

УДК 519.987

Смещение оценки энтропии для симметричных мер Бернулли и слабой метрики

Тимофеева Н.Е.

ЯФ НОУ ВПО "Институт управления"

e-mail: snifiska@rambler.ru

получена 16 января 2012

Ключевые слова: энтропия, непараметрическая оценка, смещение, мера Бернулли

Для симметричной меры Бернулли и новой слабой метрики найдено смещение оценки энтропии.

Введение

В настоящей статье предлагается новая (слабая) метрика для повышения эффективности непараметрической оценки энтропии h случайного процесса, принимающего значения в конечном множестве $\mathcal{A}$, если заданы n его независимых реализаций.

В работе [2] предложена оценка энтропии случайного процесса, которая задается формулами (8)–(9). Эта оценка зависит от метрики ρ на пространстве последовательностей $\Omega = \mathcal{A}^{\mathbb{N}}$. Если метрика ρ билипшицево эквивалентна метрике (2), то оценка (8)–(9) будет оценивать энтропию. В [7] применялась метрика (2) и результаты вычислительных экспериментов подтвердили теоретически полученные в статье результаты и показали, что эта оценка энтропии имеет точность 0.01 для $n = 10^4$. Однако в статье [8] показано, что смещение оценки при использовании метрики (2) в случае симметричной меры Бернулли является периодической функцией с периодом, пропорциональным $\log n$. Подчеркнем, что хорошее согласование теоретических значений энтропии и результатов вычислительного эксперимента в статье [7] можно объяснить тем, что для небольших значений энтропии ($h < 3$) смещение настолько мало, что его нельзя заметить в вычислительном эксперименте. Так, например, для алфавита из двух символов амплитуда смещения меньше, чем 10^{-6} [8].

Для метрики (2) мера шара является разрывной функцией от радиуса (шары совпадают с цилиндрами). В работе [4] предложен ряд слабых метрик, которые билипшицево эквивалентны метрике (2) и для которых мера шара будет более гладкой функцией. В статье найдена метрика, для которой можно явно определить меру шара для симметричной меры Бернулли. Исследуются свойства оценки, и хотя смещение остается, но для симметричной меры Бернулли оно будет на порядок меньше по сравнению с метрикой (2). Отметим, что симметричная мера Бернулли является той мерой, для которой смещение непараметрических оценок максимально [3].

1. Постановка задачи и обозначения

Задача оценивания энтропии будет рассматриваться в следующей постановке.

Пусть $\Omega = \mathcal{A}^{\mathbb{N}}$ — пространство правосторонних последовательностей ($\mathcal{A}$ — конечный алфавит).

Пусть даны $n + 1$ независимых случайных величин $\xi_0, \dots, \xi_n$, принимающих значения в Ω и одинаково распределенных по мере μ . Будем считать, что μ — борелевская эргодическая стационарная (инвариантная относительно сдвига) вероятностная мера.

Требуется оценить энтропию меры μ .

Напомним (см., напр., [1]), что *энтропия* (энтропия на символ) определяется как

$$h = - \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E} \ln \mu(C_n(\xi)). \quad (1)$$

Через

$$C_n(\mathbf{x}) = \{\mathbf{y} \in \mathcal{A}^{\mathbb{N}} : y_i = x_i, i = 1, \dots, n\}, \quad C_0(x) = \Omega,$$

будем обозначать цилиндры в пространстве последовательностей $\Omega = \mathcal{A}^{\mathbb{N}}$.

Отметим, что выбор натурального логарифма ($\ln$) позволяет избежать дополнительных множителей при построении оценок.

Введем на Ω две метрики

$$\rho_0(\mathbf{x}, \mathbf{y}) = \frac{1}{\min \{k : x_k \neq y_k\}}, \quad (2)$$

$$\rho_1(\mathbf{x}, \mathbf{y}) = \frac{1}{\alpha(\mathbf{x}, \mathbf{y})}, \quad (3)$$

где

$$\alpha(a\mathbf{x}, b\mathbf{y}) = \begin{cases} \alpha(\mathbf{x}, \mathbf{y}) + 1, & a = b; \\ \lambda(\alpha(\mathbf{x}, \mathbf{y})), & a \neq b; \end{cases} \quad (4)$$

$$\lambda(t) = \begin{cases} (1 - \beta)t, & 0 \leq t \leq 1; \\ 1 - \beta + \beta\lambda(t - 1), & t \geq 1, \end{cases} \quad (5)$$

где $0 \leq \beta < 1$ — некоторый параметр.

Через $B(\mathbf{x}, r, \rho) = \{\mathbf{y} \in \Omega : \rho(\mathbf{x}, \mathbf{y}) < r\}$ будем обозначать открытый шар радиуса r с центром в точке $\mathbf{x}$.

Подчеркнем, что для метрики (2) цилиндры совпадают с шарами

$$C_n(\mathbf{x}) = B(\mathbf{x}, r, \rho_0), \quad \frac{1}{n+1} < r \leq \frac{1}{n}. \quad (6)$$

Отметим, что метрика (3) является *слабой метрикой* [6], т.е. неравенство треугольника выполняется только с некоторой константой. Более того, справедливы оценки

$$\rho_0(\mathbf{x}, \mathbf{y})^{-1} \leq \alpha(\mathbf{x}, \mathbf{y}) \leq \rho_0(\mathbf{x}, \mathbf{y})^{-1} + 1. \quad (7)$$

2. Непараметрические оценки энтропии

Для каждой метрики ρ статистическая оценка $\eta_n(\rho)$ энтропии строится следующими простыми вычислениями:

- находим вспомогательную случайную величину

$$r_n^{(k)} = \frac{1}{n+1} \sum_{j=0}^n \left(\min_{i:i \neq j}^{(k)} \rho(\xi_i, \xi_j) \right)^{-1}, \quad (8)$$

где $\min^{(k)}\{x_1, \dots, x_N\} = x_k$, если $x_1 \leq x_2 \leq \dots \leq x_N$;

- полагаем оценкой обратной энтропии случайную величину

$$\eta_n^{(k)}(\rho) = k \left(r_n^{(k)}(\rho) - r_n^{(k+1)}(\rho) \right). \quad (9)$$

Отметим, что для упрощения дальнейших вычислений, случайные величины $\xi_0, \dots, \xi_n$ считаются бесконечными последовательностями. Нетрудно показать (см. [7]), что достаточно использовать только $O(\ln n)$ первых символов этих последовательностей, если мера удовлетворяет следующему условию:

$$\exists \gamma > 0 : \mu(C_n(\omega)) = O(e^{-n\gamma}), \text{ для } \mu - \text{почти всех } \omega \in \Omega. \quad (10)$$

3. Симметричная мера Бернулли

Чтобы показать основные трудности и возникающие преимущества при применении новой метрики, рассмотрим меру Бернулли с равновероятными символами. Подчеркнем, что в [3] показано, что смещение оценки для этой меры будет максимальным среди всех марковских мер.

Для этой меры энтропия $h = \ln |\mathcal{A}|$.

3.1. Нахождение меры шара

Как видно из (8), более удобно иметь дело с величиной обратной к радиусу. Поэтому положим

$$F(t) = \mu(B(\mathbf{x}, 1/t, \rho_1)) = \mu\{\xi : \alpha(\mathbf{x}, \xi) > t\}. \quad (11)$$

Ясно, что для меры Бернулли с равновероятными символами F не зависит от $\mathbf{x}$.

Утверждение 1. Для $\beta = \frac{1}{\mathcal{A}}$

$$F(t) = 1 - \lambda(t).$$

Доказательство. Применяя формулу полной вероятности по совпадению первого символа у $\mathbf{x}, \xi$ в формуле (11), получим

$$F(t) = \frac{1}{\mathcal{A}} F(t-1) + \frac{\mathcal{A}-1}{\mathcal{A}} F(\lambda^{-1}(t)).$$

Подставив $\beta = \frac{1}{A}$, перепишем это уравнение в следующем виде:

$$F(t) = \begin{cases} \beta + (1 - \beta)F(\lambda^{-1}(t)), & 0 \leq t \leq 1; \\ \beta F(t - 1), & t \geq 1. \end{cases} \quad (12)$$

Сравнивая уравнения (12) и (5), получаем требуемое утверждение. $\square$

Решив рекуррентное уравнение (12) в явном виде, получим

$$F(t) = \begin{cases} 1 - (1 - \beta)t, & 0 \leq t \leq 1; \\ \beta^k - (1 - \beta)\beta^k(t - k), & k \leq t \leq k + 1. \end{cases} \quad (13)$$

Итак, мера шара является непрерывной кусочно-линейной функцией от радиуса.

3.2. Нахождение $\text{Er}_n^{(1)}(\rho_1)$

Для нахождения смещения статистики (8) найдем асимптотику вспомогательной величины

$$\phi(n) = \text{Er}_n^{(1)}(\rho_1). \quad (14)$$

Для математического ожидания максимума из n независимых случайных величин в обозначениях (11) имеем

$$\phi(n) = \int_0^\infty (1 - (1 - F(t))^n) dt.$$

Подставляя функцию F из (13), получим

$$\phi(n) = \sum_{k=0}^\infty \left[1 - \frac{1}{(n+1)(1-\beta)\beta^k} ((1 - \beta^{k+1})^{n+1} - (1 - \beta^k)^{n+1}) \right]. \quad (15)$$

Введем функцию

$$f(x) = \frac{1}{(n+1)x} ((1-x)^{n+1} - 1 + (n+1)x). \quad (16)$$

Тогда $\phi(n)$ можно переписать в следующем виде:

$$\phi(n) = \frac{n\beta}{(n+1)(1-\beta)} + \sum_{k=0}^\infty f(\beta^k). \quad (17)$$

Так же как и в [8] для нахождения суммы (17) применим формулу суммирования Пуассона [5, (13.8.4)] для функции $f(\beta^x)$

$$\sum_{k=0}^\infty f(\beta^k) = \int_0^\infty f(\beta^t) dt + \frac{1}{2}f(1) - \sum_{k=1}^\infty \frac{\ln \beta}{\pi k} \int_0^\infty f'(\beta^t) \beta^t \sin 2\pi kt dt.$$

Сделав замену переменных $x = \beta^t$, заменив синус на разность экспонент и используя обозначение $h = \ln |\mathcal{A}| = -\ln \beta$, получим

$$\sum_{k=0}^{\infty} f(\beta^k) = \frac{1}{h} \int_0^1 \frac{f(x)}{x} dx + \frac{1}{2} f(1) + \sum_{m \neq 0} \frac{1}{2\pi i m} \int_0^1 x^{-\frac{2\pi i m}{h}} f'(x) dx.$$

Подставляя в (17), получим

$$\phi(n) = \frac{(\beta + 1)n}{2(1 - \beta)(n + 1)} + \frac{1}{h} \int_0^1 \frac{f(x)}{x} dx + \sum_{m \neq 0} \frac{1}{2\pi i m} \int_0^1 x^{-\frac{2\pi i m}{h}} f'(x) dx.$$

Для упрощения вычислений проинтегрируем второй интеграл по частям и получим

$$\phi(n) = \frac{(\beta + 1)n}{2(1 - \beta)(n + 1)} + \frac{1}{h} \int_0^1 \frac{f(x)}{x} dx + \sum_{m \neq 0} \frac{1}{h} \int_0^1 x^{-\frac{2\pi i m}{h} - 1} f(x) dx. \quad (18)$$

Воспользовавшись легко проверяемым тождеством

$$1 - (n + 1)x - (1 - x)^{n+1} = -x^2 \sum_{k=1}^n (n + 1 - k)(1 - x)^{k-1},$$

найдем первый интеграл в (18)

$$\int_0^1 \frac{f(x)}{x} dx = \sum_{k=1}^n \frac{n + 1 - k}{k(n + 1)} = H_n - \frac{n}{n + 1},$$

где через H_s обозначаются гармонические числа

$$H_s = \sum_{i=1}^s \frac{1}{i}. \quad (19)$$

Интегрируя по частям, нетрудно получить следующее тождество (при $\Re a > -1$):

$$\int_0^1 x^{a-2} ((1 - x)^b - 1 + bx) dx = \frac{b - 1}{a} - \frac{1}{a(a - 1)} + \frac{b(b - 1)}{a(a - 1)} B(a + 1, b - 1),$$

применяя которое, найдем второй интеграл в (18)

$$(n + 1) \int_0^1 x^{-\frac{2\pi i m}{h} - 1} f(x) dx = -\frac{nh}{2\pi i m} - \frac{h^2}{2\pi i m(2\pi i m + h)} + \\ + \frac{\Gamma\left(-\frac{2\pi i m}{h} - 1\right) \Gamma(n + 2)}{\Gamma\left(n - \frac{2\pi i m}{h} + 1\right)}.$$

Подставляя найденные интегралы в (18), получим

$$\begin{aligned} \phi(n) = \frac{H_n}{h} + \frac{(\beta+1)n}{2(1-\beta)(n+1)} - \frac{n}{(n+1)h} \\ - \frac{1}{n+1} \sum_{m \neq 0} \frac{h}{2\pi i m (2\pi i m + h)} \\ + \sum_{m \neq 0} \frac{\Gamma\left(\frac{2\pi i m}{h} - 1\right) \Gamma(n+1)}{\ln 2 \Gamma\left(n + \frac{2\pi i m}{h} + 1\right)}. \end{aligned}$$

Для некоторой константы C_0 имеем

$$\phi(n) = \frac{H_n - 1}{h} + \frac{(\beta+1)}{2(1-\beta)} + \sum_{m \neq 0} \frac{\Gamma\left(\frac{2\pi i m}{h} - 1\right) \Gamma(n+1)}{h \Gamma\left(n + \frac{2\pi i m}{h} + 1\right)} + \frac{C_0}{n+1}. \quad (20)$$

Интересно сравнить этот результат с полученным в [8] для метрики (3)

$$\phi(n, \rho_0) = \frac{H_n}{h} + \frac{1}{2} - \sum_{m \neq 0} \frac{\Gamma\left(\frac{2\pi m i}{h}\right) \Gamma(n+1)}{h \Gamma\left(n + 1 + \frac{2\pi i m}{h}\right)}. \quad (21)$$

Гамма-функция убывает по мнимой оси довольно быстро, поэтому основной вклад в периодическое смещение дают только первые члены ряда ($m = \pm 1$), а он уменьшается в $\left|\frac{2\pi i}{h} - 1\right|$ раз.

3.3. Асимптотические и приближенные формулы

Применяя формулу Стирлинга, нетрудно вывести

$$\frac{\Gamma(s+1)}{\Gamma(s+1+\alpha)} = s^{-\alpha} + \mathcal{O}(s^{-1}). \quad (22)$$

Отметим, что оценка является равномерной по α .

Подставляя эту оценку в (20), получим асимптотическое разложение

$$\phi(n) = \frac{H_n - 1}{h} + \frac{(\beta+1)}{2(1-\beta)} + \frac{1}{h} \sum_{k \neq 0} \Gamma\left(\frac{2\pi k i}{h} - 1\right) n^{-2\pi k i/h} + \mathcal{O}(n^{-1}). \quad (23)$$

Первое приближение получается из (23) отбрасыванием суммы

$$\phi(n) \approx \frac{H_n - 1}{h} + \frac{(\beta+1)}{2(1-\beta)}.$$

Второе приближение получается из (23) отбрасыванием слагаемых с $|k| > 1$.

Математическое ожидание случайной величины $r_n^{(k)}$ находится так же, как и в [8]

$$\begin{aligned} \mathbb{E} r_n^{(k)} = \frac{H_n - H_{k-1} - 1}{h} + \frac{(\beta+1)}{2(1-\beta)} + \\ + \sum_{m \neq 0} \frac{\Gamma\left(\frac{2\pi i m}{h} + k\right) \Gamma\left(\frac{2\pi i m}{h} - 1\right) \Gamma(n+1)}{(k-1)! h \Gamma\left(n + \frac{2\pi i m}{h} + 1\right) \Gamma\left(\frac{2\pi i m}{h} + 1\right)} + \mathcal{O}(n^{-1}). \end{aligned} \quad (24)$$

А его асимптотика имеет следующий вид:

$$Er_n^{(k)} = \frac{H_n - H_{k-1} - 1}{h} + \frac{(\beta + 1)}{2(1 - \beta)} + \sum_{m \neq 0} \frac{\Gamma\left(\frac{2\pi im}{h} + k\right) \Gamma\left(\frac{2\pi im}{h} - 1\right)}{(k-1)! h \Gamma\left(\frac{2\pi im}{h} + 1\right)} n^{-2\pi im/h} + \mathcal{O}(n^{-1}). \quad (25)$$

Асимптотические формулы для математических ожиданий оценок (9)

$$E\eta_n^{(k)}(\rho_1) = \frac{1}{h} - \frac{1}{(k-1)!h} \sum_{m \neq 0} \frac{\Gamma\left(\frac{2\pi im}{h} + k\right)}{\frac{2\pi im}{h} - 1} n^{-2\pi im/h} + \mathcal{O}(n^{-1}). \quad (26)$$

Список литературы

1. Мартин Н., Ингленд Дж. Математическая теория энтропии. М. : Мир, 1988.
2. Тимофеев Е.А. Статистически оцениваемые инварианты мер //Алгебра и анализ. Т. 17, №3. 2005. С. 204–236.
3. Тимофеев Е.А. Смещение непараметрической оценки энтропии для марковской меры //Записки научных семинаров ПОМИ. Т. 377. 2010. С. 134–158.
4. Тимофеев Е.А. Выбор метрики для непараметрических оценок энтропии //Математика в Ярославском университете: сборник обзор- ных статей: к 35-летию математического факультета. Ярославль: ЯрГУ, 2011. С. 187–198.
5. Харди Г. Расходящиеся ряды. М.: ИЛ, 1951.
6. Deza M., Deza E. Encyclopedia of Distances. Springer, 2009.
7. Kaltchenko A., Timofeeva N. Entropy Estimators with Almost Sure Convergence and an $O(n^{-1})$ Variance // Advances in Mathematics of Communications. V. 2, №1. 2008. P. 1–13.
8. Kaltchenko A., Timofeeva N. Rate of convergence of the nearest neighbor entropy estimator // AEU - International Journal of Electronics and Communications. V. 64, №.1. 2010. P. 75–79.

Bias of the Entropy Estimator for Symmetric Bernoulli Measures and the Weak Metric

Timofeeva N.E.

Keywords: entropy, nonparametric statistics, bias, Bernoulli measure

We consider symmetric Bernoulli measures and new weak metrics and obtain a closed-form expression of the entropy estimator bias.

Сведения об авторе:

Тимофеева Нина Евгеньевна,
ЯФ НОУ ВПО "Институт управления",
канд. физ.-мат. наук, заведующий кафедрой информатики