Information-Theoretic Modeling

Lecture 9: The MDL Principle

Teemu Roos

Department of Computer Science, University of Helsinki

Fall 2009

UNIVERSITY OF HELSINKI

Lecture 9: MDL Principle

Jorma Rissanen (left) receiving the IEEE Information Theory Society Best Paper Award from Claude Shannon in 1986.

IEEE Golden Jubilee Award for Technological Innovation (for the invention of arithmetic coding) 1998; **IEEE Richard W. Hamming Medal** (for fundamental contribution to information theory, statistical inference, control theory, and the theory of complexity) 1993; **Kolmogorov Medal** 2006; **IBM Outstanding Innovation Award** (for work in statistical inference, information theory, and the theory of complexity) 1988; **IEEE Claude E. Shannon Award** 2009; ...

1. Occam's Razor
   - House
   - Visual Recognition
   - Astronomy
   - Razor

1. Occam's Razor
   - House
   - Visual Recognition
   - Astronomy
   - Razor

2. MDL Principle
   - Rules & Exceptions
   - Probabilistic Models
   - Old-Style MDL
   - Modern MDL

House

House

Brandon has

1. cough,
2. severe abdominal pain,
3. nausea,
4. low blood pressure,
5. fever.

House

Brandon has

1. cough,
2. severe abdominal pain,
3. nausea,
4. low blood pressure,
5. fever.

No single disease causes all of these.

House

Brandon has

1. cough,
2. severe abdominal pain,
3. nausea,
4. low blood pressure,
5. fever.

No single disease causes all of these.

Each symptom can be caused by *some* (possibly different) disease...

House

Brandon has

1. cough,
2. severe abdominal pain,
3. nausea,
4. low blood pressure,
5. fever.

1. pneumonia,

No single disease causes all of these.

Each symptom can be caused by *some* (possibly different) disease...

House

Brandon has

1. cough,
2. severe abdominal pain,
3. nausea,
4. low blood pressure,
5. fever.

1. pneumonia,
2. appendicitis,

No single disease causes all of these.

Each symptom can be caused by *some* (possibly different) disease...

House

Brandon has

1. cough,
2. severe abdominal pain,
3. nausea,
4. low blood pressure,
5. fever.

1. pneumonia,
2. appendicitis,
3. food poisoning,

No single disease causes all of these.

Each symptom can be caused by *some* (possibly different) disease...

House

Brandon has

1. cough,
2. severe abdominal pain,
3. nausea,
4. low blood pressure,
5. fever.

1. pneumonia,
2. appendicitis,
3. food poisoning,
4. hemorrhage,

No single disease causes all of these.

Each symptom can be caused by *some* (possibly different) disease...

House

Brandon has

1. cough,
2. severe abdominal pain,
3. nausea,
4. low blood pressure,
5. fever.

1. pneumonia,
2. appendicitis,
3. food poisoning,
4. hemorrhage,
5. meningitis.

No single disease causes all of these.

Each symptom can be caused by *some* (possibly different) disease...

House

Brandon has

1. cough,
2. severe abdominal pain,
3. nausea,
4. low blood pressure,
5. fever.

1. pneumonia,
2. appendicitis,
3. food poisoning,
4. hemorrhage,
5. meningitis.

No single disease causes all of these.

Each symptom can be caused by *some* (possibly different) disease...

Dr. House explains the symptoms with two simple causes:

House

Brandon has

1. cough,
2. severe abdominal pain,
3. nausea,
4. low blood pressure,
5. fever.

1. common cold,
2. appendicitis,
3. food poisoning,
4. hemorrhage,
5. common cold.

No single disease causes all of these.

Each symptom can be caused by *some* (possibly different) disease...

Dr. House explains the symptoms with two simple causes:

1. common cold, causing the cough and fever,

House

Brandon has

1. cough,
2. severe abdominal pain,
3. nausea,
4. low blood pressure,
5. fever.

1. common cold,
2. gout medicine,
3. gout medicine,
4. gout medicine,
5. common cold.

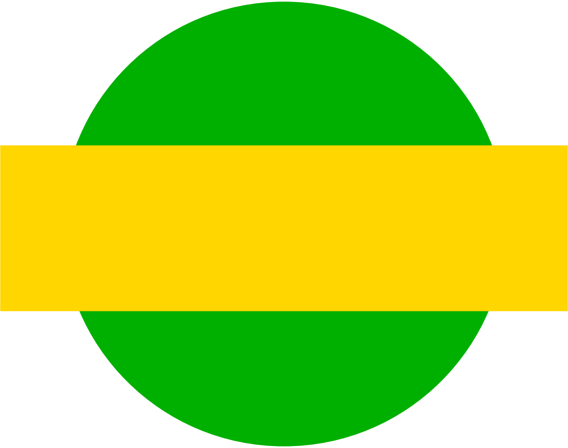
No single disease causes all of these.

Each symptom can be caused by *some* (possibly different) disease...

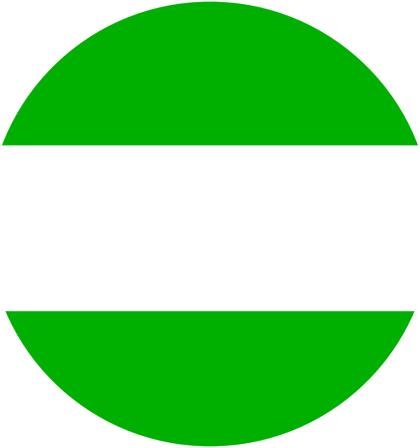
Dr. House explains the symptoms with two simple causes:

1. common cold, causing the cough and fever,
2. pharmacy error: cough medicine replaced by gout medicine.

Visual Recognition

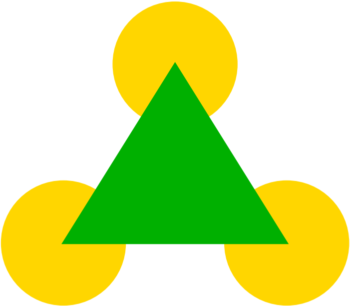


Visual Recognition

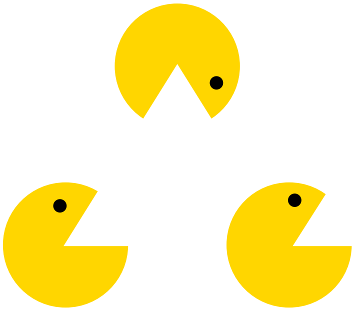


Visual Recognition

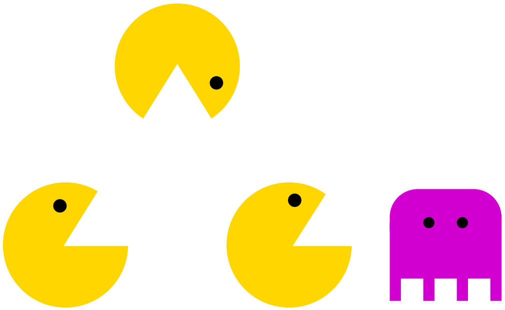
Visual Recognition



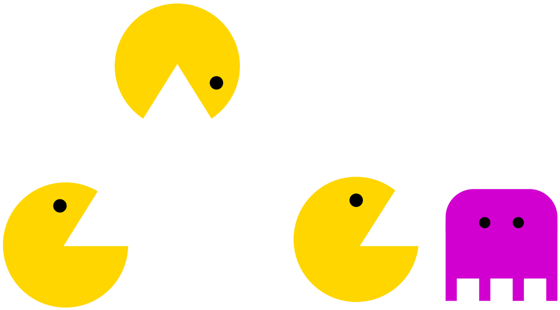
Visual Recognition



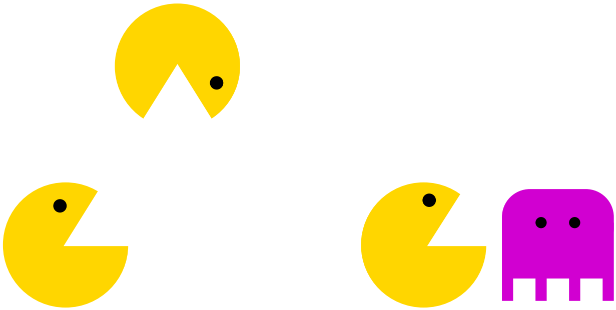
Visual Recognition



Visual Recognition

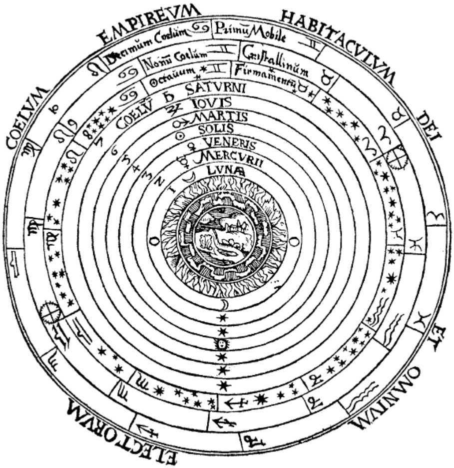


Visual Recognition

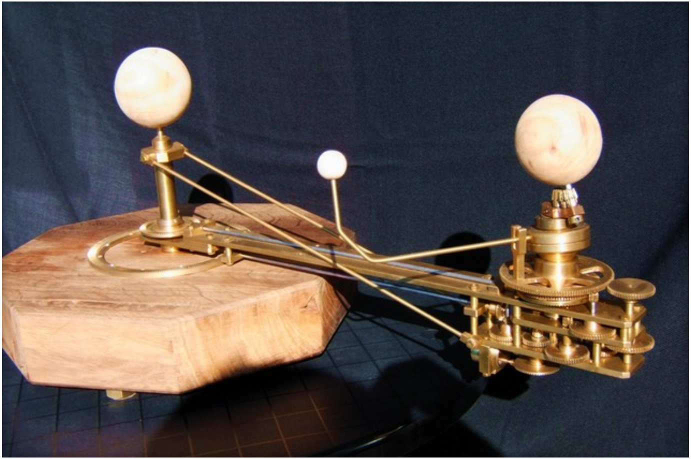


Astronomy

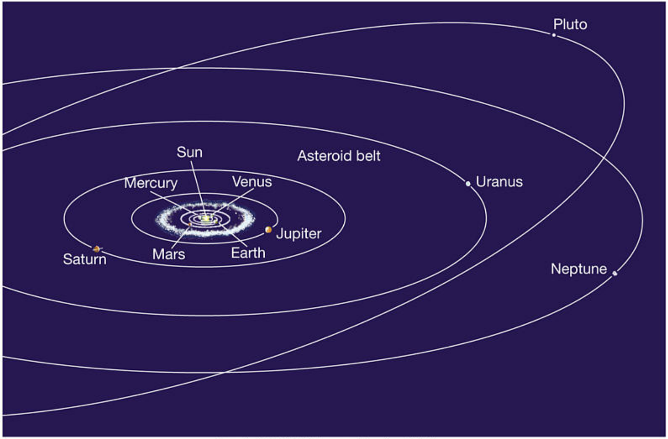
Schema huius præmissæ diuisionis Sphærarum.



Astronomy



Astronomy


Pluto
Sun
Asteroid belt
Mercury
Venus
Uranus
Jupiter
Saturn
Mars
Earth
Neptune

William of Ockham (c. 1288–1348)

Occam's Razor

Occam's Razor

Entities should not be multiplied beyond necessity.

Occam's Razor

Occam's Razor

Entities should not be multiplied beyond necessity.

Isaac Newton: "We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances."

Occam's Razor

Occam's Razor

Entities should not be multiplied beyond necessity.

Isaac Newton: "We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances."

Diagnostic parsimony: Find the fewest possible causes that explain the symptoms.

Occam's Razor

Occam's Razor

Entities should not be multiplied beyond necessity.

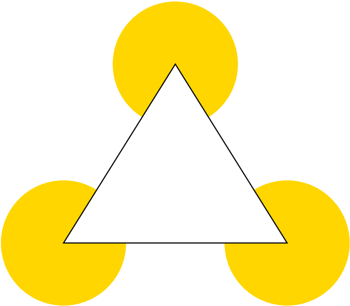
Isaac Newton: "We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances."

Diagnostic parsimony: Find the fewest possible causes that explain the symptoms.

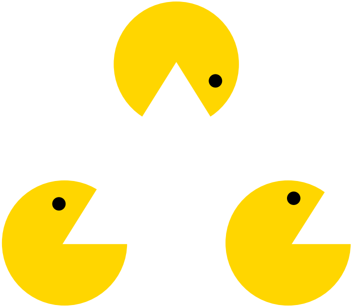
(**Hickam's dictum:** "Patients can have as many diseases as they damn well please.")

Visual Recognition

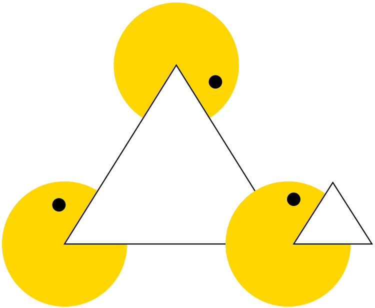
Visual Recognition



Visual Recognition



Visual Recognition



1. Occam's Razor
   - House
   - Visual Recognition
   - Astronomy
   - Razor

2. MDL Principle
   - Rules & Exceptions
   - Probabilistic Models
   - Old-Style MDL
   - Modern MDL

MDL Principle

Minimum Description Length (MDL) Principle (2-part)

Choose the hypothesis which minimizes the sum of

1. the codelength of the hypothesis, and
2. the codelength of the data with the help of the hypothesis.

MDL Principle

Minimum Description Length (MDL) Principle (2-part)

Choose the hypothesis which minimizes the sum of

1. the codelength of the hypothesis, and
2. the codelength of the data with the help of the hypothesis.

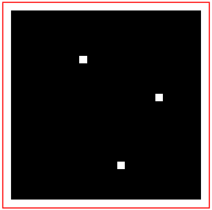
How to encode data *with the help of a hypothesis*?

Encoding Data: Rules & Exceptions

Idea 1: Hypothesis = rule; encode exceptions.

Encoding Data: Rules & Exceptions

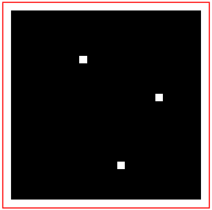
Idea 1: Hypothesis = rule; encode exceptions.



Black box of size $25 \times 25 = 625$, white dots at $(x_1, y_1), (x_2, y_2), (x_3, y_3)$.

Encoding Data: Rules & Exceptions

Idea 1: Hypothesis = rule; encode exceptions.



Black box of size $25 \times 25 = 625$, white dots at $(x_1, y_1), (x_2, y_2), (x_3, y_3)$.

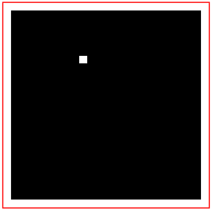
For image of size $n = 625$, there are 2^n different images, and

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

different groups of k exceptions.

Encoding Data: Rules & Exceptions

Idea 1: Hypothesis = rule; encode exceptions.



Black box of size $25 \times 25 = 625$, white dots at $(x_1, y_1), (x_2, y_2), (x_3, y_3)$.

For image of size $n = 625$, there are 2^n different images, and

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

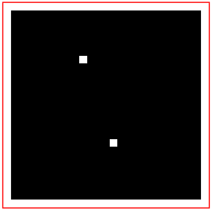
different groups of k exceptions.

$$k = 1 : \binom{n}{1} = 625 \ll 2^{625} \approx 1.4 \times 10^{188}.$$

Codelength $\log_2(n+1) + \log_2 \binom{n}{k} \approx 19$ vs. $\log_2 2^{625} = 625$

Encoding Data: Rules & Exceptions

Idea 1: Hypothesis = rule; encode exceptions.



Black box of size $25 \times 25 = 625$, white dots at $(x_1, y_1), (x_2, y_2), (x_3, y_3)$.

For image of size $n = 625$, there are 2^n different images, and

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

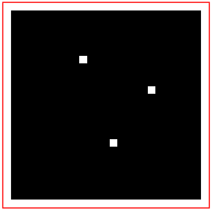
different groups of k exceptions.

$$k = 2 : \binom{n}{2} = 195\,000 \ll 2^{625} \approx 1.4 \times 10^{188}.$$

Codelength $\log_2(n+1) + \log_2 \binom{n}{k} \approx 27$ vs. $\log_2 2^{625} = 625$

Encoding Data: Rules & Exceptions

Idea 1: Hypothesis = rule; encode exceptions.



Black box of size $25 \times 25 = 625$, white dots at $(x_1, y_1), (x_2, y_2), (x_3, y_3)$.

For image of size $n = 625$, there are 2^n different images, and

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

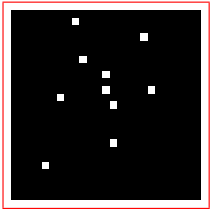
different groups of k exceptions.

$$k = 3 : \binom{n}{3} = 40\,495\,000 \ll 2^{625} \approx 1.4 \times 10^{188}.$$

Codelength $\log_2(n+1) + \log_2 \binom{n}{k} \approx 35$ vs. $\log_2 2^{625} = 625$

Encoding Data: Rules & Exceptions

Idea 1: Hypothesis = rule; encode exceptions.



Black box of size $25 \times 25 = 625$, white dots at $(x_1, y_1), (x_2, y_2), (x_3, y_3)$.

For image of size $n = 625$, there are 2^n different images, and

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

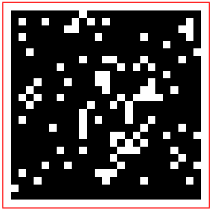
different groups of k exceptions.

$$k = 10 : \binom{n}{10} = 2\,331\,354\,000\,000\,000\,000\,000 \ll 2^{625}.$$

Codelength $\log_2(n+1) + \log_2 \binom{n}{k} \approx 80$ vs. $\log_2 2^{625} = 625$

Encoding Data: Rules & Exceptions

Idea 1: Hypothesis = rule; encode exceptions.



Black box of size $25 \times 25 = 625$, white dots at $(x_1, y_1), (x_2, y_2), (x_3, y_3)$.

For image of size $n = 625$, there are 2^n different images, and

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

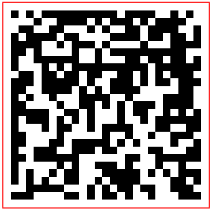
different groups of k exceptions.

$$k = 100 : \binom{n}{100} \approx 9.5 \times 10^{117} \ll 2^{625} \approx 1.4 \times 10^{188}.$$

$$\text{Codelength } \log_2(n+1) + \log_2 \binom{n}{k} \approx 401 \text{ vs. } \log_2 2^{625} = 625$$

Encoding Data: Rules & Exceptions

Idea 1: Hypothesis = rule; encode exceptions.



Black box of size $25 \times 25 = 625$, white dots at $(x_1, y_1), (x_2, y_2), (x_3, y_3)$.

For image of size $n = 625$, there are 2^n different images, and

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

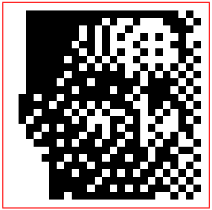
different groups of k exceptions.

$$k = 300 : \binom{n}{300} \approx 2.7 \times 10^{186} < 2^{625} \approx 1.4 \times 10^{188}.$$

Codelength $\log_2(n+1) + \log_2 \binom{n}{k} \approx 629$ vs. $\log_2 2^{625} = 625$

Encoding Data: Rules & Exceptions

Idea 1: Hypothesis = rule; encode exceptions.



Black box of size $25 \times 25 = 625$, white dots at $(x_1, y_1), (x_2, y_2), (x_3, y_3)$.

For image of size $n = 625$, there are 2^n different images, and

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

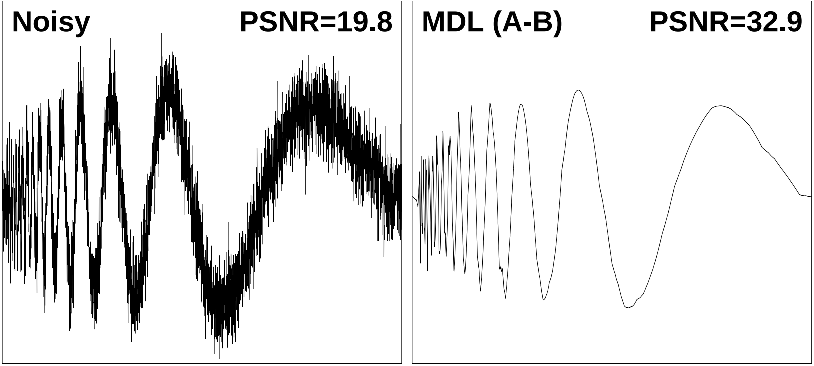
different groups of k exceptions.

$$k = 372 : \binom{n}{372} \approx 5.1 \times 10^{181} \ll 2^{625} \approx 1.4 \times 10^{188}.$$

Codelength $\log_2(n+1) + \log_2 \binom{n}{k} \approx 613$ vs. $\log_2 2^{625} = 625$

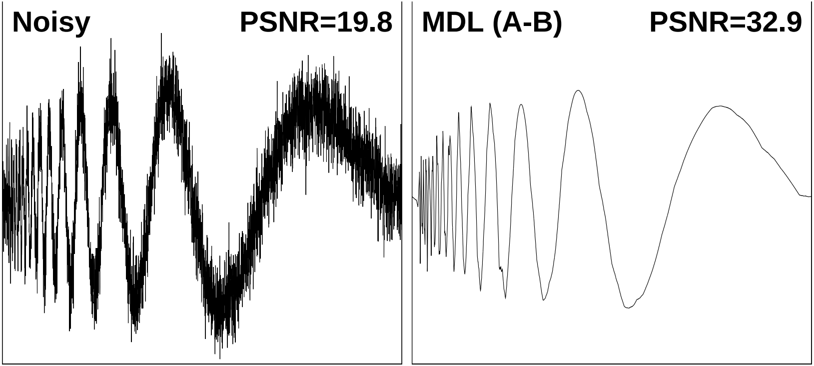
Encoding Data: Probabilistic Models

Idea 2: Hypothesis = probability distribution.



Encoding Data: Probabilistic Models

Idea 2: Hypothesis = probability distribution.



Rissanen & Shannon: $\log_2 \frac{1}{p_{\hat{\theta}}(D)} + \frac{k}{2} \log_2 n.$

Polynomials

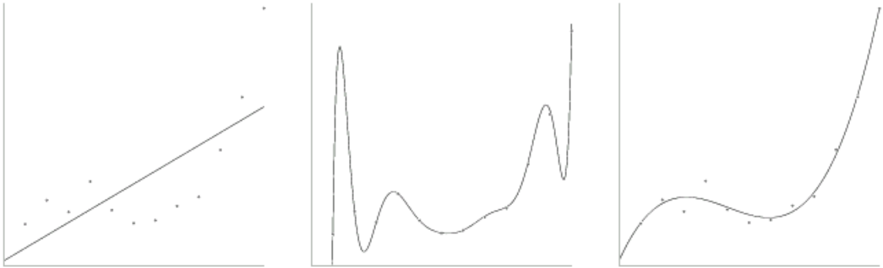


Figure 1: A simple (1.1), complex (1.2) and a trade-off (3rd degree) polynomial.

From P. Grünwald

Old-Style MDL

With the precision $\frac{1}{\sqrt{n}}$ the codelength for data is almost optimal:

$$\min_{\theta^q \in \{\theta^{(1)}, \theta^{(2)}, \ldots\}} \ell_{\theta^q}(D) \approx \min_{\theta \in \Theta} \ell_{\theta}(D) = \log_2 \frac{1}{p_{\hat{\theta}}(D)} \ .$$

Old-Style MDL

With the precision $\frac{1}{\sqrt{n}}$ the codelength for data is almost optimal:

$$\min_{\theta^q \in \{\theta^{(1)}, \theta^{(2)}, \ldots\}} \ell_{\theta^q}(D) \;\approx\; \min_{\theta \in \Theta} \ell_\theta(D) = \log_2 \frac{1}{p_{\hat{\theta}}(D)} \; .$$

This gives the total codelength formula:

"Steam MDL"

$$\ell_{\theta^q}(D) + \ell(\theta^q) \approx \log_2 \frac{1}{p_{\hat{\theta}}(D)} + \frac{k}{2} \log_2 n \; .$$

Old-Style MDL

The $\frac{k}{2} \log_2 n$ formula is only a rough approximation, and works well only for very large samples.

Old-Style MDL

The $\frac{k}{2} \log_2 n$ formula is only a rough approximation, and works well only for very large samples.

MDL in the 21st century:

- More advanced codes: mixtures, normalized maximum likelihood, etc.

Flavours of MDL

1. "Pedestrian"
   Asymptotic two-part code-length same as BIC.

2. "Sophisticated"
   Bayesian marginal likelihood.

3. "Champions League"
   Modern (minimax regret optimal) code

   normalized maximum likelihood (NML)

Problem: NML computationally very hard.

MDL Model Selection

Recall (from Lecture 7) the multi-part codes used when multiple model classes, $\mathcal{M}_1, \mathcal{M}_2, \ldots$ are available:

MDL Model Selection

Recall (from Lecture 7) the multi-part codes used when multiple model classes, $\mathcal{M}_1, \mathcal{M}_2, \ldots$ are available:

1. Encoding of the model class: $\ell(\mathcal{M}_i)$, $i \in \mathbb{N}$.

MDL Model Selection

Recall (from Lecture 7) the multi-part codes used when multiple model classes, $\mathcal{M}_1, \mathcal{M}_2, \ldots$ are available:

1. Encoding of the model class: $\ell(\mathcal{M}_i)$, $i \in \mathbb{N}$.
2. Encoding of the parameter (vector): $\ell_1(\theta)$, $\theta \in \Theta_i$.

MDL Model Selection

Recall (from Lecture 7) the multi-part codes used when multiple model classes, $\mathcal{M}_1, \mathcal{M}_2, \ldots$ are available:

1. Encoding of the model class: $\ell(\mathcal{M}_i)$, $i \in \mathbb{N}$.
2. Encoding of the parameter (vector): $\ell_1(\theta)$, $\theta \in \Theta_i$.
3. Encoding of the data: $\log_2 \dfrac{1}{p_\theta(D)}$, $D \in \mathcal{D}$.

MDL Model Selection

Recall (from Lecture 7) the multi-part codes used when multiple model classes, $\mathcal{M}_1, \mathcal{M}_2, \ldots$ are available:

1. Encoding of the model class: $\ell(\mathcal{M}_i)$, $i \in \mathbb{N}$.
2. Encoding of the parameter (vector): $\ell_1(\theta)$, $\theta \in \Theta_i$.
3. Encoding of the data: $\log_2 \dfrac{1}{p_\theta(D)}$, $D \in \mathcal{D}$.

If we are interested in choosing a model class (and not the parameters), we can improve parts 2 & 3 by combining them into a better universal code than two-part:

MDL Model Selection

Recall (from Lecture 7) the multi-part codes used when multiple model classes, $\mathcal{M}_1, \mathcal{M}_2, \ldots$ are available:

1. Encoding of the model class: $\ell(\mathcal{M}_i)$, $i \in \mathbb{N}$.
2. Encoding of the parameter (vector): $\ell_1(\theta)$, $\theta \in \Theta_i$.
3. Encoding of the data: $\log_2 \frac{1}{p_\theta(D)}$, $D \in \mathcal{D}$.

If we are interested in choosing a model class (and not the parameters), we can improve parts 2 & 3 by combining them into a better universal code than two-part:

1. Encoding of the model class index: $\ell(\mathcal{M}_i)$, $i \in \mathbb{N}$.

MDL Model Selection

Recall (from Lecture 7) the multi-part codes used when multiple model classes, $\mathcal{M}_1, \mathcal{M}_2, \ldots$ are available:

1. Encoding of the model class: $\ell(\mathcal{M}_i)$, $i \in \mathbb{N}$.
2. Encoding of the parameter (vector): $\ell_1(\theta)$, $\theta \in \Theta_i$.
3. Encoding of the data: $\log_2 \dfrac{1}{p_\theta(D)}$, $D \in \mathcal{D}$.

If we are interested in choosing a model class (and not the parameters), we can improve parts 2 & 3 by combining them into a better universal code than two-part:

1. Encoding of the model class index: $\ell(\mathcal{M}_i)$, $i \in \mathbb{N}$.
2. Encoding of the data: $\ell_{\mathcal{M}_i}(D)$, $D \in \mathcal{D}$, where $\ell_{\mathcal{M}_i}$ is a universal code-length (e.g., mixture, NML) based on model class $\mathcal{M}_i$.

MDL Model Selection

MDL Explanation of MDL

The success in extracting the structure from data can be measured by the codelength.

MDL Model Selection

MDL Explanation of MDL

The success in extracting the structure from data can be measured by the codelength.

In practice, we only find the structure that is "visible" to the used model class(es). For instance, the Bernoulli (coin flipping) model only sees the number of 1s.

MDL & Bayes

The MDL model selection criterion

$$\text{minimize } \ell(\theta) + \ell_\theta(D)$$

can be interpreted (via $p = 2^{-\ell}$) as

$$\text{maximize } p(\theta) \times p_\theta(D) \quad .$$

MDL & Bayes

The MDL model selection criterion

$$\text{minimize } \ell(\theta) + \ell_\theta(D)$$

can be interpreted (via $p = 2^{-\ell}$) as

$$\text{maximize } p(\theta) \times p_\theta(D) \quad .$$

In Bayesian probability, this is equivalent to **maximization of posterior probability**:

$$p(\theta \mid D) = \frac{p(\theta)\, p(D \mid \theta)}{p(D)} \quad ,$$

where the term $p(D)$ (the marginal probability of D) is constant wrt. θ and doesn't affect model selection.

MDL & Bayes

The MDL model selection criterion

$$\text{minimize } \ell(\theta) + \ell_\theta(D)$$

can be interpreted (via $p = 2^{-\ell}$) as

$$\text{maximize } p(\theta) \times p_\theta(D) \quad .$$

In Bayesian probability, this is equivalent to **maximization of posterior probability**:

$$p(\theta \mid D) = \frac{p(\theta)\, p(D \mid \theta)}{p(D)} \quad ,$$

where the term $p(D)$ (the marginal probability of D) is constant wrt. θ and doesn't affect model selection.

$\Rightarrow$ **Three Conceps: Probability**

Example: Denoising

$$\begin{aligned} \textit{Complexity} &= \textit{Information} + \textit{Noise} \\ &= \textit{Regularity} + \textit{Randomness} \\ &= \textit{Algorithm} + \textit{Compressed file} \end{aligned}$$

Example: Denoising

$$\begin{aligned} \textit{Complexity} &= \textit{Information} + \textit{Noise} \\ &= \textit{Regularity} + \textit{Randomness} \\ &= \textit{Algorithm} + \textit{Compressed file} \end{aligned}$$

Denoising means the process of removing noise from a signal.

Example: Denoising

Complexity	=	*Information*	+	*Noise*
	=	*Regularity*	+	*Randomness*
	=	*Algorithm*	+	*Compressed file*

Denoising means the process of removing noise from a signal.

The MDL principle gives a natural method for denoising since the very idea of MDL is to separate the total complexity of a signal into information and noise.

Example: Denoising

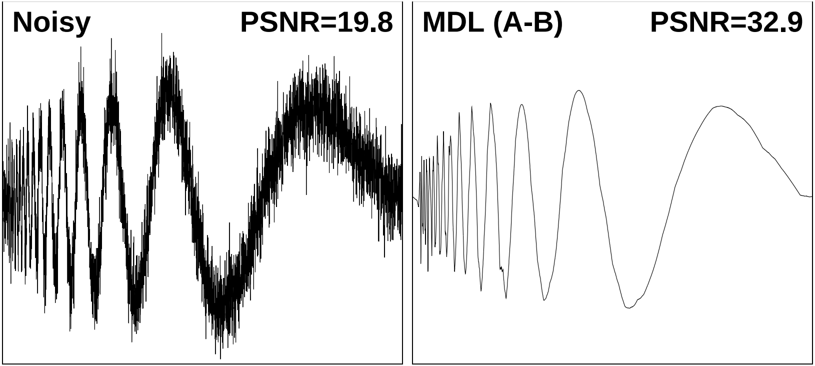
$$
\begin{aligned}
\color{red}{\textit{Complexity}} &= \color{red}{\textit{Information}} + \color{red}{\textit{Noise}} \\
&= \textit{Regularity} + \textit{Randomness} \\
&= \textit{Algorithm} + \textit{Compressed file}
\end{aligned}
$$

Denoising means the process of removing noise from a signal.

The MDL principle gives a natural method for denoising since the very idea of MDL is to separate the total complexity of a signal into information and noise.

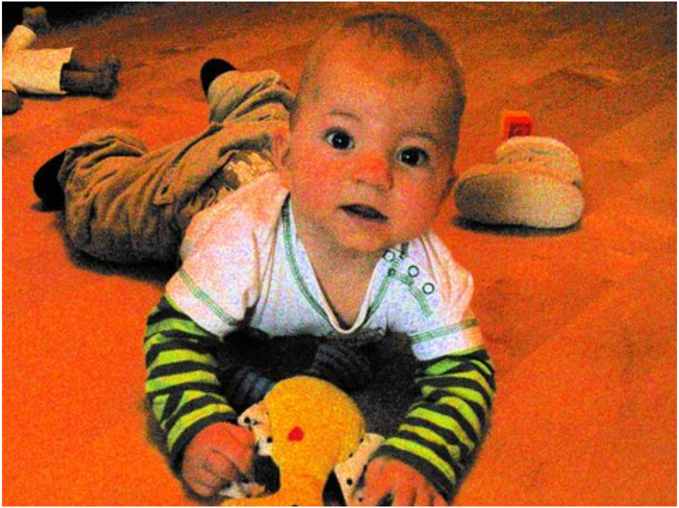
First encode a smooth signal (information), and then the difference to the observed signal (noise).

Example: Denoising

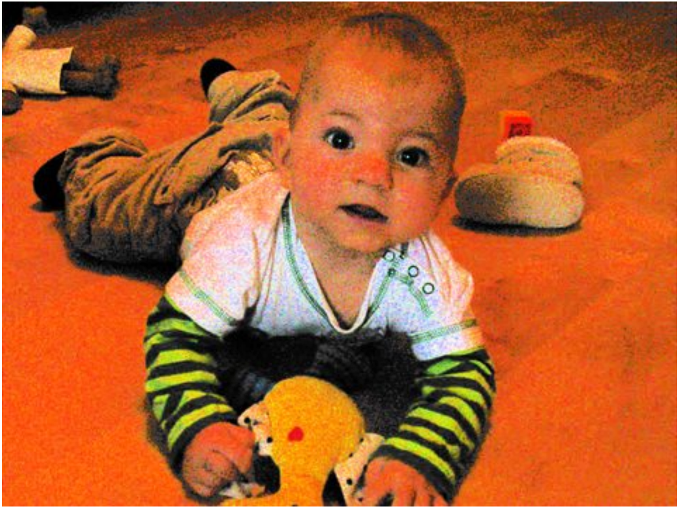


Example: Denoising

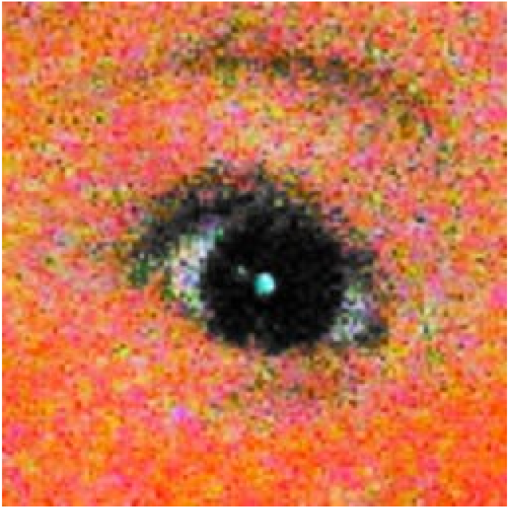
Example: Denoising



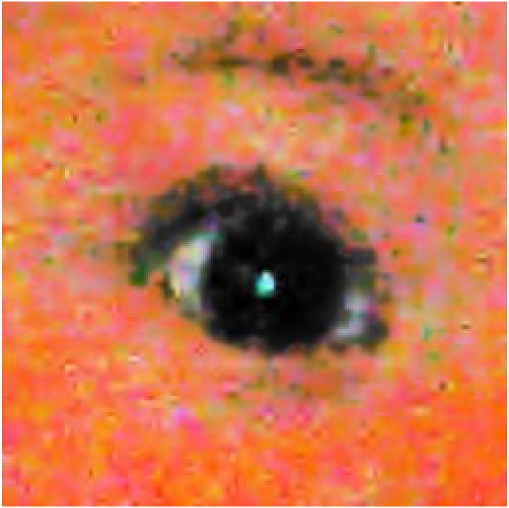
Example: Denoising



Example: Denoising



Example: Denoising



Next Lecture

Friday's lecture:

- Real examples of MDL in action.