Praktyczne i teoretyczne problemy statystycznego modelowania języka naturalnego

Część III: Twierdzenie o faktach i słowach

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Instytut Podstaw Informatyki
Polskiej Akademii Nauk

XLIII Konferencja „Statystyka Matematyczna”
Będlewo, 04–08.12.2017

Statystyczne prawa językowe

Teksty w języku naturalnym spełniają przybliżone prawa ilościowe:

1. **Prawo Zipfa:** częstość słowa jest odwrotnie proporcjonalna do rangi słowa.
2. **Prawo Heapsa:** liczba różnych słów w tekście rośnie potęgowo z długością tekstu.
3. **Intensywność entropii Shannona:** jest rzędu 1 bita na literę.
4. **Hipoteza Hilberga:** entropia warunkowa litery maleje potęgowo z długością kontekstu.
5. **Prawo kodu PPM:** liczba różnych „słów" wykrywanych przez algorytm PPM w tekście rośnie potęgowo z długością tekstu.
6. **Prawo maksymalnego powtórzenia:** długość maksymalnego powtórzenia rośnie jak sześcian logarytmu długości tekstu.

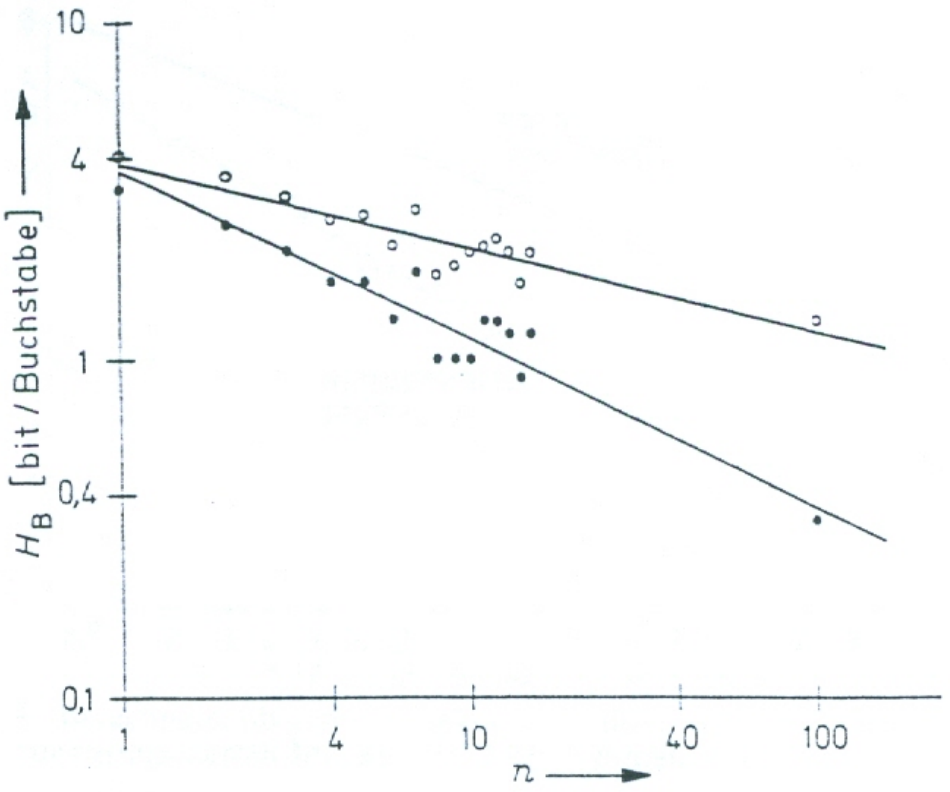
Czy można coś wywnioskować o języku jako procesie stochastycznym na podstawie tych obserwacji/hipotez?

Pytania matematyka

1. Czy istnieje idealny probabilistyczny model języka?
2. Czy model ten może być modelem Markowa?
3. Czy model ten może być ukrytym modelem Markowa?
4. Czy model ten jest ergodyczny?
5. Czy model ten jest stacjonarny?
6. Czy model ten jest asymptotycznie średnio stacjonarny?
7. Czy model ten jest kodem uniwersalnym?
8. Czy model ten jest efektywnie obliczalny?

Hipoteza Hilberga dla entropii Shannona (1990)

$$H(X) := \mathbb{E}\left[-P(X)\right] \qquad H(X|Y) := \mathbb{E}\left[-P(X|Y)\right]$$



$$H(X_n|X_1^{n-1}) \approx Bn^{\beta-1} + h, \quad \beta \approx 1/2, \quad n \leq 100$$

Prawa potęgowe dla entropii i informacji wzajemnej

Informacja wzajemna dla procesu stacjonarnego:

$$I(X_1^n; X_{n+1}^{2n}) := H(X_1^n) + H(X_{n+1}^{2n}) - H(X_1^{2n})$$
$$= 2H(X_1^n) - H(X_1^{2n})$$

Hipoteza Hilberga:

$$H(X_n|X_1^{n-1}) \approx Bn^{\beta-1} + h$$
$$\Downarrow$$
$$H(X_1^n) = \sum_{i=1}^{n} H(X_i|X_1^{i-1}) \approx B'n^{\beta} + hn$$
$$\Downarrow$$
$$I(X_1^n; X_{n+1}^{2n}) \approx B''n^{\beta}$$

Wykładnik Hilberga

Wykładnik Hilberga definiujemy jako

$$\operatorname*{hilb}_{n\to\infty} s(n) := \limsup_{n\to\infty} \frac{\log^+ s(2^n)}{\log 2^n}, \quad \log^+ x = \begin{cases} \log(x+1), & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Na przykład $\operatorname*{hilb}_{n\to\infty} n^\beta = \beta$.

Twierdzenie

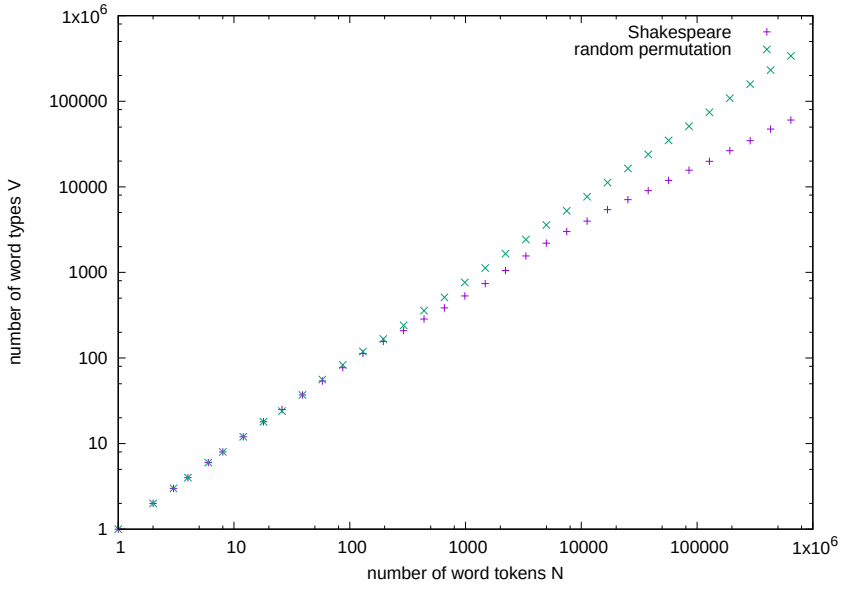
Niech $\mathfrak{J}(n) := 2\mathfrak{G}(n) - \mathfrak{G}(2n)$. Jeżeli $\lim_{n\to\infty} \mathfrak{G}(n)/n = \mathfrak{g}$, to

$$\operatorname*{hilb}_{n\to\infty} (\mathfrak{G}(n) - n\mathfrak{g}) \leq \operatorname*{hilb}_{n\to\infty} \mathfrak{J}(n),$$

gdzie równość zachodzi, gdy $\mathfrak{J}(n) \geq 0$.

Stąd $\operatorname{hilb}_{n\to\infty} \left(H(X_1^n) - nh\right) = \operatorname{hilb}_{n\to\infty} I(X_1^n; X_{n+1}^{2n})$.

Prawo Heapsa (liczba różnych słów ortograficznych)



Kuraszkiewicz i Łukaszewicz 1951; Herdan 1964; Heaps 1978:

$$V \propto N^{\gamma}, \quad \gamma < 1$$

V — liczba różnych słów w tekście (typów/types).
N — liczba wszystkich słów tekście (okazów/tokens).

Kod PPM (Prediction by Partial Matching)

Definiujemy

$$\mathbf{PPM}_k(x_i|x_1^{i-1}) := \begin{cases} \dfrac{1}{D}, & i \leq k, \\ \dfrac{N(x_{i-k}^{i}|x_1^{i-1}) + 1}{N(x_{i-k}^{i-1}|x_1^{i-2}) + D}, & i > k, \end{cases}$$

$$\mathbf{PPM}_k(x_1^n) := \prod_{i=1}^{n} \mathbf{PPM}_k(x_i|x_1^{i-1}),$$

$$\mathbf{PPM}(x_1^n) := \frac{6}{\pi^2} \sum_{k=-1}^{\infty} \frac{\mathbf{PPM}_k(x_1^n)}{(k+2)^2}.$$

Wielkość $\mathbf{PPM}(x_1^n)$ nazywa się p-stwem PPM napisu x_1^n.

Zauważmy, że $\mathbf{PPM}_k(x_1^n) = D^{-n}$ dla $k > L(x_1^n)$.

Uniwersalność p-stwa PPM

Entropia bloku: $H(X_1^n) = \mathbb{E}\left[-\log P(X_1^n)\right]$

Intensywność entropii: $h = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[-\log P(X_1^n)\right]$

Twierdzenie

P-stwo PPM jest p-stwem uniwersalnym, tzn. zachodzi

$$\mathbb{E}\left[-\log \mathrm{PPM}(X_1^n)\right] \geq \mathbb{E}\left[-\log P(X_1^n)\right]$$

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[-\log \mathrm{PPM}(X_1^n)\right] = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[-\log P(X_1^n)\right]$$

dla dowolnego procesu stacjonarnego $(X_i)_{i=1}^{\infty}$ o skończ. alfabecie.

Rząd PPM i słownik PPM

- Rząd PPM $G_{\mathrm{PPM}}(x_1^n)$ to najmniejsza liczba G taka, że

  $-\log \mathrm{PPM}_G(x_1^n) \leq -\log \mathrm{PPM}_k(x_1^n)$ dla każdego $k \geq -1$.

- Zbiór wszystkich podsłów długości m w napisie x_1^n to

  $$V(m|x_1^n) := \left\{ y_1^m : x_{t+1}^{t+m} = y_1^m \text{ dla pewnego } 0 \leq t \leq n-m \right\}.$$

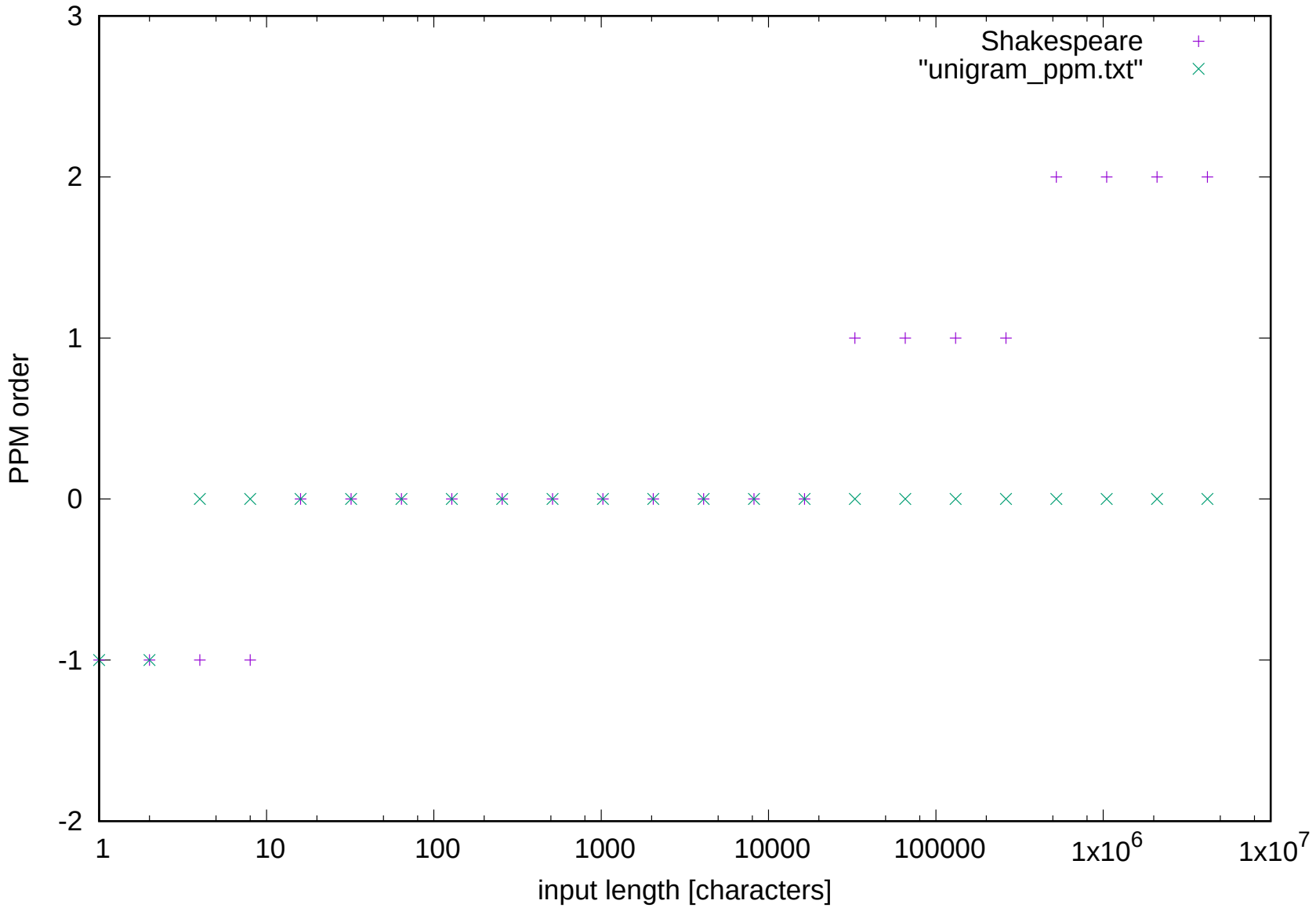
- Zbiór różnych słów PPM w napisie X_1^n to

  $$V_{\mathrm{PPM}}(x_1^n) := V(G_{\mathrm{PPM}}(x_1^n)|x_1^n).$$

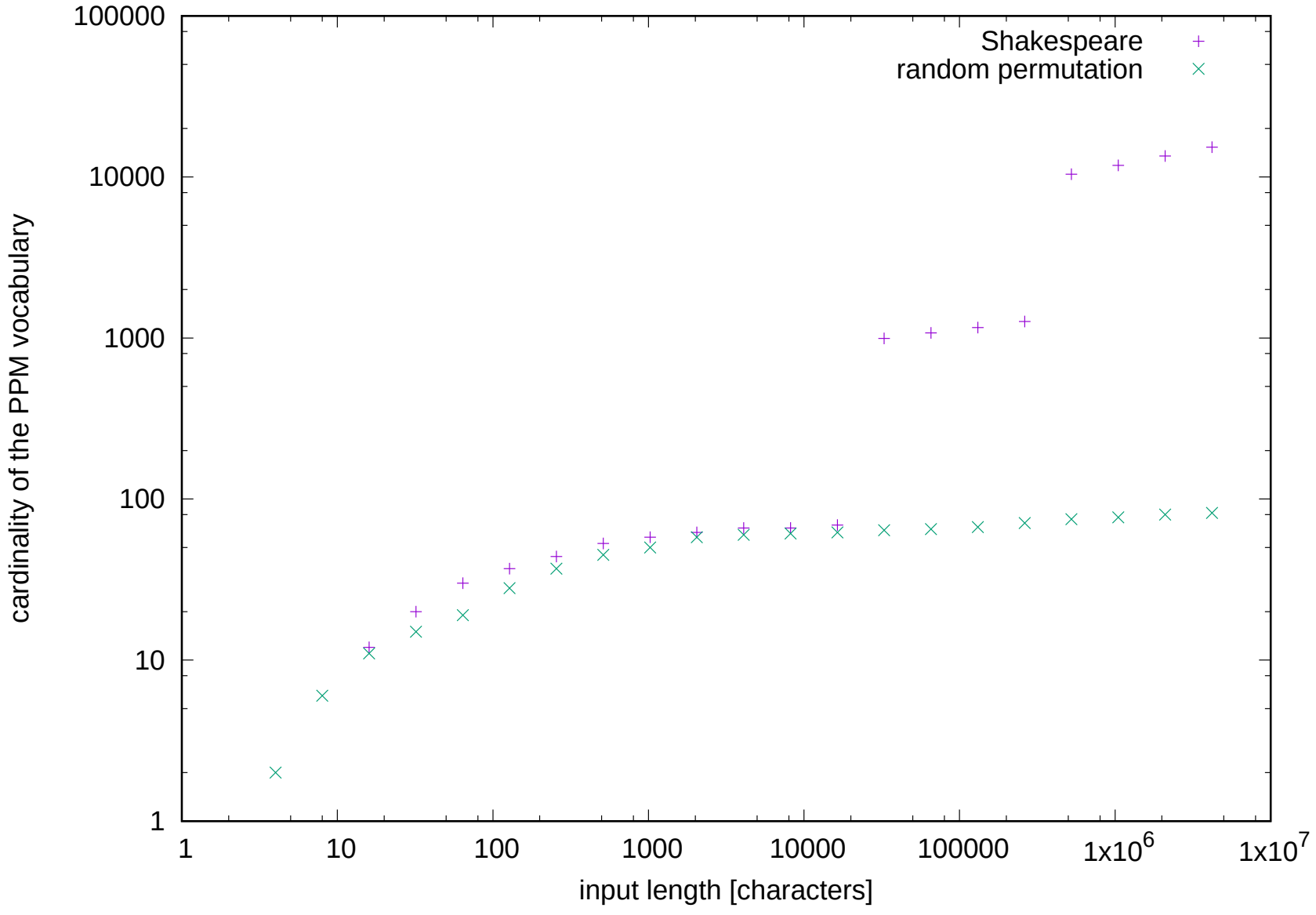
- Ogólnie zachodzi nierówność

  $$\operatorname{card} V_{\mathrm{PPM}}(x_1^n) \leq \min \left\{ D^{G_{\mathrm{PPM}}(x_1^n)}, n - G_{\mathrm{PPM}}(x_1^n) + 1 \right\}.$$

Rząd PPM na wykresie



Moc słownika PPM na wykresie



Wykładnik Hilberga dla mocy słownika PPM

Dla procesów Markowa k-tego rzędu nad skończonym alfabetem

$$\lim_{n\to\infty} G_{\mathrm{PPM}}(X_1^n) \leq k, \text{ czyli } \operatorname{hilb}_{n\to\infty} \operatorname{card} V_{\mathrm{PPM}}(X_1^n) = 0.$$

Dla języka naturalnego mamy prawdopodobnie

$$\operatorname{hilb}_{n\to\infty} \operatorname{card} V_{\mathrm{PPM}}(X_1^n) = \beta, \quad \beta \in (0,1).$$

Czyli język naturalny nie jest procesem Markowa.

W poprzednim wykładzie wykazaliśmy mocniejszy fakt, że język nie jest ukrytym procesem Markowa.

Twierdzenie o słowach i informacji wzajemnej

Entropia krzyżowa PPM:

$$H_{\mathrm{PPM}}(X_1^n) = \mathbb{E}\left[-\log \mathrm{PPM}(X_1^n)\right]$$

Informacja wzajemna PPM dla procesu stacjonarnego:

$$\begin{aligned} I_{\mathrm{PPM}}(X_1^n; X_{n+1}^{2n}) &:= H_{\mathrm{PPM}}(X_1^n) + H_{\mathrm{PPM}}(X_{n+1}^{2n}) - H_{\mathrm{PPM}}(X_1^{2n}) \\ &= 2H_{\mathrm{PPM}}(X_1^n) - H_{\mathrm{PPM}}(X_1^{2n}) \end{aligned}$$

Twierdzenie (informacja wzajemna i słowa PPM)

Dla dowolnego procesu stacjonarnego $(X_i)_{i=1}^{\infty}$ o skończ. alfabecie

$$\begin{aligned} \operatorname*{hilb}_{n\to\infty} I(X_1^n; X_{n+1}^{2n}) &\le \operatorname*{hilb}_{n\to\infty} I_{\mathrm{PPM}}(X_1^n; X_{n+1}^{2n}) \\ &\le \operatorname*{hilb}_{n\to\infty} \mathbb{E}\left[G_{\mathrm{PPM}}(X_1^n) + \operatorname{card} V_{\mathrm{PPM}}(X_1^n)\right]. \end{aligned}$$

Twierdzenie ergodyczne i procesy ergodyczne

1. Rozpatrzmy proces dyskretny $(X_i)_{i=1}^{\infty} = (X_1, X_2, X_3, ...)$.
2. Dla napisu $w = (x_1, ..., x_n)$ określmy zmienną losową

$$Y_i^w := \begin{cases} 1 & \text{jeżeli } X_i = x_1, ..., X_{i+n-1} = x_n, \\ 0 & \text{inaczej.} \end{cases}$$

3. Proces $(X_i)_{i=1}^{\infty}$ nazywa się stacjonarnym, gdy wartości oczekiwane $\mathbb{E}\, Y_i^w$ nie zależą od i dla wszystkich w.

Twierdzenie (ergodyczne)

Dla dowolnego procesu stacjonarego $(X_i)_{i=1}^{\infty}$, istnieją granice

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Y_i^w = Y^w \text{ prawie na pewno.}$$

4. Proces $(X_i)_{i=1}^{\infty}$ nazywa się ergodycznym, gdy granice Y^w są stałe dla każdego w.

Przykłady ergodycznych procesów stacjonarnych

1. Proces $(X_i)_{i=1}^{\infty}$ nazywa się IID, jeżeli

$$P(X_1 = x_1, ..., X_n = x_n) = \pi(x_1)...\pi(x_n).$$

— Procesy IID są ergodyczne.

2. Proces $(X_i)_{i=1}^{\infty}$ nazywa się Markowa, jeżeli

$$P(X_1 = x_1, ..., X_n = x_n) = \pi(x_1)p(x_2|x_1)...p(x_n|x_{n-1}).$$

— Proces Markowa jest ergodyczny, jeśli $p(x_i|x_{i-1}) > c > 0$.

3. Proces $(Y_i)_{i=1}^{\infty}$ nazywa się ukrytym Markowa, jeżeli $Y_i = f(X_i)$ dla pewnego procesu Markowa $(X_i)_{i=1}^{\infty}$.

— Ukryty proces Markowa jest ergodyczny, jeśli jego proces Markowa jest ergodyczny.

Czy język naturalny jest nieergodyczny?

1. Proces jest ergodyczny, jeżeli częstości napisów w dostatecznie długiej próbie zbiegają do wartości stałych.
2. Wyobraźmy sobie teraz, że sięgamy po tekst będący losowo wybraną książką z biblioteki.
3. Spróbujmy policzyć częstości słowa kluczowego, np. słowa *bijekcja* czy słowa *skamielina*.
4. Spodziewamy się, że częstości słów kluczowych są zmiennymi losowymi o wartościach zależych od tematu losowego tekstu.
5. Ponieważ słowa kluczowe są pewnymi napisami, to proces stochastyczny modelujący język naturalny powinien być nieergodyczny.

Zliczając słowa kluczowe, możemy rozpoznać temat tekstu.

Procesy nieergodyczne — inna perspektywa

Intuicja: Proces jest nieergodyczny $\iff$ istnieją $\geq$ dwa tematy.

Twierdzenie

Proces $(X_i)_{i=1}^{\infty}$ jest nieergodyczny wtedy i tylko wtedy, gdy istnieje funkcja $f(x_1, \ldots, x_n)$ ciągów symboli i binarna zmienna losowa Z takie, że $0 < P(Z = 0) < 1$ oraz

$$\lim_{n \to \infty} P(f(X_{t+1}, \ldots, X_{t+n}) = Z) = 1 \qquad (1)$$

dla dowolnej pozycji t.

Definicja

Zmienną losową Z spełniającą (1) nazywam losowym faktem.

Zatem proces jest nieergodyczny, jeśli istnieje $\geq$ jeden losowy fakt. Losowy fakt informuje, na który z dwóch tematów jest tekst.

Proces Santa Fe — przykład procesu nieergodycznego

1. Niech $(Z_k)_{k=1}^{\infty}$ będzie procesem IID o $Z_k \in \{0, 1\}$ oraz
$$P(Z_k = 0) = P(Z_k = 1) = 1/2.$$
2. Niech $(K_i)_{i=1}^{\infty}$ będzie procesem IID o $K_i \in \{1, 2, 3, ...\}$ oraz
$$P(K_i = k) \propto \frac{1}{k^{\alpha}}, \quad \alpha > 1.$$
3. Proces Santa Fe to proces $(X_i)_{i=1}^{\infty}$, gdzie
$$X_i = (K_i, Z_{K_i}).$$
4. Proces Santa Fe jest nieergodyczny, gdyż wszystkie Z_k są probabilistycznie niezależnymi faktami losowymi.

Mocna nieergodyczność

Intuicja: Proces Santa Fe process jest mocno nieergodyczny, gdyż istnieje nieskończenie wiele probab. niezależnych faktów losowych.

Definicja

Proces $(X_i)_{i=1}^{\infty}$ nazywa się mocno nieergodycznym, gdy istnieją funkcje $f_k(x_1, ..., x_n)$ ciągów symboli i binarny proces IID $(Z_k)_{k=1}^{\infty}$ takie, że $P(Z_k = 0) = P(Z_k = 1) = 1/2$ oraz

$$\lim_{n\to\infty} P(f_k(X_{t+1}, ..., X_{t+n}) = Z_k) = 1$$

dla dowolnej pozycji t i dowolnego $k = 1, 2, 3, ...$

(liczba tematów) $\approx 2^{(\text{liczba niezależnych faktów losowych})}$

mocna nieergodyczność $\iff$ kontinuum tematów

Potęgowy wzrost liczby przewidywalnych faktów

Zbiór niezależnych losowych faktów przewidywalnych z X_1^n to

$$U(X_1^n) := \left\{ l \in \{1, 2, ...\} : f_k(X_1^n) = Z_k \text{ for all } k \leq l \right\}.$$

Dla procesów Santa Fe mamy dokładne równości

$$\operatorname*{hilb}_{n\to\infty} \mathbb{E} \operatorname{card} U(X_1^n) = \operatorname*{hilb}_{n\to\infty} I(X_1^n; X_{n+1}^{2n}) = \beta,$$

gdzie $\beta = 1/\alpha \in (0, 1)$.

W ogólnym przypadku wykładniki Hilberga mogą być różne.

Twierdzenie o faktach i słowach

Twierdzenie (fakty i informacja wzajemna)

Dla dowolnego mocno nieergodycznego procesu stacjonarnego $(X_i)_{i=1}^{\infty}$ o skończ. alfabecie

$$\operatorname*{hilb}_{n\to\infty} \mathbb{E}\ \operatorname{card} U(X_1^n) \leq \operatorname*{hilb}_{n\to\infty} I(X_1^n; X_{n+1}^{2n}).$$

Twierdzenie (informacja wzajemna i słowa PPM)

Dla dowolnego procesu stacjonarnego $(X_i)_{i=1}^{\infty}$ o skończ. alfabecie

$$\begin{aligned}\operatorname*{hilb}_{n\to\infty} I(X_1^n; X_{n+1}^{2n}) &\leq \operatorname*{hilb}_{n\to\infty} I_{\mathrm{PPM}}(X_1^n; X_{n+1}^{2n}) \\ &\leq \operatorname*{hilb}_{n\to\infty} \mathbb{E}\left[G_{\mathrm{PPM}}(X_1^n) + \operatorname{card} V_{\mathrm{PPM}}(X_1^n)\right].\end{aligned}$$

Konkluzje

Czy język naturalny jest mocno nieergodyczny?

Nie możemy tego wykluczyć, gdyż:

1. Liczba różnych słów PPM w tekstach zdaje się rosnąć potęgowo z długością tekstu.
2. W świetle twierdzenia o faktach i słowach liczba niezależnych faktów przewidywalnych na podstawie tekstu może także rosnąć potęgowo z długością tekstu.

Problem otwarty

Udowodnić, że zachodzi prawie na pewno

$$\operatorname*{hilb}_{n\to\infty} \operatorname{card} U(X_1^n) \leq \operatorname*{hilb}_{n\to\infty} \left[G_{\mathrm{PPM}}(X_1^n) + \operatorname{card} V_{\mathrm{PPM}}(X_1^n) \right].$$

Podejście w duchu złożoności Kołmogorowa

1. Modyfikacje definicji mocnej nieergodyczności:

   $(Z_k)_{k=1}^{\infty} \longrightarrow (z_k)_{k=1}^{\infty}$ — ciąg algorytmicznie losowy

   $(f_k)_{k=1}^{\infty} \longleftarrow$ funkcja efektywnie obliczalna

2. Modyfikacja zbioru przewidywalnych faktów:

$$U_a(X_1^n) := \left\{ l \in \{1, 2, \ldots\} : f_k(X_1^n) = z_k \text{ for all } k \leq l \right\}.$$

3. Dla zmodyfikowanego procesu Santa Fe postaci $X_i = (K_i, z_{K_i})$, gdzie K_i jak poprzednio, mamy

$$\operatorname*{hilb}_{n\to\infty} \mathbb{E} \operatorname{card} U_a(X_1^n) = \beta,$$

   gdzie $\beta = 1/\alpha \in (0, 1)$.

4. Informację wzajemną $I(X_1^n; X_{n+1}^{2n})$ zastępujemy przez oczekiwaną algorytmiczną informację wzajemną $I_a(X_1^n; X_{n+1}^{2n})$

Algorytmiczne twierdzenie o faktach i słowach

Twierdzenie (fakty i informacja wzajemna)

Dla dowolnego procesu stacjonarnego $(X_i)_{i=1}^{\infty}$ o skończ. alfabecie

$$\operatorname*{hilb}_{n\to\infty} \mathbb{E}\ \mathbf{card}\, U_a(X_1^n) \leq \operatorname*{hilb}_{n\to\infty} I_a(X_1^n; X_{n+1}^{2n}).$$

Twierdzenie (informacja wzajemna i słowa PPM)

Dla dowolnego procesu stacjonarnego $(X_i)_{i=1}^{\infty}$ o skończ. alfabecie

$$\begin{aligned}\operatorname*{hilb}_{n\to\infty} I_a(X_1^n; X_{n+1}^{2n}) &\leq \operatorname*{hilb}_{n\to\infty} I_{\mathrm{PPM}}(X_1^n; X_{n+1}^{2n}) \\ &\leq \operatorname*{hilb}_{n\to\infty} \mathbb{E}\left[G_{\mathrm{PPM}}(X_1^n) + \mathbf{card}\, V_{\mathrm{PPM}}(X_1^n)\right].\end{aligned}$$