Decision Tree

COEN140

Santa Clara University

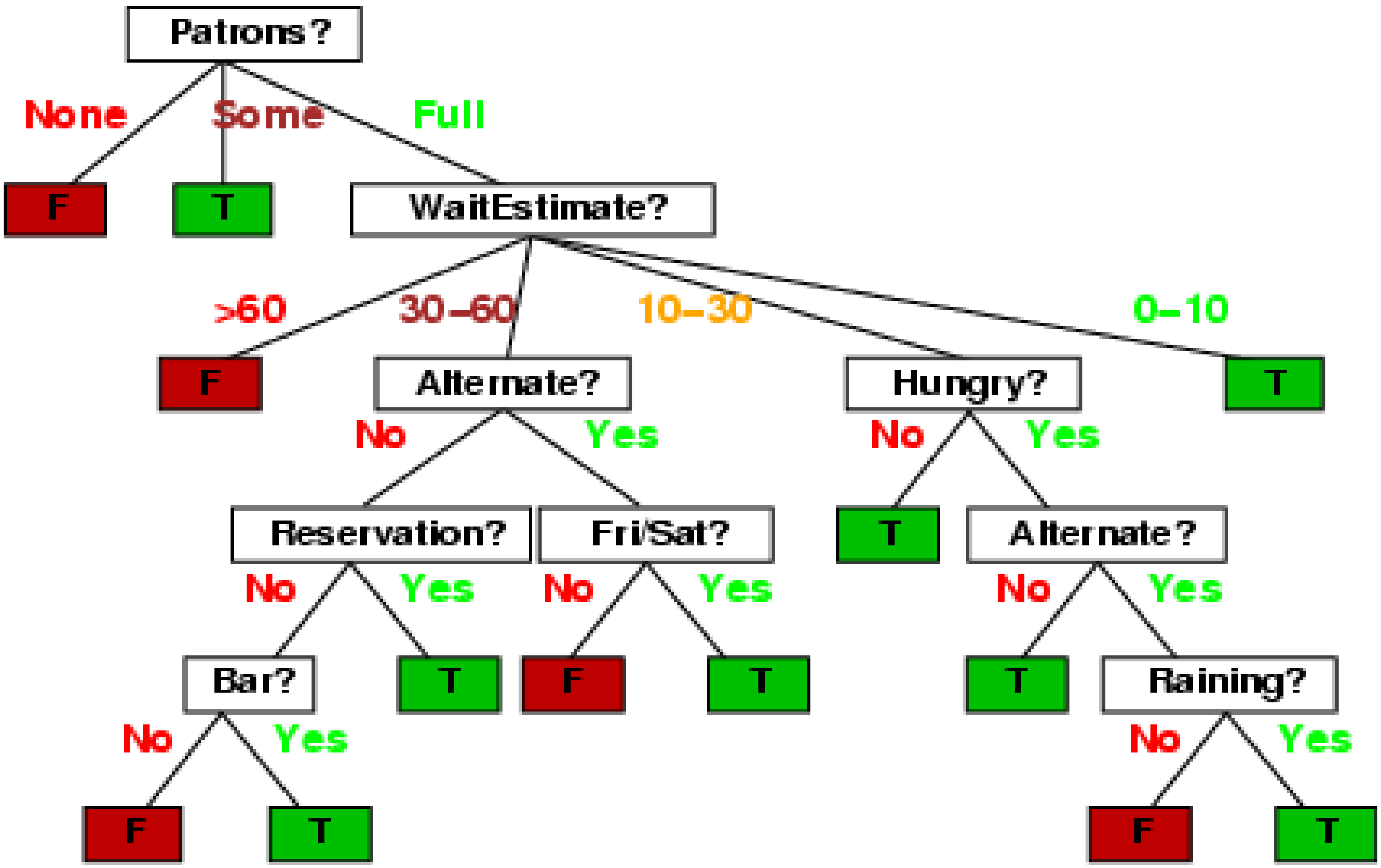
Decide whether to wait in a restaurant?

- Ask yourself questions
  - Alternate: is there an alternative restaurant nearby?
  - Bar: is there a comfortable bar area to wait in?
  - Fri/Sat: is today Friday or Saturday?
  - Hungry: are we hungry?
  - Patrons: number of people in the restaurant (None, Some, Full)

Decide whether to wait in a restaurant?

- Ask yourself questions
  - Price: price range ($, $$, $$$)
  - Raining: is it raining outside?
  - Reservation: have we made a reservation?
  - Type: kind of restaurant (French, Italian, Thai, Burger)
  - WaitEstimate: estimated waiting time (0-10, 10-30, 30-60, >60)

Ask questions one by one



Ask questions one by one

- What should be the first question to ask?
- What should be the next question to ask?
- …
- When can you get to a decision?

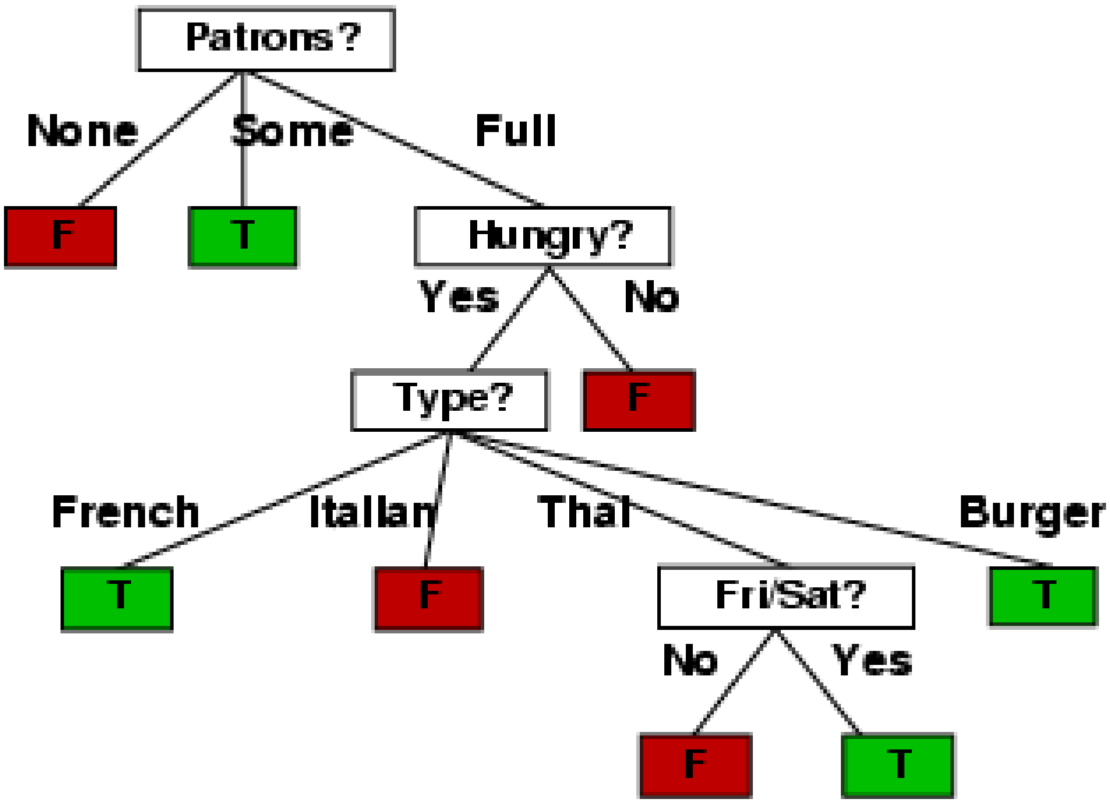
Data Samples: a set of examples

- Classification of examples is positive (T) or negative (F)
- General form for data: N samples, each with attributes $(x_1, x_2, x_3, \ldots x_d)$ and target value y.

Example	Input Attributes										Goal
	Alt	Bar	Fri	Hun	Pat	Price	Rain	Res	Type	Est	*WillWait*
$\mathbf{x}_1$	*Yes*	*No*	*No*	*Yes*	*Some*	$$$	*No*	*Yes*	*French*	*0–10*	y_1 = *Yes*
$\mathbf{x}_2$	*Yes*	*No*	*No*	*Yes*	*Full*	$	*No*	*No*	*Thai*	*30–60*	y_2 = *No*
$\mathbf{x}_3$	*No*	*Yes*	*No*	*No*	*Some*	$	*No*	*No*	*Burger*	*0–10*	y_3 = *Yes*
$\mathbf{x}_4$	*Yes*	*No*	*Yes*	*Yes*	*Full*	$	*Yes*	*No*	*Thai*	*10–30*	y_4 = *Yes*
$\mathbf{x}_5$	*Yes*	*No*	*Yes*	*No*	*Full*	$$$	*No*	*Yes*	*French*	*> 60*	y_5 = *No*
$\mathbf{x}_6$	*No*	*Yes*	*No*	*Yes*	*Some*	$$	*Yes*	*Yes*	*Italian*	*0–10*	y_6 = *Yes*
$\mathbf{x}_7$	*No*	*Yes*	*No*	*No*	*None*	$	*Yes*	*No*	*Burger*	*0–10*	y_7 = *No*
$\mathbf{x}_8$	*No*	*No*	*No*	*Yes*	*Some*	$$	*Yes*	*Yes*	*Thai*	*0–10*	y_8 = *Yes*
$\mathbf{x}_9$	*No*	*Yes*	*Yes*	*No*	*Full*	$	*Yes*	*No*	*Burger*	*> 60*	y_9 = *No*
$\mathbf{x}_{10}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	$$$	*No*	*Yes*	*Italian*	*10–30*	y_{10} = *No*
$\mathbf{x}_{11}$	*No*	*No*	*No*	*No*	*None*	$	*No*	*No*	*Thai*	*0–10*	y_{11} = *No*
$\mathbf{x}_{12}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	$	*No*	*No*	*Burger*	*30–60*	y_{12} = *Yes*

Decision Tree

- You want to "learn" a tree from those training examples
  - a small tree consistent with the training examples



Decision Tree

- Decision tree: is a classifier
  - An input-output mapping
  - $y = f(\mathbf{x})$
  - Input: $\mathbf{x} = [x_1, x_2, \dots, x_d]^T$
    d attributes/features
  - Output: y, the decision

- It performs classification (makes decisions) by:
  - Executing a sequence of tests
  - Each test: test the value of an attribute

Decision Tree

- Decision tree: is a classifier
  - An input-output mapping
  - $y = f(\mathbf{x})$
  - Input: $\mathbf{x} = [x_1 \ , x_2 \ , \ldots , x_d \ ]^T$
    d attributes
  - Output: y, the decision

- We are given a set of training samples
  - Learn a decision tree (i.e. learn a classifier)
- Then we can apply this decision tree to a new instance to make a decision

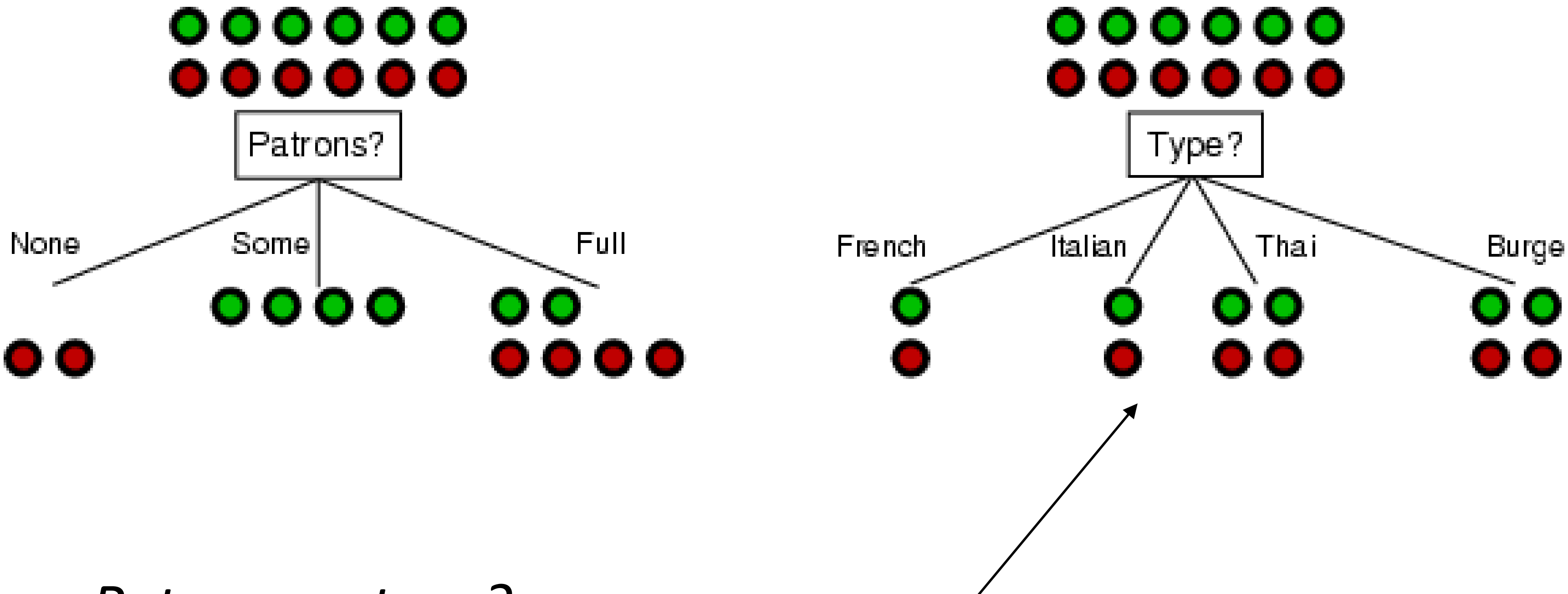
Data Samples

- Classification of examples is positive (T) or negative (F)
- General form for data: N instances, each with attributes $(x_1, x_2, x_3, \ldots x_d)$ and target value y.

Example	Input Attributes										Goal
	Alt	Bar	Fri	Hun	Pat	Price	Rain	Res	Type	Est	*WillWait*
$\mathbf{x}_1$	*Yes*	*No*	*No*	*Yes*	*Some*	*$$$*	*No*	*Yes*	*French*	*0–10*	y_1 = *Yes*
$\mathbf{x}_2$	*Yes*	*No*	*No*	*Yes*	*Full*	*$*	*No*	*No*	*Thai*	*30–60*	y_2 = *No*
$\mathbf{x}_3$	*No*	*Yes*	*No*	*No*	*Some*	*$*	*No*	*No*	*Burger*	*0–10*	y_3 = *Yes*
$\mathbf{x}_4$	*Yes*	*No*	*Yes*	*Yes*	*Full*	*$*	*Yes*	*No*	*Thai*	*10–30*	y_4 = *Yes*
$\mathbf{x}_5$	*Yes*	*No*	*Yes*	*No*	*Full*	*$$$*	*No*	*Yes*	*French*	*> 60*	y_5 = *No*
$\mathbf{x}_6$	*No*	*Yes*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Italian*	*0–10*	y_6 = *Yes*
$\mathbf{x}_7$	*No*	*Yes*	*No*	*No*	*None*	*$*	*Yes*	*No*	*Burger*	*0–10*	y_7 = *No*
$\mathbf{x}_8$	*No*	*No*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Thai*	*0–10*	y_8 = *Yes*
$\mathbf{x}_9$	*No*	*Yes*	*Yes*	*No*	*Full*	*$*	*Yes*	*No*	*Burger*	*> 60*	y_9 = *No*
$\mathbf{x}_{10}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$$$*	*No*	*Yes*	*Italian*	*10–30*	y_{10} = *No*
$\mathbf{x}_{11}$	*No*	*No*	*No*	*No*	*None*	*$*	*No*	*No*	*Thai*	*0–10*	y_{11} = *No*
$\mathbf{x}_{12}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$*	*No*	*No*	*Burger*	*30–60*	y_{12} = *Yes*

How to choose an attribute?

- Idea: a good attribute splits the examples into subsets that are (ideally) “all positive” or “all negative”



- *Patrons or type?*

To wait or not to wait is still at 50%.

Information Theory

- Consider a discrete random source S, which takes on symbols from a fixed finite alphabet

$$\mathcal{S} = \{s_0, s_1, \dots, s_{K-1}\}$$

with probabilities

$$P(S = s_k) = p_k, \qquad k = 0, 1, \dots, K-1$$

- The set of probabilities satisfy

$$\sum_{k=0}^{K-1} p_k = 1$$

Information Theory

- Measure how much information is produced by such a random source S?

- Information: related to “uncertainty”

Example

- Let random source S represent tomorrow will rain or not rain.
- $S = 1$: tomorrow will rain
- $S = 0$: tomorrow will not rain

- I told you:

$$p_1 = P(S = 1) = 0.98$$

- When tomorrow arrives, it does rain.
  - Is it surprising or not?
- What if $p_1 = P(S = 1) = 0.01$?

Uncertainty, Surprise, Information

- Uncertainty: before the event $S = s_k$ occurs, there is an amount of uncertainty

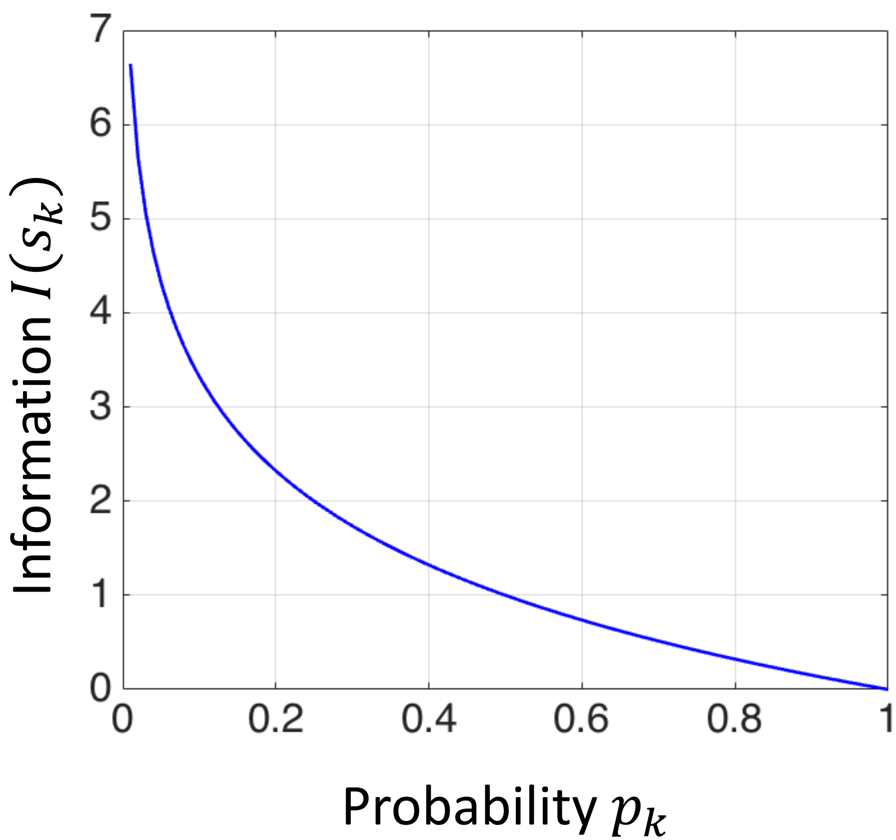
- Surprise: when the event $S = s_k$ occurs, there is an amount of surprise

- Information: after the occurrence of the event $S = s_k$, there is gain in the amount of information.

  The amount of information is related to the inverse of p_k (probability of occurrence)

Information

- Define the amount of information gained after observing the event $S = s_k$, which occurs with probability p_k

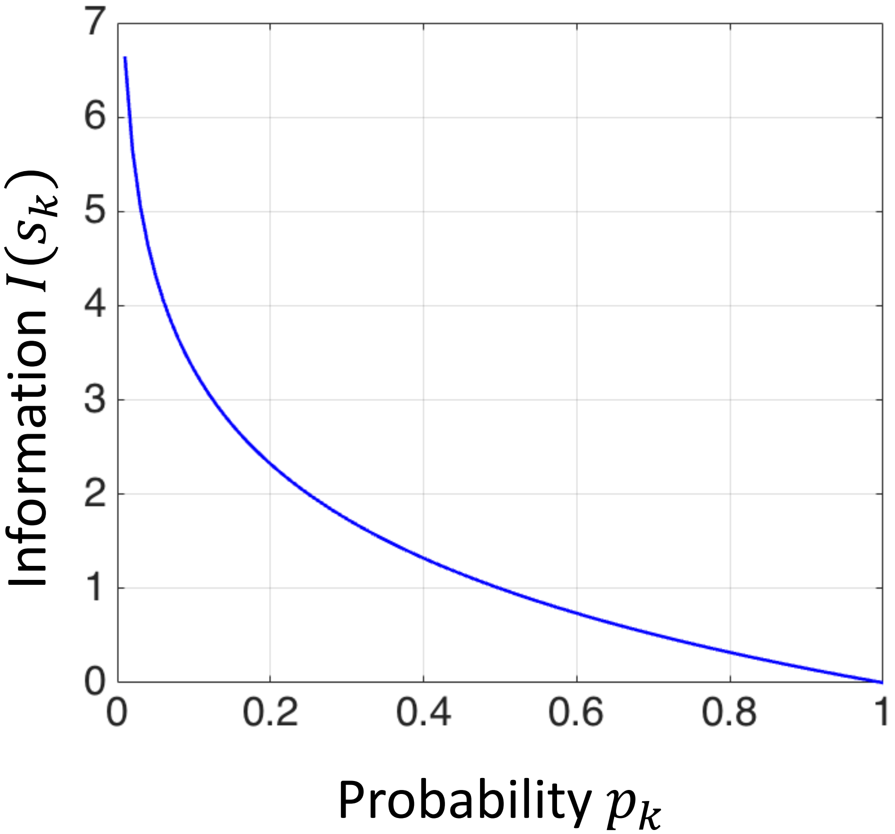


$$I(s_k) = \log_2 \frac{1}{p_k}$$

Information

$$I(s_k) = \log \frac{1}{p_k}$$

- $I(s_k) = 0$ for $p_k = 1$
- $I(s_k) \geq 0$ for $0 \leq p_k \leq 1$
- $I(s_k) > I(s_i)$ for $p_k < p_i$



Information

$$I(s_k) = \log \frac{1}{p_k}$$

- base of the logarithm: arbitrary
- we usually use $\log_2$
- The resulting unit of information is called the bit

$$I(s_k) = \log_2 \frac{1}{p_k} = -\log_2 p_k, k = 0, 1, \cdots, K-1$$

e.g. If $k = 0, 1$, and $p_k = 1/2$, then $I(s_k) = 1$ bit (one bit is the amount of information that we gain when one of two possible and equally likely events occurs)

Example

- You have a message to send to a friend
- S: tomorrow's weather condition

k	s_k	$P(S = s_k)$	$I(s_k) = \log_2 \frac{1}{p_k}$
1	sunny	1/4	2 bits
2	rainy	1/4	2 bits
3	windy	1/4	2 bits
4	cloudy	1/4	2 bits

- Encode these messages in a sequence of binary "bits".
  - How many bits do you need to represent each of the four messages?

Entropy

- The expectation of $I(s_k)$ over the source alphabet $\mathcal{S}$ is
- $H(S) = E[I(s_k)] = \sum_{k=0}^{K-1} p_k\, I(p_k) = \sum_{k=0}^{K-1} p_k \log_2 \frac{1}{p_k}$

- The entropy of a discrete random source S

$$H(S) = \sum_{k=0}^{K-1} p_k \log_2 \frac{1}{p_k}$$

Or $H(S) = -\sum_{k=0}^{K-1} p_k \log_2 p_k$

Information

- If there are K symbols in the source alphabet, then

$$0 \leq H(S) \leq \log_2 K$$

- $H(S) = 0$ if and only if $p_k = 1$ for some k, and the remaining probabilities in the set are all zero

  $\Rightarrow$ the lower bound of entropy (corresponds to no uncertainty)

- $H(S) = \log_2 K$ if and only if $p_k = 1/K$ for all k (i.e. all symbols in the alphabet are equiprobable)

  $\Rightarrow$ upper bound of entropy (corresponds to maximum uncertainty)

Entropy of Binary Source (Classes)

- Symbol 0 occurs with probability p_0, symbol 1 occurs with probability $p_1 = 1 - p_0$

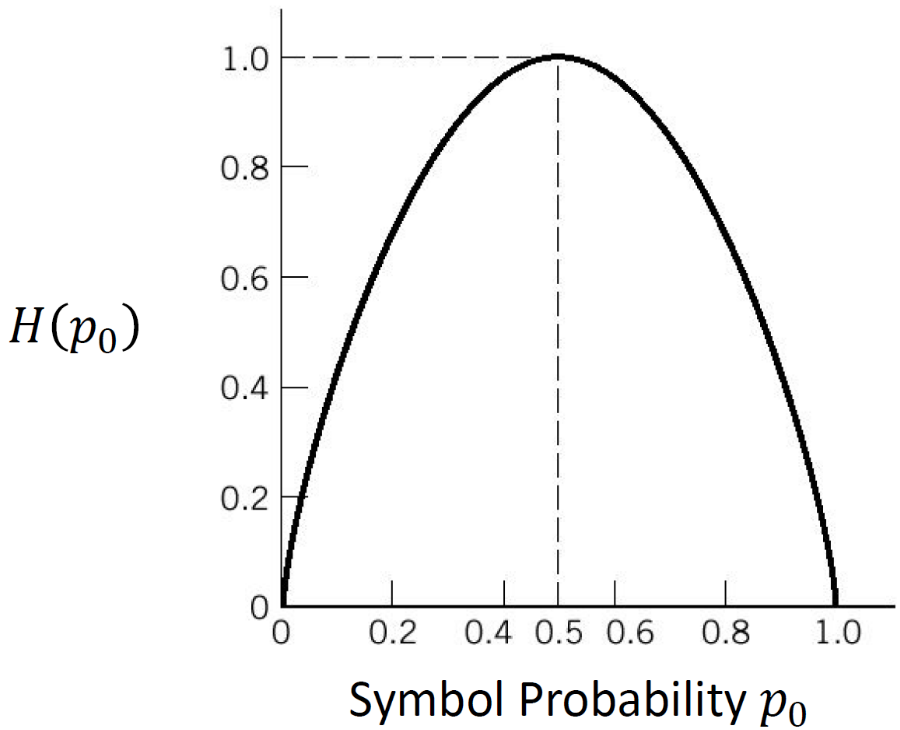
- The entropy of such a source equals

$$\begin{aligned} H(S) &= -p_0 \log_2 p_0 - p_1 \log_2 p_1 \\ &= -p_0 \log_2 p_0 - (1 - p_0) \log_2 (1 - p_0) \;\; \text{bits} \end{aligned}$$

- Use a special notation for such $H(S)$, which is

$$H(p_0) = -p_0 \log_2 p_0 - (1 - p_0) \log_2 (1 - p_0) \;\; \text{bits}$$

Entropy of Binary Source (Classes)



- When $p_0 = 0$ or $p_0 = 1$, $H(p_0) = 0$ (no information)
- When $p_0 = p_1 = \frac{1}{2}$, $H(p_0) = 1$ (maximum information)

Example

- The distribution of a discrete random source X is the following, calculate the entropy $H(X)$.
- $P(X = x_1) = \frac{1}{4}, P(X = x_2) = \frac{3}{4}$

- Answer
- $H(X) = -\sum_{i=1}^{2} p_i \times \log_2 p_i = -\frac{1}{4}\log_2 \frac{1}{4} - \frac{3}{4}\log_2 \frac{3}{4}$
  $= 0.8113$ bits

Example

- The distribution of a discrete random source X is the following, calculate the entropy $H(X)$.
- $P(X = x_1) = \frac{1}{2}, P(X = x_2) = \frac{1}{4}, P(X = x_3) = \frac{1}{4}$

- Answer
- $H(X) = -\sum_{i=1}^{3} p_i \times \log_2 p_i$
- $= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4}$

  $= 1.5$ bits

Example – Compare different distributions

- Source S_1: $p_1 = 0, p_2 = 1$
- Source S_2: $p_1 = \frac{1}{2}, p_2 = \frac{1}{2}$
- Source S_3: $p_1 = \frac{1}{3}, p_2 = \frac{1}{3}, p_3 = \frac{1}{3}$
- Source S_4: $p_1 = \frac{1}{2}, p_2 = \frac{1}{3}, p_3 = \frac{1}{6}$
- Compare $H(S_1), H(S_2), H(S_3), H(S_4)$ without computation?

Example – Compare different distributions

- Source S_1: $p_1 = 0, p_2 = 1$
- Source S_2: $p_1 = \frac{1}{2}, p_2 = \frac{1}{2}$
- Source S_3: $p_1 = \frac{1}{3}, p_2 = \frac{1}{3}, p_3 = \frac{1}{3}$
- Source S_4: $p_1 = \frac{1}{2}, p_2 = \frac{1}{3}, p_3 = \frac{1}{6}$
- Answer:
- H(S1): smallest
- H(S2)<H(S3)
- H(S4)<H(S3)
- H(S2) vs H(S4)?
- H(S2)<H(S4)

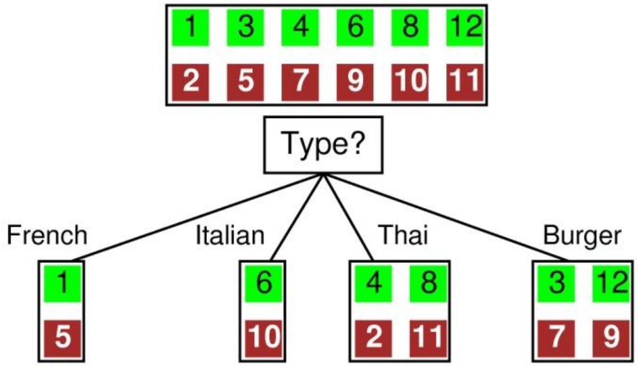
Decision Tree

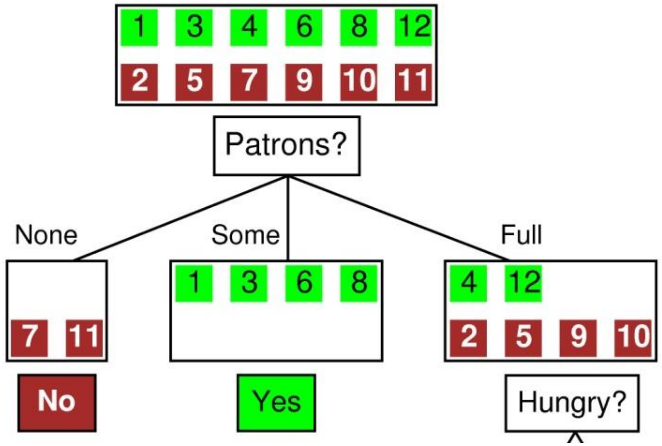
- What is the uncertainty of the outcome if we disclose the value of some attribute?

- Information Gain:

the uncertainty before testing an attribute − the uncertainty after testing an attribute

Example





Uncertainty Before:

Entropy(Y) = $-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = \log_2 2$ = 1 bit:

There is "1 bit of information to be discovered".

Uncertainty After testing the attribute Type:

If we go into branch "French", the uncertainty is still 1 bit, similarly for Italian, Thai, and Burger.

French: 1bit

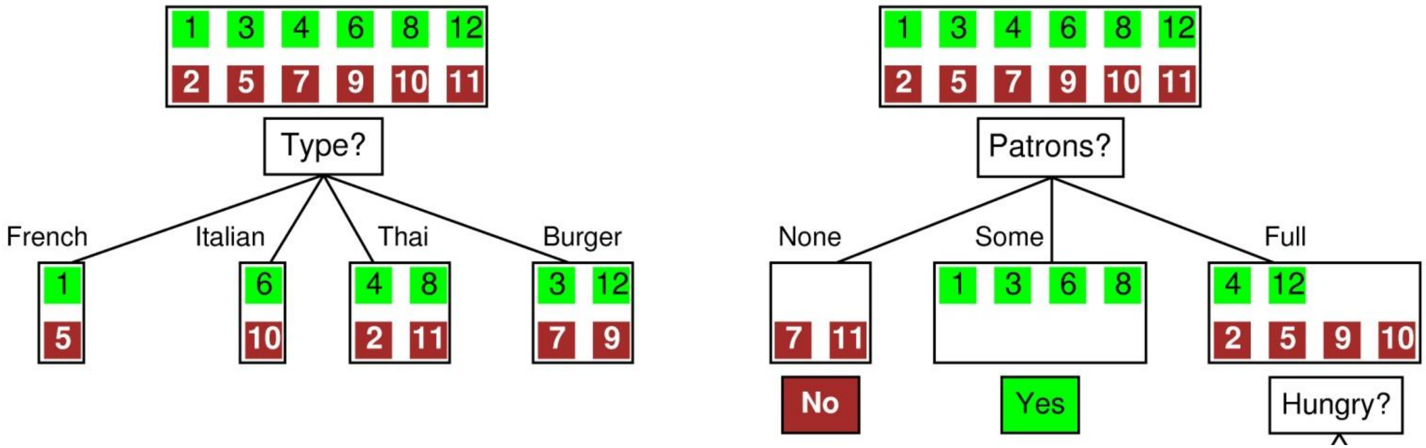
Italian: 1 bit

Thai: 1 bit

Burger: 1bit

On average: 1 bit ! We gained no information!

Example



Uncertainty Before: entropy = $-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}$=$\log_2 2$ = 1 bit:

There is "1 bit of information to be discovered".

Uncertainty After testing attribute Patrons:

In branches "None" and "Some": entropy = 0,

In branch "Full" entropy = $-\frac{1}{3}\times\log_2\frac{1}{3}-\frac{2}{3}\times\log_2\frac{2}{3} = 0.918$ bits

Uncertainty is reduced!

So attribute **Patrons** gains more information!

Conditional Entropy

- Consider two RVs X and Y
  - X has N possible values: $x_1, x_2, \ldots, x_N$
  - Y also has a set of possible values
- The conditional entropy of Y under X (or given X) is defined as
- $H(Y|X) = \sum_{i=1}^{N} P(X = x_i) \times H(Y|X = x_i)$

Conditional Entropy

- Combine branches to obtain the entropy (of the outcome) when testing a certain attribute $Patrons$:

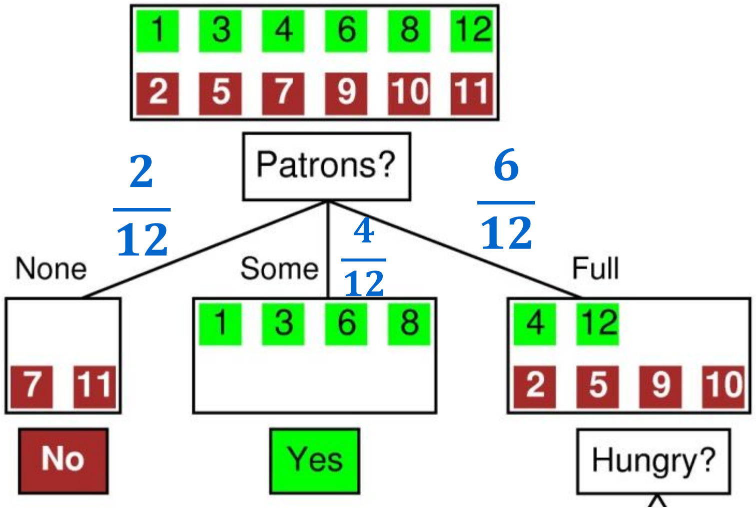
$$H(outcome|Patrons) = \sum_{i=1}^{3} \frac{p_i + n_i}{p + n} H\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right)$$

weight for the i^{th} branch

Conditional entropy for the i^{th} branch.

$Patrons$: has 3 possible values, indexed by i

p_i: the number of positive outcomes when $Patrons =$ the i^{th} value



Conditional Entropy

- Combine branches to obtain the entropy (of the outcome) when testing a certain attribute $Patrons$:

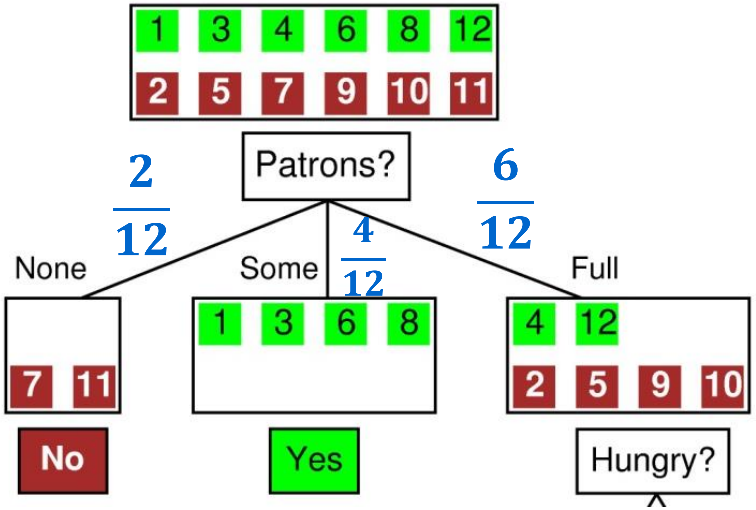
$$H(outcome|Patrons) = \sum_{i=1}^{3} \frac{p_i + n_i}{p + n} H\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right)$$

weight for the i^{th} branch

Conditional entropy for the i^{th} branch.

$Patrons$: has 3 possible values, indexed by i

n_i: the number of negative outcomes when $Patrons =$ the i^{th} value



Conditional Entropy

- Combine branches to obtain the entropy (of the outcome) when testing a certain attribute $Patrons$:

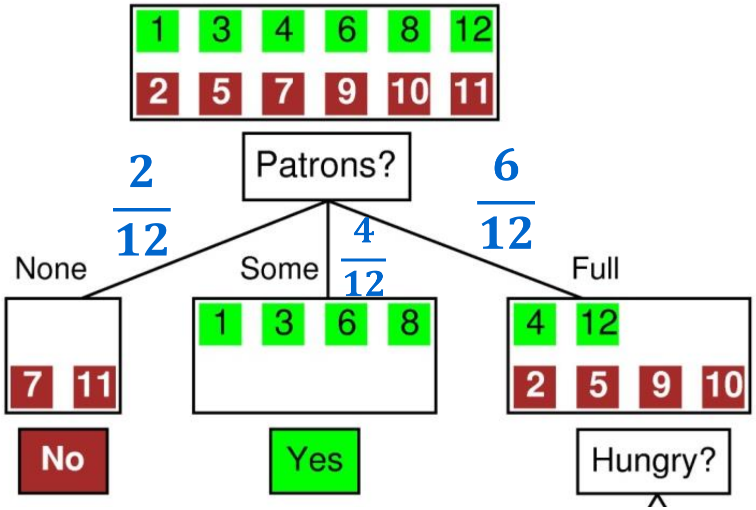
$$H(outcome|Patrons) = \sum_{i=1}^{3} \frac{p_i + n_i}{p + n} H\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right)$$

weight for the i^{th} branch

Conditional entropy for the i^{th} branch.

p: the total number of positive outcomes in the training examples

n: the total number of negative outcomes in the training examples



Conditional Entropy

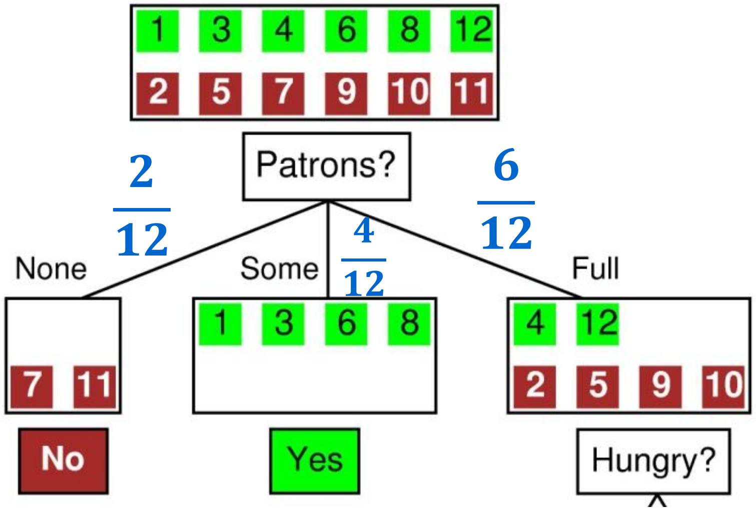
- Combine branches to obtain the entropy (of the outcome) when testing a certain attribute $Patrons$:

$$H(outcome|Patrons) = \sum_{i=1}^{3} \frac{p_i + n_i}{p + n} H\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right)$$

weight for the i^{th} branch

Conditional entropy for the i^{th} branch.

$\frac{p_i+n_i}{p+n}$: proportion of training examples when $Patrons =$ the i^{th} value



Conditional Entropy

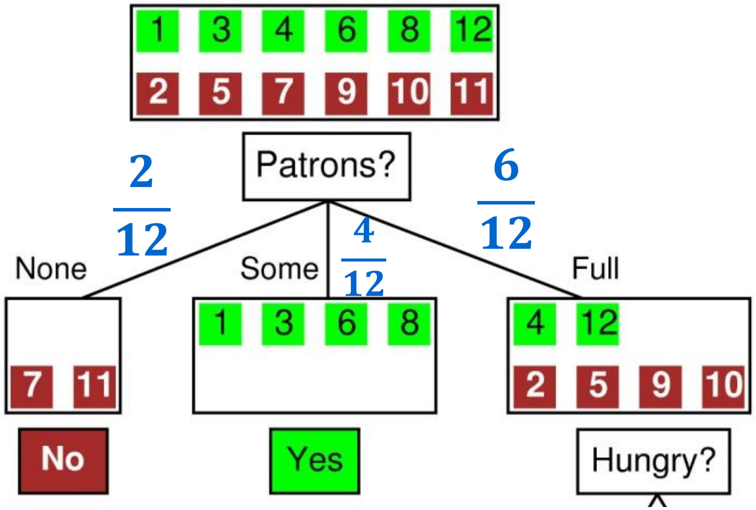
- Combine branches to obtain the entropy (of the outcome) when testing a certain attribute $Patrons$:

$$H(outcome|Patrons) = \sum_{i=1}^{3} \frac{p_i + n_i}{p + n} H\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right)$$

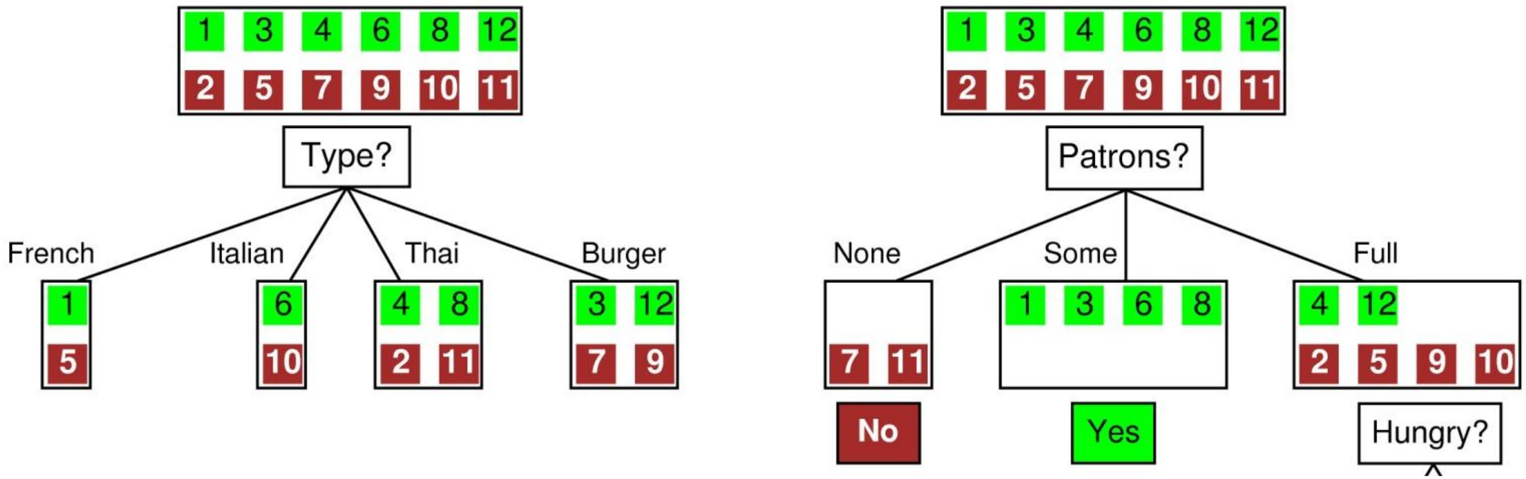
weight for the i^{th} branch

Conditional entropy for the i^{th} branch.

$\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i}$: distribution of positive and negative outcomes when $Patrons = the\ i^{th}$ value

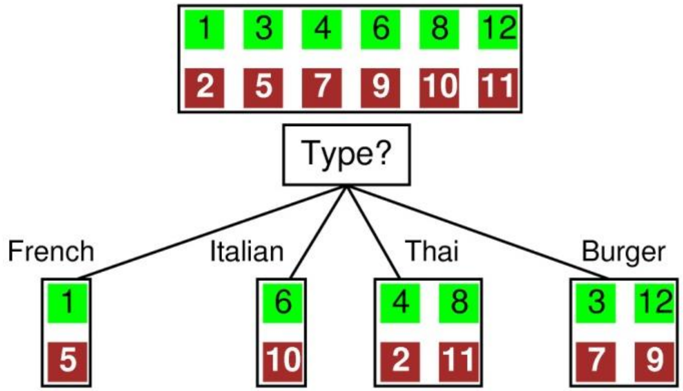


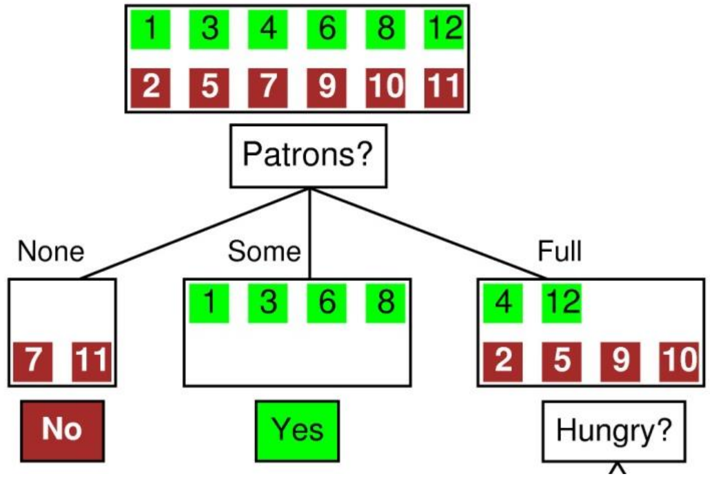
Minimum Conditional Entropy



- Find the Attribute that leads to the minimum conditional entropy of the outcome
  - Find the attribute A such that $H(Outcome|A)$ is the minimum.
- $H(Outcome|Patrons) = \frac{2}{12}H(0,1) + \frac{4}{12}H(1,0) + \frac{6}{12}H\left(\frac{2}{6},\frac{4}{6}\right)$ =0.459 bits

Minimum Conditional Entropy

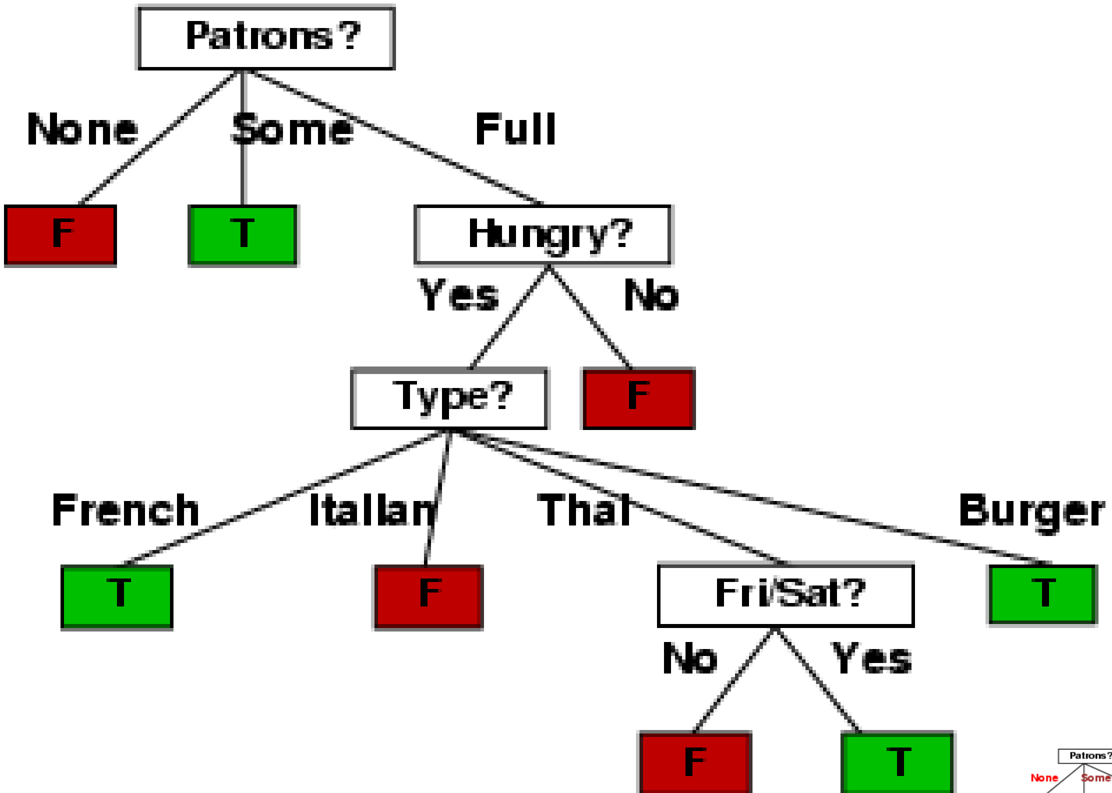




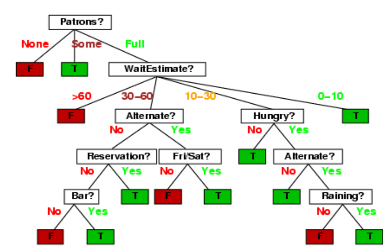
- Find the Attribute that leads to the minimum conditional entropy of the outcome
  - Find the attribute A such that $H(Outcome|A)$ is the minimum.
- $H(Outcome|Patrons) = \frac{2}{12}H(0,1) + \frac{4}{12}H(1,0) + \frac{6}{12}H\left(\frac{2}{6},\frac{4}{6}\right)$ =0.459 bits
- $H(Outcome|Type) = \frac{2}{12}H\left(\frac{1}{2},\frac{1}{2}\right) + \frac{2}{12}H\left(\frac{1}{2},\frac{1}{2}\right) + \frac{4}{12}H\left(\frac{2}{4},\frac{2}{4}\right) + \frac{4}{12}H\left(\frac{2}{4},\frac{2}{4}\right) =$ 1 bit

Example contd.

- Decision tree learned from the 12 training examples:



- Substantially simpler than “true” tree



Example

- You are a robot in the aquarium section of a pet store, and must learn to discriminate Red fish from Blue fish. You will learn to discriminate them by body parts. You choose to learn a Decision Tree classifier. You are given the following examples:

Example	Fins	Tail	Body	Class
Example #1	Thin	Small	Slim	Red
Example #2	Wide	Large	Slim	Red
Example #3	Thin	Large	Slim	Red
Example #4	Wide	Small	Medium	Red
Example #5	Thin	Small	Medium	Blue
Example #6	Wide	Large	Fat	Blue
Example #7	Thin	Large	Fat	Blue
Example #8	Wide	Small	Fat	Blue

Example

Example	Fins	Tail	Body	Class
Example #1	Thin	Small	Slim	Red
Example #2	Wide	Large	Slim	Red
Example #3	Thin	Large	Slim	Red
Example #4	Wide	Small	Medium	Red
Example #5	Thin	Small	Medium	Blue
Example #6	Wide	Large	Fat	Blue
Example #7	Thin	Large	Fat	Blue
Example #8	Wide	Small	Fat	Blue

- What is the entropy of Class before testing any attribute?

$$H\left(\frac{4}{8},\frac{4}{8}\right) = H\left(\frac{1}{2},\frac{1}{2}\right) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1 \text{ bit}$$

Example

Example	Fins	Tail	Body	Class
Example #1	Thin	Small	Slim	Red
Example #2	Wide	Large	Slim	Red
Example #3	Thin	Large	Slim	Red
Example #4	Wide	Small	Medium	Red
Example #5	Thin	Small	Medium	Blue
Example #6	Wide	Large	Fat	Blue
Example #7	Thin	Large	Fat	Blue
Example #8	Wide	Small	Fat	Blue

- What is the conditional entropy of Class under attribute Fins?

$$H(C|Fins) \ = \frac{4}{8} \times H(C|Fins = Thin) + \frac{4}{8} \times H(C|Fins = Wide)$$

$$= \frac{1}{2} \times H\left(\frac{2}{4}, \frac{2}{4}\right) + \frac{1}{2} \times H\left(\frac{2}{4}, \frac{2}{4}\right)$$

$$= \frac{1}{2} \times H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} \times H\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{1}{2} \times 1 + \frac{1}{2} \times 1 = 1 \text{ bit}$$

Example

Example	Fins	Tail	Body	Class
Example #1	Thin	Small	Slim	Red
Example #2	Wide	Large	Slim	Red
Example #3	Thin	Large	Slim	Red
Example #4	Wide	Small	Medium	Red
Example #5	Thin	Small	Medium	Blue
Example #6	Wide	Large	Fat	Blue
Example #7	Thin	Large	Fat	Blue
Example #8	Wide	Small	Fat	Blue

- What is the conditional entropy of Class under attribute Tail ?

$$H(C|Tail) = \frac{4}{8} \times H(C|Tail = Small) + \frac{4}{8} \times H(C|Tail = Large)$$

$$= \frac{1}{2} \times H\left(\frac{2}{4}, \frac{2}{4}\right) + \frac{1}{2} \times H\left(\frac{2}{4}, \frac{2}{4}\right)$$

$$= \frac{1}{2} \times H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} \times H\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{1}{2} \times 1 + \frac{1}{2} \times 1 = 1 \text{ bit}$$

Example

Example	Fins	Tail	Body	Class
Example #1	Thin	Small	Slim	Red
Example #2	Wide	Large	Slim	Red
Example #3	Thin	Large	Slim	Red
Example #4	Wide	Small	Medium	Red
Example #5	Thin	Small	Medium	Blue
Example #6	Wide	Large	Fat	Blue
Example #7	Thin	Large	Fat	Blue
Example #8	Wide	Small	Fat	Blue

- What is the conditional entropy of Class under attribute Body ?

$$H(C|Body) = \frac{3}{8} \times H(C|Body = Slim) + \frac{2}{8} \times H(C|Body = Medium) + \frac{3}{8} \times H(C|Body = Fat)$$

$$= \frac{3}{8} \times H\left(\frac{3}{3}, \frac{0}{3}\right) + \frac{2}{8} \times H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{3}{8} \times H\left(\frac{0}{3}, \frac{3}{3}\right)$$

$$= \frac{3}{8} \times H(1,0) + \frac{2}{8} \times H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{3}{8} \times H(0,1)$$

$$= \frac{3}{8} \times 0 + \frac{2}{8} \times 1 + \frac{3}{8} \times 0 = 0.25 \text{ bits}$$

Example

Example	Fins	Tail	Body	Class
Example #1	Thin	Small	Slim	Red
Example #2	Wide	Large	Slim	Red
Example #3	Thin	Large	Slim	Red
Example #4	Wide	Small	Medium	Red
Example #5	Thin	Small	Medium	Blue
Example #6	Wide	Large	Fat	Blue
Example #7	Thin	Large	Fat	Blue
Example #8	Wide	Small	Fat	Blue

- Which attribute will you select as the root attribute, and why?

  Body, because the conditional entropy of Class is the smallest under attribute Body.

Example

Example	Fins	Tail	Body	Class
Example #1	Thin	Small	Slim	Red
Example #2	Wide	Large	Slim	Red
Example #3	Thin	Large	Slim	Red
Example #4	Wide	Small	Medium	Red
Example #5	Thin	Small	Medium	Blue
Example #6	Wide	Large	Fat	Blue
Example #7	Thin	Large	Fat	Blue
Example #8	Wide	Small	Fat	Blue

- What is the entropy of Class under Fins when Body=Medium?
- Answer:
- When Body=Medium, if Fins=Wide, then Class=Red; if Fins=Thin, then Class=Blue. Hence, there is no uncertainty.
- H(C|Fins, Body=Medium) = 0 bit

Example

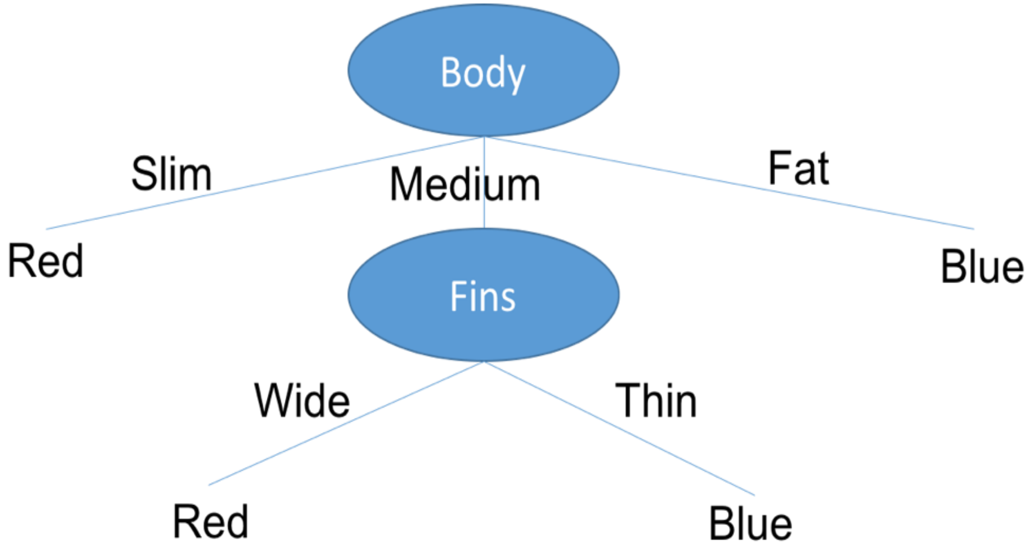
Example	Fins	Tail	Body	Class
Example #1	Thin	Small	Slim	Red
Example #2	Wide	Large	Slim	Red
Example #3	Thin	Large	Slim	Red
Example #4	Wide	Small	Medium	Red
Example #5	Thin	Small	Medium	Blue
Example #6	Wide	Large	Fat	Blue
Example #7	Thin	Large	Fat	Blue
Example #8	Wide	Small	Fat	Blue

- What is the entropy of Class under Tail when Body=Medium?
- Answer:
- When Body=Medium, Example #4 and Example #5 show that Tail=Small, and corresponding result is Class=Red and Class=Blue, respectively.
- Hence, H(C|Tail, Body=Medium) = $\frac{2}{2} \times H\left(\frac{1}{2}, \frac{1}{2}\right) = 1$

Example

- Draw the complete decision tree.
- Answer:

From the previous results, the first attribute to test is Body. If Body=Slim, then Class=Red; if Body=Fat, then Class=Blue; if Body=Medium, then we test attribute Fins, because the entropy of Class under Fins given that Body=Medium is 0.



Example

- Consider the following data set comprised of three binary input attributes (A_1, A_2, A_3), and one binary output:

Example	A_1	A_2	A_3	Output y
$\mathbf{x}_1$	1	0	0	0
$\mathbf{x}_2$	1	0	1	0
$\mathbf{x}_3$	0	1	0	0
$\mathbf{x}_4$	1	1	1	1
$\mathbf{x}_5$	1	1	0	1

- Learn a decision tree for these data.

Example

Example	A_1	A_2	A_3	Output y
$\mathbf{x}_1$	1	0	0	0
$\mathbf{x}_2$	1	0	1	0
$\mathbf{x}_3$	0	1	0	0
$\mathbf{x}_4$	1	1	1	1
$\mathbf{x}_5$	1	1	0	1

- Before testing any attribute, $H(y) = ?$
- $H(y) = H\left(\frac{2}{5}, \frac{3}{5}\right) = 0.971$ bits

Example

Example	A_1	A_2	A_3	Output y
$\mathbf{x}_1$	1	0	0	0
$\mathbf{x}_2$	1	0	1	0
$\mathbf{x}_3$	0	1	0	0
$\mathbf{x}_4$	1	1	1	1
$\mathbf{x}_5$	1	1	0	1

- What is the entropy of y under attribute A_1?
- $H(y|A_1) = \frac{4}{5} \times H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{5} \times H(0,1)$
- $= \frac{4}{5} \times 1 + 0 = 0.8$ bits

Example

Example	A_1	A_2	A_3	Output y
$\mathbf{x}_1$	1	0	0	0
$\mathbf{x}_2$	1	0	1	0
$\mathbf{x}_3$	0	1	0	0
$\mathbf{x}_4$	1	1	1	1
$\mathbf{x}_5$	1	1	0	1

- What is the entropy of y under attribute A_2?
- $H(y|A_2) = \frac{3}{5} \times H\left(\frac{2}{3}, \frac{1}{3}\right) + \frac{2}{5} \times H(0,1) =$ 0.6 ×0.918=0.551 bits

Example

Example	A_1	A_2	A_3	Output y
$\mathbf{x}_1$	1	0	0	0
$\mathbf{x}_2$	1	0	1	0
$\mathbf{x}_3$	0	1	0	0
$\mathbf{x}_4$	1	1	1	1
$\mathbf{x}_5$	1	1	0	1

- What is the entropy of y under attribute A_3?
- $H(y|A_3) = \frac{2}{5} \times H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{3}{5} \times H\left(\frac{1}{3}, \frac{2}{3}\right) = 0.4 + 0.6 \times 0.918 = 0.951$ bits

- Which attribute to test first?
- Test A_2 first!

Example

Example	A_1	A_2	A_3	Output y
$\mathbf{x}_1$	1	0	0	0
$\mathbf{x}_2$	1	0	1	0
$\mathbf{x}_3$	0	1	0	0
$\mathbf{x}_4$	1	1	1	1
$\mathbf{x}_5$	1	1	0	1

- Test A_2 first!
- If $A_2 = 0$, do you need to test another attribute?
- If $A_2 = 0$, Output $y = 0$, finished.
- If $A_2 = 1$, test A_1 or A_3?

Example

Example	A_1	A_2	A_3	Output y
$\mathbf{x}_1$	1	0	0	0
$\mathbf{x}_2$	1	0	1	0
$\mathbf{x}_3$	0	1	0	0
$\mathbf{x}_4$	1	1	1	1
$\mathbf{x}_5$	1	1	0	1

- If $A_2 = 1$, test A_1 or A_3?
- H(y|$A_1, A_2 = 1$)=?
- H(y|$A_1, A_2 = 1$)=$\frac{2}{3} \times H(1,0) + \frac{1}{3} \times H(0,1) = 0$ bit

Example

Example	A_1	A_2	A_3	Output y
$\mathbf{x}_1$	1	0	0	0
$\mathbf{x}_2$	1	0	1	0
$\mathbf{x}_3$	0	1	0	0
$\mathbf{x}_4$	1	1	1	1
$\mathbf{x}_5$	1	1	0	1

- If $A_2 = 1$, test A_1 or A_3?
- H(y|$A_3, A_2 = 1$)=?
- H(y|$A_3, A_2 = 1$)=$\frac{1}{3} \times H(1,0) + \frac{2}{3} \times H\left(\frac{1}{2}, \frac{1}{2}\right) > 0$ bit

Example

Example	A_1	A_2	A_3	Output y
$\mathbf{x}_1$	1	0	0	0
$\mathbf{x}_2$	1	0	1	0
$\mathbf{x}_3$	0	1	0	0
$\mathbf{x}_4$	1	1	1	1
$\mathbf{x}_5$	1	1	0	1

- If $A_2 = 1$, test A_1 or A_3?
- H(y|$A_1, A_2 = 1$)= 0 bit
- H(y|$A_3, A_2 = 1$)> 0 bit
- Hence, test A_1!

Example

- Draw the decision tree