Применение MDL (Minimal Descripton length) принципа Риссанена для марковских процессов.

Ремизова Анна Петровна

23 апреля 2023 г.

Что нового

1. Оказалось, что я неверно считала диграммы: я использовала в Pyhton метод строки `s.count(substring)`, который считает непересекающиеся вхождения подстроки. Пример: для строки `'00000'` при подсчёте $n(00)$ данный метод выдавал ответ 2 при правильном ответе 4. Т.о., $n(00)$ и $n(11)$ занижались. Эти моменты я исправила и обновила таблицы (1, 2, 3), а также в Таблицу (4) выписала количества диграмм для каждого из 3 случаев.

2. Также в таблицах (1, 2, 3) в 1 строку каждой ячейки добавила значение MDL для данных k, l, на второй строке везде логарифм, затем p, q.

3. Добавила Утверждение (1) про независимые оптимальные значения p и q и доказательство к нему.

4. Предложение (1) о том, что можем просто найти p и q по формулам $p = \dfrac{n(01)}{n(01) + n(00))}$ и $q = \dfrac{n(10)}{n(10) + n(11)}$ почему-то подтверждается по таблицам только для π, а для $\sqrt{2}, \sqrt{3}$ верно только q, а p немного отличается от найденного по формуле. Возможно, я в программе как-то неверно ищу p, но ошибки я пока там не нашла.

5. Добавила Таблицу (5) с оптимальными значениями k для $n(00) = 1$ и различных $n(01)$.

Введение

Есть данные, мы хотим подобрать марковскую цепь, для которой наибольшая вероятность получить заданную траекторию. По Риссанену, если мы хотим предсказать, что будет дальше, то должны сравнивать друг с другом гипотезы по их сложности, причём даём преимущество простым гипотезам. Выражения для Description length будет выглядеть следующим образом:

$$C(\mu) + \log_2 \frac{1}{\mu(x)} \tag{1}$$

где $C(\mu)$ – complexity, μ – распределение вероятности.

Марковские цепи с 2 состояниями

Для начала рассмотрим простые марковские цепи. Пусть марковская цепь состоит из 2 состояний. Дана последовательность состояний Марковской цепи из 2 состояний: 0 и 1. Найти оптимальные переходные вероятности p из 0 в 1 и q из 1 в 0 по принципу Риссанена MDL.

Для решения этой задачи запишем вероятность получения заданной реализации: пусть $n(ij)$ – число переходов из состояния i в состояние j, тогда:

$$P_c(x) = p^{n(01)} \cdot (1-p)^{n(00)} \cdot q^{n(10)} \cdot (1-q)^{n(11)} \to max \tag{2}$$

$$\log_2 \frac{1}{P_c(x)} = -(n(01) \cdot \log_2 p + n(00) \cdot \log_2 (1-p) + n(10) \cdot \log_2 q + n(11) \cdot \log_2 (1-q)) \tag{3}$$

Сложность $C(\mu)$ будем определять как суммарную длину записи p и q в двоичной системе счисления. Пусть вероятность p имеет k знаков в двоичной системе, q – l знаков, тогда $C(\mu) = k + l$. Далее рассмотрим несколько реализаций Марковских цепей и исследуем, как меняются значения в зависимости от k и l.

Таблица с двоичными значениями

В Таблицах (1, 2, 3) в каждой ячейке представлены сначала оптимальные (минимальные, т.к. ищем минимальную описательную длину) значения $\log_2 \frac{1}{\mu(x)} = -(n_{01} \log_2 p + n_{00} \log_2 (1-p) + n_{10} \log_2 q + n_{11} \log_2 (1-q))$, затем сложность по Риссанену, а после - значения p и q, при которых оно достигается, представленные в двоичной системе счисления, для марковских цепей с траекториями, соответствующими 30 первым знакам $\pi, sqrt(2), sqrt(3)$ соответственно. По горизонтали отмечены значения l - длина перебираемых q в двоичной системе, по вертикали - значения k - длина перебираемых p в двоичной системе.

Выводы к Таблице (1) для π: заметим, что при фиксированной длине l (по столбцам) двоичной записи переходной вероятности q оптимальное значение q неизменно, но при этом с увеличением k оптимальное значение логарифма уменьшается. Аналогично для фиксированного k (по строкам).

Выводы к Таблице (2): для $\sqrt{2}$ практически то же, что и для π.

Выводы к Таблице (3): для $\sqrt{3}$ результаты уже отличаются от π, но наблюдаются те же закономерности. Отличие $\sqrt{3}$ от π и $\sqrt{2}$ в количестве диграмм в их двоичной записи, были рассмотрены первые 30 знаков для каждого числа, не считая точки. Если для π и $\sqrt{2}$ распределение количества диграмм близко к равномерному, то для $\sqrt{3}$ оно менее сбалансированно: количество диграмм 00 меньше остальных, а диграмм 11 - больше (см.Таблицу 4).

Утверждение 1 *Оптимальное значение p не зависит от q и наоборот, оптимальное значение q не зависит от p.*

Доказательство. Рассмотрим выражение (3) для логарифма. Значения $n(00), n(01), n(10), n(11)$ – постоянные, и данное выражения можно представить в виде линейной комбинации двух функций $f_1(p) + f_2(q)$. Соотвественно, при максимизации всего выражения (логарифм (3) должен быть маленьким, а так как перед всем выражением стоит минус, то выражение в скобках должно быть большим), так как переменные p и q содержатся в отдельных слагаемых, необходимо найти минимум отдельно для $f_1(p)$ и $f_2(q)$, друг на друга их значения при минимизации не влияют. ■

Таблица 1: Таблица оптимальных зн-й p и q в двоичной записи для π

k / l	1	2	3	4	5	6
1	31.0 29.0 0.1 0.1	32.0 29.0 0.1 0.10	33.0 29.0 0.1 0.100	33.9891 28.9891 0.1 0.1001	34.9521 28.9521 0.1 0.10001	35.9521 28.9521 0.1 0.100010
2	32.0 29.0 0.10 0.1	33.0 29.0 0.10 0.10	34.0 29.0 0.10 0.100	34.9891 28.9891 0.10 0.1001	35.9521 28.9521 0.10 0.10001	36.9521 28.9521 0.10 0.100010
3	33.0 29.0 0.100 0.1	34.0 29.0 0.100 0.10	35.0 29.0 0.100 0.100	35.9891 28.9891 0.100 0.1001	36.9521 28.9521 0.100 0.10001	37.9521 28.9521 0.100 0.100010
4	34.0 29.0 0.1000 0.1	35.0 29.0 0.1000 0.10	36.0 29.0 0.1000 0.100	36.9891 28.9891 0.1000 0.1001	37.9521 28.9521 0.1000 0.10001	38.9521 28.9521 0.1000 0.100010
5	35.0 29.0 0.10000 0.1	36.0 29.0 0.10000 0.10	37.0 29.0 0.10000 0.100	37.9891 28.9891 0.10000 0.1001	38.9521 28.9521 0.10000 0.10001	39.9521 28.9521 0.10000 0.100010
6	36.0 29.0 0.100000 0.1	37.0 29.0 0.100000 0.10	38.0 29.0 0.100000 0.100	38.9891 28.9891 0.100000 0.1001	39.9521 28.9521 0.100000 0.10001	40.9521 28.9521 0.100000 0.100010

Анализ диграмм

В Таблицу (4) представлены количества диграмм по рассмотренным примерам - их сумма в каждом случае равна 29, так как рассматриваемые числа округлялись до 30 знаков в двоичной записи суммарно, далее оптимальные значения $k, l, \log_2 \frac{1}{\mu x}, MDL$, найденные при $k, l \in [1, 6]$ для минимизации MDL.

Предложение 1 *Значения* $p = \dfrac{n(01)}{n(01) + n(00))}$ *и* $q = \dfrac{n(10)}{n(10) + n(11)}$ *являются точкой максимума функции* $\log_2 \dfrac{1}{P_c(x)}$ *(3). Их значения при заданных длинах двоичной записи* k *и* l *- это соответственно первые* k *и* l *знаков их двоичного представления.*

Обоснование

1. Найдём точку максимума функции $f_1(p) = n(01) \log_2 p + n(00) \log_2 (1 - p)$. Её производная: $f_1'(p) = \dfrac{n(01)}{p \ln 2} - \dfrac{n(00)}{(1-p) \ln 2}$, критические точки $p = \dfrac{n(01)}{n(01) + n(00)}, p = 0, p = 1$. Т.к.

Таблица 2: Таблица оптимальных зн-й p и q в двоичной записи для $\sqrt{2}$

k / l	1	2	3	4	5	6
1	31.0 29.0 0.1 0.1	32.0 29.0 0.1 0.10	32.9148 28.9148 0.1 0.101	33.7965 28.7965 0.1 0.1001	34.7965 28.7965 0.1 0.10010	35.795 28.795 0.1 0.100101
2	32.0 29.0 0.10 0.1	33.0 29.0 0.10 0.10	33.9148 28.9148 0.10 0.101	34.7965 28.7965 0.10 0.1001	35.7965 28.7965 0.10 0.10010	36.795 28.795 0.10 0.100101
3	33.0 29.0 0.100 0.1	34.0 29.0 0.100 0.10	34.9148 28.9148 0.100 0.101	35.7965 28.7965 0.100 0.1001	36.7965 28.7965 0.100 0.10010	37.795 28.795 0.100 0.100101
4	33.9891 28.9891 0.0111 0.1	34.9891 28.9891 0.0111 0.10	35.9039 28.9039 0.0111 0.101	36.7856 28.7856 0.0111 0.1001	37.7856 28.7856 0.0111 0.10010	38.7842 28.7842 0.0111 0.100101
5	34.9521 28.9521 0.01111 0.1	35.9521 28.9521 0.01111 0.10	36.8669 28.8669 0.01111 0.101	37.7485 28.7485 0.01111 0.1001	38.7485 28.7485 0.01111 0.10010	39.7471 28.7471 0.01111 0.100101
6	35.9521 28.9521 0.011110 0.1	36.9521 28.9521 0.011110 0.10	37.8669 28.8669 0.011110 0.101	38.7485 28.7485 0.011110 0.1001	39.7485 28.7485 0.011110 0.10010	40.7471 28.7471 0.011110 0.100101

$0 \leq \dfrac{n(01)}{n(01)+n(00)} \leq 1$, то $f_1'(p)$ отрицательна на $p \in (-\infty; 0) \cup \left(\dfrac{n(01)}{n(01)+n(00)}; 1\right)$, положительная на остальных промежутках, а значит точка максимума $p = \dfrac{n(01)}{n(01)+n(00)}$, если это значение отлично от 0 и 1, и $p = 0$ иначе. Аналогично для $f_2(q)$ точкой максимума является $q = \dfrac{n(10)}{n(10)+n(11)}$, либо $q = 0$.

2. Для π оптимальные p и q, вычисленные по указанным формулам, выглядят следующим образом: $p_0 = 0.5_{10} = 0.1_2, q_0 = 0.5(3)_{10} = 0.(1000)_2$, чему удовлетворяют значения из Таблицы (1).

   Для $\sqrt{2}$ имеем: $p_0 = 0.4(6)_{10} = 0.(0111)_2, q_0 = 0.(571428)_{10} = 0.(100)_2$ – по Таблице (2) совпадает q, но не совпадает p.

   Для $\sqrt{3}$ имеем: $p_0 = 0.(63)_{10} = 0.(1010001011)_2, q_0 = 0.(4)_{10} = 0.(011100)_2$ – по Таблице (3) также совпадает q, но не совпадает p.

Заметим по Таблице (4), что для рассмотренных трёх случаев высокая точность переход-

Таблица 3: Таблица оптимальных зн-й p и q в двоичной записи для $\sqrt{3}$

k / l	1	2	3	4	5	6
1	31.0 29.0 0.1 0.1	32.0 29.0 0.1 0.10	33.0 29.0 0.1 0.100	33.8419 28.8419 0.1 0.0111	34.8419 28.8419 0.1 0.01110	35.8419 28.8419 0.1 0.011100
2	31.9053 28.9053 0.11 0.1	32.9053 28.9053 0.11 0.10	33.9053 28.9053 0.11 0.100	34.7472 28.7472 0.11 0.0111	35.7472 28.7472 0.11 0.01110	36.7472 28.7472 0.11 0.011100
3	32.4067 28.4067 0.101 0.1	33.4067 28.4067 0.101 0.10	34.4067 28.4067 0.101 0.100	35.2486 28.2486 0.101 0.0111	36.2486 28.2486 0.101 0.01110	37.2486 28.2486 0.101 0.011100
4	33.4067 28.4067 0.1010 0.1	34.4067 28.4067 0.1010 0.10	35.4067 28.4067 0.1010 0.100	36.2486 28.2486 0.1010 0.0111	37.2486 28.2486 0.1010 0.01110	38.2486 28.2486 0.1010 0.011100
5	34.4067 28.4067 0.10100 0.1	35.4067 28.4067 0.10100 0.10	36.4067 28.4067 0.10100 0.100	37.2486 28.2486 0.10100 0.0111	38.2486 28.2486 0.10100 0.01110	39.2486 28.2486 0.10100 0.011100
6	35.4029 28.4029 0.101001 0.1	36.4029 28.4029 0.101001 0.10	37.4029 28.4029 0.101001 0.100	38.2448 28.2448 0.101001 0.0111	39.2448 28.2448 0.101001 0.01110	40.2448 28.2448 0.101001 0.011100

Таблица 4: Числа, количество диграмм в них, оптимальные k и l

Число	$n(00)$	$n(01)$	$n(10)$	$n(11)$	k	l	$\log_2 \frac{1}{\mu(x)}$	MDL
π	7	7	8	7	1	1	29.0	31.0
$\sqrt{2}$	8	7	8	6	1	1	29.0	31.0
$\sqrt{3}$	4	7	8	10	1	1	29.0	31.0

ных вероятностей p и q не выгодна по Риссанену, Minimal description length достигается при $k = l = 1$. Это будет не так, если при увеличении на 1 бит точности переходной вероятности логарифм будет уменьшаться больше, чем на 1. Т.е. количество диграмм $n(00)$ и $n(01)$ должно быть сильно не сбалансированно.

Рассмотрим различные значения $n(01)$ для $n(00) = 1$ и найдём, при каком k достигается MDL. Алгоритм: берём $p = \dfrac{n(01)}{n(01) + n(00))}$, переводим в двоичую систему с k знаками после запятой, считаем для каждого $Descriptionlength = k - (n(01)log_2(p) + n(00)log_2(1-p))$ и ищем минимальное такое при различных $k \in [1, 100]$ В Таблице (5) видно, что результаты k уже нетривиальные. Симметричная ситуация будет наблюдаться и для l.

Таблица 5: Оптимальные k для разных n(01)

n(01)	k	MDL
2	1	4.0
4	2	5.6601
8	2	7.3203
16	3	9.0823
32	4	10.9795
64	5	12.9314
128	6	14.9082
256	7	16.8967
512	8	18.891
1024	9	20.8882