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

Ремизова Анна Петровна

21 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)

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. Добавила Таблицу (4) с количеством диграмм в рассмотренных случаях $pi, \sqrt{2}, \sqrt{3}$, а также (в процессе) оптимальные значения $k, l, \log \frac{1}{\mu(x)}, MDL$ к ним. При поиске оптимального значения уже учитывала тот факт, что p и q можно искать независимо друг от друга.

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:

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

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

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 $\sqrt{2}$ 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	32.0 28.0 0.1 0.1	32.0 28.0 0.1 0.10	32.0 28.0 0.1 0.100	31.9891 27.9891 0.1 0.1001	31.9521 27.9521 0.1 0.10001	31.9521 27.9521 0.1 0.100010
2	34.0 28.0 0.10 0.1	34.0 28.0 0.10 0.10	34.0 28.0 0.10 0.100	33.9891 27.9891 0.10 0.1001	33.9521 27.9521 0.10 0.10001	33.9521 27.9521 0.10 0.100010
3	36.0 28.0 0.100 0.1	36.0 28.0 0.100 0.10	36.0 28.0 0.100 0.100	35.9891 27.9891 0.100 0.1001	35.9521 27.9521 0.100 0.10001	35.9521 27.9521 0.100 0.100010
4	37.9664 27.9664 0.1001 0.1	37.9664 27.9664 0.1001 0.10	37.9664 27.9664 0.1001 0.100	37.9555 27.9555 0.1001 0.1001	37.9185 27.9185 0.1001 0.10001	37.9185 27.9185 0.1001 0.100010
5	39.9464 27.9464 0.10001 0.1	39.9464 27.9464 0.10001 0.10	39.9464 27.9464 0.10001 0.100	39.9355 27.9355 0.10001 0.1001	39.8985 27.8985 0.10001 0.10001	39.8985 27.8985 0.10001 0.100010
6	41.9464 27.9464 0.100010 0.1	41.9464 27.9464 0.100010 0.10	41.9464 27.9464 0.100010 0.100	41.9355 27.9355 0.100010 0.1001	41.8985 27.8985 0.100010 0.10001	41.8985 27.8985 0.100010 0.100010

Анализ диграмм

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

k / l	1	2	3	4	5	6
1	32.0 28.0 0.1 0.1	32.0 28.0 0.1 0.10	31.9148 27.9148 0.1 0.101	31.7965 27.7965 0.1 0.1001	31.7965 27.7965 0.1 0.10010	31.795 27.795 0.1 0.100101
2	34.0 28.0 0.10 0.1	34.0 28.0 0.10 0.10	33.9148 27.9148 0.10 0.101	33.7965 27.7965 0.10 0.1001	33.7965 27.7965 0.10 0.10010	33.795 27.795 0.10 0.100101
3	36.0 28.0 0.100 0.1	36.0 28.0 0.100 0.10	35.9148 27.9148 0.100 0.101	35.7965 27.7965 0.100 0.1001	35.7965 27.7965 0.100 0.10010	35.795 27.795 0.100 0.100101
4	38.0 28.0 0.1000 0.1	38.0 28.0 0.1000 0.10	37.9148 27.9148 0.1000 0.101	37.7965 27.7965 0.1000 0.1001	37.7965 27.7965 0.1000 0.10010	37.795 27.795 0.1000 0.100101
5	40.0 28.0 0.10000 0.1	40.0 28.0 0.10000 0.10	39.9148 27.9148 0.10000 0.101	39.7965 27.7965 0.10000 0.1001	39.7965 27.7965 0.10000 0.10010	39.795 27.795 0.10000 0.100101
6	42.0 28.0 0.100000 0.1	42.0 28.0 0.100000 0.10	41.9148 27.9148 0.100000 0.101	41.7965 27.7965 0.100000 0.1001	41.7965 27.7965 0.100000 0.10010	41.795 27.795 0.100000 0.100101

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

k / l	1	2	3	4	5	6
1	32.0 28.0 0.1 0.1	32.0 28.0 0.1 0.10	32.0 28.0 0.1 0.100	31.8419 27.8419 0.1 0.0111	31.8419 27.8419 0.1 0.01110	31.8419 27.8419 0.1 0.011100
2	32.9053 26.9053 0.11 0.1	32.9053 26.9053 0.11 0.10	32.9053 26.9053 0.11 0.100	32.7472 26.7472 0.11 0.0111	32.7472 26.7472 0.11 0.01110	32.7472 26.7472 0.11 0.011100
3	34.9053 26.9053 0.110 0.1	34.9053 26.9053 0.110 0.10	34.9053 26.9053 0.110 0.100	34.7472 26.7472 0.110 0.0111	34.7472 26.7472 0.110 0.01110	34.7472 26.7472 0.110 0.011100
4	36.8182 26.8182 0.1011 0.1	36.8182 26.8182 0.1011 0.10	36.8182 26.8182 0.1011 0.100	36.6601 26.6601 0.1011 0.0111	36.6601 26.6601 0.1011 0.01110	36.6601 26.6601 0.1011 0.011100
5	38.8182 26.8182 0.10110 0.1	38.8182 26.8182 0.10110 0.10	38.8182 26.8182 0.10110 0.100	38.6601 26.6601 0.10110 0.0111	38.6601 26.6601 0.10110 0.01110	38.6601 26.6601 0.10110 0.011100
6	40.8132 26.8132 0.101101 0.1	40.8132 26.8132 0.101101 0.10	40.8132 26.8132 0.101101 0.100	40.6552 26.6552 0.101101 0.0111	40.6552 26.6552 0.101101 0.01110	40.6552 26.6552 0.101101 0.011100

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

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