Three Concepts: Information

Lecture 5: MDL Principle

Teemu Roos

Complex Systems Computation Group
Department of Computer Science, University of Helsinki

Fall 2007

UNIVERSITY OF HELSINKI

Lecture 5: 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 Corporate Award** (for the MDL/PMDL principles and stochastic complexity) 1991; **IBM Outstanding Innovation Award** (for work in statistical inference, information theory, and the theory of complexity) 1988; ...

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

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

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

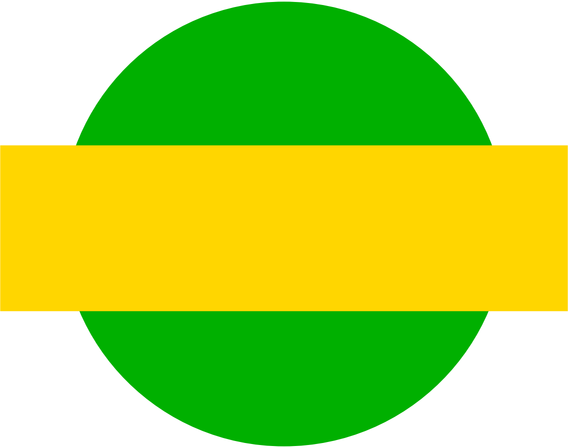
No single disease causes all of these.

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

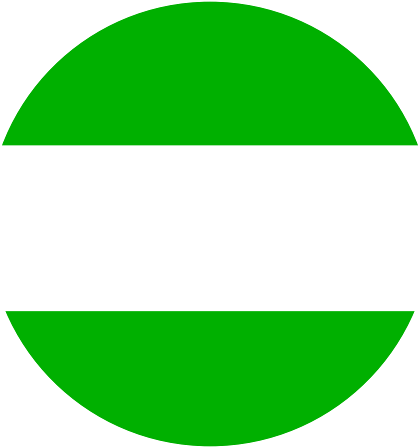
Dr. House explains the symptoms with two simple causes:

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

Visual Recognition

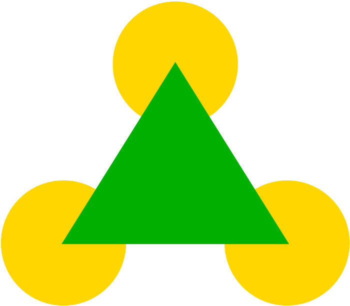


Visual Recognition

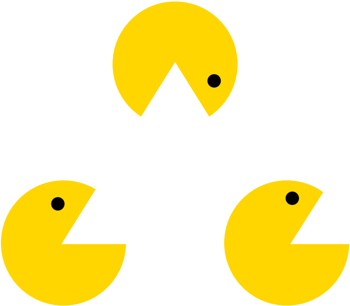


Visual Recognition

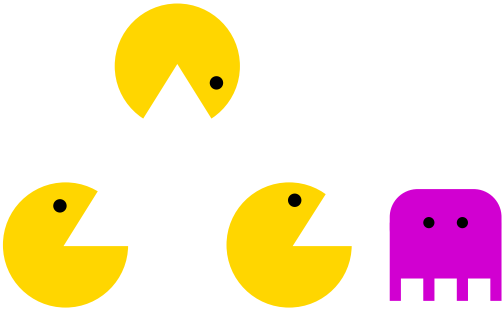
Visual Recognition



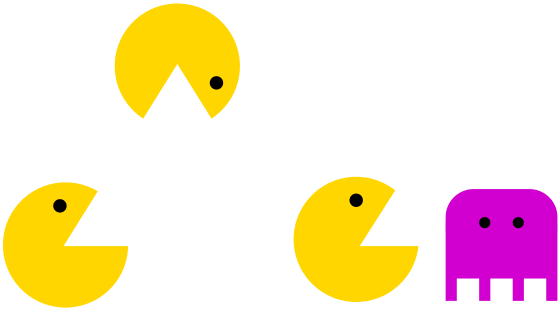
Visual Recognition



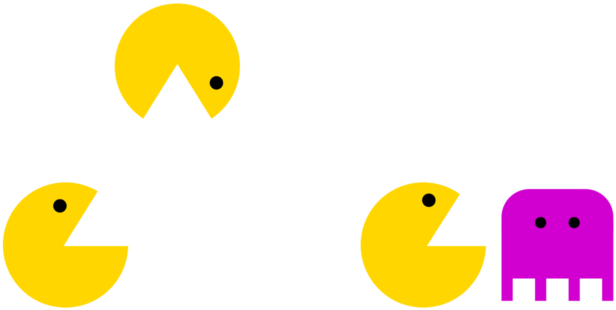
Visual Recognition



Visual Recognition



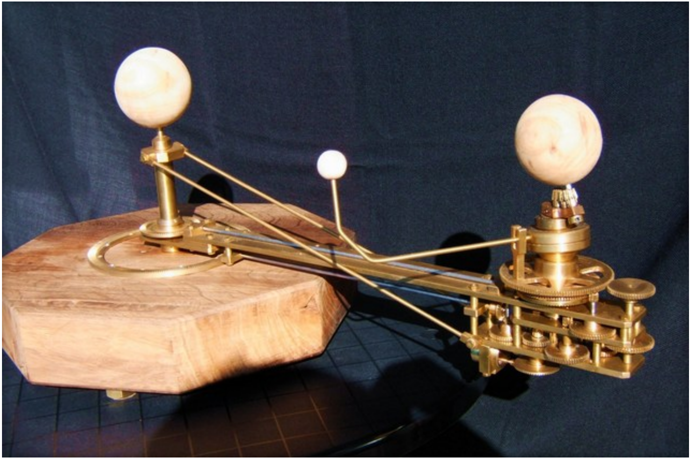
Visual Recognition



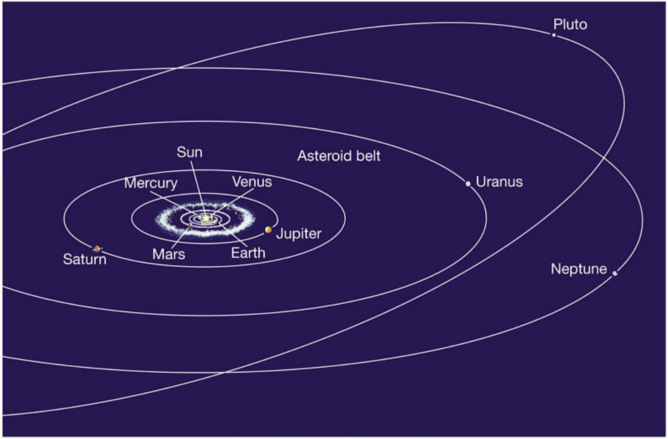
Astronomy


Schema huius præmissæ diuisionis Sphærarum.
COELVM EMPIREVM HABITACVLVM DEI ET OMNIVM ELECTORVM
Decimum Coelum
Primu Mobile
Nonu Coelum
Crystallinum
Octauum
Firmamentum
SATVRNI
IOVIS
MARTIS
SOLIS
VENERIS
MERCVRII
LVNÆ

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.

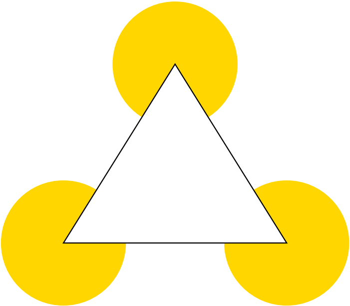
Isaac Newton: "We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances."

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

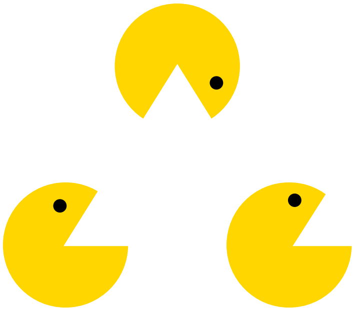
(**Hickam's dictum:** "Patients can have as many diseases as they damn well please.")

Visual Recognition

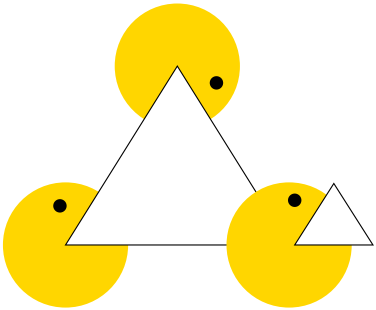
Visual Recognition



Visual Recognition



Visual Recognition



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

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

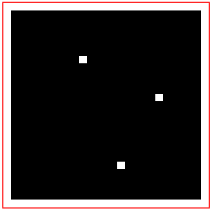
How to encode data *with the help of a hypothesis*?

Encoding Data: Rules & Exceptions

Idea 1: Hypothesis = rule; encode exceptions.

Encoding Data: Rules & Exceptions

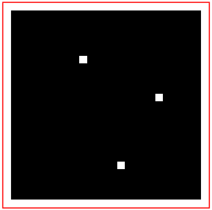
Idea 1: Hypothesis = rule; encode exceptions.



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 $(x_1, y_1), (x_2, y_2), (x_3, y_3)$.

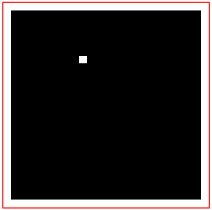
For image of size $n = 625$, there are 2^n different images, and

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

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

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

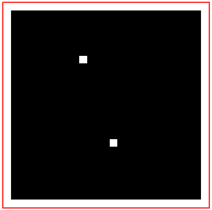
different groups of k exceptions.

$$k = 1 : \binom{n}{1} = 625 \ll 2^{625}.$$

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

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

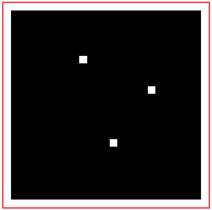
different groups of k exceptions.

$$k = 2 : \binom{n}{2} = 195\,000 \ll 2^{625}.$$

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

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

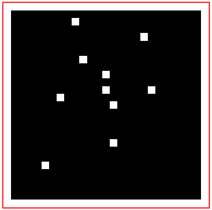
different groups of k exceptions.

$$k = 3 : \binom{n}{3} = 40\,495\,000 \ll 2^{625}.$$

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

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

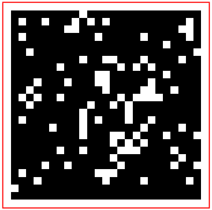
different groups of k exceptions.

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

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

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

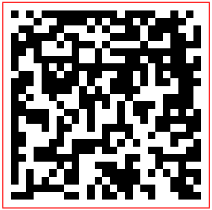
different groups of k exceptions.

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

Codelength $\log_2(n+1) + \log_2 \binom{n}{k} \approx 401$ vs. 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

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

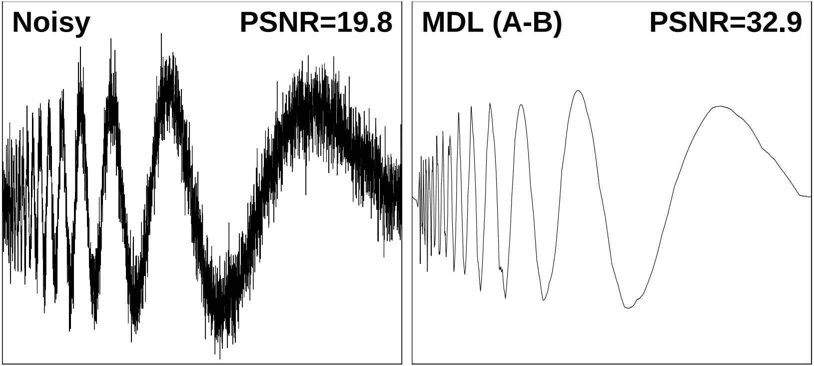
different groups of k exceptions.

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

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

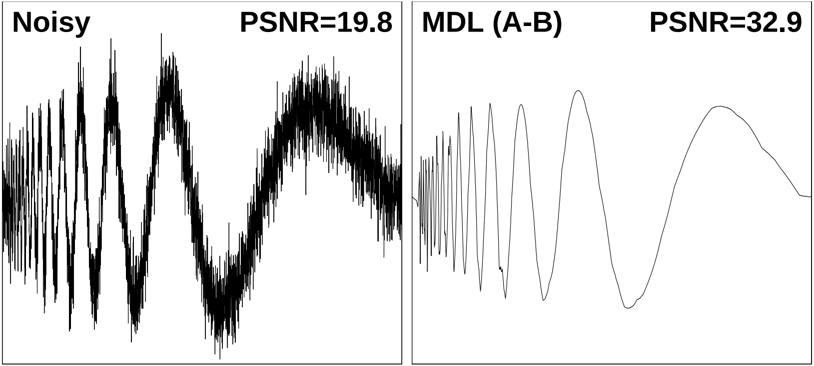
Encoding Data: Probabilistic Models

Idea 2: Hypothesis = probability distribution.



Encoding Data: Probabilistic Models

Idea 2: Hypothesis = probability distribution.



How to encode a distribution?

Two-Part Codes

Let $\mathcal{M} = \{p_\theta \ : \ \theta \in \Theta\}$ be a parametric probabilistic model class, i.e., a set of distributions p_θ indexed by parameter θ.

Two-Part Codes

Let $\mathcal{M} = \{p_\theta \ : \ \theta \in \Theta\}$ be a parametric probabilistic model class, i.e., a set of distributions p_θ indexed by parameter θ.

If the parameter space Θ is discrete, we can construct a (prefix) code $C_1 \ : \ \Theta \to \{0,1\}^*$ which maps each parameter value to a codeword.

Two-Part Codes

Let $\mathcal{M} = \{p_\theta \ : \ \theta \in \Theta\}$ be a parametric probabilistic model class, i.e., a set of distributions p_θ indexed by parameter θ.

If the parameter space Θ is discrete, we can construct a (prefix) code $C_1 \ : \ \Theta \to \{0,1\}^*$ which maps each parameter value to a codeword.

For each distribution p_θ there is a prefix code $C_\theta \ : \ \mathcal{D} \to \{0,1\}^*$ where $D \in \mathcal{D}$ is a data-set to be encoded, such that the codeword lengths satisty

$$\ell_\theta(D) \approx \log_2 \frac{1}{p_\theta(D)} \quad .$$

Two-Part Codes

Let $\mathcal{M} = \{p_\theta : \theta \in \Theta\}$ be a parametric probabilistic model class, i.e., a set of distributions p_θ indexed by parameter θ.

If the parameter space Θ is discrete, we can construct a (prefix) code $C_1 : \Theta \to \{0,1\}^*$ which maps each parameter value to a codeword.

For each distribution p_θ there is a prefix code $C_\theta : \mathcal{D} \to \{0,1\}^*$ where $D \in \mathcal{D}$ is a data-set to be encoded, such that the codeword lengths satisty

$$\ell_\theta(D) \approx \log_2 \frac{1}{p_\theta(D)} \quad .$$

Using parameter value θ, the total codelength becomes ($\approx$)

$$\ell_1(\theta) + \log_2 \frac{1}{p_\theta(D)} \quad .$$

Two-Part Codes

The parameter value minimizing the codelength is given by the **maximum likelihood** parameter $\hat{\theta}$:

$$\min_{\theta \in \Theta} \log_2 \frac{1}{p_\theta(D)} = \log_2 \frac{1}{\max_{\theta \in \Theta} p_\theta(D)} = \log_2 \frac{1}{p_{\hat{\theta}}(D)} \quad .$$

Two-Part Codes

The parameter value minimizing the codelength is given by the **maximum likelihood** parameter $\hat{\theta}$:

$$\min_{\theta \in \Theta} \log_2 \frac{1}{p_\theta(D)} = \log_2 \frac{1}{\max_{\theta \in \Theta} p_\theta(D)} = \log_2 \frac{1}{p_{\hat{\theta}}(D)} \quad .$$

It could of course be that $\ell_1(\hat{\theta})$ is so large that some other parameter value gives a shorter total codelength.

Multi-Part Codes

If there are more than one model classes, $\mathcal{M}_1, \mathcal{M}_2, \ldots$ it is possible to construct **multi-part codes** where the parts are

1. Encoding of the model class index: $C_0(i)$, $i \in \mathbb{N}$.

Multi-Part Codes

If there are more than one model classes, $\mathcal{M}_1, \mathcal{M}_2, \ldots$ it is possible to construct **multi-part codes** where the parts are

1. Encoding of the model class index: $C_0(i)$, $i \in \mathbb{N}$.
2. Encoding of the parameter (vector): $C_i(\theta)$, $\theta \in \Theta_i$.

Multi-Part Codes

If there are more than one model classes, $\mathcal{M}_1, \mathcal{M}_2, \ldots$ it is possible to construct **multi-part codes** where the parts are

1. Encoding of the model class index: $C_0(i)$, $i \in \mathbb{N}$.
2. Encoding of the parameter (vector): $C_i(\theta)$, $\theta \in \Theta_i$.
3. Encoding of the data: $C_\theta(D)$, $D \in \mathcal{D}$.

Multi-Part Codes

If there are more than one model classes, $\mathcal{M}_1, \mathcal{M}_2, \ldots$ it is possible to construct **multi-part codes** where the parts are

1. Encoding of the model class index: $C_0(i)$, $i \in \mathbb{N}$.
2. Encoding of the parameter (vector): $C_i(\theta)$, $\theta \in \Theta_i$.
3. Encoding of the data: $C_\theta(D)$, $D \in \mathcal{D}$.

For instance, the models could be polynomials with different degrees, the parameters are the coefficients

$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \ldots + \theta_k x^k \quad .$$

Multi-Part Codes

If there are more than one model classes, $\mathcal{M}_1, \mathcal{M}_2, \ldots$ it is possible to construct **multi-part codes** where the parts are

1. Encoding of the model class index: $C_0(i)$, $i \in \mathbb{N}$.
2. Encoding of the parameter (vector): $C_i(\theta)$, $\theta \in \Theta_i$.
3. Encoding of the data: $C_\theta(D)$, $D \in \mathcal{D}$.

For instance, the models could be polynomials with different degrees, the parameters are the coefficients

$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \ldots + \theta_k x^k \quad .$$

The more complex the model class (the higher the degree), the better it fits the data but the longer the second part $C_i(\theta)$ becomes.

Polynomials

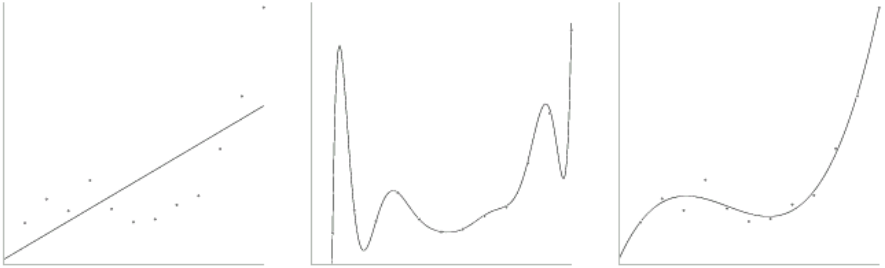


Figure 1: A simple (1.1), complex (1.2) and a trade-off (3rd degree) polynomial.

Continuous Parameters

What if the parameters are continuous (like polynomial coefficients)? How to encode continuous values?

Continuous Parameters

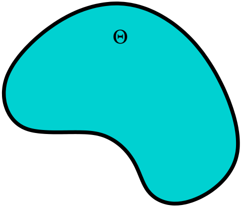
What if the parameters are continuous (like polynomial coefficients)? How to encode continuous values?

Solution: Quantization. Choose a discrete subset of points, $\theta^{(1)}, \theta^{(2)}, \ldots$, and use only them.

Continuous Parameters

What if the parameters are continuous (like polynomial coefficients)? How to encode continuous values?

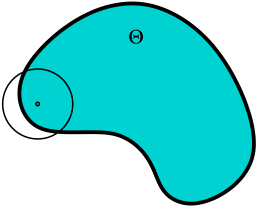
Solution: Quantization. Choose a discrete subset of points, $\theta^{(1)}, \theta^{(2)}, \ldots$, and use only them.



Continuous Parameters

What if the parameters are continuous (like polynomial coefficients)? How to encode continuous values?

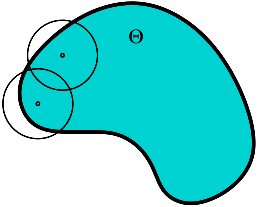
Solution: Quantization. Choose a discrete subset of points, $\theta^{(1)}, \theta^{(2)}, \ldots$, and use only them.



Continuous Parameters

What if the parameters are continuous (like polynomial coefficients)? How to encode continuous values?

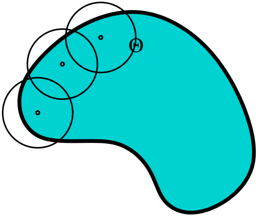
Solution: Quantization. Choose a discrete subset of points, $\theta^{(1)}, \theta^{(2)}, \ldots$, and use only them.



Continuous Parameters

What if the parameters are continuous (like polynomial coefficients)? How to encode continuous values?

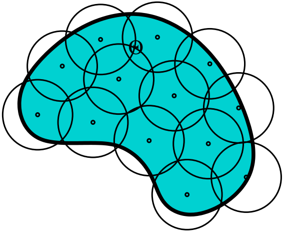
Solution: Quantization. Choose a discrete subset of points, $\theta^{(1)}, \theta^{(2)}, \ldots$, and use only them.



Continuous Parameters

What if the parameters are continuous (like polynomial coefficients)? How to encode continuous values?

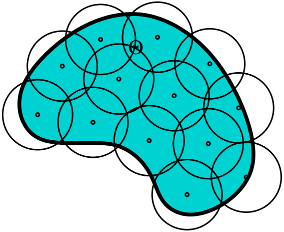
Solution: Quantization. Choose a discrete subset of points, $\theta^{(1)}, \theta^{(2)}, \ldots$, and use only them.



Continuous Parameters

What if the parameters are continuous (like polynomial coefficients)? How to encode continuous values?

Solution: Quantization. Choose a discrete subset of points, $\theta^{(1)}, \theta^{(2)}, \ldots$, and use only them.

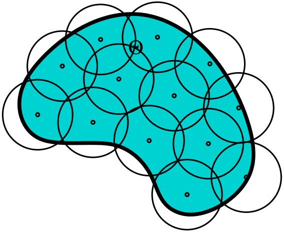


If the points are sufficiently *dense* (in a codelength sense) then the codelength for data is still almost as short as $\min_{\theta \in \Theta} \ell_\theta(D)$.

Continuous Parameters

What if the parameters are continuous (like polynomial coefficients)? How to encode continuous values?

Solution: Quantization. Choose a discrete subset of points, $\theta^{(1)}, \theta^{(2)}, \ldots$, and use only them.



Information Geometry!

If the points are sufficiently *dense* (in a codelength sense) then the codelength for data is still almost as short as $\min_{\theta \in \Theta} \ell_\theta(D)$.

About Quantization

How many points should there be in the subset $\theta^{(1)}, \theta^{(2)}, \ldots$?

About Quantization

How many points should there be in the subset $\theta^{(1)}, \theta^{(2)}, \ldots$?

Intuition: Estimation accuracy of order $\frac{1}{\sqrt{n}}$.

About Quantization

How many points should there be in the subset $\theta^{(1)}, \theta^{(2)}, \ldots$?

Intuition: Estimation accuracy of order $\dfrac{1}{\sqrt{n}}$.

Theorem

Optimal quantization accuracy is of order $\dfrac{1}{\sqrt{n}}$.

$\Rightarrow$ number of points $\approx \sqrt{n}^k = n^{k/2}$, where $k = \dim(\Theta)$.

About Quantization

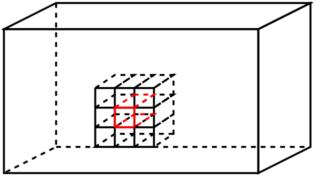
How many points should there be in the subset $\theta^{(1)}, \theta^{(2)}, \ldots$?

Intuition: Estimation accuracy of order $\dfrac{1}{\sqrt{n}}$.

Theorem

Optimal quantization accuracy is of order $\dfrac{1}{\sqrt{n}}$.

$\Rightarrow$ number of points $\approx \sqrt{n}^k = n^{k/2}$, where $k = \dim(\Theta)$.



About Quantization

How many points should there be in the subset $\theta^{(1)}, \theta^{(2)}, \ldots$?

Intuition: Estimation accuracy of order $\frac{1}{\sqrt{n}}$.

Theorem

Optimal quantization accuracy is of order $\frac{1}{\sqrt{n}}$.

$\Rightarrow$ number of points $\approx \sqrt{n}^k = n^{k/2}$, where $k = \dim(\Theta)$.

The codelength for the quantized parameters becomes

$$\ell(\theta^q) \approx \log_2 n^{k/2} = \frac{k}{2} \log_2 n \ .$$

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:

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

This gives the total codelength formula:

"Steam MDL"

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

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.

Next week:

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