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

Ремизова Анна Петровна

14 апреля 2023 г.

Введение

Для начала рассмотрим простые марковские цепи. Пусть марковская цепь состоит из 2 состояний. Есть данные, мы хотим подобрать марковскую цепь, для которой наибольшая вероятность получить '001'*300. По Риссанену, если мы хотим предсказать, что будет дальше, то должны сравнивать друг с другом гипотезы по их сложности, причём даём преимущество простым гипотезам.

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

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

Задача 1

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:

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

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

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 Таблице 1 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 p и q, при которых оно достигается, представленные в двоичной системе счисления. По горизонтали отмечены значения l - длина перебираемых q в двоичной системе, по вертикали - значения k - длина перебираемых p в двоичной системе.

Выводы: 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: для $\sqrt{2}$ то же, что и для π.

Выводы: для $\sqrt{3}$ результаты уже отличаются от π, но наблюдаются те же закономерности.

Таблица с десятичными значениями

В таблице в каждой ячейке представлены сначала оптимальные значения $C(\mu) + \log_2 \frac{1}{\mu(x)}$, затем p и q, при которых оно достигается, округлённые до десятитысячных. По горизонтали отмечены значения l - длина перебираемых q в двоичной системе, по вертикали - значения k - длина перебираемых p в двоичной системе.

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

k / l	1	2	3	4	5	6
1	0.1 0.1 28.0	0.1 0.10 28.0	0.1 0.100 28.0	0.1 0.1001 27.9891	0.1 0.10001 27.9521	0.1 0.100010 27.9521
2	0.10 0.1 28.0	0.10 0.10 28.0	0.10 0.100 28.0	0.10 0.1001 27.9891	0.10 0.10001 27.9521	0.10 0.100010 27.9521
3	0.100 0.1 28.0	0.100 0.10 28.0	0.100 0.100 28.0	0.100 0.1001 27.9891	0.100 0.10001 27.9521	0.100 0.100010 27.9521
4	0.1001 0.1 27.9664	0.1001 0.10 27.9664	0.1001 0.100 27.9664	0.1001 0.1001 27.9555	0.1001 0.10001 27.9185	0.1001 0.100010 27.9185
5	0.10001 0.1 27.9464	0.10001 0.10 27.9464	0.10001 0.100 27.9464	0.10001 0.1001 27.9355	0.10001 0.10001 27.8985	0.10001 0.100010 27.8985
6	0.100010 0.1 27.9464	0.100010 0.10 27.9464	0.100010 0.100 27.9464	0.100010 0.1001 27.9355	0.100010 0.10001 27.8985	0.100010 0.100010 27.8985

Таблица 2: Таблица оптимальных зн-й p и q в двоичной записи для π

k / l	7	8	9	10	11
7	0.1001011 0.1001111 24.25487169	0.1001011 0.10011110 24.25487169	0.1001011 0.100111011 24.25474365	0.1001011 0.1001110110 24.25474365	0.1001011 0.10011101100 24.25474365
8	0.10010101 0.1001111 24.25469022	0.10010101 0.10011110 24.25469022	0.10010101 0.100111011 24.25456218	0.10010101 0.1001110110 24.25456218	0.10010101 0.10011101100 24.25456218
9	0.100101011 0.1001111 24.25464497	0.100101011 0.10011110 24.25464497	0.100101011 0.100111011 24.25451694	0.100101011 0.1001110110 24.25451694	0.100101011 0.10011101100 24.25451694
10	0.1001010101 0.1001111 24.25463365	0.1001010101 0.10011110 24.25463365	0.1001010101 0.100111011 24.25450561	0.1001010101 0.1001110110 24.25450561	0.1001010101 0.10011101100 24.25450561
11	0.10010101011 0.1001111 24.25463082	0.10010101011 0.10011110 24.25463082	0.10010101011 0.100111011 24.25450278	0.10010101011 0.1001110110 24.25450278	0.10010101011 0.10011101100 24.25450278

Таблица 3: Таблица оптимальных зн-й p и q в двоичной записи для $\sqrt{2}$

k / l	1	2	3	4	5	6
1	0.1 0.1 25.0	0.1 0.10 25.0	0.1 0.101 24.4998	0.1 0.1010 24.4998	0.1 0.10100 24.4998	0.1 0.100111 24.4975
2	0.10 0.1 25.0	0.10 0.10 25.0	0.10 0.101 24.4998	0.10 0.1010 24.4998	0.10 0.10100 24.4998	0.10 0.100111 24.4975
3	0.101 0.1 24.8217	0.101 0.10 24.8217	0.101 0.101 24.3215	0.101 0.1010 24.3215	0.101 0.10100 24.3215	0.101 0.100111 24.3192
4	0.1001 0.1 24.7738	0.1001 0.10 24.7738	0.1001 0.101 24.2735	0.1001 0.1010 24.2735	0.1001 0.10100 24.2735	0.1001 0.100111 24.2713
5	0.10011 0.1 24.7623	0.10011 0.10 24.7623	0.10011 0.101 24.2621	0.10011 0.1010 24.2621	0.10011 0.10100 24.2621	0.10011 0.100111 24.2598
6	0.100101 0.1 24.7594	0.100101 0.10 24.7594	0.100101 0.101 24.2592	0.100101 0.1010 24.2592	0.100101 0.10100 24.2592	0.100101 0.100111 24.2569

Таблица 4: Таблица оптимальных зн-й p и q в двоичной записи для $\sqrt{3}$

k / l	1	2	3	4	5	6
1	0.1 0.1 24.0	0.1 0.10 24.0	0.1 0.100 24.0	0.1 0.1001 23.9891	0.1 0.10001 23.9521	0.1 0.100010 23.9521
2	0.11 0.1 21.9053	0.11 0.10 21.9053	0.11 0.100 21.9053	0.11 0.1001 21.8944	0.11 0.10001 21.8573	0.11 0.100010 21.8573
3	0.110 0.1 21.9053	0.110 0.10 21.9053	0.110 0.100 21.9053	0.110 0.1001 21.8944	0.110 0.10001 21.8573	0.110 0.100010 21.8573
4	0.1100 0.1 21.9053	0.1100 0.10 21.9053	0.1100 0.100 21.9053	0.1100 0.1001 21.8944	0.1100 0.10001 21.8573	0.1100 0.100010 21.8573
5	0.11001 0.1 21.8783	0.11001 0.10 21.8783	0.11001 0.100 21.8783	0.11001 0.1001 21.8674	0.11001 0.10001 21.8304	0.11001 0.100010 21.8304
6	0.110010 0.1 21.8783	0.110010 0.10 21.8783	0.110010 0.100 21.8783	0.110010 0.1001 21.8674	0.110010 0.10001 21.8304	0.110010 0.100010 21.8304

Таблица 5: Таблица оптимальных значений p и q для π

k / l	1	2	3	4	5	6
1	0.5 0.5 25.0	0.5 0.5 25.0	0.5 0.625 24.4998	0.5 0.625 24.4998	0.5 0.625 24.4998	0.5 0.6094 24.4975
2	0.5 0.5 25.0	0.5 0.5 25.0	0.5 0.625 24.4998	0.5 0.625 24.4998	0.5 0.625 24.4998	0.5 0.6094 24.4975
3	0.625 0.5 24.8217	0.625 0.5 24.8217	0.625 0.625 24.3215	0.625 0.625 24.3215	0.625 0.625 24.3215	0.625 0.6094 24.3192
4	0.5625 0.5 24.7738	0.5625 0.5 24.7738	0.5625 0.625 24.2735	0.5625 0.625 24.2735	0.5625 0.625 24.2735	0.5625 0.6094 24.2713
5	0.5938 0.5 24.7623	0.5938 0.5 24.7623	0.5938 0.625 24.2621	0.5938 0.625 24.2621	0.5938 0.625 24.2621	0.5938 0.6094 24.2598
6	0.5781 0.5 24.7594	0.5781 0.5 24.7594	0.5781 0.625 24.2592	0.5781 0.625 24.2592	0.5781 0.625 24.2592	0.5781 0.6094 24.2569