Decision Trees

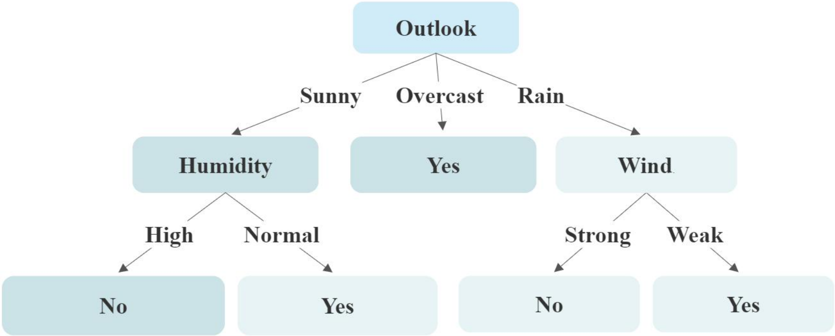
Mengta Chung, PhD

Department of Management Sciences
Tamkang University

Goal for Today

#	Outlook	Temp	Humidity	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rainy	Mild	High	Weak	Yes
5	Rainy	Cool	Normal	Weak	Yes
6	Rainy	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rainy	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rainy	Mild	High	Strong	No

A Decision Tree



Different Algorithms

- A decision tree is a supervised learning algorithm.
- Ross Quinlan invented the Iterative Dichotomizer 3 (ID3) algorithm to generate decision trees in 1986.
- C4.5 and C5.0 are successors of ID3.

Parsimony

- ID3 attempts to create the smallest decision tree possible.
- ID3 considers only attributes never selected before.
- Occam's razor (or law of parsimony) : the simplest explanation is usually the right one.
- If a smaller decision tree classifies observations as good as larger trees, then why use the larger one?

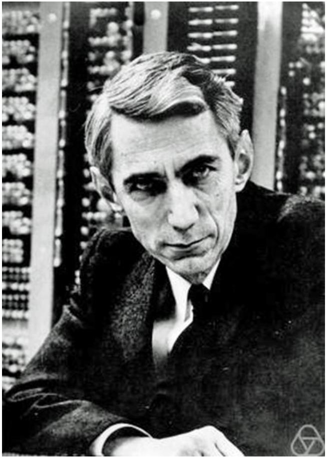
Information Theory

- Information theory was proposed by Claude Shannon in 1949 to find fundamental limits on signal processing and communication operations such as data compression.

  Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. University of Illinois Press.

- A key measure in information theory is information entropy (or Shannon entropy) invented by Shannon.

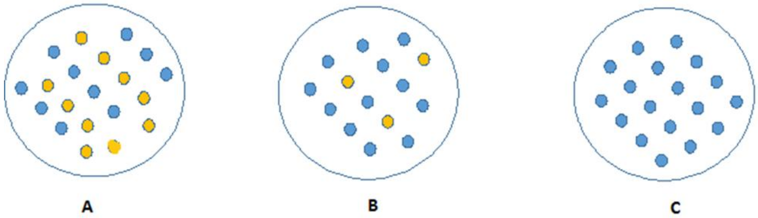
Claude Elwood Shannon (1916 - 2001)



Entropy

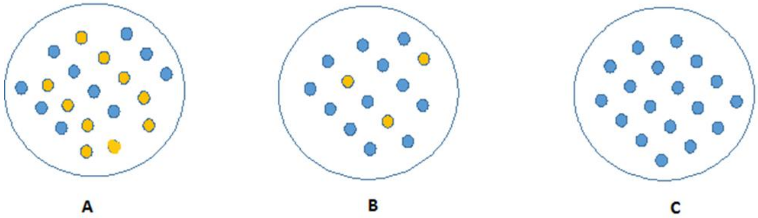
- In thermodynamics, entropy is a measure of the molecular disorder of a system.
- In information theory, entropy measures the impurity in the system.
- Entropy is 0 if all the members in the system belong to the same class (meaning NO impure).

Entropy



- C: all dots are blue (pure)
- B: the majority of dots are blue and 3 other dots are yellow
- A: A half of the dots are blue and the other half of the dots are yellow (most impure)

Entropy



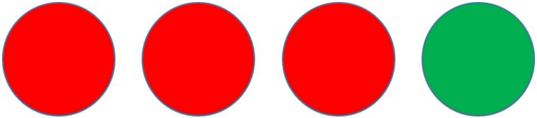
H is used to denote entropy:

- C is a pure node: $H(C) = 0$
- B is less impure: $0 < H(B) < 1$
- A is the most impure: $H(A) = 1$

Probability

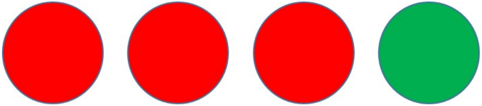
Entropy is used to measure impurity. Probability is used to measure how likely a particular event is to occur. Shannon entropy combines entropy with probability.

Probability



- There are 3 red balls and 1 green ball in the box (box 1).
- Suppose we sample with replacement. What is the probability that we select 3 balls first and then 1 green ball?

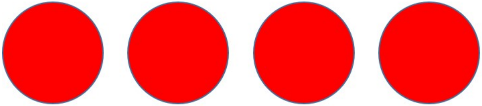
Probability



0.75 * 0.75 * 0.75 * 0.25 = 0.105

- If we sample with replacement, the probability of obtaining 3 red balls before a green ball is 0.105.

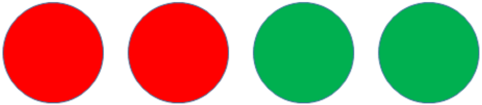
Probability



1 * 1 * 1 * 1 = 1

- Suppose there are 4 red balls in the box (box 2).
- If we sample with replacement, the probability of selecting 4 red balls is 1.

Probability



0.5 * 0.5 * 0.5 * 0.5 = 0.0625

- Suppose there are 2 red balls and 2 green balls in the box (box 3).
- If we sample with replacement, the probability of selecting 2 red balls before 2 green balls is 0.0625.

Probability

The purer the condition, the higher the probability.

Shannon Entropy

Suppose probabilities of I possible outcomes for an event X are $p_1, \cdots, p_I$. The Shannon entropy $H(X)$ is defined as

$$H(X) = -\sum_{i=1}^{I} p_i \log_2(p_i)$$

With 2 possible outcomes,

$$H(X) = -\sum_{i=1}^{2} p_i \log_2(p_i) = -p_1 \log_2(p_1) - p_2 \log_2(p_2)$$

Shannon Entropy

Let p_1 be the probability of selecting red balls and p_2 be the probability of selecting green balls:

- $H\left(\text{box } 1\right) = -0.75\log_2\left(0.75\right) - 0.25\log_2\left(0.25\right) = 0.811$
- $H\left(\text{box } 2\right) = -1\log_2\left(1\right) - 0\log_2\left(0\right) = 0$
- $H\left(\text{box } 3\right) = -0.5\log_2\left(0.5\right) - 0.5\log_2\left(0.5\right) = 1$

$0 \log 0 = ?$

Using L'Hôpital's rule, we have

$$\begin{aligned}
\lim_{x\to 0} x \log x &= \lim_{x\to 0} \frac{\log x}{\frac{1}{x}} \\
&= \lim_{x\to 0} \frac{\frac{1}{x}}{-x^{-2}} \\
&= -\lim_{x\to 0} x \\
&= 0
\end{aligned}$$

Example

- Our purpose is to see whether Gender (male vs female) and Class (class A vs class B) are useful variables in predicting whether a student likes to play basketball (or classifying students).
- There are a total of 30 students. Among them, 15 like to play basketball:

LLLLLLLLLLLLLLL NNNNNNNNNNNNNNN

$$H = -\frac{15}{30}\log_2\left(\frac{15}{30}\right) - \frac{15}{30}\log_2\left(\frac{15}{30}\right) = 1$$

Using Gender

Suppose 13 out of 20 male and 2 out of 10 female students like to play basketball:

MMMMMMMMMMMMMMMMMMMM FFFFFFFFFF

- $H\left(\text{male}\right) = -\frac{13}{20}\log_2\left(\frac{13}{20}\right) - \frac{7}{20}\log_2\left(\frac{7}{20}\right) = 0.93$
- $H\left(\text{female}\right) = -\frac{2}{10}\log_2\left(\frac{2}{10}\right) - \frac{8}{10}\log_2\left(\frac{8}{10}\right) = 0.72$

$$H\left(\text{Gender}\right) = \frac{20}{30} \times 0.93 + \frac{10}{30} \times 0.72 = 0.86$$

Using Class

6 students out of 14 in class A, and 9 students out of 16 in class B like to play basketball:

AAAAAAAAAAAAAA BBBBBBBBBBBBBBBB

- $H\left(\text{class A}\right) = -\frac{6}{14}\log_2\left(\frac{6}{14}\right) - \frac{8}{14}\log_2\left(\frac{8}{14}\right) = 0.99$
- $H\left(\text{class B}\right) = -\frac{9}{16}\log_2\left(\frac{9}{16}\right) - \frac{7}{16}\log_2\left(\frac{7}{16}\right) = 0.99$

$$H\left(\text{Class}\right) = \frac{14}{30} \times 0.99 + \frac{16}{30} \times 0.99 = 0.99$$

Gender vs Class

Since the result from Gender is less impure (Gender $0.86 < $ Class 0.99), Gender is more effective than Class in classification. Therefore we use Gender before Class.

Information Gain

Information gain is the amount of reduction in entropy:

- Information gain for Gender: $1 - 0.86 = 0.14$
- Information gain for Class: $1 - 0.99 = 0.01$

Gini Impurity

$$Gini = 1 - \sum_{i=1}^{I} p_j^2$$

- when the node is pure:

$$Gini_{min} = 1 - \left(1^2 + 0^2\right) = 0$$

- when the node is most chaotic:

$$Gini_{max} = 1 - \left(0.5^2 + 0.5^2\right) = 0.5$$

Back to our Goal

#	Outlook	Temp	Humidity	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rainy	Mild	High	Weak	Yes
5	Rainy	Cool	Normal	Weak	Yes
6	Rainy	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rainy	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rainy	Mild	High	Strong	No

Initial Entropy

- Probability of playing tennis $p_1 = \frac{9}{14}$
- Probability of not playing tennis $p_2 = \frac{5}{14}$ $(= 1 - p_1 = 1 - \frac{9}{14})$
- The initial entropy (without using any variable) is

$$H = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.94$$

4 Attributes (Variables)

We need to decide the root from these 4 attributes:

- Outlook
- Temp
- Wind
- Humidity

Outlook: Sunny

There are 5 sunny days. Among those 5 days, tennis was played on 2 days and tennis was not played on 3 days:

$$H\left(\text{sunny}\right) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.97$$

Outlook: Overcast

There are 4 overcast days. Among those 4 days, tennis was played on all of the 4 days:

$$H\left(\text{overcast}\right) = -\frac{4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right) = 0$$

Outlook: Rain

There are 5 rainy days. Among those 5 days, tennis was played on 3 days and tennis was not played on 2 days:

$$H\left(\text{rainy}\right) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) = 0.97$$

Outlook

- Sunny: $H\left(\text{sunny}\right) = 0.97$ (5 days)
- Overcast: $H\left(\text{overcast}\right) = 0$ (4 days)
- Rain: $H\left(\text{rainy}\right) = 0.97$ (5 days)
- Entropy of Outlook is the weighted average:

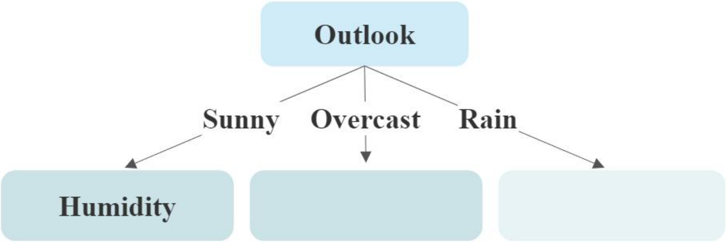
$$H\left(\text{Outlook}\right) = \frac{5}{14} \times 0.97 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.97 = 0.69$$

Outlook as Root

Outlook has the largest information gain and is chosen as the root of the decision tree, in that

- Information gain for Outlook $= 0.25$ $(= 0.94 - 0.69)$
- Information gain for Temp $= 0.029$
- Information gain for Wind $= 0.048$
- Information gain for Humidity $= 0.152$

Sunny



Sunny: Temp

#	Outlook	Temp	Humidity	Wind	Play
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
8	sunny	mild	high	weak	no
9	sunny	cold	normal	weak	yes
11	sunny	mild	normal	strong	yes

- $H\left(\text{hot}\right) = -\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right) = 0$
- $H\left(\text{mild}\right) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$
- $H\left(\text{cold}\right) = -\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right) = 0$

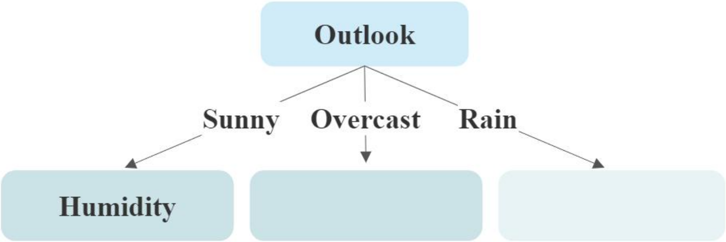
$$H\left(\text{Temp}\right) = \frac{2}{5} \times 0 + \frac{2}{5} \times 1 + \frac{1}{5} \times 0 = 0.4$$

Sunny

We choose humidity as the branch for sunny because

- Information gain for Temp $= 0.57\,(= 0.970 - 0.4)$
- Information gain for Humidity $= 0.970$
- Information gain for Wind $= 0.019$

Sunny



Lab

Analyze playtennis.csv, regtree.csv and possum.csv using decision trees in R and Python