

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

Ремизова Анна Петровна

30 апреля 2023 г.

## Что нового

1. Оказалось, что я неверно считала диграммы: я использовала в Python 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 (2) о том, что можем просто найти  $p$  и  $q$  по формулам  $p = \frac{n(01)}{n(01) + n(00)}$  и  $q = \frac{n(10)}{n(10) + n(11)}$  почему-то подтверждается по таблицам только для  $\pi$ , 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:

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

где  $C(\mu)$  – complexity,  $\mu$  – 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 Рissanena 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 Рissanену, 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  $\pi$ : 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  $\pi$ .

Выводы к Таблице (3): для  $\sqrt{3}$  результаты уже отличаются от  $\pi$ , но наблюдаются те же закономерности. Отличие  $\sqrt{3}$  от  $\pi$  и  $\sqrt{2}$  в количестве диграмм в их двоичной записи, были рассмотрены первые 30 знаков для каждого числа, не считая точки. Если для  $\pi$  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. ■

Таблица 1: Таблица оптимальных зн-й  $p$  и  $q$  в двоичной записи для  $\pi$

$k / l$	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

## 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  $MDL$ .

**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 соответственно первые  $k$  и  $l$  знаков их двоичного представления.

## 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.

Таблица 2: Таблица оптимальных зн-й р 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

$0 \leq \frac{n(01)}{n(01) + n(00)} \leq 1$ , 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  $q = 0$ .

- 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  $k - 1$  знаками после запятой (так как это вероятности, и 1 бит занят под целую часть, равную 0), то  $x = \frac{n}{2^{k-1}}$ . 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  $P(x)$ .

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

3. Для  $\pi$  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 переходных 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  $p = \frac{n(01)}{n(01) + n(00)}$ , 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

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
$\pi$	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

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

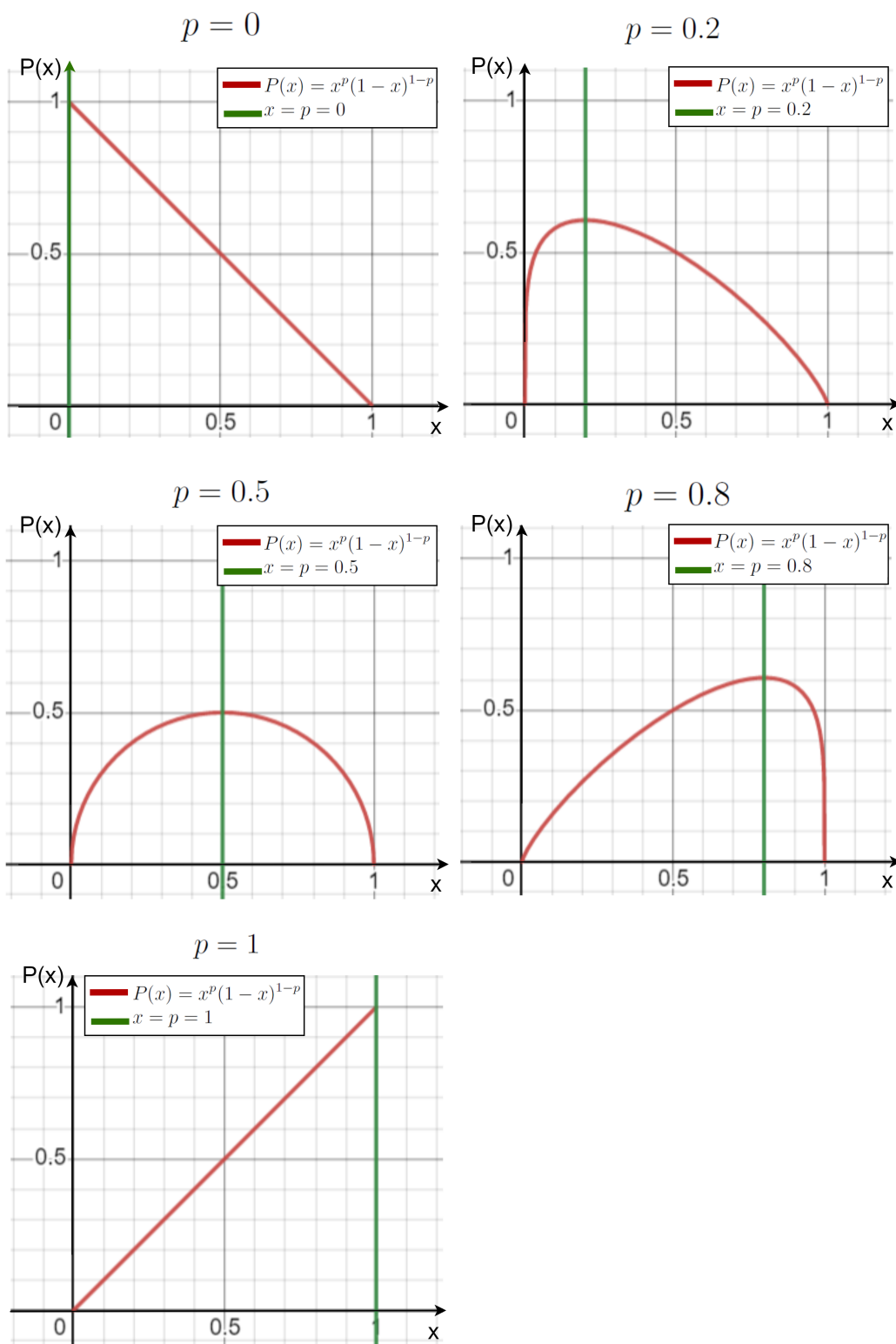


Рис. 1: График функции  $P(x)$  при различных значениях  $p$