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 правило для упорядочивания 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:

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

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:

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

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 распределния

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 распределния

Определение

Пусть задана дискретная случайная величина x с вероятностным распределением p, принимающая значения $x_1, \ldots, x_n$, 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 распределния

Определение

Пусть задана дискретная случайная величина x с вероятностным распределением p, принимающая значения $x_1, \ldots, x_n$, 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.

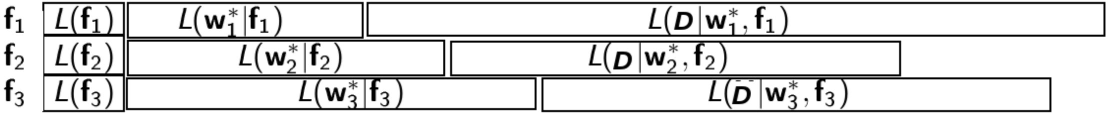
Принцип минимальной длины описания

$$\text{MDL}(\boldsymbol{f}, \mathfrak{D}) = L(\boldsymbol{f}) + L(\mathfrak{D}|\boldsymbol{f}),$$

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

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

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



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}')`
- $L(\boldsymbol{f}_1) = 0$, $L(\mathfrak{D}|\boldsymbol{f}_1) = 100505$;
- $L(\boldsymbol{f}_2) = 0$, $L(\mathfrak{D}|\boldsymbol{f}_2) = 45$;
- $L(\boldsymbol{f}_3) \gg 0$, $L(\mathfrak{D}|\boldsymbol{f}_3) = 38$;

MDL и Колмогоровская сложность

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

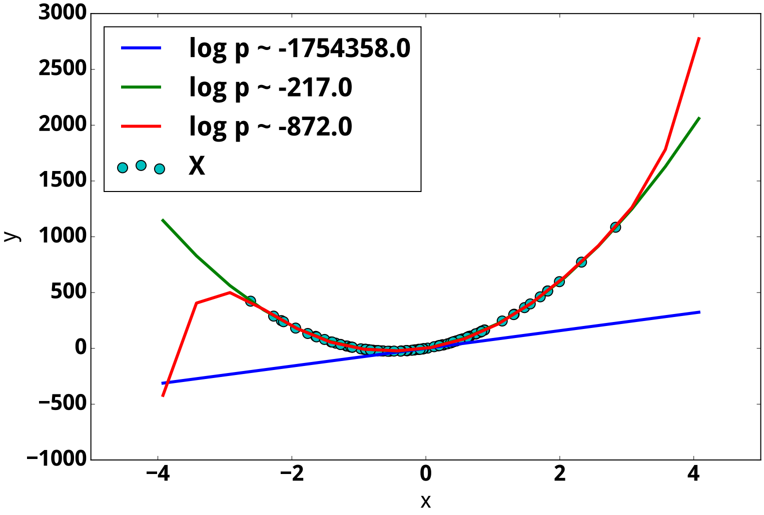
Теорема инвариантности

Для двух сводимых по Тьюрингу языков колмогоровская сложность отличается не более чем на константу, не зависяющую 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 $\boldsymbol{X}, x \in \boldsymbol{X}$.

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

Критерий качества вероятностного кодирования с помощью смеси кодов:

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

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:

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

MDL и аппроксимация Лапласа

Утверждение

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

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

где $h, g, \boldsymbol{T}$ — некоторые функции, $\boldsymbol{\eta}$ — некоторый параметр распределения.
Пусть в качестве априорного распределения выступает распределение Джеффриса:

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

Тогда при $\boldsymbol{w} \to \infty$ регрет отличается от Evidence на константу:

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

MDL и Evidence

Evidence	MDL
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 работат бритва Оккама: https://hsm.stackexchange.com/questions/26/was-occam-s-razor-ever-wrong