

# Байесовское мультимоделирование: принцип минимальной длины описания

Московский Физико-Технический Институт

2021

# Бритва Оккама



- Уильям из Оккама: «Что может быть сделано на основе меньшего числа, не следует делать, исходя из большего» и «Многообразие не следует предполагать без необходимости».

# Бритва Оккама



- Уильям из Оккама: «Что может быть сделано на основе меньшего числа, не следует делать, исходя из большего» и «Многообразие не следует предполагать без необходимости».
- Современная интерпретация: Не следует множить сущее без необходимости.
- Поль Дирак: A theory with mathematical beauty is more likely to be correct than an ugly one that fits some experimental data.
- Альберт Эйнштейн: Всё следует упрощать до тех пор, пока это возможно, но не более того.

# Когда бритва Оккама не работает

Бритва Оккама — эмпирическое правило, предлагающее упорядочивание гипотез при исследовании.

Это правило может быть неверным:

- Эрнст Мах: молекулы являются мыслительными конструктами, т.к. их существование не может быть проверено прямым наблюдением.

# Длина описания программы

## Задача

Задана строка: 001011001011001011... 001011, где повторение строки 001011 производится 100500 раз.

Каким способом лучше всего ее описать?

- `s == "001011...001011001011001011"`
- `s == (''.join('001011' for _ in range(100500)))`

# Колмогоровская сложность

## Определение

Пусть задано вычислимое частично определенное отображение из множества бинарных слов в себя:

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

Колмогоровской сложностью бинарной строки  $x$  назовем минимальную длину описания относительно  $T$ :

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

# Колмогоровская сложность

В общем виде, Колмогоровская сложность невычислима.

## Определение

Пусть задано вычислимое частично определенное отображение из множества бинарных слов в себя:

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

Колмогоровской сложностью бинарной строки  $x$  назовем минимальную длину описания относительно  $T$ :

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

# Энтропия дискретного распределения

## Определение

Пусть задана дискретная случайная величина  $x$  с вероятностным распределением  $p$ , принимающая значения  $x_1, \dots, x_n$ . Энтропией распределения случайной величины  $x$  назовем:

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



# Энтропия дискретного распределения

## Определение

Пусть задана дискретная случайная величина  $x$  с вероятностным распределением  $p$ , принимающая значения  $x_1, \dots, x_n$ , Энтропией распределения случайной величины  $x$  назовем:

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

- интерпретация: мера беспорядка в распределении;
- максимум: равномерное распределение;
- минимум: распределение, сконцентрированное в одном событии ( $x_i = 1, x_j = 0, i \neq j$ ).

# Энтропия дискретного распределения

## Определение

Пусть задана дискретная случайная величина  $x$  с вероятностным распределением  $p$ , принимающая значения  $x_1, \dots, x_n$ . Энтропией распределения случайной величины  $x$  назовем:

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

- интерпретация: мера беспорядка в распределении;
- максимум: равномерное распределение;
- минимум: распределение, сконцентрированное в одном событии ( $x_i = 1, x_j = 0, i \neq j$ ).
- связь с Колмогоровской сложностью:

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

для бинарных строк длины  $n$ .

# Принцип минимальной длины описания

$$\text{MDL}(\mathbf{f}, \mathcal{D}) = L(\mathbf{f}) + L(\mathcal{D}|\mathbf{f}),$$

где  $\mathbf{f}$  — модель,  $\mathcal{D}$  — выборка,  $L$  — длина описания в битах.

$$\text{MDL}(\mathbf{f}, \mathcal{D}) \sim L(\mathbf{f}) + L(\mathbf{w}^*|\mathbf{f}) + L(\mathcal{D}|\mathbf{w}^*, \mathbf{f}),$$

$\mathbf{w}^*$  — оптимальные параметры модели.

$\mathbf{f}_1$	$L(\mathbf{f}_1)$	$L(\mathbf{w}_1^* \mathbf{f}_1)$	$L(\mathcal{D} \mathbf{w}_1^*, \mathbf{f}_1)$
$\mathbf{f}_2$	$L(\mathbf{f}_2)$	$L(\mathbf{w}_2^* \mathbf{f}_2)$	$L(\mathcal{D} \mathbf{w}_2^*, \mathbf{f}_2)$
$\mathbf{f}_3$	$L(\mathbf{f}_3)$	$L(\mathbf{w}_3^* \mathbf{f}_3)$	$L(\mathcal{D} \mathbf{w}_3^*, \mathbf{f}_3)$

# MDL: пример

## Задача

Задана строка: 001011001011001011... 001011, где повторение строки 001011 производится 100500 раз.

Каким способом лучше всего ее описать?

- `s == "001011...001011001011001011"`
- `s == (''.join('001011' for _ in range(100500)))`
- `import re; re.match('(001011){100500}')`
- $L(\mathbf{f}_1) = 0, L(\mathcal{D}|\mathbf{f}_1) = 100505;$
- $L(\mathbf{f}_2) = 0, L(\mathcal{D}|\mathbf{f}_2) = 45;$
- $L(\mathbf{f}_3) \gg 0, L(\mathcal{D}|\mathbf{f}_3) = 38;$

# MDL и Колмогоровская сложность

**Колмогоровская сложность** — длина минимального кода для выборки на предварительно заданном языке.

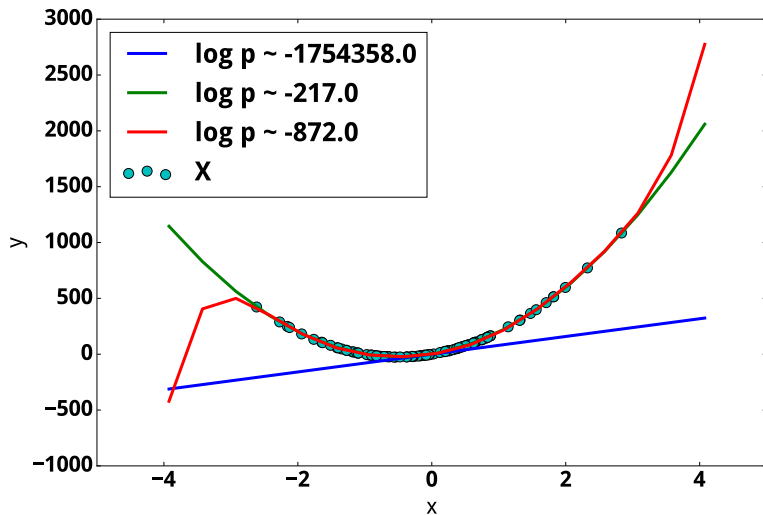
## **Теорема инвариантности**

Для двух сводимых по Тьюрингу языков колмогоровская сложность отличается не более чем на константу, не зависящую от мощности выборки.

## **Отличия от MDL:**

- Колмогоровская сложность невычислима.
- Длина кода может зависеть от выбранного языка. Для небольших выборок теорема инвариантности не дает адекватных результатов.

# Задача вероятностного кодирования: полиномы



# Вероятностный MDL

Задача выбора модели — задача передачи информации от кодировщика декодировщику.

Задана выборка  $\mathbf{X}$ ,  $x \in \mathbf{X}$ .

- Кодировщик кодирует информацию о выборке  $\mathbf{X}$  с помощью некоторого кода  $\mathbf{f}$  и передает ее декодировщику.
- Декодировщик декодирует код  $\mathbf{f}(\mathbf{X})$ , полученный от кодировщика и восстанавливает исходную выборку  $\mathbf{X}$  (возможно, с некоторой потерей информации).
- Требуется выбрать оптимальный способ кодирования  $x$
- Длина кода:  $-\log p(x)$

Критерий качества вероятностного кодирования с помощью смеси кодов:

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

Регрет характеризует разницу между длиной рассматриваемого  $\log P(x)$  кода для  $x$  в сравнении с наилучшим кодом из некоторого множества  $\mathfrak{F}$ .

Регрет для выборки с использованием параметрического распределения:

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

# MDL и аппроксимация Лапласа

## Утверждение

Пусть правдоподобие  $p(\mathbf{X}|\mathbf{w}, \mathbf{f})$  соответствует экспоненциальному семейству распределений, т.е.

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

где  $h, g, \mathbf{T}$  — некоторые функции,  $\boldsymbol{\eta}$  — некоторый параметр распределения.

Пусть в качестве априорного распределения выступает распределение Джеффриса:

$$p(\mathbf{w}|\mathbf{f}) = \frac{\sqrt{I}}{\int_{\mathbf{w}} \sqrt{I(\mathbf{w})}}, \text{ где } I \text{ — определить матрицы Фишера:}$$

Тогда при  $\mathbf{w} \rightarrow \infty$  регрет отличается от Evidence на константу:

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



# MDL и Evidence

<b>Evidence</b>	<b>MDL</b>
Использует априорные знания	Независима от априорных знаний
Основывается на гипотезе о порождении выборки вне зависимости от их природы	Минимизирует длину описания выборки

## Литература и прочие ресурсы

- 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.
- Когда не работают бритва Оккама:  
<https://hsm.stackexchange.com/questions/26/was-occam-s-razor-ever-wrong>