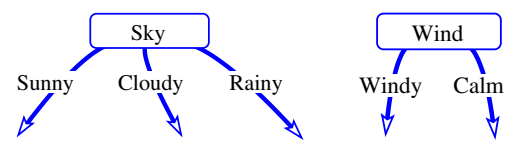


# Decision Trees

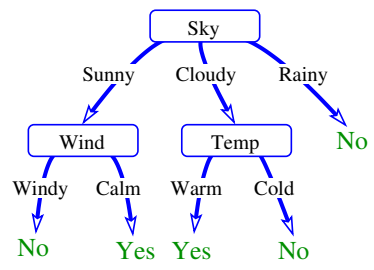
- 1 Decision Trees
  - Using Trees
  - Learning
- 2 Unpredictability
  - Entropy
  - Entropy for datasets
  - Information Gain
- 3 Bias
  - Bias
  - Occam's principle
  - Overfitting
- 4 Improvements

- 1 Decision Trees
  - Using Trees
  - Learning
- 2 Unpredictability
  - Entropy
  - Entropy for datasets
  - Information Gain
- 3 Bias
  - Bias
  - Occam's principle
  - Overfitting
- 4 Improvements

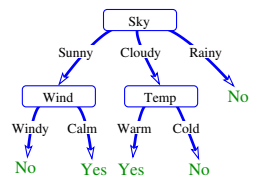
Basic Idea: Test the attributes sequentially



The whole analysis strategy can be seen as a tree.



The results (classifications) are coded by the leaves



What does the tree encode?

$$(Sunny \wedge Calm) \vee (Cloudy \wedge Warm)$$

Works as a *disjunction of conjunctions*

Normal Form for boolean functions  
 Arbitrary boolean functions can be represented!

Decision Trees ● Unpredictability ○○○○○○○○ Bias ○○○○ Improvements

Learning

How can a decision tree be constructed automatically?

- 1 Choose an attribute to test
- 2 Branches with a unique class become leaves
- 3 Other branches are extended recursively

Remaining question: how do we choose attributes?

Greedy approach:  
Choose the attribute which *tells us most* about the answer

Decision Trees ○○○○ Unpredictability ●○○○○○○○ Bias ○○○○ Improvements

- 1 Decision Trees
  - Using Trees
  - Learning
- 2 Unpredictability
  - Entropy
  - Entropy for datasets
  - Information Gain
- 3 Bias
  - Bias
  - Occam's principle
  - Overfitting
- 4 Improvements

Decision Trees ○○○○ Unpredictability ●○○○○○○○ Bias ○○○○ Improvements

Entropy

Entropy

Entropy — measure of **unpredictability**

$$\text{Entropy} = \sum_i -p_i \log_2 p_i$$

$p_i$  probability for event  $i$

Decision Trees ○○○○ Unpredictability ●○○○○○○○ Bias ○○○○ Improvements

Entropy

Entropy

Example: tossing a coin  
 $p_{\text{head}} = 0.5; \quad p_{\text{tail}} = 0.5$

$$\begin{aligned} \text{Entropy} &= \sum_i -p_i \log_2 p_i = \\ &= -0.5 \log_2 0.5 + -0.5 \log_2 0.5 = -0.5 \underbrace{\log_2 0.5}_{-1} + -0.5 \underbrace{\log_2 0.5}_{-1} \\ &= 1 \end{aligned}$$

The result of a coin-toss has **1 bit** of information

Decision Trees ○○○○ Unpredictability ●○○○○○○○ Bias ○○○○ Improvements

Entropy

Entropy

Example: rolling a dice  
 $p_1 = \frac{1}{6}; \quad p_2 = \frac{1}{6}; \dots \quad p_6 = \frac{1}{6}$

$$\begin{aligned} \text{Entropy} &= \sum_i -p_i \log_2 p_i = \\ &= 6 \times -\frac{1}{6} \log_2 \frac{1}{6} = \\ &= -\log_2 \frac{1}{6} = \log_2 6 \approx 2.58 \end{aligned}$$

The result of a dice-roll has **2.58 bit** of information

Decision Trees ○○○○ Unpredictability ●○○○○○○○ Bias ○○○○ Improvements

Entropy

Entropy

Example: rolling a **fake dice**  
 $p_1 = 0.1; \dots \quad p_5 = 0.1; \quad p_6 = 0.5$

$$\begin{aligned} \text{Entropy} &= \sum_i -p_i \log_2 p_i = \\ &= -5 \cdot 0.1 \log_2 0.1 - 0.5 \log_2 0.5 = \\ &\approx 2.16 \end{aligned}$$

A real dice is **more unpredictable** (2.58 bit) than a fake (2.16 bit)



Decision Trees 0000 Unpredictability 00000000 Bias 0000 Improvements

- Decision Trees
  - Using Trees
  - Learning
- Unpredictability
  - Entropy
  - Entropy for datasets
  - Information Gain
- Bias
  - Bias
  - Occam's principle
  - Overfitting
- Improvements

Decision Trees 0000 Unpredictability 00000000 Bias 0000 Improvements

Bias

Which Bias does this learning algorithm have?

- Restriction Bias?**  
No, all hypotheses can be represented
- Preference Bias?**  
Yes, some trees are found before others

Which hypotheses (here: trees) are preferred?

- Shallow trees
- "Important attributes" early

Decision Trees 0000 Unpredictability 00000000 Bias 0000 Improvements


Occam's principle

Which hypothesis should be preferred when several are compatible with the data?

**Occam's principle** (*Occam's razor*, "Occam's rakkniv")

William from Ockham, Theologian and Philosopher (1288–1348)

*"Entia non sunt multiplicanda praeter necessitatem"*



translated:

Decision Trees 0000 Unpredictability 00000000 Bias 0000 Improvements

Occam's principle

Why are simple hypotheses more likely to be correct?

It is more likely that the reality from which the examples come have a simple generating mechanism.

Simple hypotheses tends to generalize better.

Decision Trees 0000 Unpredictability 00000000 Bias 0000 Improvements

Overfitting

Overfitting, (*överträning*)

When the hypotheses are overly specialized for the available training examples.

Good results on training data, but generalizes badly

When does this occur?

- Non-representative sample
- Noisy examples

What can be done about it?

Choose a simpler hypothesis and accept some errors for the training examples

Decision Trees 0000 Unpredictability 00000000 Bias 0000 Improvements

Possible ways of improving the decision trees

- Avoid overfitting
  - Limit the tree's height
  - Pruning (*Beskärning*)
- Attributes with graded values
- Missing attribute values
- Variable cost for different attributes