0.1 Колмогоровская сложность моделей

TODO: обозначения

Одним из фундаментальных способов определить сложность произвольного математического объекта является колмогоровская сложность. Ниже представлено формальеное определине колмогоровской сложности и основные ее соуйства.

Определение 1. *Способом описания назовем вычислимое частинчо определенное отображение из множества бинарных слов в себя:*

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

Определение 2. *Пусть задан некоторый способ описания D. Колмогоровской сложностью бинарной строки x назовем минимальную длину описания относительно D:*

$$K_D(x) = \min_{p \in \{0,1\}^*} \{|p| : D(p) = x\},$$

Перечислим некоторые свойства колмогоровской сложности [1].

Независимости от способа написания.

Теорема (без доказательств) 1. *Пусть заданы отображения D_1, D_2, такие что существуют константы c_1, c_2 такие что для любого другого отображения D' и для любой строки x:*

$$K_{D_1}(x) \le K_{D'}(x) + c_1, \quad K_{D_2}(x) \le K_{D'}(x) + c_2.$$

Тогда $K_{D_1}(x) = K_{D_2}(x) + O(1)$.

Т.к. колмогоровская сложность независима от способа написания, зафиксируем некоторый способ описания D и положим $K(x) = K_D(x)$.

Невычислимость

Теорема (без доказательств) 2. *Пусть k — произвольная вычислимая функция. Если $k(x) \le K(x)$ для всех x, для которых определена k, то k — ограничена.*

Из теоремы следует, что колмогровская сложность в общем случае невычислима: любая оценка сложности будет ограничена, и потому тривиальна.

Условная сложность Обобщим понятие колмогоровскорй сложности на случай двух бинарных строк.

Определение 3. *Пусть задано вычисилмое и частично определенное отображение из декартового произведения двух множеств бинарных слов в себя:*

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

Условной колмогоровской сложностью бинарной строки y при условии x назовем минимальную длину описания относительно D:

$$K_D(y|x) = \min_{p \in \{0,1\}^*} \{|p| : D(p, y) = x\},$$

Оценка условной Колмогоровской сложности [1]

$$K(x, y) \le K(x) + K(y|x) + O(\log K(x, y)).$$

Разность $I(x : y) = K(y) - K(y|x)$ задает количество информации в x об объекте y. **Количество информации в паре x,y симметрично с точностью до константы**:

$$I(x : y) = I(y : x) + O(\log K(x, y)).$$

Отметим, что схожими свойствами обладает взаимная информация, определение которой дано ниже.

Определение 4. *Пусть задана дискретная случайная величина x с вероятностным распределением p, принимающая значения $x_1, \dots, x_n$, Энтропией распределения случайной величины x назовем:*

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

Взаимной информацией I двух случайных величин x, y назовем следующее выражение:

$$I(x, y) = H(x) - H(x|y), \quad H(x) = -\sum_i p_x(x_i) \log p_x(x_i)$$

$$I(x, y) = I(y, x).$$

Таким образом, свойства количества информации $I(x : y)$ и взаимной информации, во многом совпадают. Докажем теорему о связи колмогоровской сложности и энтропии распределения, подытоживающую связь этих двух математических объектов.

Теорема 1. *[2] Пусть задано семейство частично-определенных отображений $\mathfrak{D} = \{D : \{0,1\}^* \to \{0,1\}^*\}$, такое что для любого отображения $D \in \mathfrak{D}$ и элемента из области определения D в области определения не содержится префиксов этого элемента. Пусть f — вычислимая функция вероятности на пространстве бинарных векторов произвольной длины. Тогда*

$$0 \leq \big(\mathsf{E}_f K(X) - H(x)\big) \leq K(f) + O(1). \tag{0.1}$$

Для доказательства предварительно приведем две теоремы из [2] без доказательства.

Теорема (без доказательств) 3. *Пусть задано семейство частично-определенных отображений $\mathfrak{D} = \{D : \{0,1\}^* \to \{0,1\}^*\}$, такое что для любого отображения $D \in \mathfrak{D}$ и элемента из области определения D в области определения не содержится префиксов этого элемента.*

Тогда для минимальной средней длины описания слова:

$$L = \min_{D \in \mathfrak{D}} \sum_i |D(x_i)| p(x = x_i)$$

справедливо неравенсвто:

$$H(x) \leq L \leq H(x) + 1.$$

Теорема (без доказательств) 4. *Пусть f — вычислимое распределение на бинарных словах. Тогда справедлива следующие оценки:*

$$2^{K(f) \pm O(1) - K(x)} \geq f(x),$$

где $O(1)$ — длина некоторой программы, не зависящей от f, x.

Перейдем к доказательству основной теоремы.

Доказательство. Т.к. $K(X)$ — это длина кода для x, то по теореме 3:

$$H(X) \leq L \leq \mathsf{E}_f K(X).$$

Таким образом левая часть неравенства (**??**) доказана.

По теореме 4:

$$f(x) \leq 2^{K(f) \pm O(1) - K(x)}.$$

Тогда

$$\log \frac{1}{f(x)} \geq K(f) - O(1) - K(x) :$$

Посчитаем матожидание данной величины по всем x:

$$H(x) \geq \sum_x f(x)K(f) - \sum_x O(1) - \sum_x K(x).$$

Пользуясь тем, что $\sum_x f(x) = 1$ получим итоговую формулу для правой части неравенства:

$$H(x) + O(1) + K(x) \geq \sum_x f(x)K(f),$$

что и т.д. □

0.2 Колмогоровская сложность и принцип минимальной длины описания

Рассмотрим задачу выбора модели для заданной выборки. Будем полагать что заданная выборка описывается в виде некоторй бинарной строки x. В дальнейшем будем отождествлять выборки и ее бинарное описание x.

Для этого рассмотрим частный случай колмогоровской сложности, называемый префиксной колмогоровской сложностью. Эта сложность задается машиной Тьюринга специального вида, имеющей две ленты: однонаправленную ленту для чтения и двунаправленную рабочую ленту. Будем полагать что машина Тьюринга T останавливается на p с выводом x: $T(p) = x$, если вся запись p осталась слева от читающей каретки, x осталась слева от пишушщей каретки и T остановлена.

Определение 5. *Префиксная Колмогоровская сложность:*

$$K(x) = \min_{p\in\{0,1\}^*, i\in\mathcal{N}} \{|i| + |p| : T_i(p) = x\},$$

где $|i|$ — длина описания i-й префиксной машины Тьюринга.

Задачу выбора модели для выборки можно рассматривать как задачу нахождения префиксной колмогоровской сложности для выборки. В случае, если модель является дискриминативной, то вместо колмогоровсокй сложности можно использовать условную колмогоровскую сложность.Т.к. колмогоровская сложность невычислима, рассмотрим упрощенный подход к выбору модели: вместо колмогоровской сложности строки x будем искать некоторое множество S, в которое входит x, и чья сложность описания при помощи машины Тьюринга невелика. Таким образом, мы сможем найти "хорошую" машину Тьюрингу не для конкретной строки, а для некоторого семейства строк (или выборок), обладающих некоторыми общими свойствами или регулярностью.

Определение 6. *Сложностью конечного множества S назовем следующей величину:*

$$K(S) = \min_{p\in\{0,1\}^*, i\in\mathcal{N}} \{|i| + |p| : T_i(p) \text{ перечисляет все элементы множества } S\}.$$

Вместо задачи нахождения минимальной сложности для выборки x будем искать множество S, которое описывается некоторой машиной Тьюринга, и в которое входит заданная строка x. Приведем формулу для оценки разности между сложность ю выборки x и множества S, в которое входит данная выборка.

Теорема (без доказательств) 5. *Для любого $x \in S$ справедливо неравенство [3]:*

$$K(x) - K(S) \geq +\log|S| + O(1).$$

На практике задача выбора модели подразумевает, что мы можем выбрать модель, которая описывает выборку (или множество выборок) S неидеально, а с некоторым допустимым

уровнем потери информации. Тогда задача выбора модели для заданной выборки ставится следующим образом:

$$\arg\min_{S}\{\log|S| + K(S) : x \in S, K(s) \leq \alpha, \tag{0.2}$$

где α — максимально допустимая сложность множества S.

Заметим, что решение задачи выбора модели в приведенном выше виде является вычислимой, то есть можно предложить алгоритм, вычисляющий данную задачу. Приведем схему данного алгоритма:

1. Положим $\hat{p}, \hat{S}$ неопределенным.
2. Для всех $S, p : T(p) = S, K(S) \leq \alpha$:
3. Если $\hat{S}$ неопределен или $|p| + \log(S) \leq \hat{p} + \log\hat{S}$, то $\hat{p}, \hat{S} = p, S$.

Т.к. множество пар (S, p) конечно (TODO: почему?), то алгоритм остановится, а потому вычислим. По построению он также доставлячет решение оптимизационной задачи (**??**).

TODO: оценка ошибки

TODO: про то что это MDL

0.3 Вероятностная интерпретация минимальной длины описания

Список литературы

[1] Успенский В., Шень А., Верещагин Н. Колмогоровская сложность и алгоритмическая случайность. – Litres, 2017

[2] Grunwald P., Vitányi P. Shannon information and Kolmogorov complexity //arXiv preprint cs/0410002. – 2004.

[3] 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.