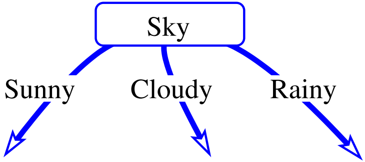
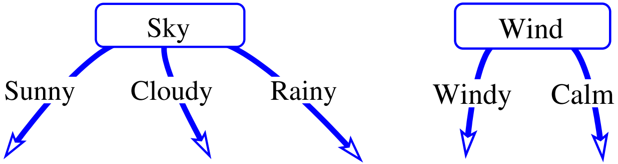
Decision Trees

Basic Idea: Test the attributes sequentially

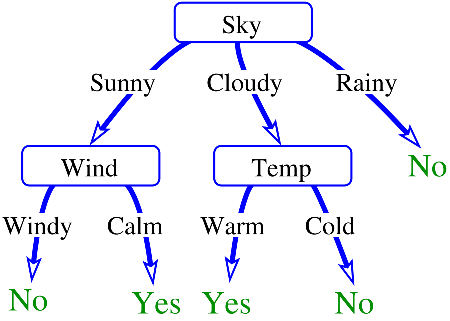
Basic Idea: Test the attributes sequentially



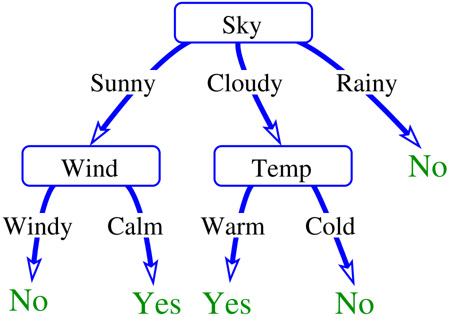
Basic Idea: Test the attributes sequentially



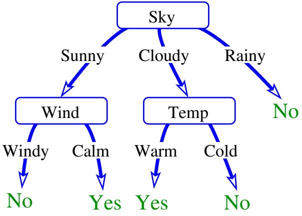
The whole analysis strategy can be seen as a tree.



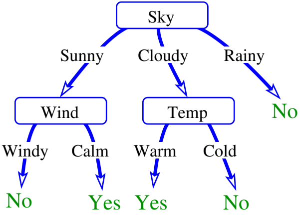
The whole analysis strategy can be seen as a tree.



The results (classifications) are coded by the *leaves*

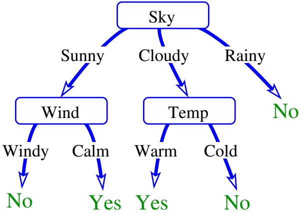


What does the tree encode?



What does the tree encode?

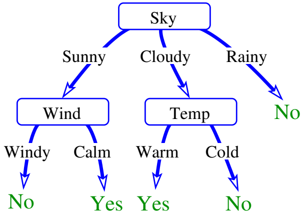
$$(\text{Sunny} \wedge \text{Calm}) \vee (\text{Cloudy} \wedge \text{Warm})$$



What does the tree encode?

$$(\text{Sunny} \wedge \text{Calm}) \vee (\text{Cloudy} \wedge \text{Warm})$$

Works as a *disjunction of conjunctions*



What does the tree encode?

$$(\text{Sunny} \wedge \text{Calm}) \vee (\text{Cloudy} \wedge \text{Warm})$$

Works as a *disjunction of conjunctions*

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

How can a decision tree be constructed automatically?

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

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?

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

Entropy

Entropy — measure of unpredictability

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

p_i probability for event i

Entropy

Example: tossing a coin

Entropy

Example: tossing a coin

$p_{\text{head}} = 0.5; \qquad p_{\text{tail}} = 0.5$

Entropy

Example: tossing a coin

$p_{\text{head}} = 0.5; \qquad p_{\text{tail}} = 0.5$

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

Entropy

Example: tossing a coin

$p_{\text{head}} = 0.5; \qquad p_{\text{tail}} = 0.5$

$$\text{Entropy} = \sum_i -p_i \log_2 p_i =$$
$$= -0.5 \log_2 0.5 + -0.5 \log_2 0.5$$

Entropy

Example: tossing a coin

$p_{\text{head}} = 0.5; \qquad p_{\text{tail}} = 0.5$

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

Entropy

Example: tossing a coin

$p_{\text{head}} = 0.5; \qquad p_{\text{tail}} = 0.5$

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

Entropy

Example: tossing a coin

$p_{\text{head}} = 0.5; \qquad p_{\text{tail}} = 0.5$

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

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

Entropy

Example: rolling a dice

Entropy

Example: rolling a dice

$p_1 = \frac{1}{6}; \quad p_2 = \frac{1}{6}; \ldots \quad p_6 = \frac{1}{6}$

Entropy

Example: rolling a dice

$p_1 = \frac{1}{6}; \quad p_2 = \frac{1}{6}; \ldots \quad p_6 = \frac{1}{6}$

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

Entropy

Example: rolling a dice

$p_1 = \frac{1}{6}; \quad p_2 = \frac{1}{6}; \ldots \quad p_6 = \frac{1}{6}$

$$\text{Entropy} = \sum_i -p_i \log_2 p_i = $$
$$= 6 \times -\frac{1}{6} \log_2 \frac{1}{6}$$

Entropy

Example: rolling a dice

$p_1 = \frac{1}{6}; \quad p_2 = \frac{1}{6}; \ldots \quad p_6 = \frac{1}{6}$

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

Entropy

Example: rolling a dice

$p_1 = \frac{1}{6}; \quad p_2 = \frac{1}{6}; \ldots \quad p_6 = \frac{1}{6}$

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

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

Entropi

Example: rolling a fake dice

Entropi

Example: rolling a fake dice

$p_1 = 0.1; \ldots \quad p_5 = 0.1; \quad p_6 = 0.5$

Entropi

Example: rolling a fake dice

$p_1 = 0.1; \ldots \quad p_5 = 0.1; \quad p_6 = 0.5$

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

Entropi

Example: rolling a fake dice

$p_1 = 0.1; \ldots \quad p_5 = 0.1; \quad p_6 = 0.5$

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

$$= -5 \cdot 0.1 \log_2 0.1 - 0.5 \log_2 0.5$$

Entropi

Example: rolling a fake dice

$p_1 = 0.1; \ldots \quad p_5 = 0.1; \quad p_6 = 0.5$

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

Entropi

Example: rolling a fake dice

$p_1 = 0.1; \ldots \quad p_5 = 0.1; \quad p_6 = 0.5$

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

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

Entropy

Unpredictability of a dataset

Entropy

Unpredictability of a dataset

- 100 examples, 42 positive

Entropy

Unpredictability of a dataset

- 100 examples, 42 positive

$$-\frac{58}{100}\log_2\frac{58}{100}-\frac{42}{100}\log_2\frac{42}{100}=0.981$$

Entropy

Unpredictability of a dataset

- 100 examples, 42 positive

$$-\frac{58}{100}\log_2\frac{58}{100}-\frac{42}{100}\log_2\frac{42}{100}=0.981$$

- 100 examples, 3 positive

Entropy

Unpredictability of a dataset

- 100 examples, 42 positive

$$-\frac{58}{100}\log_2\frac{58}{100}-\frac{42}{100}\log_2\frac{42}{100}=0.981$$

- 100 examples, 3 positive

$$-\frac{97}{100}\log_2\frac{97}{100}-\frac{3}{100}\log_2\frac{3}{100}=0.194$$

Back to the decision trees

Back to the decision trees

Smart idea:

Ask about the attribute which maximizes the expected reduction of the entropy.

Back to the decision trees

Smart idea:

Ask about the attribute which maximizes the expected reduction of the entropy.

Information Gain

Back to the decision trees

Smart idea:

Ask about the attribute which maximizes the expected reduction of the entropy.

Information Gain

Assume that we ask about attribute A for a dataset S

Back to the decision trees

Smart idea:

Ask about the attribute which maximizes the expected reduction of the entropy.

Information Gain

Assume that we ask about attribute A for a dataset S

$$\text{Gain} = \text{Ent}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Ent}(S_v)$$

Back to the decision trees

Smart idea:

Ask about the attribute which maximizes the expected reduction of the entropy.

Information Gain

Assume that we ask about attribute A for a dataset S

$$\text{Gain} = \underbrace{\text{Ent}(S)}_{\text{before}} - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Ent}(S_v)$$

Back to the decision trees

Smart idea:

Ask about the attribute which maximizes the expected reduction of the entropy.

Information Gain

Assume that we ask about attribute A for a dataset S

$$\text{Gain} = \underbrace{\text{Ent}(S)}_{\text{before}} - \underbrace{\sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|}}_{\substack{\text{weighted} \\ \text{average}}} \text{Ent}(S_v)$$

Back to the decision trees

Smart idea:

Ask about the attribute which maximizes the expected reduction of the entropy.

Information Gain

Assume that we ask about attribute A for a dataset S

$$\text{Gain} = \underbrace{\text{Ent}(S)}_{\text{before}} - \underbrace{\sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|}}_{\substack{\text{weighted}\\ \text{average}}} \underbrace{\text{Ent}(S_v)}_{\text{after}}$$

What is the entropy for this dataset?

A	B	C	D	
●	●	○	○	+
○	●	●	○	+
○	○	○	○	
●	○	○	●	+
○	●	○	○	+
●	○	●	○	
●	●	○	○	+
○	○	○	○	
○	○	●	○	
●	●	○	○	+
○	○	○	●	+
●	○	○	○	
●	●	●	○	+
○	●	○	●	
○	○	○	○	
●	○	○	○	
●	●	○	●	
○	●	○	○	+
○	○	●	○	
●	○	○	○	
○	●	○	○	+
●	○	○	●	+
○	●	●	○	+
○	○	○	○	
●	○	○	○	

What is the entropy for this dataset?

$$\text{Ent} = -\frac{12}{25}\log_2\frac{12}{25} - \frac{13}{25}\log_2\frac{13}{25} \approx \mathbf{0.9988}$$

A	B	C	D	
●	●	○	○	+
○	●	●	○	+
○	○	○	○	
●	○	○	●	+
○	●	○	○	+
●	○	●	○	
●	●	○	○	+
○	○	○	○	
○	○	●	○	
●	●	○	○	+
○	○	○	●	+
●	○	○	○	
●	●	●	○	+
○	●	○	●	
○	○	○	○	
●	○	○	○	
●	●	○	●	
○	●	○	○	+
○	○	●	○	
●	○	○	○	
○	●	○	○	+
●	○	○	●	+
○	●	●	○	+
○	○	○	○	
●	○	○	○	

What is the entropy for this dataset?

$$\text{Ent} = -\frac{12}{25}\log_2\frac{12}{25} - \frac{13}{25}\log_2\frac{13}{25} \approx \mathbf{0.9988}$$

$A = \bullet$: $\frac{6}{12}$ positive $\rightarrow 1.0$

$A = \circ$: $\frac{6}{13}$ positive $\rightarrow 0.9957$

Expected: $\frac{12}{25} \cdot 1.0 + \frac{13}{25} \cdot 0.9957 \approx \mathbf{0.9977}$

A	B	C	D	
$\bullet$	$\bullet$	$\circ$	$\circ$	+
$\circ$	$\bullet$	$\bullet$	$\circ$	+
$\circ$	$\circ$	$\circ$	$\circ$	
$\bullet$	$\circ$	$\circ$	$\bullet$	+
$\circ$	$\bullet$	$\circ$	$\circ$	+
$\bullet$	$\circ$	$\bullet$	$\circ$	
$\bullet$	$\bullet$	$\circ$	$\circ$	+
$\circ$	$\circ$	$\circ$	$\circ$	
$\circ$	$\circ$	$\bullet$	$\circ$	
$\bullet$	$\bullet$	$\circ$	$\circ$	+
$\circ$	$\circ$	$\circ$	$\bullet$	+
$\bullet$	$\circ$	$\circ$	$\circ$	
$\bullet$	$\bullet$	$\bullet$	$\circ$	+
$\circ$	$\bullet$	$\circ$	$\bullet$	
$\circ$	$\circ$	$\circ$	$\circ$	
$\bullet$	$\circ$	$\circ$	$\circ$	
$\bullet$	$\bullet$	$\circ$	$\bullet$	
$\circ$	$\bullet$	$\circ$	$\circ$	+
$\circ$	$\circ$	$\bullet$	$\circ$	
$\bullet$	$\circ$	$\circ$	$\circ$	
$\circ$	$\bullet$	$\circ$	$\circ$	+
$\bullet$	$\circ$	$\circ$	$\bullet$	+
$\circ$	$\bullet$	$\bullet$	$\circ$	+
$\circ$	$\circ$	$\circ$	$\circ$	
$\bullet$	$\circ$	$\circ$	$\circ$	

What is the entropy for this dataset?

$$\text{Ent} = -\frac{12}{25}\log_2\frac{12}{25} - \frac{13}{25}\log_2\frac{13}{25} \approx \mathbf{0.9988}$$

$A = \bullet$: $\frac{6}{12}$ positive $\rightarrow 1.0$

$A = \circ$: $\frac{6}{13}$ positive $\rightarrow 0.9957$

Expected: $\frac{12}{25} \cdot 1.0 + \frac{13}{25} \cdot 0.9957 \approx \mathbf{0.9977}$

$B = \bullet$: $\frac{9}{11}$ positive $\rightarrow 0.684$

$B = \circ$: $\frac{3}{14}$ positive $\rightarrow 0.750$

Expected: **0.721**

A	B	C	D	
●	●	○	○	+
○	●	●	○	+
○	○	○	○	
●	○	○	●	+
○	●	○	○	+
●	○	●	○	
●	●	○	○	+
○	○	○	○	
○	○	●	○	
●	●	○	○	+
○	○	○	●	+
●	○	○	○	
●	●	●	○	+
○	●	○	●	
○	○	○	○	
●	○	○	○	
●	●	○	●	
○	●	○	○	+
○	○	●	○	
●	○	○	○	
○	●	○	○	+
●	○	○	●	+
○	●	●	○	+
○	○	○	○	
●	○	○	○	

What is the entropy for this dataset?

$$\text{Ent} = -\frac{12}{25}\log_2\frac{12}{25} - \frac{13}{25}\log_2\frac{13}{25} \approx \mathbf{0.9988}$$

$A = \bullet$: $\frac{6}{12}$ positive $\rightarrow 1.0$
$A = \circ$: $\frac{6}{13}$ positive $\rightarrow 0.9957$
Expected: $\frac{12}{25}\cdot 1.0 + \frac{13}{25}\cdot 0.9957 \approx \mathbf{0.9977}$

$B = \bullet$: $\frac{9}{11}$ positive $\rightarrow 0.684$
$B = \circ$: $\frac{3}{14}$ positive $\rightarrow 0.750$
Expected: **0.721**

$C = \bullet$: $\frac{3}{6}$ positive $\rightarrow 1.0$
$C = \circ$: $\frac{9}{19}$ positive $\rightarrow 0.9980$
Expected: **0.9985**

A	B	C	D	
●	●	○	○	+
○	●	●	○	+
○	○	○	○	
●	○	○	●	+
○	●	○	○	+
●	○	●	○	
●	●	○	○	+
○	○	○	○	
○	○	●	○	
●	●	○	○	+
○	○	○	●	+
●	○	○	○	
●	●	●	○	+
○	●	○	●	
○	○	○	○	
●	○	○	○	
●	●	○	●	
○	●	○	○	+
○	○	●	○	
●	○	○	○	
○	●	○	○	+
●	○	○	●	+
○	●	●	○	+
○	○	○	○	
●	○	○	○	

What is the entropy for this dataset?

$\text{Ent} = -\frac{12}{25}\log_2\frac{12}{25} - \frac{13}{25}\log_2\frac{13}{25} \approx \mathbf{0.9988}$

$A = \bullet$: $\frac{6}{12}$ positive $\rightarrow 1.0$
$A = \circ$: $\frac{6}{13}$ positive $\rightarrow 0.9957$
Expected: $\frac{12}{25} \cdot 1.0 + \frac{13}{25} \cdot 0.9957 \approx \mathbf{0.9977}$

$B = \bullet$: $\frac{9}{11}$ positive $\rightarrow 0.684$
$B = \circ$: $\frac{3}{14}$ positive $\rightarrow 0.750$
Expected: **0.721**

$C = \bullet$: $\frac{3}{6}$ positive $\rightarrow 1.0$
$C = \circ$: $\frac{9}{19}$ positive $\rightarrow 0.9980$
Expected: **0.9985**

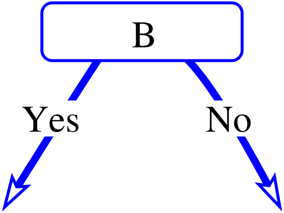
$D = \bullet$: $\frac{3}{5}$ positive $\rightarrow 0.9710$
$D = \circ$: $\frac{9}{20}$ positive $\rightarrow 0.9928$
Expected: **0.9884**

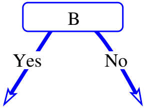
A	B	C	D	
●	●	○	○	+
○	●	●	○	+
○	○	○	○	
●	○	○	●	+
○	●	○	○	+
●	○	●	○	
●	●	○	○	+
○	○	○	○	
○	○	●	○	
●	●	○	○	+
○	○	○	●	+
●	○	○	○	
●	●	●	○	+
○	●	○	●	
○	○	○	○	
●	○	○	○	
●	●	○	●	
○	●	○	○	+
○	○	●	○	
●	○	○	○	
○	●	○	○	+
●	○	○	●	+
○	●	●	○	+
○	○	○	○	
●	○	○	○	

$$\text{Gain}(A) = 0.9988 - 0.9977 = \mathbf{0.0011}$$
$$\text{Gain}(B) = 0.9988 - 0.7210 = \mathbf{0.2778}$$
$$\text{Gain}(C) = 0.9988 - 0.9985 = \mathbf{0.0003}$$
$$\text{Gain}(D) = 0.9988 - 0.9884 = \mathbf{0.0104}$$

$$\mathrm{Gain}(A) = 0.9988 - 0.9977 = \mathbf{0.0011}$$
$$\mathrm{Gain}(B) = 0.9988 - 0.7210 = \mathbf{0.2778}$$
$$\mathrm{Gain}(C) = 0.9988 - 0.9985 = \mathbf{0.0003}$$
$$\mathrm{Gain}(D) = 0.9988 - 0.9884 = \mathbf{0.0104}$$

Attribute B gives most information





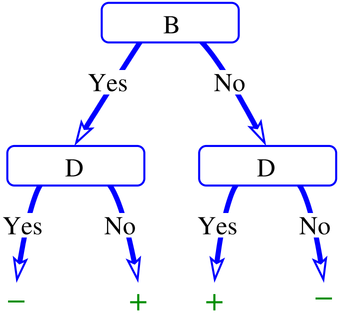
Examples where $B = \bullet$

A	B	C	D	
●	●	○	○	+
○	●	●	○	+
○	●	○	○	+
●	●	○	○	+
●	●	○	○	+
●	●	●	○	+
○	●	○	●	
●	●	○	●	
○	●	○	○	+
○	●	○	○	+
○	●	●	○	+

Examples where $B = \circ$

A	B	C	D	
○	○	○	○	
●	○	○	●	+
●	○	●	○	
○	○	○	○	
○	○	●	○	
○	○	○	●	+
●	○	○	○	
○	○	○	○	
●	○	○	○	
○	○	●	○	
●	○	○	○	
●	○	○	●	+
○	○	○	○	
●	○	○	○	

Information Gain



Which Bias does this learning algorithm have?

Which Bias does this learning algorithm have?

- **Restriction Bias?**
- **Preference Bias?**

Which Bias does this learning algorithm have?

- **Restriction Bias?**
  No, all hypotheses can be represented
- **Preference 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 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?

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

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

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

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

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"

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:

"Entities should not be multiplied beyond necessity"

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:

"Entities should not be multiplied beyond necessity"

All things being equal,
the simplest explanation tends to be the right one.

Why are simple hypotheses more likely to be correct?

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.

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.

Overfitting, (överträning)

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

Overfitting, (överträning)

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

Good results on training data, but generalizes badly

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?

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

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?

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

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

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

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

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

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

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