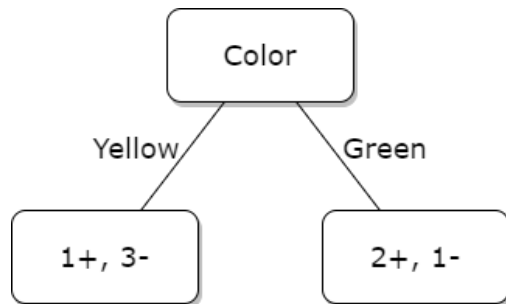
$$E(Color\_yellow) = -\frac{3}{4}\log(\frac{3}{4}) - \frac{1}{4}\log(\frac{1}{4})$$
$$= 0.811$$

$$E(Color\_green) = -\frac{1}{3}\log(\frac{1}{3}) - \frac{2}{3}\log(\frac{2}{3})$$
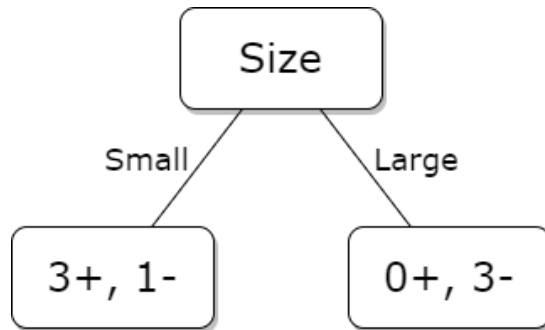$$= 0.918$$

Information gained, IG(S,Color)
$$= 0.985 - \frac{4}{7}\times 0.811 - \frac{3}{7}\times 0.918$$
$$= 0.128$$

Information Gain Calculation

SIZE

SN	Color	Size	Shape	Edible
1	Yellow	Small	Round	No (−)
2	Yellow	Large	Irregular	No (−)
3	Green	Large	Round	No (−)
4	Yellow	Large	Round	No (−)
5	Green	Small	Round	Yes (+)
6	Green	Small	Irregular	Yes (+)
7	Yellow	Small	Irregular	Yes (+)

$E(Size\_small)$

$= -\frac{3}{4}\log(\frac{3}{4}) - \frac{1}{4}\log(\frac{1}{4})$

$= 0.811$
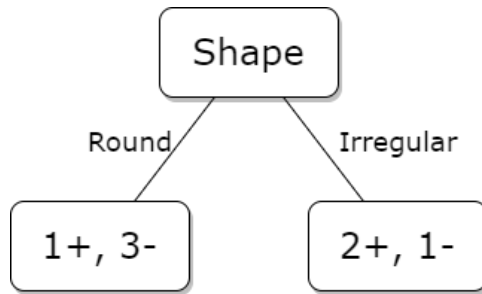
$E(Size\_large)$

$= -\frac{3}{3}\log(\frac{3}{3})$

$= 0$

Information gained, $IG(S,Size)$

$= 0.985 - \frac{4}{7} \times 0.811 - \frac{3}{7} \times 0$

$= 0.521$

Information Gain Calculation

SHAPE

SN	Color	Size	Shape	Edible
1	Yellow	Small	Round	No (–)
2	Yellow	Large	Irregular	No (–)
3	Green	Large	Round	No (–)
4	Yellow	Large	Round	No (–)
5	Green	Small	Round	Yes (+)
6	Green	Small	Irregular	Yes (+)
7	Yellow	Small	Irregular	Yes (+)

E(Shape_round)

$$= -\frac{3}{4}\log(\frac{3}{4}) - \frac{1}{4}\log(\frac{1}{4})$$

$$= 0.811$$

E(Shape_irregular)

$$= -\frac{1}{3}\log(\frac{1}{3}) - \frac{2}{3}\log(\frac{2}{3})$$

$$= 0.918$$

Information gained, IG(S,Shape)

$$= 0.985 - \frac{4}{7} \times 0.811 - \frac{3}{7} \times 0.918$$

$$= 0.128$$

Information Gain Calculation

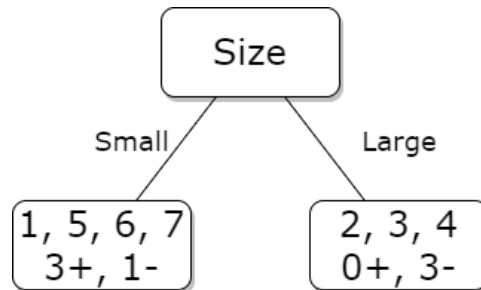| SN	Color	Size	Shape	Edible
1	Yellow	Small	Round	No (−)
2	Yellow	Large	Irregular	No (−)
3	Green	Large	Round	No (−)
4	Yellow	Large	Round	No (−)
5	Green	Small	Round	Yes (+)
6	Green	Small	Irregular	Yes (+)
7	Yellow	Small	Irregular	Yes (+)

- Information gain for Color attribute is: 0.128

- Information gain for Size attribute is: 0.521

- Information gain for Shape attribute is: 1.28

- We can see Information Gain for Size is the highest among all the training attributes. So **Size** should be the training attribute and root.
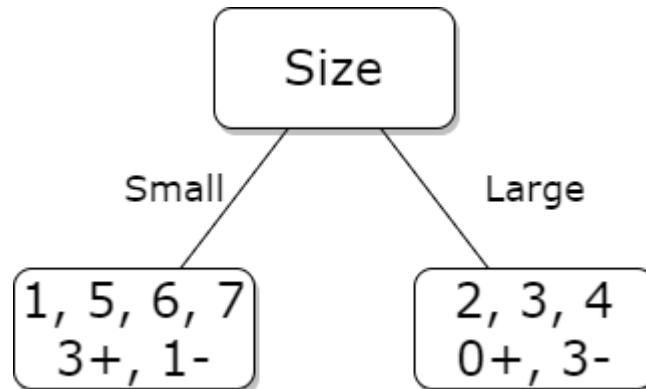
Information GainCalculation

SIZE

SN	Color	Size	Shape	Edible
1	Yellow	Small	Round	No (−)
2	Yellow	Large	Irregular	No (−)
3	Green	Large	Round	No (−)
4	Yellow	Large	Round	No (−)
5	Green	Small	Round	Yes (+)
6	Green	Small	Irregular	Yes (+)
7	Yellow	Small	Irregular	Yes (+)

$$E(S1) = -\frac{3}{4}\log(\frac{3}{7}) - \frac{1}{4}\log(\frac{1}{4})$$
$$= 0.811$$

$$E(Color, Y) = -\frac{1}{2}\log(\frac{1}{2}) - \frac{1}{2}\log(\frac{1}{2}) \qquad E(Color,G) = 0$$
$$= 1$$

$$\text{Information Gain, } IG(S,Color) = 0.811 - \frac{2}{4}\text{x1} - 0$$
$$= 0.311$$

Accordingly,
$$\text{Information Gain, } IG(S,Shape) = 0.311$$

Here, we get same information gain for both Color and shape attributes. So, any one of the nodes can be chosen as decision node

Decision Tree

SN	Color	Size	Shape	Edible
1	Yellow	Small	Round	No (–)
2	Yellow	Large	Irregular	No (–)
3	Green	Large	Round	No (–)
4	Yellow	Large	Round	No (–)
5	Green	Small	Round	Yes (+)
6	Green	Small	Irregular	Yes (+)
7	Yellow	Small	Irregular	Yes (+)

Size

SMALL   LARGE

Color   NO

YELLOW   GREEN

Shape   YES

ROUND   IRREGULAR

NO   YES
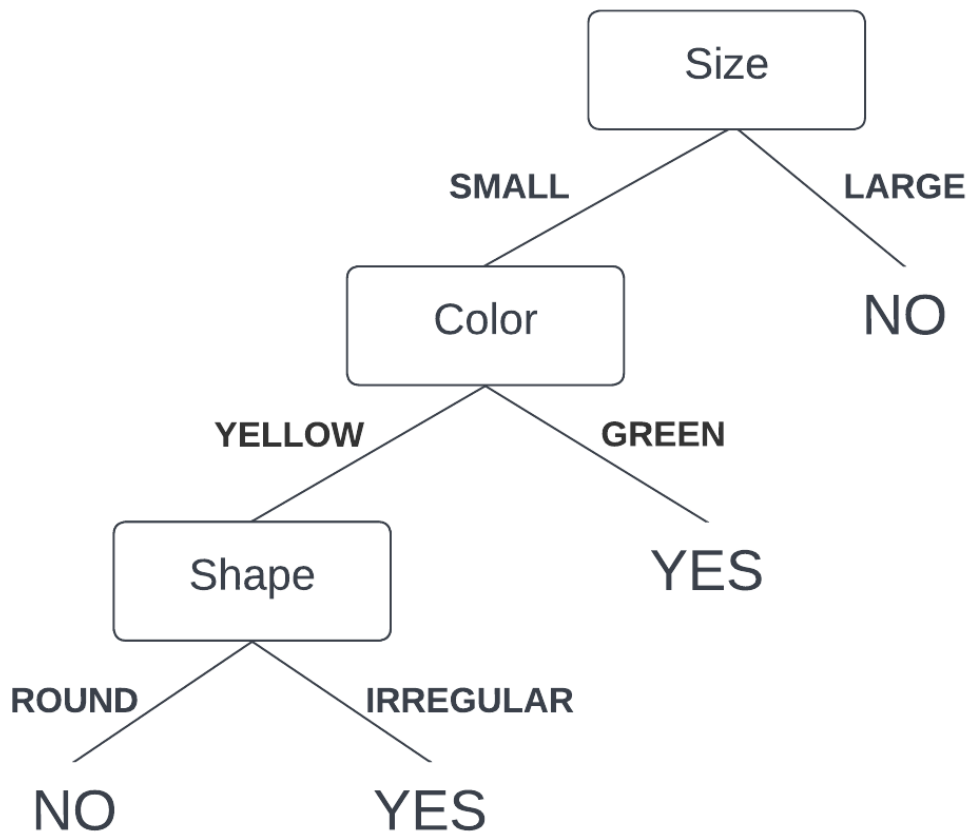
Construction of Decision Tree

- ID3
- CART
- Regression Tree

ID3 - Algorithm

- Calculate the entropy of every attribute using the data set.
- Split the set into subsets using the attribute for which entropy is minimum (or, equivalently, information gain is maximum).
- Make a Decision Tree node containing that attribute.
- Recurse on subsets using remaining attributes.

ID3 - Formulas

Formula for Entropy is:

$$\text{Entropy}(S) = -p(I)\log 2\, p(I)$$

where $p(I)$ is the proportion of S belonging to class I. Σ is over total outcomes. Log2 is log base 2.

Formula for calculating information Gain is,

$$\text{Gain}(S,A) = \text{Entropy}(S) - ((|S_v| / |S|) * \text{Entropy}(S_v))$$

- Application
- Advantages and disadvantages