

Энтропии в предсказании последовательностей

Обобщённые энтропии и асимптотическая сложность языков

Ю. А. Калнишкан

Department of Computer Science
and Computer Learning Research Centre
Royal Holloway, University of London

2008

- в докладе будет сформулирован результат из работы
Y. Kalnishkan, V. Vovk, and M. V. Vyugin. Generalised Entropy and Asymptotic Complexities of Languages. In Learning Theory, 20th Annual Conference on Learning Theory, COLT 2007, volume 4539 of Lecture Notes in Computer Science, pages 293-307, Springer 2007.
- мы также поговорим о разных смежных вопросах, касающихся предсказания последовательностей

Содержание

1. Предсказание последовательностей
2. Определение сложности
3. Выпуклые игры
4. Основной результат
5. Схема доказательства

1. Предсказание последовательностей
2. Определение сложности
3. Выпуклые игры
4. Основной результат
5. Схема доказательства

Протокол

- мы наблюдаем элементы последовательности $\omega_1, \omega_2, \omega_3, \dots \in \Omega$
- и выдаём предсказания $\gamma_1, \gamma_2, \gamma_3, \dots \in \Gamma$
- протокол:
FOR $t = 1, 2, \dots$
 (1) \mathfrak{A} выдаёт предсказание $\gamma_t \in \Gamma$
 (2) \mathfrak{A} наблюдает исход $\omega_t \in \Omega$
END FOR
- качество предсказаний измеряется функцией потерь $\lambda(\omega, \gamma)$
— кумулятивные потери (Loss) это сумма потерь за T шагов:

$$\text{Loss}_{\mathfrak{A}}(\omega_1, \omega_2, \dots, \omega_T) = \sum_{i=1}^T \lambda(\omega_i, \gamma_i)$$

Формализация

- игра* \mathfrak{G} это тройка $\langle \Omega, \Gamma, \lambda \rangle$, где
 - Ω это *пространство исходов*
 - Γ это *пространство предсказаний*
 - $\lambda : \Omega \times \Gamma \rightarrow [0, +\infty]$ это *функция потерь*
- мы будем считать, что
 - пространство исходов конечно, т.е., $\Omega = \{\omega^{(0)}, \omega^{(1)}, \dots, \omega^{(M-1)}\}$
 - пространство предсказаний Γ компактно
 - функция потерь λ непрерывна
- важный частный случай: двоичные игры
 - $\Omega = \mathbb{B} = \{0, 1\}$
 - $\Gamma = [0, 1]$

Примеры

- квадратичная игра: $\Omega = \{0, 1\}$, $\Gamma = [0, 1]$,
 $\lambda(\omega, \gamma) = (\omega - \gamma)^2$
- абсолютная игра: $\Omega = \{0, 1\}$, $\Gamma = [0, 1]$, $\lambda(\omega, \gamma) = |\omega - \gamma|$
- логарифмическая игра: $\Omega = \{0, 1\}$, $\Gamma = [0, 1]$

$$\lambda(\omega, \gamma) = \begin{cases} -\log_2(1 - \gamma) & \text{если } \omega = 0 \\ -\log_2 \gamma & \text{если } \omega = 1 \end{cases}$$

— потери могут принимать значение $+\infty$

- простая предсказательная игра: $\Omega = \Gamma = \{0, 1\}$

$$\lambda(\omega, \gamma) = \begin{cases} 0 & \text{if } \omega = \gamma \\ 1 & \text{if } \omega \neq \gamma \end{cases}$$

Предсказательная стратегия

- $\mathfrak{A} : \Omega^* \rightarrow \Gamma$ отображает конечные последовательности (предыдущих) исходов в предсказания
- можно рассматривать различные классы стратегий, напр., вычислимые, вычислимы за полиномиальное время и т.д.
 - но наша основная цель — изучение предсказуемости, а не вычислимости

Потери как сложность

1. Предсказание последовательностей

2. Определение сложности

3. Выпуклые игры

4. Основной результат

5. Схема доказательства

- потери стратегии \mathcal{A} на последовательности $\mathbf{x} = (\omega_1, \omega_2, \dots, \omega_n)$ это

$$\text{Loss}_{\mathcal{A}}(\mathbf{x}) = \sum_{i=1}^n \lambda(\omega_i, \mathcal{A}(\omega_1, \omega_2, \dots, \omega_{i-1}))$$

- можно считать эти потери сложностью \mathbf{x} по отношению к \mathcal{A}
- можно ли определить сложность, не зависящую от выбора стратегии \mathcal{A} ?
 - идея: возьмём «лучшую» стратегию \mathcal{A} и определим сложность \mathbf{x} как её потери на \mathbf{x}

Диагональный аргумент

- всякая нетривиальная стратегия где-нибудь работает намного хуже другой стратегии
- с другой стороны, для всякой последовательности есть стратегия, которая знает её заранее
 - если только мы не вводим ограничений по вычислимости
- определить сложность не просто

Предсказательная сложность (1)

- возможное решение: *предсказательная сложность* [Vovk and Watkins, 1998]
 - берётся класс перечислимых не-совсем-стратегий
 - там обычно есть оптимальный элемент
 - можно определить сложность конечной последовательности с точностью до константы
 - можно также рассматривать сложности с точностью до $o(n)$ [Kalnishkan and M. V. Vyugin, 2002]
 - предсказательная сложность не вычислима
- предсказательная сложность во многом аналогична колмогоровской
 - логарифмическая функция потерь задаёт «минус логарифм левинской априорной полумеры», т.е. один из вариантов колмогоровской сложности

Предсказательная сложность (2)

- некоторые свойства колмогоровской сложности можно обобщить на предсказательную
- несжимаемость → непредсказуемость [Kalnishkan, Vovk and M. V. Vyugin, 2003]
 - большинство строк не может быть сжато → большинство строк нельзя успешно предсказать
 - доля тех, которые можно сжать/предсказать, убывает экспоненциально
- при изучении предсказательной сложности возникает много технических трудностей
 - вопрос о существовании сложности частично открыт
 - вычислимость часто создаёт проблемы, не связанные с сутью процесса предсказания

Асимптотическая сложность

- мы будем рассматривать сложности *языков* (= множеств последовательностей) а не индивидуальных последовательностей
- возьмём удельные потери (потери, делённые на длину последовательности)
- и перейдём к пределу по длине
- мы получим что-то вроде

$$AC(L) = \inf_{\mathfrak{A}} \lim_{n \rightarrow +\infty} \max_{x \in L \cap \Omega^n} \frac{Loss_{\mathfrak{A}}(x)}{n}$$

Вопросы

- наша «рабочая гипотеза»

$$AC(L) = \inf_{\mathfrak{A}} \lim_{n \rightarrow +\infty} \max_{x \in L \cap \Omega^n} \frac{Loss_{\mathfrak{A}}(x)}{n}$$

- что если нет строк x длины n ?
 - пропустим это значение n
- что если нет строк x длины n и больше?
 - у таких языков сложности нет
- что если предел не существует?
 - возьмём верхний и нижний пределы; они всегда существуют
- если последовательность бесконечна, мы можем сначала взять предел $\lim_{n \rightarrow +\infty}$ вдоль последовательности а потом $\sup_{x \in L}$ по последовательностям
 - отлично, получаем ещё два варианта сложности

Конечные последовательности

- пусть $L \subseteq \Omega^*$ (т.е. L это множество конечных последовательностей)
 - пусть L бесконечно
- *верхняя* (униформная) сложность:

$$\overline{AC}(L) = \inf_{\mathfrak{A}} \limsup_{n \rightarrow +\infty} \max_{x \in L \cap \Omega^n} \frac{Loss_{\mathfrak{A}}(x)}{n}$$

- *нижняя* (униформная) сложность:

$$\underline{AC}(L) = \inf_{\mathfrak{A}} \liminf_{n \rightarrow +\infty} \max_{x \in L \cap \Omega^n} \frac{Loss_{\mathfrak{A}}(x)}{n}$$

- в первом определении мы предполагаем $\max \emptyset = 0$, а во втором $\max \emptyset = +\infty$

Бесконечные последовательности

- пусть $L \subseteq \Omega^\infty$ (т.е. L это множество бесконечных последовательностей)
- обозначим через $x|_n$ префикс x длины n
- мы можем рассмотреть множество всех конечных префиксов строк из L ; у него есть верхняя и нижняя сложности; назовём их *верхней равномерной* сложностью $\overline{AC}(L)$ и *нижней равномерной* сложностью $\underline{AC}(L)$
- *верхняя неuniformная* сложность:

$$\overline{AC}(L) = \inf_{\mathfrak{A}} \sup_{x \in L} \limsup_{n \rightarrow +\infty} \frac{\text{Loss}_{\mathfrak{A}}(x|_n)}{n}$$

- *нижняя неuniformная* сложность:

$$\underline{AC}(L) = \inf_{\mathfrak{A}} \sup_{x \in L} \liminf_{n \rightarrow +\infty} \frac{\text{Loss}_{\mathfrak{A}}(x|_n)}{n}$$

Различающие примеры

- все определённые нами сложности различны
- возьмём множество бесконечных последовательностей с чередующимися постоянными и случайными отрезками; её верхняя сложность высока, а нижняя низка
 - на конце постоянного отрезка удельные потери малы, а на конце случайного велики
 - длины отрезков должны довольно быстро расти
- рассмотрим множество бесконечных последовательностей, которые стабилизируются к нулю с какого-то места; у него низкая (\approx нулевая) неuniformная сложность и высокая (\approx максимальная) uniformная
 - на каждой последовательности удельные потери становятся маленькими, но сколь угодно поздно

Постановка задачи

- пусть у нас есть две функции потерь, λ_1 и λ_2 , задающие сложности AC_1 и AC_2
- какие отношения существуют между AC_1 и AC_2 ?
- аналогичный вопрос для предсказательной сложности рассматривается в [Kalnishkan 1999, 2002]
- ответ будет дан в следующей форме:
 - мы опишем множество точек $\{(AC_1(L), AC_2(L))\}$ в \mathbb{R}^2

1. Предсказание последовательностей

2. Определение сложности

3. Выпуклые игры

4. Основной результат

5. Схема доказательства

Ограничение

- наш результат выполняется только для игр из некоторого класса
- сейчас мы вкратце обсудим задачу *предсказания с использованием советов экспертов*

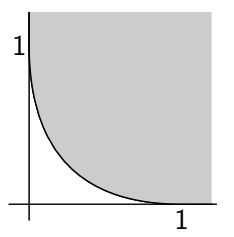
Геометрический образ игры

- рассмотрим игру $\langle \Omega, \Gamma, \lambda \rangle$ с конечным пространством исходов $\Omega = \{\omega^{(0)}, \omega^{(1)}, \dots, \omega^{(M-1)}\}$
 - построим множество образов предсказаний Γ в \mathbb{R}^M
- $$P = \{(\lambda(\omega^{(0)}, \gamma), \lambda(\omega^{(1)}, \gamma), \dots, \lambda(\omega^{(M-1)}, \gamma)) \mid \gamma \in \Gamma\} \subseteq \mathbb{R}^M$$
- назовём точку $(s_0, s_1, \dots, s_{M-1}) \in \mathbb{R}^M$ *суперпредсказанием*, если найдётся $p = (p_0, p_1, \dots, p_{M-1}) \in P$ such that

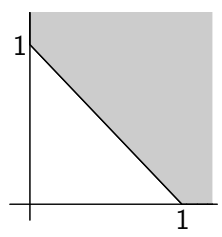
$$\begin{aligned}
 p_0 &\leq s_0 \\
 p_1 &\leq s_1 \\
 &\dots \\
 p_{M-1} &\leq s_{M-1}
 \end{aligned}$$

- суперпредсказания лежат «к северо-востоку» от предсказаний из P

Примеры (1)

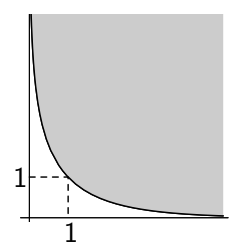


квадратичная игра
 $\lambda(\omega, \gamma) = (\omega - \gamma)^2$

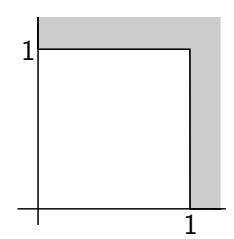


абсолютная игра
 $\lambda(\omega, \gamma) = |\omega - \gamma|$

Примеры (2)



логарифмическая игра
 $\lambda(\omega, \gamma) = \begin{cases} -\log_2(1 - \gamma), & \omega = 0 \\ -\log_2 \gamma, & \omega = 1 \end{cases}$



простая предсказательная игра
 $\lambda(\omega, \gamma) = \begin{cases} 0, & \omega = \gamma \\ 1, & \omega \neq \gamma \end{cases}$

- если множество суперпредсказаний выпукло, назовём игру *выпуклой*
 - квадратичная, абсолютная и логарифмическая игры выпуклы
 - простая предсказательная игра не выпукла
- это геометрическое свойство влечёт нетривиальные следствия...

- пусть N экспертов $\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_N$ предсказывают ту же самую последовательность
 - мы узнаём их предсказания до того, как выдаём наше
 - наша цель – предсказывать (почти) так же успешно, как лучший эксперт, в терминах кумулятивных потерь
- расширенный протокол:
 - (1) FOR $t = 1, 2, \dots$
 - (2) \mathfrak{M} считывает $\gamma_t^{(1)}, \gamma_t^{(2)}, \dots, \gamma_t^{(N)} \in \Gamma$
 - (3) \mathfrak{M} выдаёт предсказание $\gamma_t \in \Gamma$
 - (4) \mathfrak{M} наблюдает исход $\omega_t \in \Omega$
 - (5) END FOR

- мы хотим найти смешивающую стратегию, несущую потери, которые удовлетворяют условию

$$\text{Loss}_{\mathfrak{M}} \leq f(\text{Loss}_{\mathcal{E}_i})$$

где \mathcal{E}_i это наиболее успешный на данный момент эксперт

- на экспертов не накладывается никаких ограничений
 - вычислительные возможности экспертов не ограничены
 - по сути «эксперт» это метафора последовательности предсказаний, подаваемых на вход протоколу
 - мы рассматриваем антагонистическую игру «предсказатель» против «природа и эксперты»
- эта проблема изучалась с 1980-х годов; см. монографию Prediction, learning, and games, Nicolò Cesa-Bianchi and Gábor Lugosi, Cambridge University Press, 2006

- в работе [Kalnishkan and M. V. Vyugin, 2005] показано, что для выпуклых игр и только для них существует смешивающая стратегия, несущая потери

$$\text{Loss}_{\mathfrak{M}}(\mathbf{x}) \leq \text{Loss}_{\mathcal{E}_i}(\mathbf{x}) + o(|\mathbf{x}|)$$

- для всякой последовательности \mathbf{x} и набора экспертов \mathcal{E}_i ($|\mathbf{x}|$ это длина последовательности \mathbf{x})
 - если функция потерь к тому же ограничена, слагаемое $o(|\mathbf{x}|)$ можно заменить на $O(\sqrt{|\mathbf{x}|})$- результаты о квадратном корне были независимо получены в работах [Hutter and Poland, 2005] и [Cesa-Bianchi and Lugosi, 2006]

- понятие *слабой смешиваемости* введено по аналогии с понятием *смешиваемости*
- применим преобразование $x \rightarrow e^{-\eta x}$ ко всем координатам множества суперпредсказаний
- образ S под действием данного преобразования будет выпуклым для некоторого $\eta > 0$ тогда и только тогда, когда мы можем получить константный (по длине последовательности) добавочный член [Vovk 1991, 1998]

1. Предсказание последовательностей

2. Определение сложности

3. Выпуклые игры

4. Основной результат

5. Схема доказательства

Энтропия

- рассмотрим распределение p^* на Ω
 - т.е. $p^* = (p_0, p_1, \dots, p_{M-1})$, где $p_i \in [0, 1]$ и $\sum p_i = 1$
- *обобщённая энтропия*

$$H(p^*) = \min_{\gamma \in \Gamma} \mathbf{E}_{p^*} \lambda(\omega, \gamma) = \min_{\gamma \in \Gamma} \sum_{i=0}^{M-1} p_i \lambda(\omega^{(i)}, \gamma)$$

- допустим, что следующий исход случаен и имеет распределение p^*
- будем искать предсказание $\gamma \in \Gamma$, минимизирующее ожидание потерь
- этот минимум составляет $H(p^*)$

Двоичный случай

- пусть в двоичном случае p это вероятность исхода 1
 - тогда $(1 - p)$ это вероятность исхода 0
 - распределение можно отождествить с $p \in [0, 1]$
- энтропия даётся выражением

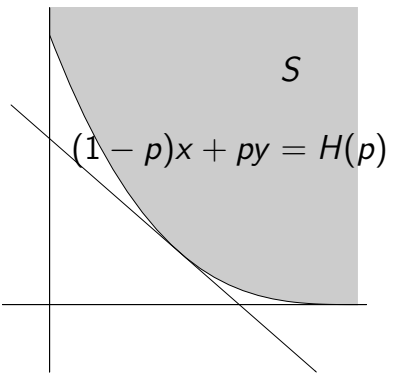
$$H(p) = \min_{\gamma \in [0,1]} [(1-p)\lambda(0, \gamma) + p\lambda(1, \gamma)]$$

- для логарифмической игры

$$H(p) = \min_{\gamma \in [0,1]} [-(1-p) \log(1-\gamma) - p \log \gamma]$$

- можно проверить (напр., дифференцированием) что минимум достигается на $\gamma = p$
- тогда $H(p) = -(1-p) \log(1-p) - p \log p$
- это шенноновская энтропия

Геометрическая интерпретация



- энтропия соответствует касательной с данным наклоном

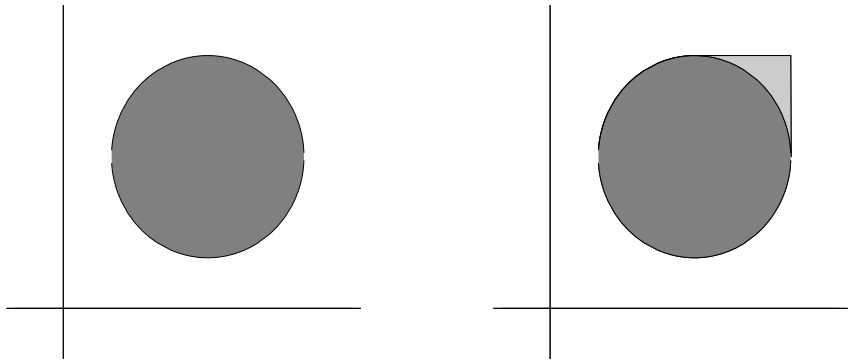
Энтропийная оболочка

- пусть даны две игры \mathcal{G}_1 и \mathcal{G}_2 (с одним и тем же Ω) — они задают энтропии H_1 и H_2
- рассмотрим множество $\{(H_1(p^*), H_2(p^*)) \mid p^* \text{ — распределение}\}$
- назовём его выпуклое замыкание $\mathcal{G}_1/\mathcal{G}_2$ -энтропийной оболочкой
- это почти решение нашей проблемы...

Некоторые типы множеств на плоскости

- будем говорить, что $M \subseteq \mathbb{R}^2$ это *звездолёт* если
 - для любой пары точек $(x_1, y_1), (x_2, y_2) \in M$
 - имеем $(\max(x_1, x_2), \max(y_1, y_2)) \in M$
- назовём выпуклое множество, не являющееся звездолётом, *репой*
- *звездолётное замыкание* множества это наименьший содержащий его звездолёт

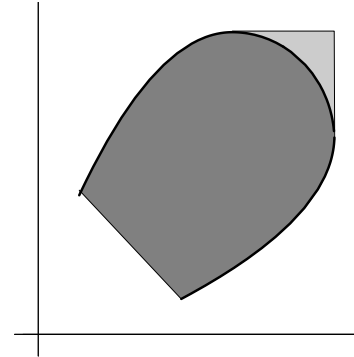
Пример



репа T

звездолётное замыкание T

- пусть \mathcal{G}_1 и \mathcal{G}_2 – две выпуклые игры с одним пространством исходов
- тогда множество точек $(AC_1(L), AC_2(L))$, где
 - AC это любая из сложностей \overline{AC} , \underline{AC} , $\overline{\underline{AC}}$, или $\underline{\overline{AC}}$
 - L пробегает множество всех непустых языков из бесконечных последовательностей или всех бесконечных языков из конечных последовательностей
- совпадает со звездолётным замыканием $\mathcal{G}_1/\mathcal{G}_2$ -энтропийной оболочки



энтропийная кривая → энтропийная оболочка → звездолёт

Обсуждение

- от условия выпуклости нельзя отказаться; в самом деле, рассмотрим простую предсказательную игру
 - $H(p) = \min(p, 1 - p) \leq 1/2$
 - $AC(\mathbb{B}^*) = 1$ (по диагональному аргументу)
- AC_1 и AC_2 должны быть тем же типом сложности; в самом деле, пусть $\mathcal{G}_2 = \mathcal{G}_1$
 - $\mathcal{G}_1/\mathcal{G}_1$ -энтропийная оболочка лежит на биссектрисе $x = y$
 - а мы знаем, что сложности различаются

1. Предсказание последовательностей

2. Определение сложности

3. Выпуклые игры

4. Основной результат

5. Схема доказательства

Сложности лежат в звездолёте

- доказательство основано на *лемме о рекалибровке*
- пусть имеется стратегия \mathcal{A} для игры \mathcal{G}_1
- мы можем построить \mathcal{A}_1 для \mathcal{G}_1 и \mathcal{A}_2 для \mathcal{G}_2 со следующими свойствами:
 - для каждой конечной x найдётся точка (u_x, v_x) из $\mathcal{G}_1/\mathcal{G}_1$ -энтропийной оболочки, такая что

$$\text{Loss}_{\mathcal{A}_1}^{\mathcal{G}_1}(x) \lesssim u_x |x|$$

$$\text{Loss}_{\mathcal{A}_2}^{\mathcal{G}_2}(x) \lesssim v_x |x|$$

— при этом \mathcal{A}_1 улучшает \mathcal{A} :

$$u_x |x| \lesssim \text{Loss}_{\mathcal{A}}^{\mathcal{G}_1}(x)$$

Доказательство леммы

- стратегия \mathcal{A} может быть дискретизована с небольшими дополнительными потерями
- дискретная стратегия опознаёт конечное множество случаев и выдаёт для каждого предсказание
 - но может быть предсказание она выдаёт неправильно; давайте её улучшим, заменив выдаваемые предсказания
 - но как?
 - рассмотрим все перестановки на конечном множестве предсказаний; каждая даёт нам новую стратегию
 - воспользуемся выпуклостью игры и смешаем их все

Сложности заполняют звездолёт

- элементарный кирпичик – множество слов длины n с примерно p^n элементами $\omega^{(i)}$ (плюс-минус)
- из этих кирпичиков можно построить язык со сложностями $H_1(p^*)$ и $H_2(p^*)$
- комбинируя кирпичики, мы заполним энтропийную оболочку
- взяв объединения, заполним звездолёт

Благодарность

- авторов вдохновили идеи Ивана Поликарова (химфак МГУ) о репе и звездолётах