Back to decision trees

- Which attribute do we choose at each level?
- The one with the highest information gain
  - The one that reduces the unpredictability the most

outlook	temperature	humidity	windy	play
overcast	cool	normal	TRUE	yes
overcast	hot	high	FALSE	yes
overcast	hot	normal	FALSE	yes
overcast	mild	high	TRUE	yes
rainy	cool	normal	TRUE	no
rainy	mild	high	TRUE	no
rainy	cool	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Before: 14 records, 9 are “yes”

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

If we choose **outlook**:

overcast : 4 records, 4 are “yes”

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

rainy : 5 records, 3 are “yes”

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

sunny : 5 records, 2 are “yes”

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

Expected new entropy:

$$\frac{4}{14}\times 0.0+\frac{5}{14}\times 0.97+\frac{5}{14}\times 0.97$$

= <u>0.69</u>

outlook	temperature	humidity	windy	play
overcast	cool	normal	TRUE	yes
overcast	hot	high	FALSE	yes
overcast	hot	normal	FALSE	yes
overcast	mild	high	TRUE	yes
rainy	cool	normal	TRUE	no
rainy	mild	high	TRUE	no
rainy	cool	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Before: 14 records, 9 are “yes”

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

If we choose **temperature**:

cool : 4 records, 3 are “yes”

0.81

rainy : 4 records, 2 are “yes”

1.0

sunny : 6 records, 4 are “yes”

0.92

Expected new entropy:

0.81(4/14) + 1.0(4/14) + 0.92(6/14)

= <u>0.91</u>

outlook	temperature	humidity	windy	play
overcast	cool	normal	TRUE	yes
overcast	hot	high	FALSE	yes
overcast	hot	normal	FALSE	yes
overcast	mild	high	TRUE	yes
rainy	cool	normal	TRUE	no
rainy	mild	high	TRUE	no
rainy	cool	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Before: 14 records, 9 are “yes”

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

If we choose **humidity**:

normal : 7 records, 6 are “yes”

0.59

high : 7 records, 2 are “yes”

0.86

Expected new entropy:

0.59(7/14) + 0.86(7/14)

= <u>0.725</u>

outlook	temperature	humidity	windy	play
overcast	cool	normal	TRUE	yes
overcast	hot	high	FALSE	yes
overcast	hot	normal	FALSE	yes
overcast	mild	high	TRUE	yes
rainy	cool	normal	TRUE	no
rainy	mild	high	TRUE	no
rainy	cool	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Before: 14 records, 9 are “yes”

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

If we choose **windy**:

TRUE : 8 records, 6 are “yes”

0.81

FALSE : 5 records, 3 are “yes”

0.97

Expected new entropy:

0.81(8/14) + 0.97(6/14)

= 0.87

outlook	temperature	humidity	windy	play
overcast	cool	normal	TRUE	yes
overcast	hot	high	FALSE	yes
overcast	hot	normal	FALSE	yes
overcast	mild	high	TRUE	yes
rainy	cool	normal	TRUE	no
rainy	mild	high	TRUE	no
rainy	cool	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Before: 14 records, 9 are “yes”

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

outlook

0.94 – 0.69 = 0.25 highest gain

temperature

0.94 – 0.91 = 0.03

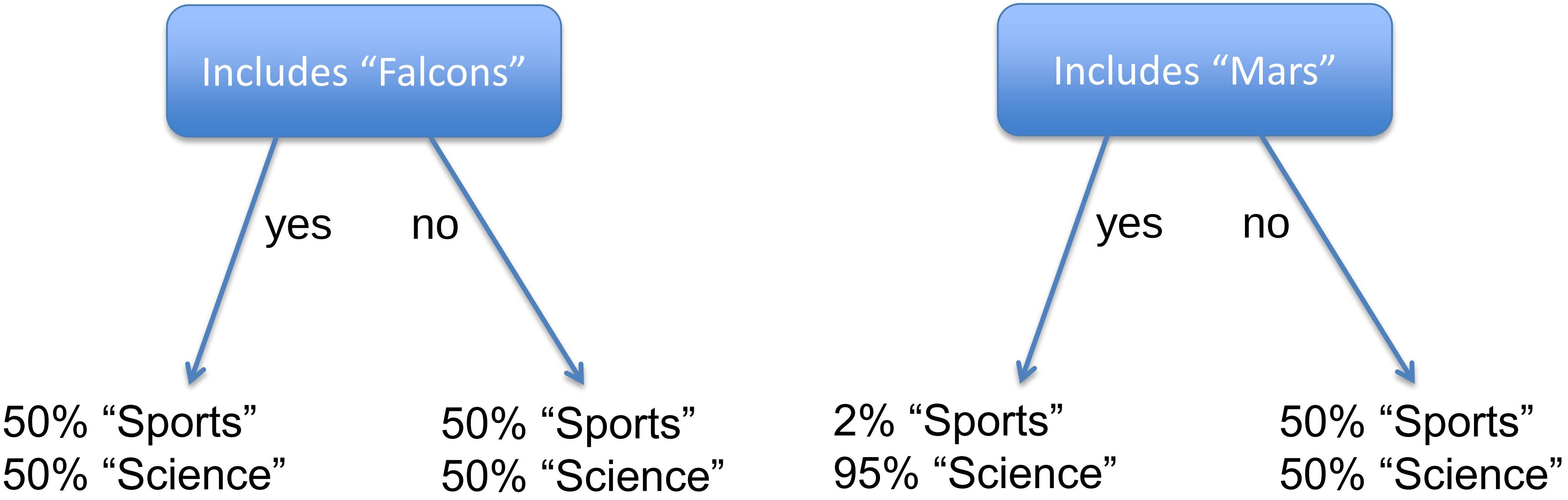
humidity

0.94 – 0.725 = 0.215

windy

0.94 – 0.87 = 0.07

Document Classification



Includes “Falcons”

yes

no

50% “Sports”
50% “Science”

50% “Sports”
50% “Science”

Includes “Mars”

yes

no

2% “Sports”
95% “Science”

50% “Sports”
50% “Science”

Building a Decision Tree (ID3 Algorithm)

- Assume attributes are discrete
  - Discretize continuous attributes
- Choose the attribute with the highest Information Gain
- Create branches for each value of attribute
- Examples partitioned based on selected attributes
- Repeat with remaining attributes
- Stopping conditions
  - All examples assigned the same label
  - No examples left

Problems

- Expensive to train
- Prone to overfitting
  - – Drive to perfection on training data, bad on test data
  - – Pruning can help: remove or aggregate subtrees that provide little discriminatory power (C45)

C4.5 Extensions

- Continuous Attributes

outlook	temperature	humidity	windy	play
overcast	cool	60	TRUE	yes
overcast	hot	80	FALSE	yes
overcast	hot	63	FALSE	yes
overcast	mild	81	TRUE	yes
rainy	cool	58	TRUE	no
rainy	mild	90	TRUE	no
rainy	cool	54	FALSE	yes
rainy	mild	92	FALSE	yes
rainy	mild	59	FALSE	yes
sunny	hot	90	FALSE	no
sunny	hot	89	TRUE	no
sunny	mild	90	FALSE	no
sunny	cool	60	FALSE	yes
sunny	mild	62	TRUE	yes

Consider every possible binary partition; choose the partition with the highest gain

outlook	temperature	humidity	windy	play
rainy	mild	54	FALSE	yes
overcast	hot	58	FALSE	yes
overcast	cool	59	TRUE	yes
rainy	cool	60	FALSE	yes
overcast	mild	60	TRUE	yes
overcast	hot	62	FALSE	yes
rainy	mild	63	TRUE	no
sunny	cool	80	FALSE	yes
rainy	mild	81	FALSE	yes
sunny	mild	89	TRUE	yes
sunny	hot	90	FALSE	no
rainy	cool	90	TRUE	no
sunny	hot	90	TRUE	no
sunny	mild	92	FALSE	no

E(6/6)

= 0.0

E(3/8) + E(5/8)

= 0.95

E(9/10)+ E(1/10)

= 0.47

E(4/4)

= 0.0

Expect = 8/14*0.95 + 6/14*0
= 0.54

Expect = 10/14*0.47 + 4/14*0
= 0.33