O problemie identyfikacji w granicy

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Instytut Podstaw Informatyki PAN

Estymacja zgodna a identyfikacja w granicy

$X_1^n = (X_1, ..., X_n)$ — dyskretne zmienne losowe, $X_i \in \mathbb{N}$.

Estymacja zgodna a identyfikacja w granicy

$X_1^n = (X_1, ..., X_n)$ — dyskretne zmienne losowe, $X_i \in \mathbb{N}$.

Estymacja zgodna parametru:

1. Mamy nieprzeliczalną rodzinę rozkładów $\{P_\theta : \theta \in \mathbb{R}\}$.
2. Szukamy takiego estymatora $\hat{\theta}_n = \hat{\theta}_n(X_1^n) \in \mathbb{R}$, że

$$\lim_{n\to\infty} \hat{\theta}_n = \theta \quad P_\theta\text{-prawie na pewno, } \forall \theta. \tag{1}$$

Estymacja zgodna a identyfikacja w granicy

$X_1^n = (X_1, ..., X_n)$ — dyskretne zmienne losowe, $X_i \in \mathbb{N}$.

Estymacja zgodna parametru:

1. Mamy nieprzeliczalną rodzinę rozkładów $\{P_\theta : \theta \in \mathbb{R}\}$.
2. Szukamy takiego estymatora $\hat{\theta}_n = \hat{\theta}_n(X_1^n) \in \mathbb{R}$, że

$$\lim_{n\to\infty} \hat{\theta}_n = \theta \quad P_\theta\text{-prawie na pewno, } \forall\theta. \tag{1}$$

Identyfikacja w granicy:

1. Mamy przeliczalną rodzinę rozkładów $\{P_a : a \in \mathbb{N}\}$.
2. Szukamy takiego estymatora $\hat{a}_n = \hat{a}_n(X_1^n) \in \mathbb{N}$, że

$$\lim_{n\to\infty} \hat{a}_n = a \quad P_a\text{-prawie na pewno, } \forall a. \tag{2}$$

Estymacja zgodna a identyfikacja w granicy

$X_1^n = (X_1, ..., X_n)$ — dyskretne zmienne losowe, $X_i \in \mathbb{N}$.

Estymacja zgodna parametru:

1. Mamy nieprzeliczalną rodzinę rozkładów $\{P_\theta : \theta \in \mathbb{R}\}$.
2. Szukamy takiego estymatora $\hat{\theta}_n = \hat{\theta}_n(X_1^n) \in \mathbb{R}$, że

$$\lim_{n\to\infty} \hat{\theta}_n = \theta \quad P_\theta\text{-prawie na pewno, } \forall\theta. \tag{1}$$

Identyfikacja w granicy:

1. Mamy przeliczalną rodzinę rozkładów $\{P_a : a \in \mathbb{N}\}$.
2. Szukamy takiego estymatora $\hat{a}_n = \hat{a}_n(X_1^n) \in \mathbb{N}$, że

$$\lim_{n\to\infty} \hat{a}_n = a \quad P_a\text{-prawie na pewno, } \forall a. \tag{2}$$

Jeżeli (2) zachodzi, to $\hat{a}_n = a$ dla dostatecznie dużych n.

Kontekst historyczny

Historycznie, problem identyfikacji w granicy związany jest z problemem wyuczalności języków formalnych:

1. Mamy rodzinę języków $\{L_a : a \in \mathbb{X}^*\}$, gdzie $L_a \subset \mathbb{X}^*$.
2. Szukamy takiego estymatora $\hat{a}_n = \hat{a}_n(X_1^n) \in \mathbb{X}^*$, że

$$\lim_{n\to\infty} d(\hat{a}_n, a) = 0 \tag{3}$$

dla dość dowolnej sekwencji $X_i \in L_a$.

Kontekst historyczny

Historycznie, problem identyfikacji w granicy związany jest z problemem wyuczalności języków formalnych:

1. Mamy rodzinę języków $\{L_a : a \in \mathbb{X}^*\}$, gdzie $L_a \subset \mathbb{X}^*$.
2. Szukamy takiego estymatora $\hat{a}_n = \hat{a}_n(X_1^n) \in \mathbb{X}^*$, że

$$\lim_{n \to \infty} d(\hat{a}_n, a) = 0 \tag{3}$$

   dla dość dowolnej sekwencji $X_i \in L_a$.

Gold (1967) badał ten problem dla kilku klas języków formalnych.

1. Dla klasy języków skończonych, tzn. gdy wszystkie L_a są skończone oraz $\{X_i : i \in \mathbb{N}\} = L_a$, problem ma rozwiązanie.
2. Jeżeli dopuścić języki nieskończone, pojawiają się kłopoty.

Kontekst historyczny

Historycznie, problem identyfikacji w granicy związany jest z problemem wyuczalności języków formalnych:

1. Mamy rodzinę języków $\{L_a : a \in \mathbb{X}^*\}$, gdzie $L_a \subset \mathbb{X}^*$.
2. Szukamy takiego estymatora $\hat{a}_n = \hat{a}_n(X_1^n) \in \mathbb{X}^*$, że

$$\lim_{n\to\infty} d(\hat{a}_n, a) = 0 \tag{3}$$

dla dość dowolnej sekwencji $X_i \in L_a$.

Gold (1967) badał ten problem dla kilku klas języków formalnych.

1. Dla klasy języków skończonych, tzn. gdy wszystkie L_a są skończone oraz $\{X_i : i \in \mathbb{N}\} = L_a$, problem ma rozwiązanie.
2. Jeżeli dopuścić języki nieskończone, pojawiają się kłopoty.

Praca Golda wywarła spory wpływ na generatywne teorie języka i koncepcję wrodzonej gramatyki uniwersalnej.

Jeden z nowszych wyników

Twierdzenie (Vitányi & Chater, 2017)

Jeżeli $\{P_a : a \in \mathbb{N}\}$ jest ciągiem rozkładów IID, gdzie $P_a \neq P_b$ dla $a \neq b$, to istnieje taki estymator $\hat{a}_n = \hat{a}_n(X_1^n) \in \mathbb{N}$, że

$$\lim_{n\to\infty} \hat{a}_n = a \quad P_a\text{-prawie na pewno, } \forall a. \tag{4}$$

Jeden z nowszych wyników

Twierdzenie (Vitányi & Chater, 2017)

Jeżeli $\{P_a : a \in \mathbb{N}\}$ jest ciągiem rozkładów IID, gdzie $P_a \neq P_b$ dla $a \neq b$, to istnieje taki estymator $\hat{a}_n = \hat{a}_n(X_1^n) \in \mathbb{N}$, że

$$\lim_{n\to\infty} \hat{a}_n = a \quad P_a\text{-prawie na pewno, } \forall a. \tag{4}$$

W dowodzie jest pewien błąd, który naprawię.

Kluczowy lemat

Oznaczmy:

- $F_n(x) := n^{-1} \sum_{i=1}^{n} 1\{X_i = x\}$,
- $p(x) := P(X_i = x)$.

Lemat

Dla P będącego rozkładem IID zachodzi

$$\sup_{x \in \mathbb{N}} |F_n(x) - p(x)| \leq C_n := n^{-1/4+\epsilon} \tag{5}$$

dla dostatecznie dużych n, P-prawie na pewno, dla każdego $\epsilon > 0$.

Kluczowy lemat

Oznaczmy:

- $F_n(x) := n^{-1} \sum_{i=1}^{n} 1\{X_i = x\}$,
- $p(x) := P(X_i = x)$.

Lemat

Dla P będącego rozkładem IID zachodzi

$$\sup_{x \in \mathbb{N}} |F_n(x) - p(x)| \leq C_n := n^{-1/4+\epsilon} \tag{5}$$

dla dostatecznie dużych n, P-prawie na pewno, dla każdego $\epsilon > 0$.

$$P\left(\sum_{x \in \mathbb{N}} (F_n(x) - p(x))^4 \geq \delta\right) \leq \delta^{-1}\mathbb{E} \sum_{x \in \mathbb{N}} (F_n(x) - p(x))^4$$
$$\leq \delta^{-1} n^{-3}(1 + 3(n-2)).$$

Wprowadzenie do konstrukcji estymatora

Oznaczmy:

- $A_n := \{x \in \mathbb{N} : F_n(x) > 0\}$,
- $B_{b,n} := \{1, ..., m\}$, gdzie m jest najmniejszą liczbą taką, że $1 - \sum_{x=1}^{m} p_b(x) \leq C_n$, $p_b(x) := P_b(X_i = x)$.
- $L_{b,n} := A_n \cup B_{b,n}$.

Wprowadzenie do konstrukcji estymatora

Oznaczmy:

- $A_n := \{x \in \mathbb{N} : F_n(x) > 0\}$,
- $B_{b,n} := \{1, ..., m\}$, gdzie m jest najmniejszą liczbą taką, że $1 - \sum_{x=1}^{m} p_b(x) \leq C_n$, $p_b(x) := P_b(X_i = x)$.
- $L_{b,n} := A_n \cup B_{b,n}$.

Mamy:

- P_a-prawie na pewno dla dostatecznie dużych n dla wszystkich $x \in L_{a,n}$ zachodzi

$$|F_n(x) - p_a(x)| \leq C_n. \tag{6}$$

- Jeżeli $b \neq a$, to P_a-prawie na pewno dla dostatecznie dużych n istnieje $x \in L_{b,n}$ i nie istnieje $x \in \mathbb{N} \setminus L_{b,n}$ takie, że

$$|F_n(x) - p_b(x)| > C_n. \tag{7}$$

Konstrukcja estymatora

Poniższy algorytm odrzuca w granicy wszystkie $b < a$ (jest ich skończenie wiele), natomiast nie odrzuca właściwego a.

1: $I \leftarrow \emptyset$
2: **for** $a \in \{1, 2, ..., n\}$ **do**
3: **if** $\max_{x \in L_{a,n}} |F_n(x) - p_a(x)| < C_n$ **then**
4: $I \leftarrow I \cup \{a\}$
5: **end if**
6: **end for**
7: $\hat{a}_n \leftarrow \min I$

Kilka uwag krytycznych

1. Vitanyi i Chater próbowali uogólniać powyższą konstrukcję na przypadek procesów Markowa. W ich rozumowaniu są luki.

Kilka uwag krytycznych

1. Vitanyi i Chater próbowali uogólniać powyższą konstrukcję na przypadek procesów Markowa. W ich rozumowaniu są luki.
2. Stosując inne rozumowanie, Vitanyi i Chater podali także rozwiązanie problemu identyfikacji w granicy dla przypadku rekurencyjnie wyliczalnej rodziny miar prawdopodobieństwa.

Kilka uwag krytycznych

1. Vitanyi i Chater próbowali uogólniać powyższą konstrukcję na przypadek procesów Markowa. W ich rozumowaniu są luki.
2. Stosując inne rozumowanie, Vitanyi i Chater podali także rozwiązanie problemu identyfikacji w granicy dla przypadku rekurencyjnie wyliczalnej rodziny miar prawdopodobieństwa.
3. Podany algorytm identyfikacji w granicy jest nieodporny na błędną specyfikację. Jeżeli właściwego rozkładu nie ma w ciągu $\{P_a : a \in \mathbb{N}\}$, to $\hat{a}_n$ rozbiega do nieskończoności, zamiast zbiegać do najbliższego P_a.

Kilka uwag krytycznych

1. Vitanyi i Chater próbowali uogólniać powyższą konstrukcję na przypadek procesów Markowa. W ich rozumowaniu są luki.
2. Stosując inne rozumowanie, Vitanyi i Chater podali także rozwiązanie problemu identyfikacji w granicy dla przypadku rekurencyjnie wyliczalnej rodziny miar prawdopodobieństwa.
3. Podany algorytm identyfikacji w granicy jest nieodporny na błędną specyfikację. Jeżeli właściwego rozkładu nie ma w ciągu $\{P_a : a \in \mathbb{N}\}$, to $\hat{a}_n$ rozbiega do nieskończoności, zamiast zbiegać do najbliższego P_a.
4. W rzeczywistych problemach prawdopodobnie mamy do czynienia z nieprzeliczalnymi rodzinami rozkładów i zmiennymi zależnymi.

Kilka uwag krytycznych

1. Vitanyi i Chater próbowali uogólniać powyższą konstrukcję na przypadek procesów Markowa. W ich rozumowaniu są luki.
2. Stosując inne rozumowanie, Vitanyi i Chater podali także rozwiązanie problemu identyfikacji w granicy dla przypadku rekurencyjnie wyliczalnej rodziny miar prawdopodobieństwa.
3. Podany algorytm identyfikacji w granicy jest nieodporny na błędną specyfikację. Jeżeli właściwego rozkładu nie ma w ciągu $\{P_a : a \in \mathbb{N}\}$, to $\hat{a}_n$ rozbiega do nieskończoności, zamiast zbiegać do najbliższego P_a.
4. W rzeczywistych problemach prawdopodobnie mamy do czynienia z nieprzeliczalnymi rodzinami rozkładów i zmiennymi zależnymi.
5. Założenie, że rozkłady są obliczalne jest (nie)realistyczne: Rzeczywistość nie musi być obliczalna, nasze modele jej muszą być obliczalne.

Pytanie otwarte

Czy istnieje algorytm identyfikacji w granicy w przypadku danych zależnych, możliwie odporny na błędną specyfikację?

Obiecujące twierdzenie

Oznaczmy:

- $F_n(w) := n^{-1} \sum_{i=1}^{n} 1\left\{X_i^{i+|w|-1} = w\right\}$,
- $p(w) := P(X_i^{i+|w|-1} = w)$.

Jednostajne twierdzenie ergodyczne (Adams & Nobel, 2010)

Dla P będącego miarą stacjonarną ergodyczną zachodzi

$$\lim_{n\to\infty} \sup_{w\in\mathbb{X}^*} |F_n(w) - p(w)| = 0 \quad P\text{-prawie na pewno.} \tag{8}$$

Obiecujące twierdzenie

Oznaczmy:

- $F_n(w) := n^{-1} \sum_{i=1}^{n} 1\left\{X_i^{i+|w|-1} = w\right\}$,
- $p(w) := P(X_i^{i+|w|-1} = w)$.

Jednostajne twierdzenie ergodyczne (Adams & Nobel, 2010)

Dla P będącego miarą stacjonarną ergodyczną zachodzi

$$\lim_{n\to\infty} \sup_{w\in\mathbb{X}^*} |F_n(w) - p(w)| = 0 \quad P\text{-prawie na pewno.} \tag{8}$$

Powyższe twierdzenie zachodzi, gdyż wymiar Vapnika-Chervonenkisa klasy cylindrów jest skończony (wynosi 2).

Obiecujące twierdzenie

Oznaczmy:

- $F_n(w) := n^{-1} \sum_{i=1}^{n} 1\left\{X_i^{i+|w|-1} = w\right\}$,
- $p(w) := P(X_i^{i+|w|-1} = w)$.

Jednostajne twierdzenie ergodyczne (Adams & Nobel, 2010)

Dla P będącego miarą stacjonarną ergodyczną zachodzi

$$\lim_{n\to\infty} \sup_{w\in\mathbb{X}^*} |F_n(w) - p(w)| = 0 \quad P\text{-prawie na pewno.} \tag{8}$$

Powyższe twierdzenie zachodzi, gdyż wymiar Vapnika-Chervonenkisa klasy cylindrów jest skończony (wynosi 2).

Dla rodziny miar, dla których (8) zachodzi jednostajnie, możemy zastosować zmodyfikowany algorytm identyfikacji w granicy.

Algorytm naiwny, jakby odporniejszy, ale czy zgodny?

Poniższy algorytm mógłby być odporny na błędną specyfikację, ale czy jest on estymatorem zgodnym?

1: $C \leftarrow \infty$
2: **for** $a \in \{1, 2, ..., n\}$ **do**
3: $\quad C_a \leftarrow \max_{x \in \mathbb{N}} |F_n(x) - p_a(x)|$
4: $\quad$ **if** $C_a < C$ **then**
5: $\quad\quad C \leftarrow C_a$
6: $\quad\quad \hat{a}_n \leftarrow a$
7: $\quad$ **end if**
8: **end for**

Algorytm naiwny, jakby odporniejszy, ale czy zgodny?

Poniższy algorytm mógłby być odporny na błędną specyfikację, ale czy jest on estymatorem zgodnym?

1: $C \leftarrow \infty$
2: **for** $a \in \{1, 2, ..., n\}$ **do**
3: $\quad C_a \leftarrow \max_{x \in \mathbb{N}} |F_n(x) - p_a(x)|$
4: $\quad$ **if** $C_a < C$ **then**
5: $\quad\quad C \leftarrow C_a$
6: $\quad\quad \hat{a}_n \leftarrow a$
7: $\quad$ **end if**
8: **end for**

Czy $\max_{x \in \mathbb{N}} |p_b(x) - p_a(x)|$ to odległość, o którą nam chodzi?

Podsumowanie

1. Problem identyfikacji w granicy wywodzi się z zainteresowania informatyków problemem uczenia się języków formalnych.

Podsumowanie

1. Problem identyfikacji w granicy wywodzi się z zainteresowania informatyków problemem uczenia się języków formalnych.
2. Jest to problem nietrywialny: Celem jest odgadnięcie rozkładu nad nieskończonym alfabetem na podstawie skończonej próby.

Podsumowanie

1. Problem identyfikacji w granicy wywodzi się z zainteresowania informatyków problemem uczenia się języków formalnych.
2. Jest to problem nietrywialny: Celem jest odgadnięcie rozkładu nad nieskończonym alfabetem na podstawie skończonej próby.
3. Jest to problem w swojej istocie dyskretny, będący szczególnym przypadkiem estymacji parametru rzeczywistego.

Podsumowanie

1. Problem identyfikacji w granicy wywodzi się z zainteresowania informatyków problemem uczenia się języków formalnych.
2. Jest to problem nietrywialny: Celem jest odgadnięcie rozkładu nad nieskończonym alfabetem na podstawie skończonej próby.
3. Jest to problem w swojej istocie dyskretny, będący szczególnym przypadkiem estymacji parametru rzeczywistego.

Podsumowanie

1. Problem identyfikacji w granicy wywodzi się z zainteresowania informatyków problemem uczenia się języków formalnych.
2. Jest to problem nietrywialny: Celem jest odgadnięcie rozkładu nad nieskończonym alfabetem na podstawie skończonej próby.
3. Jest to problem w swojej istocie dyskretny, będący szczególnym przypadkiem estymacji parametru rzeczywistego.

Spojrzenie na ten problem z boku, okiem statystyka, może przyczynić się do pewnego postępu w jego analizie.

Bibliografia

1. E. M. Gold (1967), Language Identification in the Limit, *Information and Control* 10, 447–474.
2. P. M. B. Vitányi & N. Chater (2017), Identification of Probabilities, *Journal of Mathematical Psychology* 76 A, 13–24.
3. T. M. Adams & A. B. Nobel (2010), Uniform Convergence of Vapnik-Chervonenkis Classes under Ergodic Sampling, *The Annals of Probability* 38 (4), 1345–1367.