

---

# Lecture 8 -- Decision Tree Learning (2)

Prof. Xiaowei Huang

<https://cgi.csc.liv.ac.uk/~xiaowei>

(Attendance Code: **465458**)



Decision Tree up to now,

Decision tree representation

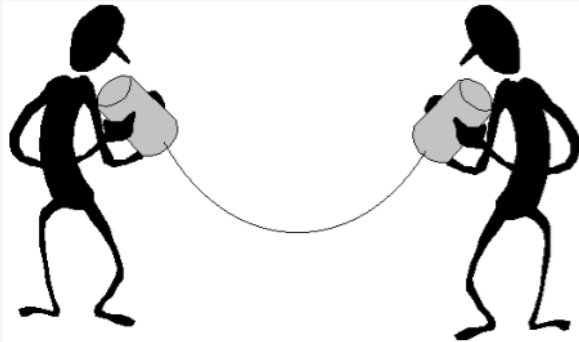
A general top-down algorithm

How to do splitting on numeric features

Occam's razor

# Today's Topics

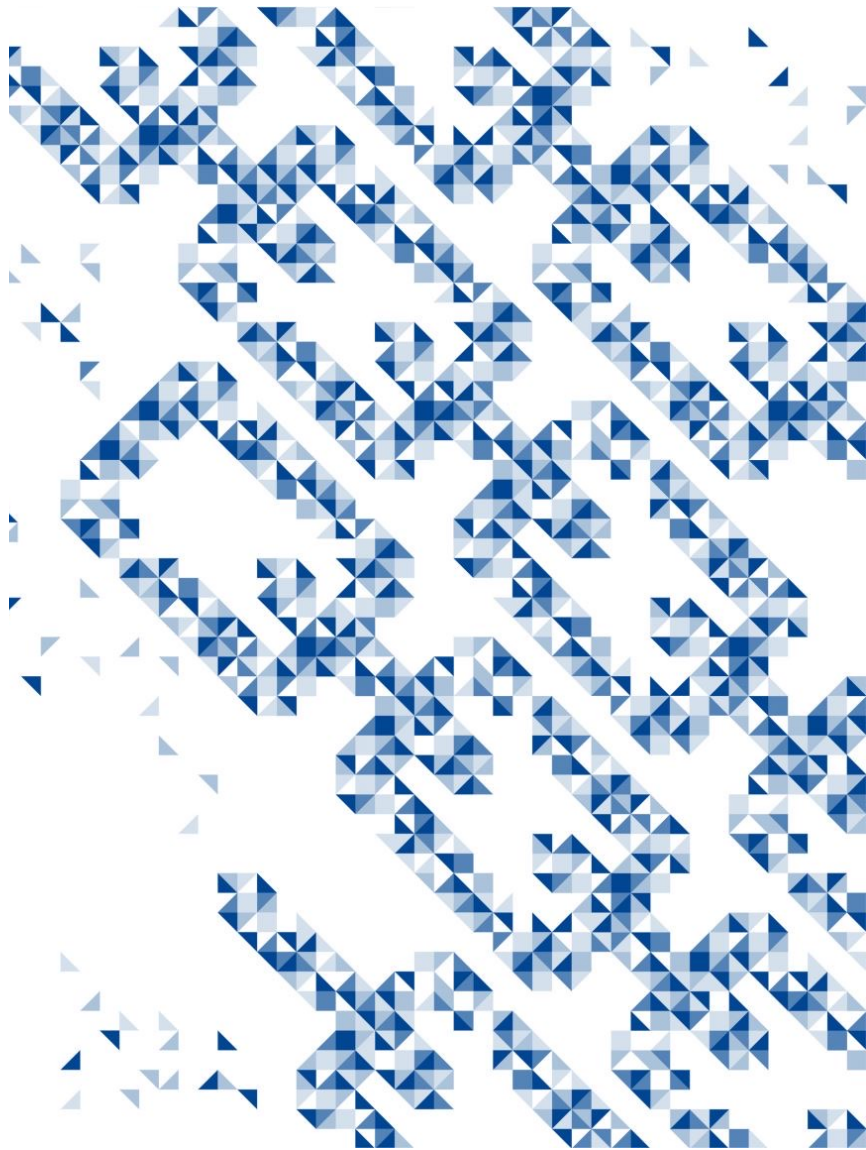
- Entropy, information gain, gain ratio -- how to determine features to split?



---

## Information theory background

- Consider a problem in which you are using a code to communicate information to a receiver
- Example: as bikes go past, you are communicating the manufacturer of each bike
- Question: What is the **most efficient (i.e., the minimum number of bits for any given information)** way of communication?



## Information theory background

- Suppose there are only four types of bikes
- We could use the following code

type	code
Trek	11
Specialized	10
Cervelo	01
Serrota	00

$$\frac{1}{4} \times 2 + \frac{1}{4} \times 2 + \frac{1}{4} \times 2 + \frac{1}{4} \times 2 = 2$$

- What is the **expected number of bits** we have to communicate?
  - 2 bits/bike



# Information theory background

- We can do better if the bike types aren't equiprobable

Type/probability	# bits	code
$P(\text{Trek}) = 0.5$	1	1
$P(\text{Specialized}) = 0.25$	2	01
$P(\text{Cervelo}) = 0.125$	3	001
$P(\text{Serrota}) = 0.125$	3	000

# Information theory background

Type/probability	# bits	code
$P(\text{Trek}) = 0.5$	1	1
$P(\text{Specialized}) = 0.25$	2	01
$P(\text{Cervelo}) = 0.125$	3	001
$P(\text{Serrota}) = 0.125$	3	000

- What is the expected number of bits we have to communicate?

$$0.5 \times 1 + 0.25 \times 2 + 0.125 \times 3 + 0.125 \times 3 = 1.75 < 2$$

# Information theory background

Type/probability	# bits	code
$P(\text{Trek}) = 0.5$	1	1
$P(\text{Specialized}) = 0.25$	2	01
$P(\text{Cervelo}) = 0.125$	3	001
$P(\text{Serrota}) = 0.125$	3	000

$$0.5 \times 1 + 0.25 \times 2 + 0.125 \times 3 + 0.125 \times 3 = 1.75 < 2$$

$$= 0.5 \times \log_2 0.5 + 0.25 \times \log_2 0.25 + 0.125 \times \log_2 0.125 + 0.125 \times \log_2 0.125$$

$$= - \sum_{y \in \text{values}(Y)} P(y) \log_2 P(y)$$



# Information theory background

$$- \sum_{y \in \text{values}(Y)} P(y) \log_2 P(y)$$

-- optimal code uses  $-\log_2 P(y)$  bits for event with probability  $P(y)$

# Entropy

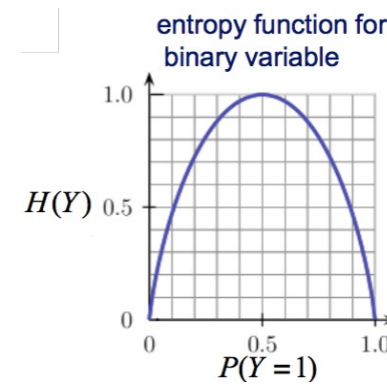
$$H(Y) = - \sum_{y \in \text{values}(Y)} P(y) \log_2 P(y)$$



entropy is a measure of uncertainty associated with a random variable

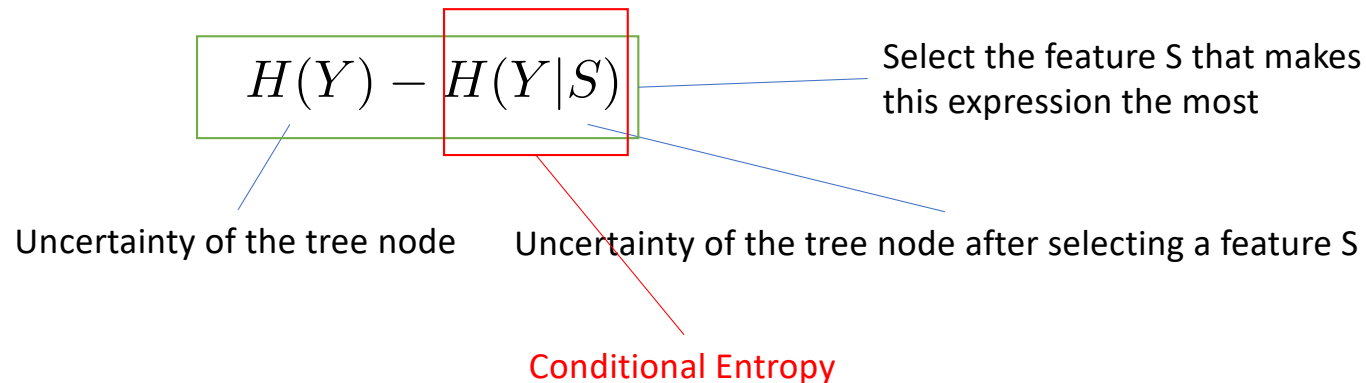


defined as the expected number of bits required to communicate the value of the variable



# General Idea of Selecting Feature for Splitting

- The training dataset has a given uncertainty.
- A set of instances with the same label has the least uncertainty 0.
- The construction of a shorter tree can be obtained by maximally reducing the uncertainty at every level of tree construction.



# Conditional entropy

- **Conditional entropy** (or equivocation) quantifies the amount of information needed to describe the outcome of a random variable given that the value of another random variable is known.
- What's the entropy of Y if we condition on some other variable X?

$$H(Y | X) = \sum_{x \in \text{values}(X)} P(X = x) H(Y | X = x)$$

similar as the expected value?

- where

$$H(Y | X = x) = - \sum_{y \in \text{values}(Y)} P(Y = y | X = x) \log_2 P(Y = y | X = x)$$

$$\text{Entropy: } H(Y) = - \sum_{y \in \text{values}(Y)} P(y) \log_2 P(y)$$

Need probability table

Similar as entropy

Need conditional probability table

Example

*PlayTennis: training examples*

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

---

# Example

- Let  $X = \text{Outlook}$  and  $Y = \text{PlayTennis}$
- Can you compute  $H(Y|X)$ ?

$$\begin{aligned} H(Y|X) = & P(X = \text{Sunny})H(Y|X = \text{Sunny}) & + \\ & P(X = \text{Overcast})H(Y|X = \text{Overcast}) & + \\ & P(X = \text{Rain})H(Y|X = \text{Rain}) \end{aligned}$$





# Example

$$H(Y|X) = P(X = \text{Sunny})H(Y|X = \text{Sunny}) + P(X = \text{Overcast})H(Y|X = \text{Overcast}) + P(X = \text{Rain})H(Y|X = \text{Rain})$$

$$H(Y|X = x) = - \sum_{y \in \text{values}(Y)} P(Y = y|X = x) \log_2 P(Y = y|X = x)$$

X = Outlook and Y = PlayTennis

*PlayTennis: training examples*

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



		Y	
		Yes	No
X	Sunny	2/14	3/14
	Overcast	4/14	0
	Rain	3/14	2/14

Need conditional probability table

## Example

- Let  $X = \text{Outlook}$  and  $Y = \text{PlayTennis}$
- Can you compute  $H(Y|X)$ ?

		Y	
		Yes	No
X	Sunny	2/14	3/14
	Overcast	4/14	0
	Rain	3/14	2/14

$$H(Y|X = \text{Sunny}) = -2/5 \log 2/5 - 3/5 \log 3/5$$

$$H(Y|X = \text{Overcast}) = 0$$

$$H(Y|X = \text{Rain}) = -3/5 \log 3/5 - 2/5 \log 2/5$$

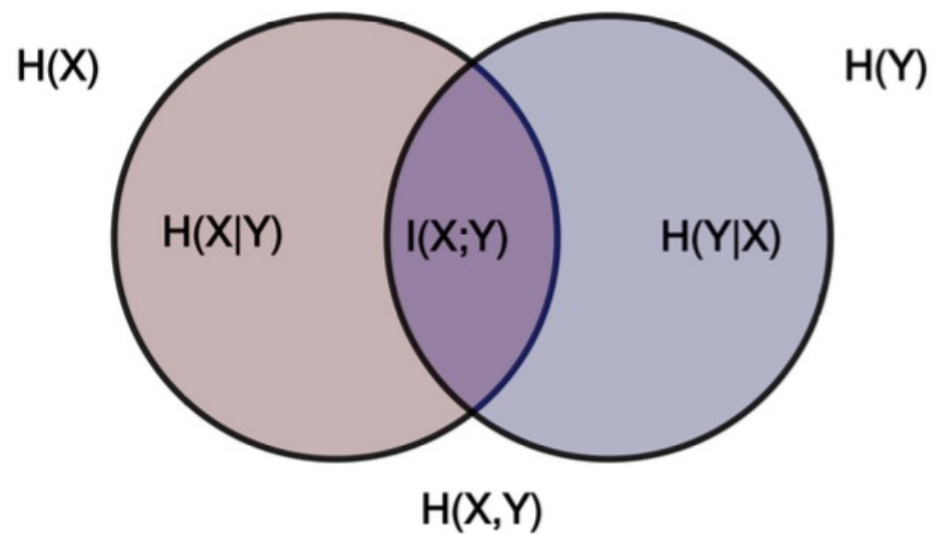
## Information gain (a.k.a. mutual information)

- Choosing splits in ID3: select the split  $S$  that **most reduces the conditional entropy** of  $Y$  for training set  $D$

$$\text{InfoGain}(D, S) = H_D(Y) - H_D(Y | S)$$

$D$  indicates that we're calculating probabilities using the specific sample  $D$

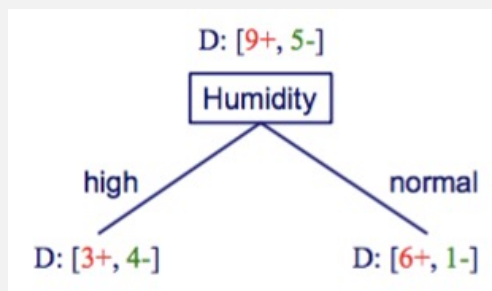
## Relations between the concepts



- [https://en.wikipedia.org/wiki/Mutual\\_information](https://en.wikipedia.org/wiki/Mutual_information)

# Information gain example

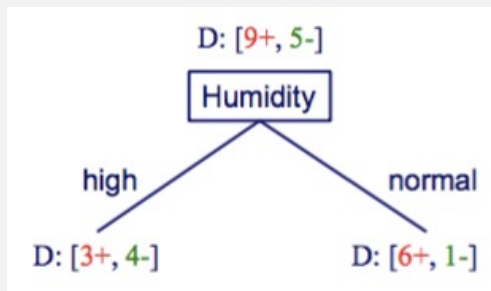
- What's the information gain of splitting on Humidity?



$$\text{InfoGain}(D, \text{Humidity}) = H_D(Y) - H_D(Y | \text{Humidity})$$

$$\text{InfoGain}(D, S) = H_D(Y) - H_D(Y | S)$$

# Information gain example

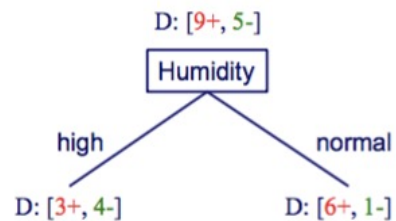


$$H_D(Y) = -\frac{9}{14} \log_2 \left( \frac{9}{14} \right) - \frac{5}{14} \log_2 \left( \frac{5}{14} \right) = 0.940$$

$$H(Y) = - \sum_{y \in \text{values}(Y)} P(y) \log_2 P(y)$$



# Information gain example



$$H_D(Y | \text{Humidity}) = P(\text{Humidity}=\text{high})H_D(Y | \text{Humidity}=\text{high}) + P(\text{Humidity}=\text{normal})H_D(Y | \text{Humidity}=\text{normal})$$

$$H(Y | X) = \sum_{x \in \text{values}(X)} P(X = x) H(Y | X = x)$$

$$H_D(Y | \text{high}) = -\frac{3}{7} \log_2 \left( \frac{3}{7} \right) - \frac{4}{7} \log_2 \left( \frac{4}{7} \right) = 0.985$$

$$H_D(Y | \text{normal}) = -\frac{6}{7} \log_2 \left( \frac{6}{7} \right) - \frac{1}{7} \log_2 \left( \frac{1}{7} \right) = 0.592$$

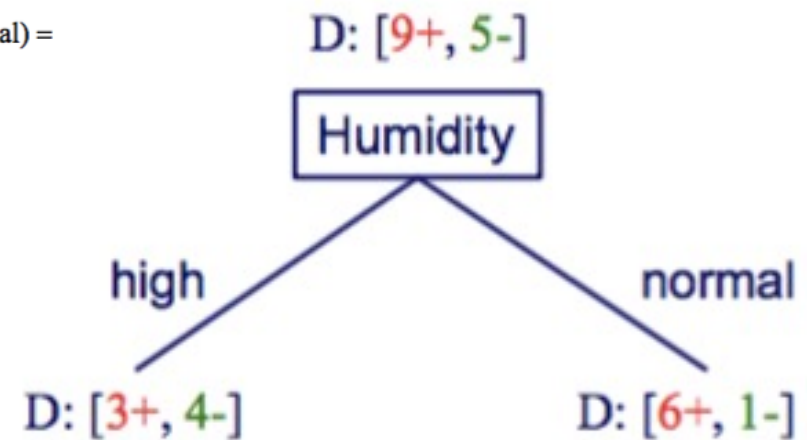
$$H(Y | X = x) = - \sum_{y \in \text{values}(Y)} P(Y = y | X = x) \log_2 P(Y = y | X = x)$$

# Information gain example

$$\begin{aligned}\text{InfoGain}(D, \text{Humidity}) &= H_D(Y) - H_D(Y | \text{Humidity}) \\ &= 0.940 - \left[ \frac{7}{14}(0.985) + \frac{7}{14}(0.592) \right] \\ &= 0.151\end{aligned}$$

$H_D(Y | \text{high}) =$


$H_D(Y | \text{normal}) =$



Where do these coefficients come from?

# Information gain example

- Is it better to split on Humidity or Wind?



The image shows two decision tree diagrams. The left diagram is for a split on 'Humidity'. The root node is 'Humidity' with a dataset label 'D: [9+, 5-]'. It has two branches: 'high' leading to a leaf node 'D: [3+, 4-]' and 'normal' leading to a leaf node 'D: [6+, 1-]'. The right diagram is for a split on 'Wind'. The root node is 'Wind' with a dataset label 'D: [9+, 5-]'. It has two branches: 'weak' leading to a leaf node 'D: [6+, 2-]' and 'strong' leading to a leaf node 'D: [3+, 3-]'. Below the 'Wind' diagram, the conditional entropies are given:  $H_D(Y | \text{weak}) = 0.811$  and  $H_D(Y | \text{strong}) = 1.0$ .

✓  $\text{InfoGain}(D, \text{Humidity}) = 0.940 - \left[ \frac{7}{14}(0.985) + \frac{7}{14}(0.592) \right]$   
 $= 0.151$

$\text{InfoGain}(D, \text{Wind}) = 0.940 - \left[ \frac{8}{14}(0.811) + \frac{6}{14}(1.0) \right]$   
 $= 0.048$

# One limitation of information gain

- 
- information gain is biased towards tests with many outcomes
  - e.g. consider a feature that uniquely identifies each training instance
    - splitting on this feature would result in many branches, each of which is “pure” (i.e., has instances of only one class,  $\text{entropy}=0$ )
    - maximal information gain!

# Gain ratio

- 
- to address this limitation, C4.5 uses a splitting criterion called *gain ratio*
  - gain ratio normalizes the information gain by the entropy of the split being considered

$$\text{GainRatio}(D, S) = \frac{\text{InfoGain}(D, S)}{H_D(S)} = \frac{H_D(Y) - H_D(Y | S)}{H_D(S)}$$

# Exercise

[Check the exercises of the lecture notes  
for answer](#)

- Compute the following:

$$\textit{GainRatio}(D, \textit{Humidity}) =$$

$$\textit{GainRatio}(D, \textit{Wind}) =$$

$$\textit{GainRatio}(D, \textit{Outlook}) =$$