

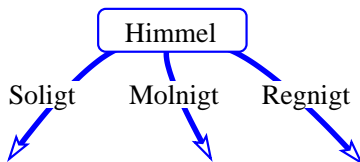
# Beslutsträd

- 1 Beslutsträd
  - Användning
  - Inläring
- 2 Oförutsägbarhet
  - Entropimåttet
  - Entropi för datamängder
  - Information Gain
- 3 Bias
  - Bias
  - Occam's princip
  - Överträning
- 4 Förbättringar

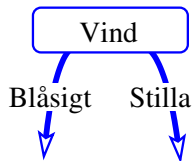
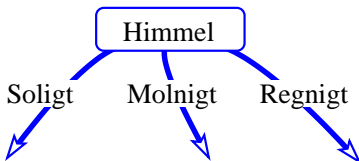
- 1 Beslutsträd
  - Användning
  - Inläring
  
- 2 Oförutsägbarhet
  - Entropimåttet
  - Entropi för datamängder
  - Information Gain
  
- 3 Bias
  - Bias
  - Occam's princip
  - Överträning
  
- 4 Förbättringar

Grundidé: Testa ett attribut i taget

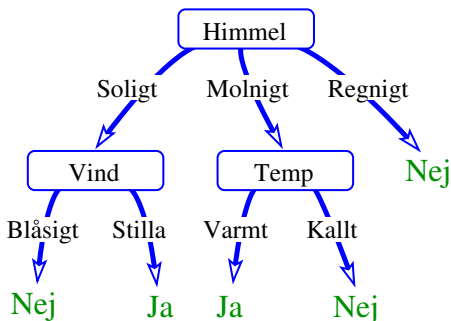
Grundidé: Testa ett attribut i taget



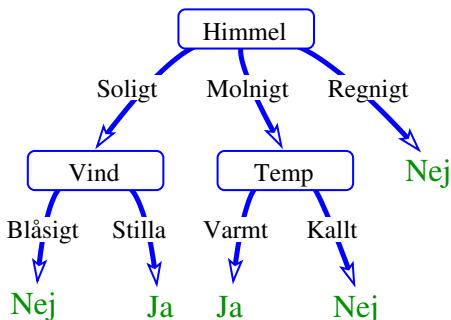
Grundidé: Testa ett attribut i taget



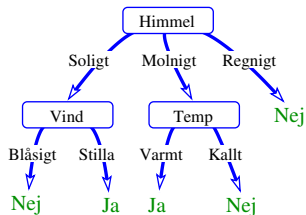
Hela analysstrategin kan betraktas som ett träd.



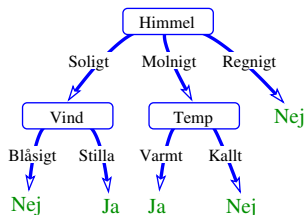
Hela analysstrategin kan betraktas som ett träd.



Svaren (kategoriseringen) beskrivs av *löven* i trädet

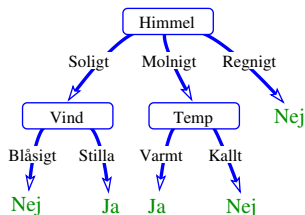


Vad representerar trädet?



Vad representerar trädet?

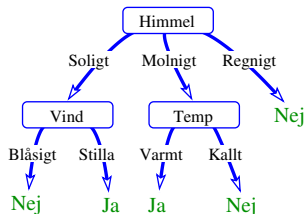
$$(\text{Soligt} \wedge \text{Stilla}) \vee (\text{Molnigt} \wedge \text{Varmt})$$



Vad representerar trädet?

$$(\text{Soligt} \wedge \text{Stilla}) \vee (\text{Molnigt} \wedge \text{Varmt})$$

Fungerar som en *disjunktion av konjunktioner*



Vad representerar trädet?

$$(\text{Soligt} \wedge \text{Stilla}) \vee (\text{Molnigt} \wedge \text{Varmt})$$

Fungerar som en *disjunktion av konjunktioner*

*Normalform* för boolska funktioner

**Godtyckliga kategoriseringar** kan göras!

Hur kan man bygga träden automatiskt?

Hur kan man bygga träden automatiskt?

- 1 Välj ett attribut att fråga om
- 2 Grenar med entydig klassning är klara
- 3 Andra grenar byggs vidare rekursivt

Hur kan man bygga träden automatiskt?

- 1 Välj ett attribut att fråga om
- 2 Grenar med entydig klassning är klara
- 3 Andra grenar byggs vidare rekursivt

Central fråga: Hur väljer vi attribut?

Hur kan man bygga träden automatiskt?

- 1 Välj ett attribut att fråga om
- 2 Grenar med entydig klassning är klara
- 3 Andra grenar byggs vidare rekursivt

Central fråga: Hur väljer vi attribut?

Girig idé:

Välj i varje läge det attribut som *säger mest* om svaret

- 1 Beslutsträd
  - Användning
  - Inläring
- 2 Oförutsägbarhet
  - Entropimåttet
  - Entropi för datamängder
  - Information Gain
- 3 Bias
  - Bias
  - Occam's princip
  - Överträning
- 4 Förbättringar

# Entropi

*Entropi* — mått på **oförutsägbarheten**

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

$p_i$  sannolikheten för händelsen  $i$

# Entropi

Exempel: singla slant

# Entropi

Exempel: singla slant

$$p_{\text{krona}} = 0.5; \quad p_{\text{klave}} = 0.5$$

# Entropi

Exempel: singla slant

$$p_{\text{krona}} = 0.5; \quad p_{\text{klave}} = 0.5$$

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

# Entropi

Exempel: singla slant

$$p_{\text{krona}} = 0.5; \quad p_{\text{klave}} = 0.5$$

$$\begin{aligned} \text{Entropin} &= \sum_i -p_i \log_2 p_i = \\ &= -0.5 \log_2 0.5 + -0.5 \log_2 0.5 \end{aligned}$$

# Entropi

Exempel: singla slant

$$p_{\text{krona}} = 0.5; \quad p_{\text{klave}} = 0.5$$

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

# Entropi

Exempel: singla slant

$$p_{\text{krona}} = 0.5; \quad p_{\text{klave}} = 0.5$$

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

# Entropi

Exempel: singla slant

$$p_{\text{krona}} = 0.5; \quad p_{\text{klave}} = 0.5$$

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

Utfallet av en slantsingling innehåller **1 bit** information

# Entropi

Exempel: kasta tärning

# Entropi

Exempel: kasta tärning

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

# Entropi

Exempel: kasta tärning

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

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

# Entropi

Exempel: kasta tärning

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

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

# Entropi

Exempel: kasta tärning

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

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

# Entropi

Exempel: kasta tärning

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

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

Utfallet av ett tärningskast innehåller **2.58 bit** information

# Entropi

Exempel: kasta en **falsk tärning**

# Entropi

Exempel: kasta en **falsk tärning**

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

# Entropi

Exempel: kasta en **falsk tärning**

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

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

# Entropi

Exempel: kasta en **falsk tärning**

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

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

# Entropi

Exempel: kasta en **falsk tärning**

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

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

# Entropi

Exempel: kasta en **falsk tärning**

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

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

En riktig tärning är **mer oförutsägbar** (2.58 bit) än en falsk (2.16 bit)

# Entropi

Oförutsägbarheten för en **datamängd**

# Entropi

Oförutsägbarheten för en **datamängd**

- 100 exempel, varav 42 positiva

# Entropi

Oförutsägbarheten för en **datamängd**

- 100 exempel, varav 42 positiva

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

# Entropi

Oförutsägbarheten för en **datamängd**

- 100 exempel, varav 42 positiva

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

- 100 exempel, varav 3 positiva

# Entropi

Oförutsägbarheten för en **datamängd**

- 100 exempel, varav 42 positiva

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

- 100 exempel, varav 3 positiva

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

Tillbaks till beslutsträden

Tillbaks till beslutsträden

Smart idé:

Fråga efter det attribut som maximerar förväntad minskning av entropin.

Tillbaks till beslutsträden

Smart idé:

Fråga efter det attribut som maximerar förväntad minskning av entropin.

Information Gain

Tillbaks till beslutsträden

Smart idé:

Fråga efter det attribut som maximerar förväntad minskning av entropin.

**Information Gain**

Antag att vi frågar om attribut  $A$  för en datamängd  $S$

Tillbaks till beslutsträden

Smart idé:

Fråga efter det attribut som maximerar förväntad minskning av entropin.

**Information Gain**

Antag att vi frågar om attribut  $A$  för en datamängd  $S$

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

Tillbaks till beslutsträden

Smart idé:

Fråga efter det attribut som maximerar förväntad minskning av entropin.

Information Gain

Antag att vi frågar om attribut  $A$  för en datamängd  $S$

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

Tillbaks till beslutsträden

Smart idé:

Fråga efter det attribut som maximerar förväntad minskning av entropin.

### Information Gain

Antag att vi frågar om attribut  $A$  för en datamängd  $S$

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

Tillbaks till beslutsträden

Smart idé:

Fråga efter det attribut som maximerar förväntad minskning av entropin.

### Information Gain

Antag att vi frågar om attribut  $A$  för en datamängd  $S$

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

Vad är entropin för denna datamängd?

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

Vad är entropin för denna datamängd?

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

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









$$\text{Gain}(A) = 0.9988 - 0.9977 = \mathbf{0.0011}$$

$$\text{Gain}(B) = 0.9988 - 0.7210 = \mathbf{0.2778}$$

$$\text{Gain}(C) = 0.9988 - 0.9985 = \mathbf{0.0003}$$

$$\text{Gain}(D) = 0.9988 - 0.9884 = \mathbf{0.0104}$$

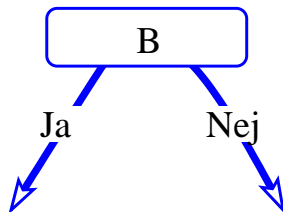
$$\text{Gain}(A) = 0.9988 - 0.9977 = \mathbf{0.0011}$$

$$\text{Gain}(B) = 0.9988 - 0.7210 = \mathbf{0.2778}$$

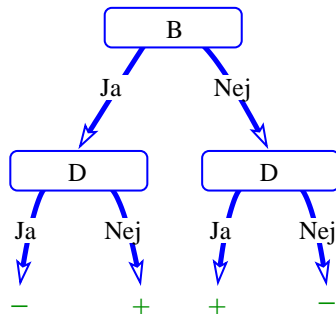
$$\text{Gain}(C) = 0.9988 - 0.9985 = \mathbf{0.0003}$$

$$\text{Gain}(D) = 0.9988 - 0.9884 = \mathbf{0.0104}$$

Attribut  $B$  ger mest information







- 1 Beslutsträd
  - Användning
  - Inlärning
- 2 Oförutsägbarhet
  - Entropimåttet
  - Entropi för datamängder
  - Information Gain
- 3 Bias
  - Bias
  - Occam's princip
  - Överträning
- 4 Förbättringar

Vilken Bias har denna inlärningsalgoritm?

Vilken Bias har denna inlärningsalgoritm?

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

Vilken Bias har denna inlärningsalgoritm?

- **Restriction Bias?**  
Nej, alla hypoteser kan erhållas
- **Preference Bias?**

Vilken Bias har denna inlärningsalgoritm?

- **Restriction Bias?**

Nej, alla hypoteser kan erhållas

- **Preference Bias?**

Ja, vissa typer av träd hittas före andra

Vilken Bias har denna inlärningsalgoritm?

- **Restriction Bias?**

Nej, alla hypoteser kan erhållas

- **Preference Bias?**

Ja, vissa typer av träd hittas före andra

Vilka hypoteser (här: träd) prioriteras?

Vilken Bias har denna inlärningsalgoritm?

- **Restriction Bias?**

Nej, alla hypoteser kan erhållas

- **Preference Bias?**

Ja, vissa typer av träd hittas före andra

Vilka hypoteser (här: träd) prioriteras?

- Grunda träd

- "Viktiga frågor" tidigt

Hur ska man veta vilka hypoteser som ska föredras när flera stämmer med exemplen?

Hur ska man veta vilka hypoteser som ska föredras när flera stämmer med exemplen?

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

Hur ska man veta vilka hypoteser som ska föredras när flera stämmer med exemplen?

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

William från Ockham, Teolog och Filosof (1285–1349)

*"Entia non sunt multiplicanda praeter necessitatem"*

Hur ska man veta vilka hypoteser som ska föredras när flera stämmer med exemplen?

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

William från Ockham, Teolog och Filosof (1285–1349)

*"Entia non sunt multiplicanda praeter necessitatem"*

fritt översatt:

"Man bör inte anta fler företeelser än vad som är nödvändigt för att förklara fenomenen"

Hur ska man veta vilka hypoteser som ska föredras när flera stämmer med exemplen?

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

William från Ockham, Teolog och Filosof (1285–1349)

*"Entia non sunt multiplicanda praeter necessitatem"*

fritt översatt:

"Man bör inte anta fler företeelser än vad som är nödvändigt för att förklara fenomenen"

Om fler hypoteser kan förklara data,  
välj då den enklaste

Varför är enkla hypoteser bättre?

Varför är enkla hypoteser bättre?

Troligare att verkligheten som genererat exemplen har en enkel genererande mekanism.

Varför är enkla hypoteser bättre?

Troligare att verkligheten som genererat exemplen har en enkel genererande mekanism.

Enkla hypoteser generaliserar normalt bättre.

## Överträning, *overfitting*

När hypoteserna är för specialiserade för de aktuella träningsexemplen.

## Överträning, *overfitting*

När hypoteserna är för specialiserade  
för de aktuella träningsexemplen.

Bra på träningsdata men generaliserar dåligt

## Överträning, *overfitting*

När hypoteserna är för specialiserade  
för de aktuella träningsexemplen.

Bra på träningsdata men generaliserar dåligt

När inträffar detta?

## Överträning, *overfitting*

När hypoteserna är för specialiserade för de aktuella träningsexemplen.

Bra på träningsdata men generaliserar dåligt

När inträffar detta?

- Icke-representativt sample
- Brus bland exemplen

## Överträning, *overfitting*

När hypoteserna är för specialiserade för de aktuella träningsexemplen.

Bra på träningsdata men generaliserar dåligt

När inträffar detta?

- Icke-representativt sample
- Brus bland exemplen

Vad kan man göra åt det?

## Överträning, *overfitting*

När hypoteserna är för specialiserade för de aktuella träningsexemplen.

Bra på träningsdata men generaliserar dåligt

När inträffar detta?

- Icke-representativt sample
- Brus bland exemplen

Vad kan man göra åt det?

Välj en enklare hypotes och acceptera fel även för träningsexemplen

## Möjliga förbättringar av beslutsträden

- Undvik överträning
  - Begränsa trädets höjd
  - Beskärning (*Pruning*)
- Attribut med graderade värden
- Saknade attributvärden
- Olika kostnad för olika attribut

## Möjliga förbättringar av beslutsträden

- Undvik överträning
  - Begränsa trädets höjd
  - Beskärning (*Pruning*)
- Attribut med graderade värden
- Saknade attributvärden
- Olika kostnad för olika attribut

## Möjliga förbättringar av beslutsträden

- Undvik överträning
  - Begränsa trädets höjd
  - Beskärning (*Pruning*)
- **Attribut med graderade värden**
- Saknade attributvärden
- Olika kostnad för olika attribut

## Möjliga förbättringar av beslutsträden

- Undvik överträning
  - Begränsa trädets höjd
  - Beskärning (*Pruning*)
- Attribut med graderade värden
- Saknade attributvärden
- Olika kostnad för olika attribut

## Möjliga förbättringar av beslutsträden

- Undvik överträning
  - Begränsa trädets höjd
  - Beskärning (*Pruning*)
- Attribut med graderade värden
- Saknade attributvärden
- Olika kostnad för olika attribut

## Möjliga förbättringar av beslutsträden

- Undvik överträning
  - Begränsa trädets höjd
  - Beskärning (*Pruning*)
- Attribut med graderade värden
- Saknade attributvärden
- Olika kostnad för olika attribut