O uniwersalnej kompresji i predykcji

Przypadek procesów stacjonarnych ergodycznych

Łukasz Dębowski [1] Tomasz Steifer [2]

[1]Instytut Podstaw Informatyki PAN

[2]Instytut Podstawowych Problemów Techniki PAN

Seminarium instytutowe IPI PAN, 17.05.2021

Tytułem wstępu

Kompresja i predykcja uniwersalna są ogólnym rozwiązaniem problemu uczenia się z danych w postaci procesu stacjonarnego.

- Słabo optymalna strategia dla obszernej klasy środowisk?
- Strategia taka istnieje i jest obliczalna w rozsądnym czasie.

Ale:

- Zadanie to daje się rozwiązać lepiej w pewnych podklasach.
- Rzeczywiste dane zawierają się w takiej podklasie, co uzasadnia byt maszynowego uczenia jako odrębnej dziedziny.

Plan referatu

- Procesy stacjonarne:
  - Przykład: procesy Markowa.
  - Twierdzenie ergodyczne Birkhoffa.
  - Procesy ergodyczne.
- Kompresja:
  - Nierówność Krafta.
  - Intensywność entropii.
  - Kompresja uniwersalna.
- Predykcja:
  - Nierówność Azumy.
  - Nieprzewidywalność.
  - Predykcja uniwersalna.
- Powiązania:
  - Predykcja indukowana przez kompresję.
  - Nierówność Pinskera.
  - Miara PPM (prediction by partial matching).

Procesy stacjonarne

Procesy stacjonarne

Procesy stochastyczne to nieskończone ciągi zmiennych losowych:

$$(X_i)_{i\in\mathbb{Z}} = (...X_{-2}, X_{-1}, X_0, X_1, X_2, ...). \tag{1}$$

Procesy stacjonarne to procesy stochastyczne, których rozkład jest niezmienniczy ze względu na przesunięcia:

$$P(X_{i+1}^{i+n} = x_1^n) = P(X_{j+1}^{j+n} = x_1^n), \tag{2}$$

gdzie $X_j^k := (X_j, X_{j+1}, ..., X_k)$ oraz $x_j^k := (x_j, x_{j+1}, ..., x_k)$.

Przykład:

Proces Markowa:

$$P(X_{i+1}^{i+n} = x_1^n) = \pi(x_1)\prod_{i=2}^{n}\sigma(x_i|x_{i-1}). \tag{3}$$

Warunek stacjonarności:

$$\sum_{x_1\in\mathbb{X}}\pi(x_1)\sigma(x_2|x_1) = \pi(x_2). \tag{4}$$

Twierdzenie ergodyczne Birkhoffa

Dla każdego procesu stacjonarnego względne częstości dowolnych słów w nieskończonym ciągu zmiennych losowych, zdefiniowane formalnie jako granice

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} 1\left\{ X_{i+1}^{i+k} = x_1^k \right\}, \tag{5}$$

istnieją prawie na pewno (= z prawdopodobieństwem 1).

Uwagi:

1. Dla procesu niestacjonarnego granice te mogą nie istnieć.
2. Dla procesu stacjonarnego granice te są zmiennymi losowymi.
3. Jeżeli granice te są stałymi, to wynoszą $P(X_1^k = x_1^k)$.

Procesy ergodyczne

Proces stacjonarny nazywa się ergodycznym, gdy względne częstości dowolnych słów w nieskończonym ciągu zmiennych losowych są równe prawdopodobieństwom tych słów, tzn. prawie na pewno

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} 1\left\{ X_{i+1}^{i+k} = x_1^k \right\} = P(X_1^k = x_1^k). \tag{6}$$

Przykłady:

Ergodyczny proces Markowa:
Wszystkie stany się bezpośrednio komunikują, tzn. $\sigma(x_2|x_1) > 0$ dla wszystkich stanów x_1, x_2.

Nieergodyczny proces Markowa:
Istnieją niekomunikujące się podzbiory stanów A i B, tzn. $\sigma(x_2|x_1) = 0$, jeżeli $x_1 \in A, x_2 \in B$ lub $x_1 \in B, x_2 \in A$.

Dekompozycja ergodyczna

- Dla każdego procesu Markowa zbiór stanów można rozbić na rozłączne podzbiory stanów komunikujących się. Podzbiory te definiują procesy ergodyczne.
- Analogicznie dowolny proces stacjonarny można rozłożyć na składowe ergodyczne i znaczącą część teorii procesów stacjonarnych można sprowadzić do teorii procesów stacjonarnych ergodycznych.

Kompresja

Kody bezprefiksowe

Kod bezprefiksowy to funkcja ze słów nad przeliczalnym alfabetem w słowa binarne

$$C : \mathbb{X}^* \to \{0,1\}^* \tag{7}$$

takie, że jeśli $C(u)$ jest prefiksem $C(w)$, to $u = w$.

Kody bezprefiksowe mają tę miłą własność, że z konkatenacji słów wyjściowych można odtworzyć ciąg słów wejściowych.

Nierówność Krafta:

Dla dowolnego kodu bezprefiksowego C zachodzi nierówność

$$\sum_{u \in \mathbb{X}^*} 2^{-|C(u)|} \leq 1. \tag{8}$$

Wielkości $2^{-|C(u)|}$ zachowują się trochę jak prawdopodobieństwa!

Problem bezstratnej kompresji danych

Entropia Shannona:

$$H(p) := \sum_{u \in \mathbb{X}^*} p(u) \left[-\log p(u) \right]. \tag{9}$$

Wniosek z nierówności Krafta:

Dla dowolnego kodu bezprefiksowego C zachodzi nierówność

$$\sum_{u \in \mathbb{X}^*} p(u) \, |C(u)| \geq H(p). \tag{10}$$

Kod Shannona-Fano:

Mając rozkład prawdopodobieństwa słów $p : \mathbb{X}^* \to [0, 1]$, możemy skonstruować kod Shannona-Fano C_p, który spełnia

$$|C_p(u)| = \lceil -\log p(u) \rceil. \tag{11}$$

Intensywność entropii

Dla dowolnego procesu stacjonarnego ergodycznego istnieje stała

$$h := \lim_{n\to\infty} \frac{1}{n} \mathbf{E}\left[-\log P(X_1^n)\right], \tag{12}$$

zwana intensywnością entropii.

Z nierówności Krafta i twierdzenia ergodycznego Birkhoffa wynika, że dla dowolnego kodu bezprefiksowego i procesu stacjonarnego ergodycznego

$$\limsup_{n\to\infty} \frac{1}{n} \left|C(X_1^n)\right| \geq h \text{ prawie na pewno.} \tag{13}$$

Dla kodu Shannona-Fano $C = C_P$ ta nierówność jest równością.

Kompresja uniwersalna

Kod C nazywa się uniwersalnym, gdy dla dowolnego procesu stacjonarnego ergodycznego zachodzi

$$\lim_{n\to\infty} \frac{1}{n} \left| C(X_1^n) \right| = h \text{ prawie na pewno.} \tag{14}$$

Inaczej mówiąc, kod uniwersalny dopasowuje się do statystycznych własności kompresowanego procesu ($\rightarrow$ maszynowe uczenie).

Kody uniwersalne istnieją $\iff$ alfabet $\mathbb{X}$ jest skończony.

Przykłady: kod Lempela-Ziva (`gzip`), kod PPM (`bzip2`).

Predykcja

Problem predykcji

Predyktor to funkcja, która ciągowi poprzednich symboli przypisuje pewnien symbol następny, czyli

$$f : \mathbb{X}^* \to \mathbb{X}. \tag{15}$$

Stopa błędu predyktora na słowie x_1^n to średnia liczba błędów:

$$\frac{1}{n} \sum_{i=0}^{n-1} 1\left\{ x_{i+1} \neq f(x_1^i) \right\}. \tag{16}$$

Predyktor indukowany przez miarę prawdopodobieństwa P to

$$f_P(x_1^i) := \arg\max_{x_{i+1} \in \mathbb{X}} P(X_{i+1} = x_{i+1} | X_1^i = x_1^i). \tag{17}$$

Czy ma on najmniejszą stopę błędu dla procesu o tej mierze?

Przydatne obserwacje

Wniosek z nierówności Azumy:

Stopa błędu dowolnego predyktora na procesie stochastycznym jest równa średniemu warunkowemu prawdopodobieństwu błędu, tzn. dla dowolnego procesu stochastycznego prawie na pewno

$$\lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=0}^{n-1}1\left\{X_{i+1}\neq f(X_1^i)\right\}-\frac{1}{n}\sum_{i=0}^{n-1}P(X_{i+1}\neq f(X_1^i)|X_1^i)\right]=0. \tag{18}$$

Proste ograniczenie:

Warunkowe prawdopodobieństwo błędu jest najmniejsze dla predyktora indukowanego:

$$\begin{aligned}P(X_{i+1}\neq f(X_1^i)|X_1^i) &\geq P(X_{i+1}\neq f_P(X_1^i)|X_1^i)\\ &= 1-\max_{x_{i+1}\in\mathbb{X}}P(X_{i+1}=x_{i+1}|X_1^i).\end{aligned} \tag{19}$$

Nieprzewidywalność

Dla dowolnego procesu stacjonarnego ergodycznego istnieje stała

$$u := \lim_{n\to\infty} \mathbf{E}\left[1 - \max_{x_0\in\mathbb{X}} P(X_0 = x_0 | X_{-n}^{-1})\right], \tag{20}$$

zwana nieprzewidywalnością.

Z nierówności Azumy i twierdzenia ergodycznego Birkhoffa wynika, że dla dowolnego predyktora i procesu stacjonarnego ergodycznego

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1\left\{X_{i+1} \neq f(X_1^i)\right\} \geq u \text{ prawie na pewno.} \tag{21}$$

Dla predyktora indukowanego $f = f_P$ ta nierówność jest równością.

Predyktor uniwersalny

Predyktor f nazywa się uniwersalnym, gdy dla dowolnego procesu stacjonarnego ergodycznego zachodzi

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1\left\{X_{i+1} \neq f(X_1^i)\right\} = u \text{ prawie na pewno.} \qquad (22)$$

Inaczej mówiąc, predyktor uniwersalny dopasowuje się do własności przewidywanego procesu ($\longrightarrow$ maszynowe uczenie).

Predyktory uniwersalne istnieją też dla alfabetu nieskończonego.

Przykłady: predyktor PPM.

Powiązania

Problem do rozwiązania

Miara prawdopodobieństwa R nazywa się uniwersalną, gdy dla dowolnego procesu stacjonarnego ergodycznego zachodzi

$$\lim_{n\to\infty} \frac{1}{n}\left[-\log R(X_1^n)\right] = h \text{ prawie na pewno.} \qquad (23)$$

Łatwo pokazać, że kod Shannona-Fano C_R jest uniwersalny.

Czy predyktor indukowany f_R jest także uniwersalny?

- Predykcja uniwersalna przypomina kompresję uniwersalną.
- Różnica polega na wzięciu innej funkcji straty:

$$\text{Strata dla kompresji} = \frac{\left|C_R(x_1^n)\right|}{n} \approx -\frac{1}{n}\sum_{i=0}^{n-1} \log R(x_{i+1}|x_1^i)$$

$$\text{Strata dla predykcji} = \frac{1}{n}\sum_{i=0}^{n-1} 1\left\{x_{i+1} \neq f_R(x_1^i)\right\}$$

Przydatne nierówności

Niech $\boldsymbol{p}$ i $\boldsymbol{q}$ będą dyskretnymi rozkładami prawdopodobieństwa.

Nierówność Pinskera:

$$\left[\sum_{x\in\mathbb{X}} |p(x)-q(x)|\right]^2 \leq \frac{2}{\log e}\sum_{x\in\mathbb{X}} p(x)\log\frac{p(x)}{q(x)}. \qquad (24)$$

Nierówność predykcji:

Dla $x_p = \arg\max_{x\in\mathbb{X}} p(x)$ oraz $x_q = \arg\max_{x\in\mathbb{X}} q(x)$ zachodzi

$$0 \leq p(x_p) - p(x_q) \leq \sum_{x\in\mathbb{X}} |p(x)-q(x)|\,. \qquad (25)$$

Rozwiązanie problemu

Miara prawdopodobieństwa R nazywa się uniwersalną, gdy dla dowolnego procesu stacjonarnego ergodycznego zachodzi

$$\lim_{n\to\infty} \frac{1}{n}\left[-\log R(X_1^n)\right] = h \text{ prawie na pewno.} \qquad (26)$$

Własności:

- Kod Shannona-Fano C_R jest uniwersalny.
- Predyktor indukowany f_R jest uniwersalny, jeżeli

$$-\log R(x_{n+1}|x_1^n) \leq \epsilon_n \sqrt{\frac{n}{\log n}}, \quad \lim_{n\to\infty} \epsilon_n = 0. \qquad (27)$$

Miara PPM (prediction by partial matching)

Weźmy alfabet $\mathbb{X} = \{a_1, .., a_D\}$, gdzie $D \geq 2$.

Częstość podsłowa w_1^k w słowie x_1^n to

$$N(w_1^k | x_1^n) := \sum_{i=1}^{n-k+1} 1\left\{x_i^{i+k-1} = w_1^k\right\}. \tag{28}$$

Miara PPM rzędu $k \geq 0$ to

$$\mathbf{PPM}_k(x_1^n) := D^{-k} \prod_{i=k+1}^{n} \frac{N(x_{i-k}^{i} | x_1^{i-1}) + 1}{N(x_{i-k}^{i-1} | x_1^{i-2}) + D}. \tag{29}$$

Całkowita miara PPM to

$$\mathbf{PPM}(x_1^n) := \sum_{k=0}^{\infty} \left[\frac{1}{k+1} - \frac{1}{k+2}\right] \mathbf{PPM}_k(x_1^n). \tag{30}$$

Miara **PPM**, kod $C_{\mathbf{PPM}}$ i predyktor $f_{\mathbf{PPM}}$ są uniwersalne.

Podsumowanie

Podsumowanie (= Wstęp)

Kompresja i predykcja uniwersalna są ogólnym rozwiązaniem problemu uczenia się z danych w postaci procesu stacjonarnego.

- Słabo optymalna strategia dla obszernej klasy środowisk?
- Strategia taka istnieje i jest obliczalna w rozsądnym czasie.

Ale:

- Zadanie to daje się rozwiązać lepiej w pewnych podklasach.
- Rzeczywiste dane zawierają się w takiej podklasie, co uzasadnia byt maszynowego uczenia jako odrębnej dziedziny.

Bibliografia

1. Dębowski, Ł.; Steifer, T. *Universal Coding and Prediction on Ergodic Martin-Löf Random Points*. `https://arxiv.org/abs/2005.03627`, 2020.
2. Morvai, G.; Weiss, B. *On universal algorithms for classifying and predicting stationary processes*. Probability Surveys 2021, 18, 77–131.