

Применение MDL (Minimal Description length) принципа Рissanena для марковских процессов.

Ремизова Анна Петровна

21 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 Рissanену, 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 Рissanена MDL.

Для решения этой задачи запишем вероятность получения заданной реализации: пусть n_{ij} - число переходов из состояния i в состояние j , тогда:

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

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

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 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 π , $\sqrt{2}$, $\sqrt{3}$ 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).

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

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

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

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

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

Таблица 5: Таблица оптимальных значений p и q для π

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