Bayesian multimodeling: minimum description length

MIPT

2022

Occam's razor

- William of Ockham: «Plurality must never be posited without necessity».

Occam's razor

- William of Ockham: «Plurality must never be posited without necessity».
- Modern interpretation: entities should not be multiplied beyond necessity.
- Paul Dirac: «A theory with mathematical beauty is more likely to be correct than an ugly one that fits some experimental data.»
- Albert Einstein: «Everything Should Be Made as Simple as Possible, But Not Simpler»

When Occam's razor does not work

Occam's razor is an empricial rule for sorting hypothesis during research.
It can be wrong:

- Ernst Mach: molecules are a fictitious construct, as they are not observable.

Minimum description length

Task

Given a string: 001011001011001011... 001011, wheere the pattern 001011 repeats 100500 times.
How can we describe this string?

- `s == "001011...0010110010110010011")`
- `s == (''.join('001011' for _ in range(100500))`

Kolmogorov complexity

Definition

Given a computable partially defined mapping from a set of binary words into itself:

$$T : \{0,1\}^* \to \{0,1\}^*.$$

Kolmogorov complexity of the binary string x is a minimal descriptiong length w.r.t. T:

$$K_T(x) = \min_{f \in \{0,1\}^*} \{|f| : T(f) = x\},$$

Kolmogorov complexity

Generally, Kolmogorov complexity is incomputable.

Definition

Given a computable partially defined mapping from a set of binary words into itself:

$$T : \{0,1\}^* \to \{0,1\}^*.$$

Kolmogorov complexity of the binary string x is a minimal descriptiong length w.r.t. T:

$$K_T(x) = \min_{f \in \{0,1\}^*} \{|f| : T(f) = x\},$$

Discrete distribution entropy

Definition

Given a discrete variable x with probability p and values $x_1, \ldots, x_n$, An entropy of x is:

$$H(x) = -\sum_{i=1}^{n} p(x = x_i) \log \; p(x = x_i).$$

Discrete distribution entropy

Definition

Given a discrete variable x with probability p and values $x_1, \ldots, x_n$, An entropy of x is:

$$H(x) = -\sum_{i=1}^{n} p(x = x_i) \log \; p(x = x_i).$$

- interpretation: measure of disorder in the distribution;
- maximal at uniform distribution;
- minimal at a disribution with concentration only at one event ($x_i = 1, x_j = 0, i \neq j$).

Discrete distribution entropy

Definition

Given a discrete variable x with probability p and values $x_1, \ldots, x_n$, An entropy of x is:

$$H(x) = -\sum_{i=1}^{n} p(x = x_i) \log \; p(x = x_i).$$

- interpretation: measure of disorder in the distribution;
- maximal at uniform distribution;
- minimal at a disribution with concentration only at one event ($x_i = 1, x_j = 0, i \neq j$).
- **relation to Kolmogorov complexity:**

$$K(x) \leq H(x) + O(\log n)$$

for binary string with length n.

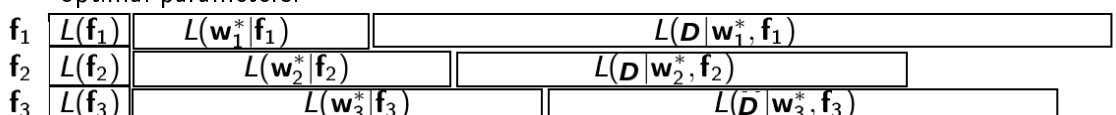
Minimum description length principle

$$\mathrm{MDL}(\mathsf{f}, \mathfrak{D}) = L(\mathsf{f}) + L(\mathfrak{D}|\mathsf{f}),$$

where f is a model, $\mathfrak{D}$ is a dataset, L is a description length in bits.

$$\mathrm{MDL}(\mathsf{f}, \mathfrak{D}) \sim L(\mathsf{f}) + L(\mathsf{w}^*|\mathsf{f}) + L(\mathfrak{D}|\mathsf{w}^*, \mathsf{f}),$$

w^* — optimal parameters.



MDL: example

Задача

Given a string: 001011001011001011... 001011, wheere the pattern 001011 repeats 100500 times.
How can we describe this string?

- `s == "001011...0010110010110010011"`)
- `s == (''.join('001011' for _ in range(100500))`
- `import re; re.match('(001011){100500}')`
- $L(\mathsf{f}_1) = 0,\ L(\mathfrak{D}|\mathsf{f}_1) = 100505;$
- $L(\mathsf{f}_2) = 0,\ L(\mathfrak{D}|\mathsf{f}_2) = 45;$
- $L(\mathsf{f}_3) \gg 0,\ L(\mathfrak{D}|\mathsf{f}_3) = 38;$

MDL and Kolmogorov complexity

Kolmogorov complexity is a minimum descriptiong length for the dataset, described by a given programming language.

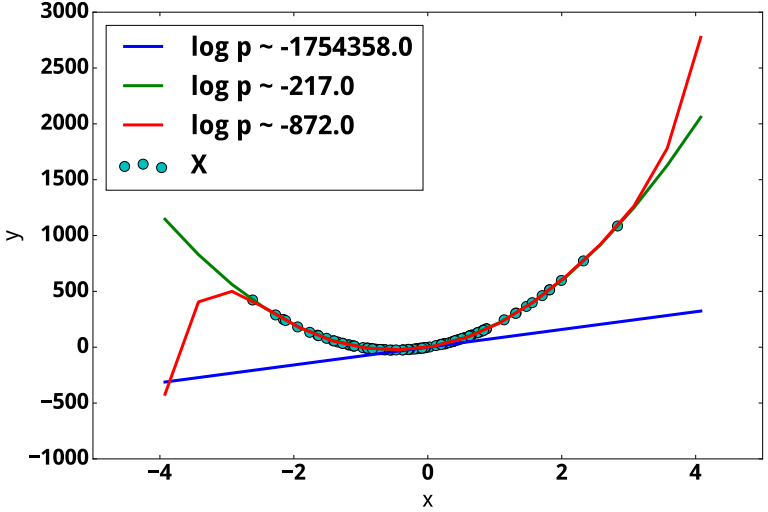
Invariance theorem

Given any description language L, the optimal description language is at least as efficient as L, with some constant overhead.

Difference from MDL:

- Incomputable.
- Code length depends on the language. For small sample size the invariance theorem can give poor results.

Probability coding



MDL + probability

Model selection problem can be viewed as a problem of information transmission, from encoder to decoder.

Given a dataset $\mathrm{X}, x \in \mathrm{X}$.

- Encoder encodes information about the dataset X with some code f and transmits it to the decoder.
- Decoder decodes the infomration f(X) ans restores the data X (with loss).
- Problem is to find optimal coding method for x
- Code length: $-\log\ p(x)$

Quality criterion is *Regret*:

$$R(x) = -\log P(x) + \min_{\mathrm{f} \in \mathfrak{F}}(\log P(x|\mathrm{f})).$$

It shows the difference between the length of the real code $\log P(x)$ for x and the best code from the set of codes $\mathfrak{F}$.

Regret for datasets with parameterized distribution:

$$R(\mathrm{X}) = \max_{x \in \mathrm{X}}(-\log P(x) + \min_{\mathrm{w}}(\log P(x|\mathrm{w})).$$

MDL and Laplace approximation

Statement

Let the likelihood function $p(\mathrm{X}|\mathrm{w},\mathrm{f})$ be from an exponential family of distributions:

$$p(x|\mathrm{w},\mathrm{f}) = h(x)g(\boldsymbol{\eta})\exp\left(\boldsymbol{\eta}\cdot\mathrm{T}(x)\right),$$

where h, g, T are some functions, $\boldsymbol{\eta}$ is a distribution parameter.
Let prior be Jeffreys prior:

$$p(\mathrm{w}|\mathrm{f}) = \frac{\sqrt{I}}{\int_w \sqrt{I(w)}},$$

Then regret and evidence differ only by some constant when $\mathrm{w}\to\infty$:

$$\lim_{\mathrm{w}\to\infty}\left(R(\mathrm{X}) - \int_{\mathrm{w}} p(\mathrm{X}|\mathrm{w},\mathrm{f})p(\mathrm{w}|\mathrm{f})d\mathrm{w}\right) = \mathrm{Const}.$$

MDL vs. Evidence

Evidence	**MDL**
Uses prior knowledge	Independent from prior
Uses data generation hypothesis	Minimizes description length

References

- MacKay D. J. C., Mac Kay D. J. C. Information theory, inference and learning algorithms. – Cambridge university press, 2003.
- Grunwald P. A tutorial introduction to the minimum description length principle //arXiv preprint math/0406077. – 2004.
- Успенский В., Шень А., Верещагин Н. Колмогоровская сложность и алгоритмическая случайность. – Litres, 2017
- Grunwald P., Vitányi P. Shannon information and Kolmogorov complexity //arXiv preprint cs/0410002. – 2004.
- Vereshchagin N. K., Vitányi P. M. B. Kolmogorov's structure functions and model selection //IEEE Transactions on Information Theory. – 2004. – Т. 50. – №. 12. – С. 3265-3290.
- Штарьков Ю. М. Универсальное последовательное кодирование отдельных сообщений //Проблемы передачи информации. – 1987. – Т. 23. – №. 3. – С. 3-17.
- When Occam's razor does not work: https://hsm.stackexchange.com/questions/26/was-occam-s-razor-ever-wrong