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 [**?**].

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 [**?**]

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

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. *[?] 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*

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

Для доказательства предварительно приведем две теоремы из [?] 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 (0.1) 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 пишущщей 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 [?]:*

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

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 $S, p : T(p) = S, |p| \leq \alpha$:
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 всех программ с длиной менее α является конечным, то алгоритм остановится, а потому вычислим. По построению он также доставлячет решение оптимизационной задачи (0.2).

Теорема (без доказательств) 6. *Обзначим за $S^*, K(S) \leq \alpha$ множество, доставляющее минимум следующей функции:*

$$\delta(S) = \log|S| - K(x|S).$$

Тогда $\delta(S^*) - \delta\hat{S} = O(\log|x|)$.

Таким образом, пара $(\hat{p}, \hat{S})$ доставляет минимуму суммы $|p| + \log(S)$ при ограничениях на длину программы $|p| \leq \alpha$. Первое слагаемое данной суммы — это длина программы, то есть сложность описания множества $\hat{S}$. Сама сумму $|p| + \log(S)$ характеризует длину кода, требуемого для двухэтапного описания строки x, где на первом этапе мы описываем множество S, а на втором этапе находим x во множестве S.

Описанный метод получения кода для описания выборки x соответствует *Принципу миниммальной длины описания* или MDL: для заданной выборки требуется найти минимальный код, который описывает выборку (возможно, с некоторой наперед заданной допустимой ошибкой(.

0.3 Вероятностная интерпретация минимальной длины описания

Рассмотрим подробнее задачу кодирования выборки. Задачу можно рассмаривать как проблему передачу информации от кодировщика декодировщику. Кодировщик кодирует информацию о выборке и передает ее декодировщику. Декодировщик раскодирует код, полученный от кодировщика и восстанавливает исходную выборку (возможно, с некоторой потерей информации, если это оговорено заранее). Допустим, для кодирования информации о выборке доступно несколько методов кодирования, при этом для разных объектов выборки наилучший (т.е. минимальный по количеству информации) метод будет отличаться. Тогда кодировщик должен передать не только информацию о выборке, но и информацию о самом методе кодировния. Перейдем к вероятностной интерпретации минимальной длины описания.

Теорема (без доказательств) 7. *Пусть задан конечное или счетное множество Z с введенной на нем вероятностью P. Тогда существует префиксный код C, такой что для любого $x \in Z : L(C(x)) = -\log P(x)$.*

Таким образом, вместо методов кодирования монжо рассматривать задачу выбора функции вероятности, введенной на признаковом описании выборки. С вероятностных позиций принцип минимальной длины описания формулируется следующим образом: задана выборка

X, требуется передать информацию о выборкем (с возможно, некоторой допустимой ошибкой) с использованием минимального количества информации. Пусть порожденная из некоторого вероятностного распределения p. Требуется построить код, доставляющий минимальную длину следующего выражения:

$$\log p(D|H) + L(H),$$

где $L(H)$ — длина описания гипотезы H.

Одним из критериев качества вероятностного кодирования является регрет:

$$R(x) = -\log P(x) + \min_H(\log P(x|H)).$$

Регрет характеризует разницу между длиной предлагаеистинного кода для x в сравнении с наилучшим кодом из некоторого множества H.

Пусть гипотеза H — зависит от параметра w:

$$\log p(D|H) = \log p(D|\hat{w}).$$

Тогда регрет выглядит следующим образом:

$$R(x) = -\log P(x) + \min_w(\log P(x|w)).$$

Для всей выборки регрет определяется следующим образом:

$$R(X) = \max_x(-\log P(x) + \min_w(\log P(x|w))).$$

Теорема 2 ((Штарьков, 1987) **??**)**.** *Пусть велеична*

$$\log \sum_x p(x|w(x))$$

конечна. Тогда следующее выражени дает единственный минимум регрета:

$$\frac{p(x|w(x))}{\sum_x p(x|w(x))}.$$

Доказательство. Рассмотрим выражение регрета для данной вероятностной меры:

$$-\log P(x) + \min_w(\log P(x|w)) = \log \sum_x p(x|w(x)).$$

Выражение регрета не зависит от объекта выборки x. Заметим, что для любых двух отличных распределений p_1, p_2 существует хотя бы один элемент, такой что

$$p_1(x) < p_2(x).$$

Действительно, пусть это неверно: $\forall x p_1(x) \geq p_2(x), \exists x' : p_1(x) > p_2(x)$. Просумсмируем вероятности p_1 и p_2 и воспользуемся тем, что сумма вероятноснетй по всем объектам будет равна единице:

$$1 = \sum p_1 > \sum p_2 = 1,$$

приходим к противоречию.

Тогда для любого распределения на выборке существует элемент x':

$$\frac{p(x'|w(x'))}{\sum_x p(x|w(x))} > p(x').$$

Тогда

$$R(P,x) \geq -\log p(x') + \log p(x'|w(x')) > \frac{p(x'|w(x'))}{\sum_x p(x|w(x))} = R(x).$$

□

Байесовской интерпертацией минимальной длины описания выступает *evidence* или *обоснованность модели*:

$$\int_w p(x|w)p(w),$$

где $p(w)$ — априорное распределения на выборке. оно задается на основе наших предположений о природе выборки и о способах ее порождения.

Теорема 3. *Пусть M — экспоненциальное семейство.*

Пусть в качестве априорного распределения выступает распределение Джефри:

$$p(w) = \frac{\sqrt{I}}{\int_w \sqrt{I(w)}},\ \textit{где } I - - - \textit{определить матрицы Фишера:}$$

$$I(w) = det\{\mathsf{E}_w \frac{-\partial^2}{\partial w_i \partial w_j} \log p(x|w)\}_{ij}.$$

Тогда регрет можно аппроксимировать следующим выражением:

$$\frac{k}{2} \log \frac{n}{2\pi} - \log p(x|w) + \log \sqrt{I(w)}.$$

Доказательство. Воспользуемся аппроксимацией Лапласа для упрощения формулы обоснованности модели. Разложим $\log p(x, w) = \log p(x|w)p(w)$ в точке локального максимума w_0 в ряд Тейлора:

$$\log p(x, w) \approx \log p(x, w_0) - \frac{1}{2}(w - w_0)^\mathsf{T} A (w - w_0),$$

$$A_{ij} = -\frac{\partial^2}{\partial w_i \partial w_j} \log p(x, w)|_{w=w_0}.$$

Полученное распределение представимо в виде ненормированного гауссового распределения. для такой аппроксимации плотности вероятности запишем нормирующий коэффициент:

$$\log p(x, w) \approx \log p(x|w_0) + p(w_0) - \log \sqrt{\frac{(2\pi)^k}{-} \det A}.$$

Подставляя в полученную формулу распределение Джеффри получим:

$$p(x, w) \approx p(x|w_0) - \log \sqrt{\frac{(2\pi)^k}{2}}.$$

TODO: там еще информация Фишера + экспоненциальное семейство. Тогда регрет будет равен:

$$R(x) \approx \log \sqrt{\frac{(2\pi)^k}{2}}.$$

□

Теорема (без доказательств) 8. *При количество параметров, стремящемся к бесконечности оптимальной модели отличается от байесовской оценки на константу.*

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. – T. 50. – №. 12. – C. 3265-3290.

[4] Grunwald P. A tutorial introduction to the minimum description length principle //arXiv preprint math/0406077. – 2004.