

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

Ремизова Анна Петровна

6 мая 2023 г.

Что нового

1. Доказательство Утверждения (1), которое обсудили при встрече, немного скорректировала формулировку: вместо "берутся первые k и l знаков p и q " при заданных длинах двоичной записи k и l - это приближения оптимальных значений p и q числами вида $x = \frac{n}{2^k}$ и $x = \frac{n}{2^l}$ соответственно добавила Рис. (2) с графиками.
2. Исправила и пояснила выводы по таблицам с учётом посчитанных по формулам оптимальных p и q (то, что написала в сообщении).
3. Начала главу про Марковские цепи с 4 состояниями, но пока не сделала моделирование и перебор.

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 будут выглядеть следующим образом:

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

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:

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

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

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 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 $\sqrt{2}$ 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 сбалансировано: 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 $f_1(p) + f_2(q)$. Соответственно, 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. ■

Анализ диграмм

В Таблице (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 .

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

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

Утверждение 2 Значения $p = \frac{n(01)}{n(01) + n(00)}$ и $q = \frac{n(10)}{n(10) + n(11)}$ являются точкой максимума функции $\log_2 \frac{1}{P_c(x)}$ (3). Их значения при заданных длинах двоичной записи k и l - это приближения оптимальных значений p и q числами вида $x = \frac{n}{2^k}$ и $x = \frac{n}{2^l}$ соответственно.

Доказательство

1. Найдём точку максимума функции $f_1(p) = n(01) \log_2 p + n(00) \log_2 (1 - p)$. Её производная: $f'_1(p) = \frac{n(01)}{p \ln 2} - \frac{n(00)}{(1 - p) \ln 2}$, критические точки $p = \frac{n(01)}{n(01) + n(00)}$, $p = 0$, $p = 1$. Т.к. $0 \leq \frac{n(01)}{n(01) + n(00)} \leq 1$, то $f'_1(p)$ отрицательна на $p \in (-\infty; 0) \cup \left(\frac{n(01)}{n(01) + n(00)}; 1\right)$, положительная на остальных промежутках, а значит точка максимума $p = \frac{n(01)}{n(01) + n(00)}$, если это значение отлично от 0 и 1, и $p = 0$ иначе. Аналогично для $f_2(q)$ точкой максимума является $q = \frac{n(10)}{n(10) + n(11)}$, либо $q = 0$.

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

2. Рассмотрим функцию вероятности $P(x) = x^p(1-x)^{1-p}$. Найдём её вторую производную: $P'(x) = (1-x)^{-p}(p-x)x^{p-1}$, $P''(x) = (p-1)p(1-x)^{-p-1}x^{p-2}$. При фиксированном p $P''(x)$ имеет нули в точках $x = 0, x = 1$ и $P''(x) \geq 0$ на $x \in [0; 1]$, а значит на этом интервале исходная функция выпукла вверх – см. Рис. (2). Кроме того, её максимальное значение достигается при $x = p$. Так как при фиксированном k мы рассматриваем двоичные числа с k знаками после запятой, то $x = \frac{n}{2^k}$. Соответственно, оптимальным будет именно приближение точки максимума функции $P(x) : x = p$, а будет это приближение с избытком или недостатком – зависит от того, какое из чисел будет ближе к p по значению функции $P(x)$.

3. 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, на первый взгляд, p . Но значение, представленное в таблице, к примеру, для $k = 6$ – это $0.011110_2 = \frac{30}{64_{10}}$, а значение, равное округ-

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

лению до 6 знаков после запятой найденного по формуле p – это $0.011101_2 = \frac{29}{64_{10}}$
И действительно, если обозначить p_0 - найденное по формуле, то $\frac{29}{64} < p_0 < \frac{30}{64}$ и
 $0.00019 \approx \left| f_1(p_0) - f_1\left(\frac{30}{64}\right) \right| < \left| f_1(p_0) - f_1\left(\frac{29}{64}\right) \right| \approx 0.00800$.

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 , и также совпадает p с верхним приближением: для $k = 6$, к примеру, $0.101001_2 = \frac{41}{64_{10}}$, $0.101000_2 = \frac{40}{64_{10}}$, $\frac{40}{64} < p_0 < \frac{41}{64}$.

Таким образом, на представленных примерах всё согласуется с утверждением. ■

Заметим по Таблице (4), что для рассмотренных трёх случаев высокая точность переходных вероятностей p и q не выгодна по Рissanену, 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.

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

Рассмотрим различные значения $n(01)$ для $n(00) = 1$ и найдём, при каком k достигается MDL. Алгоритм: берём $p = \frac{n(01)}{n(01) + n(00)}$, переводим в двоичную 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 – мы нашли те примеры последовательностей, для которых оптимальная модель подразумевает достаточно точные значения переходных вероятностей. Симметричная ситуация будет наблюдаться и для l .

Таблица 5: Оптимальные k для разных $n(01)$

$n(00)$	$n(01)$	k	p	MDL
1	2	1	0.1	4.0
1	4	2	0.11	5.6601
1	8	2	0.11	7.3203
1	16	3	0.111	9.0823
1	32	4	0.1111	10.9795
1	64	5	0.11111	12.9314
1	128	6	0.111111	14.9082
1	256	7	0.1111111	16.8967
1	512	8	0.11111111	18.891
1	1024	9	0.111111111	20.8882

Марковские цепи с 4 состояниями

Рассмотрим более сложную марковскую цепь с четырьмя состояниями, но по-прежнему над двоичным алфавитом. Пусть при посещении состояний a, b печатаются нули, при посещении состояний c, d – единицы. Для некоторых последовательностей, в частности для тех, которые имеют период 4, например 00010001...0001, такая модель будет давать более оптимальный с точки зрения MLD результат.

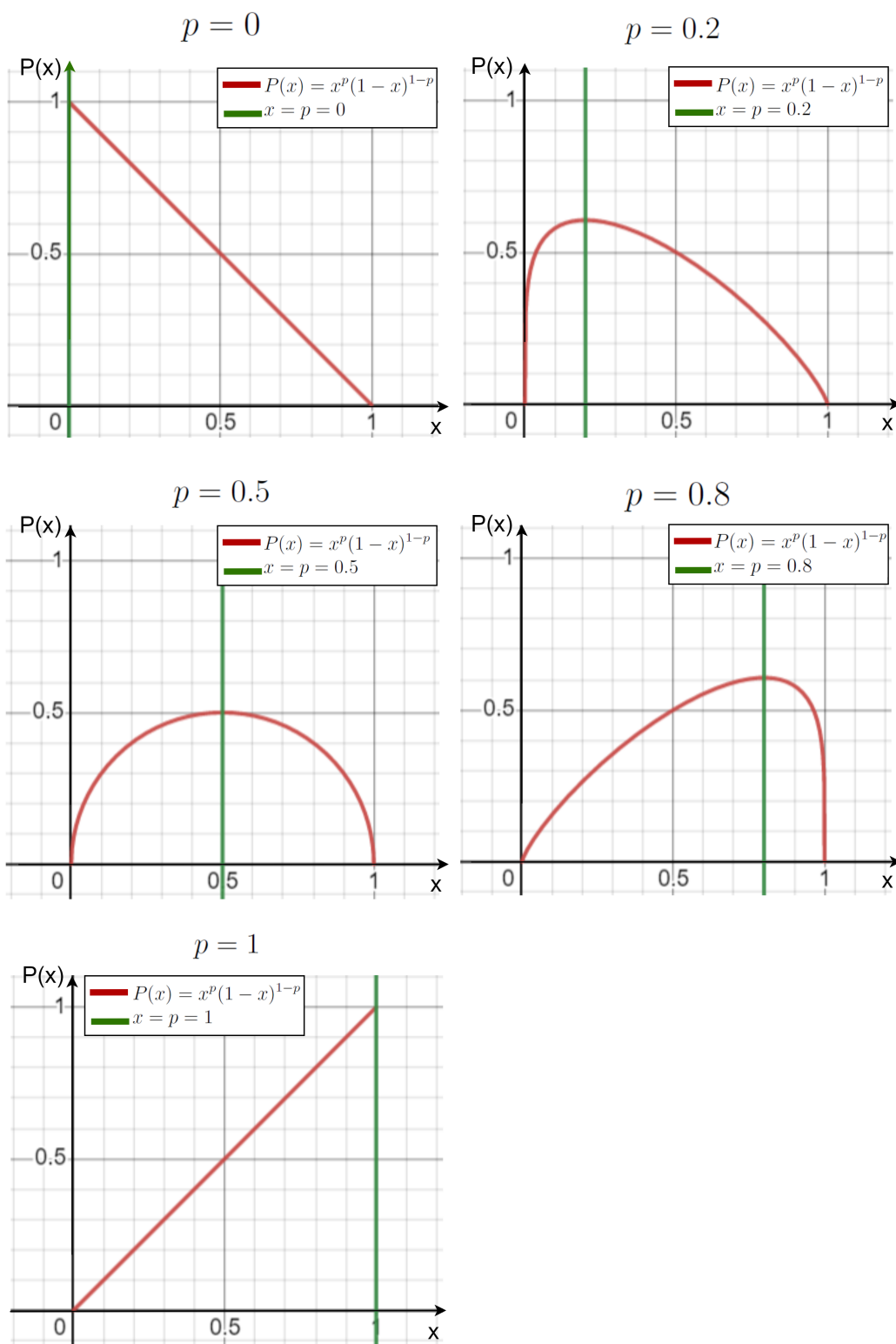


Рис. 1: График функции $P(x)$ при различных значениях p