COLT - Obliczeniowa teoria uczenia się

Hung Son Nguyen

Institute of Mathematics, Warsaw University
son@mimuw.edu.pl

2007

Spis treści

Spis treści

Spis treści

Spis treści

Spis treści

Spis treści

Źródło informacji:

Uczeń może pozyskać informacje o dziedzinie poprzez:

1. **Przykłady:** Uczeń dostaje pozytywne i/lub negatywne przykłady. Przykłady mogą być zdobywane w sposób:
   1. losowy: według pewnego znanego lub nieznanego rozkładu;
   2. arbitralny;
   3. złośliwy: (np. przez kontrolera, który chciałby poznać najgorsze zachowanie algorytmu uczenia się);
   4. specjalny przez życzliwego nauczyciela: (np., aby ułatwiać proces uczenia się)
2. **Zapytania:** uczeń zdobywa informacje o dziedzinie przez zadawanie zapytań do nauczyciela.
3. **Eksperymentowanie:** aktywne uczenie się.

Teoria uczenia się

- **Podejście indukcyjne:** wnioskowanie na podstawie skończonego zbioru obserwacji;
- Np. Pokazać, że dla każdego $n \in \mathbb{N}$

$$1^2 + 2^2 + ... + n^2 = \frac{n(n+1)(2n+1)}{6}$$

- Jakie prawa rządzą w podejściu uczenia indukcyjnego?

  Szukamy teorii dotyczącej:

  - Prawdopodobieństwa udanego uczenia się;
  - Liczby przykładów treningowych;
  - Złożoności przestrzeni hipotez;
  - Skuteczności aproksymacji;
  - Sposób reprezentacji danych treningowych;

Teoria uczenia się

- **Podejście indukcyjne:** wnioskowanie na podstawie skończonego zbioru obserwacji;
- Np. Pokazać, że dla każdego $n \in \mathbb{N}$

$$1^2 + 2^2 + ... + n^2 = \frac{n(n+1)(2n+1)}{6}$$

- Jakie prawa rządzą w podejściu uczenia indukcyjnego?

  Szukamy teorii dotyczącej:
  - Prawdopodobieństwa udanego uczenia się;
  - Liczby przykładów treningowych;
  - Złożoności przestrzeni hipotez;
  - Skuteczności aproksymacji;
  - Sposób reprezentacji danych treningowych;

Teoria uczenia się

- **Podejście indukcyjne:** wnioskowanie na podstawie skończonego zbioru obserwacji;
- Np. Pokazać, że dla każdego $n \in \mathbb{N}$

$$1^2 + 2^2 + ... + n^2 = \frac{n(n+1)(2n+1)}{6}$$

- Jakie prawa rządzą w podejściu uczenia indukcyjnego?

  Szukamy teorii dotyczącej:

  - Prawdopodobieństwa udanego uczenia się;
  - Liczby przykładów treningowych;
  - Złożoności przestrzeni hipotez;
  - Skuteczności aproksymacji;
  - Sposób reprezentacji danych treningowych;

Kryteria oceny jakości:

Skąd wiemy, czy uczeń się nauczył lub jak dobrze się nauczył?

- Miary off-line (batch) vs. on-line (interactive).
- Jakość opisu vs. jakość predykcji
- Skuteczność: obliczona na podstawie błędu klasyfikacji, dokładności opisu ...
- Efektywność uczenia: wymagana jest wielomianowa złożoność obliczeniowa.

Przykład

- Załóżmy, że chcemy nauczyć się pojęcia ”człowieka o średniej budowie ciała”. Dane – czyli osoby – są reprezentowane przez punkty $(wzrost(cm), waga(Kg))$ i są etykietowane przez $+$ dla pozytywnych przykładów i $-$ dla negatywnych.
- Dodatkowa wiedza: szukane pojęcie można wyrazić za pomocą PROSTOKĄTA
- Na przykład dany jest etykietowany zbiór: $((84, 184), +)$, $((70, 170), +)$, $((75, 163), -)$, $((80, 180), +)$, $((81, 195), -)$, $((63, 191), -)$, $((77, 187), -)$, $((68, 168), +)$
- Znajdź etykietę $((79, 183, ?)$

Problem uczenia się prostokąta

Możemy definiować problem jak następująco:

- **Cel:** Znaleźć w $\Re^2$ prostokąt R o bokach równoległych do osi.
- **Wejście:** Zbiór zawierający przykłady w postaci punktów $((x, y), +/-)$. Te punkty zostały wygenerowane losowo.
- **Wyjście:** hipotetyczny prostokąt R' będący "dobrą aproksymacją" R.
- **Dodatkowe wymagania:** Algorytm powinien być efektywny (czasowo) używając do uczenia najmniejszej liczby przykładów.

Problem uczenia się prostokąta

Możemy definiować problem jak następująco:

- **Cel:** Znaleźć w $\Re^2$ prostokąt R o bokach równoległych do osi.
- **Wejście:** Zbiór zawierający przykłady w postaci punktów $((x, y), +/-)$. Te punkty zostały wygenerowane losowo.
- **Wyjście:** hipotetyczny prostokąt R' będący "dobrą aproksymacją" R.
- **Dodatkowe wymagania:** Algorytm powinien być efektywny (czasowo) używając do uczenia najmniejszej liczby przykładów.

Problem uczenia się prostokąta

Możemy definiować problem jak następująco:

- **Cel:** Znaleźć w $\Re^2$ prostokąt R o bokach równoległych do osi.
- **Wejście:** Zbiór zawierający przykłady w postaci punktów $((x, y), +/-)$. Te punkty zostały wygenerowane losowo.
- **Wyjście:** hipotetyczny prostokąt R' będący "dobrą aproksymacją" R.
- **Dodatkowe wymagania:** Algorytm powinien być efektywny (czasowo) używając do uczenia najmniejszej liczby przykładów.

Problem uczenia się prostokąta

Możemy definiować problem jak następująco:

- **Cel:** Znaleźć w $\Re^2$ prostokąt R o bokach równoległych do osi.
- **Wejście:** Zbiór zawierający przykłady w postaci punktów $((x, y), +/-)$. Te punkty zostały wygenerowane losowo.
- **Wyjście:** hipotetyczny prostokąt R' będący "dobrą aproksymacją" R.
- **Dodatkowe wymagania:** Algorytm powinien być efektywny (czasowo) używając do uczenia najmniejszej liczby przykładów.

Ogólny model uczenia się

- **Dane są**
  - zbiór wszystkich obiektów X (skończony lub nie);
  - pojęcie $c \in \mathbb{C}$ (funkcja celu);
  - skończona próbka D obiektów $x_1, ..., x_m \in X$ wraz z wartością funkcji c na tych obiektach;
  - przestrzeń hipotez $\mathbb{H}$;
- Szukane
  - hipoteza $h \in \mathbb{H}$ będąca dobrą aproksymacją pojęcia c.
- Wymagane
  - dobra jakość aproksymacji
  - szybki czas działania.

Ogólny model uczenia się

- **Dane są**
  - zbiór wszystkich obiektów X (skończony lub nie);
  - pojęcie $c \in \mathbb{C}$ (funkcja celu);
  - skończona próbka D obiektów $x_1, ..., x_m \in X$ wraz z wartością funkcji c na tych obiektach;
  - przestrzeń hipotez $\mathbb{H}$;
- **Szukane**
  - hipoteza $h \in \mathbb{H}$ będąca dobrą aproksymacją pojęcia c.
- **Wymagane**
  - dobra jakość aproksymacji
  - szybki czas działania.

Ogólny model uczenia się

- **Dane są**
  - zbiór wszystkich obiektów X (skończony lub nie);
  - pojęcie $c \in \mathbb{C}$ (funkcja celu);
  - skończona próbka D obiektów $x_1, ..., x_m \in X$ wraz z wartością funkcji c na tych obiektach;
  - przestrzeń hipotez $\mathbb{H}$;
- **Szukane**
  - hipoteza $h \in \mathbb{H}$ będąca dobrą aproksymacją pojęcia c.
- **Wymagane**
  - dobra jakość aproksymacji
  - szybki czas działania.

Inne przykłady

- **Uczenie półosi (lub dyskretyzacji)**:

$$X = \Re; \qquad \mathbb{C} = \mathbb{H} = \{[\lambda, \infty) : \alpha \in \Re\}$$

- **Uczenie hiperpłaszczyzny**:

$$X = \Re^n; \qquad \mathbb{H} = \{f_{w_0,w_1,\dots,w_n} : \Re^n \to \{0,1\}|\}$$

gdzie $f_{w_0,\dots,w_n}(x_1,\dots,x_n) = sgn(w_0 + w_1x_1 + \dots + w_nx_n)$.

- **Uczenie jednomianów Boolowskich**:

$$X = \{0,1\}^n; \qquad c : \{0,1\}^n \to \{0,1\};$$

$\mathbb{H} = M_n =$ zbiór jednomianów Boolowskich o n zmiennych.

Błąd hipotezy

Niech

- X – zbiór wszystkich obiektów.
- $\Omega = (X, \mu)$ – przestrzeń probabilistyczna określona na X.

Błąd hipotezy $h \in \mathbb{H}$ względem pojęcia c (funkcji docelowej):

$$er_{\Omega}(h,c) = er_{\Omega}^{c}(h) = \mu\{x \in X | h(x) \neq c(x)\}$$

Z prawdopodobieństwem $(1-\varepsilon)$ możemy oszacować er_{Ω}^{c}:

$$|er_{\Omega}^{c} - er_{D}^{c}| \leqslant s_{\frac{\varepsilon}{2}} \sqrt{\frac{er_{D}^{c}(1-er_{D}^{c})}{|D|}}$$

The No Free Lunch Theorem

- Algorytm $\mathfrak{L}$ dobrze się uczy pojęcia c jeśli er^c_Ω jest mały.
- Niech $\mathbb{P}(X) = \{c : X \to \{0,1\}\}$.
  Czy można stwierdzić wiedzieć, że $\mathfrak{L}_1$ uczy się wszystkich pojęć z $\mathbb{P}(X)$ lepiej od $\mathfrak{L}_2$?
- "No Free Lunch theorem" (Wolpert, Schaffer) w wersji problemów uczenia się głosi, że:
  - Żaden algorytm nie może być najlepszy w uczeniu wszystkich pojęć.
  - Każdy algorytm jest najlepszy dla takiej samej liczby pojęć
  - Ale interesuje nas tylko pewna klasa problemów czyli klasa pojęć $\mathbb{C} \subset \mathbb{P}(X)$
  - Wniosek: Należy znaleźć odp. algorytm do każdego problemu.

Spis treści

Model uczenia się PAC

- **Dane:** dziedzina X, klasa pojęć $\mathbb{C}$ i przestrzeń hipotez ucznia $\mathbb{H}$.
- Uczeń uczy się pojęcia $c \in \mathbb{C}$ na podstawie przykładów wygenerowanych przez zm.l. $EX(\Omega, c)$ (zwaną *wyrocznią*).

rodzina wszystkich zbiorów zawierających m przykładów

$$D = \{\langle x_1, c(x_1)\rangle, ..., \langle x_m, c(x_m)\rangle\} \in \mathcal{S}(m, c)$$

- Kolejne przykłady zbioru trenującego generuje zm. losowa $EX(c, \Omega)$ (*wyrocznia*) zwracająca przykład $\langle x, c(x)\rangle$, gdzie $x \in X$ jest wylosowane zgodnie z Ω.
- Celem ucznia jest znalezienie hipotezy minimalizującej błąd rzeczywisty względem c dla rozkładu Ω, czyli er_{Ω}^{c}.
- **Zasadnicza idea modelu PAC:** określenie warunków, pod jakimi uczeń (lub algorytm uczenia się) znajdzie "dobrą hipotezę" z "dużym prawdopodobieństwem".
  (o ograniczonym błędzie rzeczywistym powyżej określonego progu)

Model uczenia się PAC

- **Dane:** dziedzina X, klasa pojęć $\mathbb{C}$ i przestrzeń hipotez ucznia $\mathbb{H}$.
- Uczeń uczy się pojęcia $c \in \mathbb{C}$ na podstawie przykładów wygenerowanych przez zm.l. $EX(\Omega, c)$ (zwaną *wyrocznią*).

rodzina wszystkich zbiorów zawierających m przykładów

$$D = \{\langle x_1, c(x_1)\rangle, ..., \langle x_m, c(x_m)\rangle\} \in \mathcal{S}(m, c)$$

- Kolejne przykłady zbioru trenującego generuje zm. losowa $EX(c, \Omega)$ (*wyrocznia*) zwracająca przykład $\langle x, c(x)\rangle$, gdzie $x \in X$ jest wylosowane zgodnie z Ω.
- Celem ucznia jest znalezienie hipotezy minimalizującej błąd rzeczywisty względem c dla rozkładu Ω, czyli er^c_Ω.
- **Zasadnicza idea modelu PAC:** określenie warunków, pod jakimi uczeń (lub algorytm uczenia się) znajdzie "dobrą hipotezę" z "dużym prawdopodobieństwem".
(o ograniczonym błędzie rzeczywistym powyżej określonego progu)

Model uczenia się PAC

- **Dane:** dziedzina X, klasa pojęć $\mathbb{C}$ i przestrzeń hipotez ucznia $\mathbb{H}$.
- Uczeń uczy się pojęcia $c \in \mathbb{C}$ na podstawie przykładów wygenerowanych przez zm.l. $EX(\Omega, c)$ (zwaną *wyrocznią*).

rodzina wszystkich zbiorów zawierających m przykładów

$$D = \{\langle x_1, c(x_1)\rangle, ..., \langle x_m, c(x_m)\rangle\} \in \mathcal{S}(m, c)$$

- Kolejne przykłady zbioru trenującego generuje zm. losowa $EX(c, \Omega)$ (*wyrocznia*) zwracająca przykład $\langle x, c(x)\rangle$, gdzie $x \in X$ jest wylosowane zgodnie z Ω.
- Celem ucznia jest znalezienie hipotezy minimalizującej błąd rzeczywisty względem c dla rozkładu Ω, czyli er^c_Ω.
- **Zasadnicza idea modelu PAC:** określenie warunków, pod jakimi uczeń (lub algorytm uczenia się) znajdzie "dobrą hipotezę" z "dużym prawdopodobieństwem".
  (o ograniczonym błędzie rzeczywistym powyżej określonego progu)

Model uczenia się PAC

- **Dane:** dziedzina X, klasa pojęć $\mathbb{C}$ i przestrzeń hipotez ucznia $\mathbb{H}$.
- Uczeń uczy się pojęcia $c \in \mathbb{C}$ na podstawie przykładów wygenerowanych przez zm.l. $EX(\Omega, c)$ (zwaną *wyrocznią*).

rodzina wszystkich zbiorów zawierających m przykładów

$$D = \{\langle x_1, c(x_1)\rangle, ..., \langle x_m, c(x_m)\rangle\} \in \mathcal{S}(m, c)$$

- Kolejne przykłady zbioru trenującego generuje zm. losowa $EX(c, \Omega)$ (*wyrocznia*) zwracająca przykład $\langle x, c(x)\rangle$, gdzie $x \in X$ jest wylosowane zgodnie z Ω.
- Celem ucznia jest znalezienie hipotezy minimalizującej błąd rzeczywisty względem c dla rozkładu Ω, czyli er^c_Ω.
- **Zasadnicza idea modelu PAC:** określenie warunków, pod jakimi uczeń (lub algorytm uczenia się) znajdzie "dobrą hipotezę" z "dużym prawdopodobieństwem".
(o ograniczonym błędzie rzeczywistym powyżej określonego progu)

Model uczenia się PAC

- **Dane:** dziedzina X, klasa pojęć $\mathbb{C}$ i przestrzeń hipotez ucznia $\mathbb{H}$.
- Uczeń uczy się pojęcia $c \in \mathbb{C}$ na podstawie przykładów wygenerowanych przez zm.l. $EX(\Omega, c)$ (zwaną *wyrocznią*).

rodzina wszystkich zbiorów zawierających m przykładów

$$D = \{\langle x_1, c(x_1)\rangle, ..., \langle x_m, c(x_m)\rangle\} \in \mathcal{S}(m, c)$$

- Kolejne przykłady zbioru trenującego generuje zm. losowa $EX(c, \Omega)$ (*wyrocznia*) zwracająca przykład $\langle x, c(x)\rangle$, gdzie $x \in X$ jest wylosowane zgodnie z Ω.
- Celem ucznia jest znalezienie hipotezy minimalizującej błąd rzeczywisty względem c dla rozkładu Ω, czyli er_Ω^c.
- **Zasadnicza idea modelu PAC:** określenie warunków, pod jakimi uczeń (lub algorytm uczenia się) znajdzie "dobrą hipotezę" z "dużym prawdopodobieństwem".
  (o ograniczonym błędzie rzeczywistym powyżej określonego progu)

PAC-owy uczeń

Definition (PAC)

Algorytm $\mathfrak{L}$ nazywamy *"prawdopodobnie aproksymacyjnie poprawnym"* wtedy i tylko wtedy, gdy

dla każdych $0 < \varepsilon, \delta < 1$, istnieje liczba $m_0 = m_0(\varepsilon, \delta)$ taka, że dla dowolnego pojęcia $c \in \mathbb{C}$ i dla dowolnego rozkładu Ω na X mamy

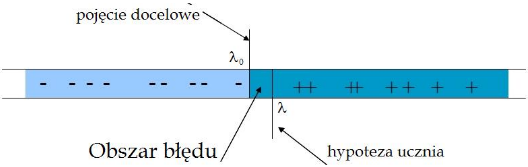
$$\mu^m\{D \in \mathcal{S}(m, c) | er_\Omega(\mathfrak{L}(D)) < \varepsilon\} > 1 - \delta$$

o ile $m > m_0$.

Wówczas mówimy w skrócie, że $\mathfrak{L}$ *jest* **PAC** (*Probably Approximately Correct*).

- ε – dopuszczalny poziom błędu
- $(1 - \delta)$ – poziom zaufania

Przykład problemu dyskretyzacji



- $X = \Re$; $\mathbb{H} = \mathbb{C} = \{f_\lambda : \Re \to \{0,1\} | f_\lambda(x) = 1 \Leftrightarrow x \geqslant \lambda\}$
- $c = f_{\lambda_0}$
- znaleźć λ_0 na podstawie losowo wygenerowanych przykładów $D = \{\langle x_1, f_{\lambda_0}(x_1)\rangle, ..., \langle x_m, f_{\lambda_0}(x_m)\rangle\}$

Algorytm:

1. Set $\lambda^* := \min_{i \in \{1,...,m\}} \{x_i : f_{\lambda_0}(x_i) = 1\}$;
2. $L(D) := f_{\lambda^*}$;

Przykład (c.d.)

Twierdzenie:

Powyższy algorytm jest PAC

Dowód

- $er^c_\Omega = \mu([\lambda_0, \lambda^*))$
- Niech $\beta_0 = \sup\{\beta | \mu([\lambda_0, \beta)) < \varepsilon\}$. Wówczas $er^c_\Omega(f_{\lambda^*}) \leqslant \varepsilon$ $\Leftrightarrow \lambda^* \leqslant \beta_0 \Leftrightarrow$ jeden z przykładów x_i znajduje się w przedziale $[\lambda_0, \beta_0]$;
- Prawdopodobieństwo tego, że żaden pośród m przykładów nie należy do $[\lambda_0, \beta_0]$ jest $\leqslant (1-\varepsilon)^m$. Stąd

$$\mu^m\{D \in \mathcal{S}(m, f_{\lambda_0}) | er_\Omega(L(D)) \leqslant \varepsilon\} \geqslant 1 - (1-\varepsilon)^m$$

- Aby to prawdopodobieństwo było $> 1 - \delta$, wystarczy wybrać

$$m \geqslant m_0 = \left\lceil \frac{1}{\varepsilon} \ln \frac{1}{\delta} \right\rceil$$

Dokładne uczenie się

- Niech Ω będzie rozkładem dyskretnym zdefiniowanym przez $\mu_1 = \mu(x_1)$, ..., $\mu_n = \mu(x_n)$ – dla pewnych $x_1, \ldots, x_n \in X$ – takich, że $\mu_1 + \ldots + \mu_n = 1$. Niech $\varepsilon_{\min} = \min_i \mu_i$.
- Jeśli $\mathfrak{L}$ jest PAC, i jeśli $\varepsilon \leqslant \varepsilon_{\min}$ to warunek $er^c_\Omega(L(D)) < \varepsilon$ jest równoważny z $er^c_\Omega(L(D)) = 0$. Stąd dla każdego δ, istnieje $m_0 = m_0(\varepsilon_{\min}, \delta)$ taka, że dla dowolnego $c \in \mathbb{C}$ i Ω

$$m > m_0 \Rightarrow \mu^m\{D \in \mathcal{S}(m,t) | er_\Omega(L(D)) = 0\} > 1 - \delta$$

- Wówczas mówimy, że prawdopodobnie $\mathfrak{L}$ jest dokładnym algorytmem (*jest PEC* – *probably exactly correct*)

Spis treści

Potencjalna wyuczalność

- Algorytm $\mathfrak{L}$ nazywamy niesprzecznym jeśli dla każdego pojęcia c i każdego zbioru D mamy $er^c_D(L(D)) = 0$
(tzn. $\mathfrak{L}(D)(x_i) = c(x_i)$ dla dowolnego przykładu $(x_i, c(x_i)) \in D$).
- $\mathbb{H}^c(D) = \{h \in \mathbb{H} | h(x_i) = c(x_i)(i = 1, ..., m)\}$.
$\mathfrak{L}$ jest niesprzeczny jeśli $\mathfrak{L}(D) \in \mathbb{H}^c(D)$ dla każdego D.
- $\mathbb{B}^c_\varepsilon = \{h \in \mathbb{H} | er_\Omega(h) \geqslant \varepsilon\}$ – zbiór słabych hipotez

Definition

Mówimy, że $\mathbb{C}$ jest potencjalnie wyuczalne za pomocą $\mathbb{H}$, jeśli dla każdego rozkładu Ω na X i dowolnego pojęcia $c \in \mathbb{C}$ oraz dla dowolnych $0 < \varepsilon, \delta < 1$ istnieje $m_0 = m_0(\varepsilon, \delta)$ takie, że

$$m \geqslant m_0 \Rightarrow \mu^m\{D \in \mathcal{S}(m, c) | \mathbb{H}^c(D) \cap \mathbb{B}^c_\varepsilon = \emptyset\} > 1 - \delta$$

Potencjalna wyuczalność

- Algorytm $\mathfrak{L}$ nazywamy niesprzecznym jeśli dla każdego pojęcia c i każdego zbioru D mamy $er_D^c(L(D)) = 0$
(tzn. $\mathfrak{L}(D)(x_i) = c(x_i)$ dla dowolnego przykładu $(x_i, c(x_i)) \in D$).
- $\mathbb{H}^c(D) = \{h \in \mathbb{H} | h(x_i) = c(x_i)(i = 1, ..., m)\}$.
$\mathfrak{L}$ jest niesprzeczny jeśli $\mathfrak{L}(D) \in \mathbb{H}^c(D)$ dla każdego D.
- $\mathbb{B}_\varepsilon^c = \{h \in \mathbb{H} | er_\Omega(h) \geqslant \varepsilon\}$ – zbiór słabych hipotez

Definition

Mówimy, że $\mathbb{C}$ jest potencjalnie wyuczalne za pomocą $\mathbb{H}$, jeśli dla każdego rozkładu Ω na X i dowolnego pojęcia $c \in \mathbb{C}$ oraz dla dowolnych $0 < \varepsilon, \delta < 1$ istnieje $m_0 = m_0(\varepsilon, \delta)$ takie, że

$$m \geqslant m_0 \Rightarrow \mu^m\{D \in \mathcal{S}(m, c) | \mathbb{H}^c(D) \cap \mathbb{B}_\varepsilon^c = \emptyset\} > 1 - \delta$$

Potencjalna wyuczalność

- Algorytm $\mathfrak{L}$ nazywamy niesprzecznym jeśli dla każdego pojęcia c i każdego zbioru D mamy $er_D^c(L(D)) = 0$ (tzn. $\mathfrak{L}(D)(x_i) = c(x_i)$ dla dowolnego przykładu $(x_i, c(x_i)) \in D$).
- $\mathbb{H}^c(D) = \{h \in \mathbb{H} | h(x_i) = c(x_i)(i = 1, ..., m)\}$.
  $\mathfrak{L}$ jest niesprzeczny jeśli $\mathfrak{L}(D) \in \mathbb{H}^c(D)$ dla każdego D.
- $\mathbb{B}_\varepsilon^c = \{h \in \mathbb{H} | er_\Omega(h) \geqslant \varepsilon\}$ – zbiór słabych hipotez

Definition

Mówimy, że $\mathbb{C}$ jest potencjalnie wyuczalne za pomocą $\mathbb{H}$, jeśli dla każdego rozkładu Ω na X i dowolnego pojęcia $c \in \mathbb{C}$ oraz dla dowolnych $0 < \varepsilon, \delta < 1$ istnieje $m_0 = m_0(\varepsilon, \delta)$ takie, że

$$m \geqslant m_0 \Rightarrow \mu^m\{D \in \mathcal{S}(m, c) | \mathbb{H}^c(D) \cap \mathbb{B}_\varepsilon^c = \emptyset\} > 1 - \delta$$

Potencjalna wyuczalność

- Algorytm $\mathfrak{L}$ nazywamy niesprzecznym jeśli dla każdego pojęcia c i każdego zbioru D mamy $er_D^c(L(D)) = 0$
(tzn. $\mathfrak{L}(D)(x_i) = c(x_i)$ dla dowolnego przykładu $(x_i, c(x_i)) \in D$).
- $\mathbb{H}^c(D) = \{h \in \mathbb{H} | h(x_i) = c(x_i)(i = 1, ..., m)\}$.
$\mathfrak{L}$ jest niesprzeczny jeśli $\mathfrak{L}(D) \in \mathbb{H}^c(D)$ dla każdego D.
- $\mathbb{B}_\varepsilon^c = \{h \in \mathbb{H} | er_\Omega(h) \geqslant \varepsilon\}$ – zbiór słabych hipotez

Definition

Mówimy, że $\mathbb{C}$ jest potencjalnie wyuczalne za pomocą $\mathbb{H}$, jeśli dla każdego rozkładu Ω na X i dowolnego pojęcia $c \in \mathbb{C}$ oraz dla dowolnych $0 < \varepsilon, \delta < 1$ istnieje $m_0 = m_0(\varepsilon, \delta)$ takie, że

$$m \geqslant m_0 \Rightarrow \mu^m\{D \in \mathcal{S}(m, c) | \mathbb{H}^c(D) \cap \mathbb{B}_\varepsilon^c = \emptyset\} > 1 - \delta$$

Potencjalna wyuczalność

Theorem

Jeśli

1. *$\mathbb{C}$ jest potencjalnie wyuczalne za pomocą $\mathbb{H}$*
2. *$\mathfrak{L}$ jest algorytmem niesprzecznym dla $\mathbb{C}$*

Wówczas $\mathfrak{L}$ jest PAC

Theorem (Haussler, 1988)

Jeśli $\mathbb{C} = \mathbb{H}$ i $|\mathbb{C}| < \infty$, to $\mathbb{C}$ jest potencjalnie wyuczalne.

Dowód: Niech $h \in \mathbb{B}_\varepsilon$ (tzn. $er_\Omega(h) \geqslant \varepsilon$). Wówczas

$$\mu^m\{D \in \mathcal{S}(m,c) | er_D(h) = 0\} \leqslant (1-\varepsilon)^m$$

$$\Rightarrow \mu^m\{D : \mathbb{H}[D] \cap \mathbb{B}_\varepsilon \neq \emptyset\} \leqslant |\mathbb{B}_\varepsilon|(1-\varepsilon)^m \leqslant |\mathbb{H}|(1-\varepsilon)^m$$

Aby $|\mathbb{H}|(1-\varepsilon)^m < \delta$ wystarczy wybrać $m \geqslant m_0 = \left\lceil \frac{1}{\varepsilon} \ln \frac{|\mathbb{H}|}{\delta} \right\rceil$

Spis treści

Wymiar Vapnika-Chervonenkisa

- Niech $\mathbf{x} = \{x_1, x_2, ...x_m\} \subset X$. Oznaczmy przez $\Pi_{\mathbb{H}}(\mathbf{x})$ liczbę podziałów zbioru $\mathbf{x}$ dokonanych przez $\mathbb{H}$, t.j. liczbę różnych wektorów postaci
$$(h(x_1), ..., h(x_m)) \in \{0, 1\}^m$$
po wszystkich $h \in H$.
- $\Pi_{\mathbb{H}}(\mathbf{x}) \leqslant 2^m$. Jeśli zachodzi równość, mówimy, że $\mathbb{H}$ rozbija $\mathbf{x}$.
- Niech $\Pi_{\mathbb{H}}(m) = \max_{\mathbf{x} \in X^m} \Pi_{\mathbb{H}}(\mathbf{x})$
- **Na przykład:** W przypadku przestrzeni półosi postaci $[\alpha, \infty)$ mamy $\Pi_{\mathbb{H}}(m) = m + 1$.
- Na ogół trudno znaleźć wzór na $\Pi_{\mathbb{H}}(m)$!!!

Wymiar Vapnika-Chervonenkisa

- Niech $\mathbf{x} = \{x_1, x_2, ...x_m\} \subset X$. Oznaczmy przez $\Pi_{\mathbb{H}}(\mathbf{x})$ liczbę podziałów zbioru $\mathbf{x}$ dokonanych przez $\mathbb{H}$, t.j. liczbę różnych wektorów postaci
$$(h(x_1), ..., h(x_m)) \in \{0,1\}^m$$
po wszystkich $h \in H$.
- $\Pi_{\mathbb{H}}(\mathbf{x}) \leqslant 2^m$. Jeśli zachodzi równość, mówimy, że $\mathbb{H}$ rozbija $\mathbf{x}$.
- Niech $\Pi_{\mathbb{H}}(m) = \max_{\mathbf{x} \in X^m} \Pi_{\mathbb{H}}(\mathbf{x})$
- **Na przykład:** W przypadku przestrzeni półosi postaci $[\alpha, \infty)$ mamy $\Pi_{\mathbb{H}}(m) = m + 1$.
- Na ogół trudno znaleźć wzór na $\Pi_{\mathbb{H}}(m)$!!!

Wymiar Vapnika-Chervonenkisa

- Niech $\mathbf{x} = \{x_1, x_2, ...x_m\} \subset X$. Oznaczmy przez $\Pi_{\mathbb{H}}(\mathbf{x})$ liczbę podziałów zbioru $\mathbf{x}$ dokonanych przez $\mathbb{H}$, t.j. liczbę różnych wektorów postaci
$$(h(x_1), ..., h(x_m)) \in \{0, 1\}^m$$
po wszystkich $h \in H$.
- $\Pi_{\mathbb{H}}(\mathbf{x}) \leqslant 2^m$. Jeśli zachodzi równość, mówimy, że $\mathbb{H}$ rozbija $\mathbf{x}$.
- Niech $\Pi_{\mathbb{H}}(m) = \max_{\mathbf{x} \in X^m} \Pi_{\mathbb{H}}(\mathbf{x})$
- **Na przykład:** W przypadku przestrzeni półosi postaci $[\alpha, \infty)$ mamy $\Pi_{\mathbb{H}}(m) = m + 1$.
- Na ogół trudno znaleźć wzór na $\Pi_{\mathbb{H}}(m)$!!!

Wymiar Vapnika-Chervonenkisa

- Niech $\mathbf{x} = \{x_1, x_2, ...x_m\} \subset X$. Oznaczmy przez $\Pi_{\mathbb{H}}(\mathbf{x})$ liczbę podziałów zbioru $\mathbf{x}$ dokonanych przez $\mathbb{H}$, t.j. liczbę różnych wektorów postaci
$$(h(x_1), ..., h(x_m)) \in \{0,1\}^m$$
po wszystkich $h \in H$.
- $\Pi_{\mathbb{H}}(\mathbf{x}) \leqslant 2^m$. Jeśli zachodzi równość, mówimy, że $\mathbb{H}$ rozbija $\mathbf{x}$.
- Niech $\Pi_{\mathbb{H}}(m) = \max_{\mathbf{x} \in X^m} \Pi_{\mathbb{H}}(\mathbf{x})$
- **Na przykład:** W przypadku przestrzeni półosi postaci $[\alpha, \infty)$ mamy $\Pi_{\mathbb{H}}(m) = m + 1$.
- Na ogół trudno znaleźć wzór na $\Pi_{\mathbb{H}}(m)$!!!

Wymiar Vapnika-Chervonenkisa

- Niech $\mathbf{x} = \{x_1, x_2, ...x_m\} \subset X$. Oznaczmy przez $\Pi_{\mathbb{H}}(\mathbf{x})$ liczbę podziałów zbioru $\mathbf{x}$ dokonanych przez $\mathbb{H}$, t.j. liczbę różnych wektorów postaci
$$(h(x_1), ..., h(x_m)) \in \{0,1\}^m$$
po wszystkich $h \in H$.
- $\Pi_{\mathbb{H}}(\mathbf{x}) \leqslant 2^m$. Jeśli zachodzi równość, mówimy, że $\mathbb{H}$ rozbija $\mathbf{x}$.
- Niech $\Pi_{\mathbb{H}}(m) = \max_{\mathbf{x} \in X^m} \Pi_{\mathbb{H}}(\mathbf{x})$
- **Na przykład:** W przypadku przestrzeni półosi postaci $[\alpha, \infty)$ mamy $\Pi_{\mathbb{H}}(m) = m + 1$.
- Na ogół trudno znaleźć wzór na $\Pi_{\mathbb{H}}(m)$!!!

Wymiar Vapnika-Chervonenkisa (c.d.)

Uwagi:

- Jeśli $\Pi_{\mathbb{H}}(m) = 2^m$, to $\mathbb{H}$ rozbija prawie każdy zbiór o mocy m (prawie zawsze można znaleźć niesprzeczną hipotezę).
- Maksymalna wartość m, dla której $\Pi_{\mathbb{H}}(m) = 2^m$ można uważać za moc wyrażania przestrzeni $\mathbb{H}$

Definition (VC dimension)

Wymiarem Vapnika-Chervonenkisa przestrzeni hipotez $\mathbb{H}$ nazywamy liczbę

$$VCdim(\mathbb{H}) = \max\{m : \Pi_{\mathbb{H}}(m) = 2^m\}$$

gdzie maksimum wynosi ∞ jeśli ten zbiór jest nieograniczony.

Wymiar Vapnika-Chervonenkisa (c.d.)

Uwagi:

- Jeśli $\Pi_{\mathbb{H}}(m) = 2^m$, to $\mathbb{H}$ rozbija prawie każdy zbiór o mocy m (prawie zawsze można znaleźć niesprzeczną hipotezę).
- Maksymalna wartość m, dla której $\Pi_{\mathbb{H}}(m) = 2^m$ można uważać za moc wyrażania przestrzeni $\mathbb{H}$

Definition (VC dimension)

Wymiarem Vapnika-Chervonenkisa przestrzeni hipotez $\mathbb{H}$ nazywamy liczbę

$$VCdim(\mathbb{H}) = \max\{m : \Pi_{\mathbb{H}}(m) = 2^m\}$$

gdzie maksimum wynosi ∞ jeśli ten zbiór jest nieograniczony.

Wymiar Vapnika-Chervonenkisa (c.d.)

Uwagi:

- Jeśli $\Pi_{\mathbb{H}}(m) = 2^m$, to $\mathbb{H}$ rozbija prawie każdy zbiór o mocy m (prawie zawsze można znaleźć niesprzeczną hipotezę).
- Maksymalna wartość m, dla której $\Pi_{\mathbb{H}}(m) = 2^m$ można uważać za moc wyrażania przestrzeni $\mathbb{H}$

Definition (VC dimension)

Wymiarem Vapnika-Chervonenkisa przestrzeni hipotez $\mathbb{H}$ nazywamy liczbę

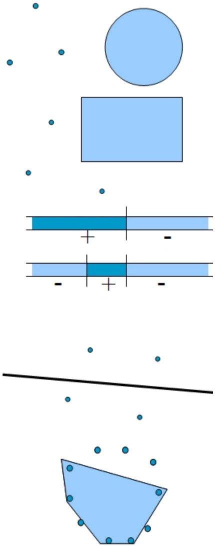
$$VCdim(\mathbb{H}) = \max\{m : \Pi_{\mathbb{H}}(m) = 2^m\}$$

gdzie maksimum wynosi ∞ jeśli ten zbiór jest nieograniczony.

Przykłady wymiaru VC-Dim

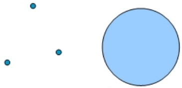
- $H = \{\text{okręgi} \ldots \} \implies VC(H) = 3$
- $H = \{\text{prostokąty} \ldots \} \implies VC(H) = 4$
- $H = \{\text{funkcje progowe} \ldots \} \implies$
  $VC(H) = 1$ jeśli "+" są zawsze po prawej stronie;
  $VC(H) = 2$ jeśli "+" mogą być po obu stronach
- $H = \{\text{przedziały} \ldots \} \implies$
  VC(H) = 2 jeśli "+" są zawsze w środku
  VC(H) = 3 jeśli w środku mogą być zarówno "+" i "-"
- $H = \{ \text{półpłaszczyzny w } \Re^2 \ldots \} \implies VC(H) = 3$
- czy istnieje H dla której $VC(H) = \infty$?
- **Tw.:** Jeśli $|\mathbb{H}| < \infty$ to $VCdim(\mathbb{H}) \leq \log |\mathbb{H}|$
- Niech M_n = zbiór jednomianów boolowskich na n zmiennych. Ponieważ, $|M_n| = 3^n$ mamy

$$VCdim(M_n) \leq n \log 3$$

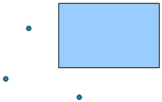


Przykłady wymiaru VC-Dim

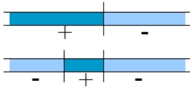
- $H = \{$okręgi ... $\} \implies VC(H) = 3$



- $H = \{$prostokąty ... $\} \implies VC(H) = 4$

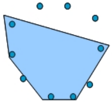


- $H = \{$funkcje progowe ... $\} \implies$
  $VC(H) = 1$ jeśli "+" są zawsze po prawej stronie;
  $VC(H) = 2$ jeśli "+" mogą być po obu stronach
- $H = \{$przedziały ... $\} \implies$
  VC(H) = 2 jeśli "+" są zawsze w środku
  VC(H) = 3 jeśli w środku mogą być zarówno "+" i "-"



- $H = \{$ półpłaszczyzny w $\Re^2$ ... $\} \implies VC(H) = 3$

- czy istnieje H dla której $VC(H) = \infty$?



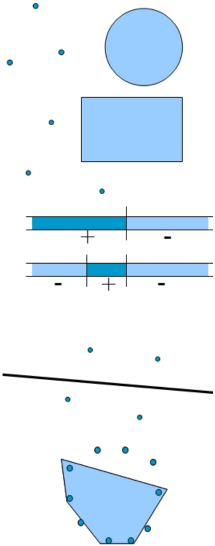
- **Tw.:** Jeśli $|\mathbb{H}| < \infty$ to $VCdim(\mathbb{H}) \leq \log |\mathbb{H}|$
- Niech M_n = zbiór jednomianów boolowskich na n zmiennych. Ponieważ, $|M_n| = 3^n$ mamy

$$VCdim(M_n) \leq n \log 3$$

Przykłady wymiaru VC-Dim

- $H = \{$okręgi ... $\} \implies VC(H) = 3$
- $H = \{$prostokąty ... $\} \implies VC(H) = 4$
- $H = \{$funkcje progowe ... $\} \implies$
  $VC(H) = 1$ jeśli "+" są zawsze po prawej stronie;
  $VC(H) = 2$ jeśli "+" mogą być po obu stronach
- $H = \{$przedziały ... $\} \implies$
  VC(H) = 2 jeśli "+" są zawsze w środku
  VC(H) = 3 jeśli w środku mogą być zarówno "+" i "-"
- $H = \{$ półpłaszczyzny w $\Re^2$ ... $\} \implies VC(H) = 3$
- czy istnieje H dla której $VC(H) = \infty$?
- **Tw.:** Jeśli $|\mathbb{H}| < \infty$ to $VCdim(\mathbb{H}) \leq \log |\mathbb{H}|$
- Niech M_n = zbiór jednomianów boolowskich na n zmiennych. Ponieważ $|M_n| = 3^n$ mamy

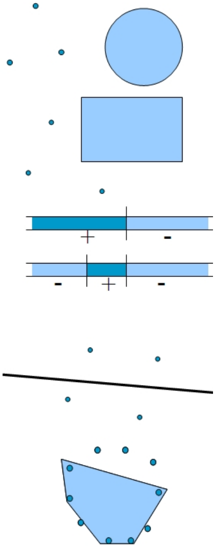
$$VCdim(M_n) \leq n \log 3$$



Przykłady wymiaru VC-Dim

- $H = \{$okręgi … $\} \implies VC(H) = 3$
- $H = \{$prostokąty … $\} \implies VC(H) = 4$
- $H = \{$funkcje progowe … $\} \implies$
  $VC(H) = 1$ jeśli "+" są zawsze po prawej stronie;
  $VC(H) = 2$ jeśli "+" mogą być po obu stronach
- $H = \{$przedziały … $\} \implies$
  VC(H) = 2 jeśli "+" są zawsze w środku
  VC(H) = 3 jeśli w środku mogą być zarówno "+" i "-"
- $H = \{$ półpłaszczyzny w $\Re^2$ … $\} \implies VC(H) = 3$
- czy istnieje H dla której $VC(H) = \infty$?
- **Tw.:** Jeśli $|\mathbb{H}| < \infty$ to $VCdim(\mathbb{H}) \leq \log |\mathbb{H}|$
- Niech M_n = zbiór jednomianów boolowskich na n zmiennych. Ponieważ $|M_n| = 3^n$ mamy

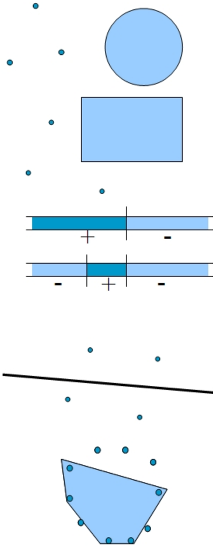
$$VCdim(M_n) \leq n \log 3$$



Przykłady wymiaru VC-Dim

- $H = \{\text{okręgi} \ldots\} \implies VC(H) = 3$
- $H = \{\text{prostokąty} \ldots\} \implies VC(H) = 4$
- $H = \{\text{funkcje progowe} \ldots\} \implies$
  $VC(H) = 1$ jeśli "+" są zawsze po prawej stronie;
  $VC(H) = 2$ jeśli "+" mogą być po obu stronach
- $H = \{\text{przedziały} \ldots\} \implies$
  VC(H) = 2 jeśli "+" są zawsze w środku
  VC(H) = 3 jeśli w środku mogą być zarówno "+" i "-"
- $H = \{\text{ półpłaszczyzny w } \Re^2 \ldots\} \implies VC(H) = 3$
- czy istnieje H dla której $VC(H) = \infty$?
- **Tw.:** Jeśli $|\mathbb{H}| < \infty$ to $VCdim(\mathbb{H}) \leqslant \log |\mathbb{H}|$
- Niech M_n = zbiór jednomianów boolowskich na n zmiennych. Ponieważ, $|M_n| = 3^n$ mamy

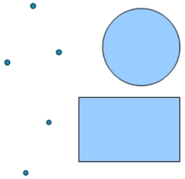
$$VCdim(M_n) \leqslant n \log 3$$

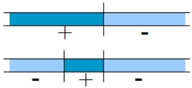


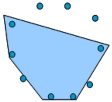
Przykłady wymiaru VC-Dim

- $H = \{$okręgi ... $\} \implies VC(H) = 3$
- $H = \{$prostokąty ... $\} \implies VC(H) = 4$
- $H = \{$funkcje progowe ... $\} \implies$
  $VC(H) = 1$ jeśli "+" są zawsze po prawej stronie;
  $VC(H) = 2$ jeśli "+" mogą być po obu stronach
- $H = \{$przedziały ... $\} \implies$
  VC(H) = 2 jeśli "+" są zawsze w środku
  VC(H) = 3 jeśli w środku mogą być zarówno "+" i "-"
- $H = \{$ półpłaszczyzny w $\Re^2$ ... $\} \implies VC(H) = 3$
- czy istnieje H dla której $VC(H) = \infty$?
- **Tw.:** Jeśli $|\mathbb{H}| < \infty$ to $VCdim(\mathbb{H}) \leqslant \log |\mathbb{H}|$
- Niech M_n = zbiór jednomianów boolowskich na n zmiennych. Ponieważ, $|M_n| = 3^n$ mamy

$$VCdim(M_n) \leqslant n \log 3$$



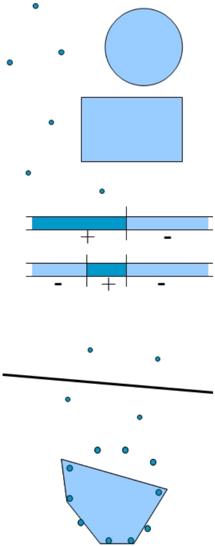




Przykłady wymiaru VC-Dim

- $H = \{\text{okręgi} \ldots\} \implies VC(H) = 3$
- $H = \{\text{prostokąty} \ldots\} \implies VC(H) = 4$
- $H = \{\text{funkcje progowe} \ldots\} \implies$
  $VC(H) = 1$ jeśli "+" są zawsze po prawej stronie;
  $VC(H) = 2$ jeśli "+" mogą być po obu stronach
- $H = \{\text{przedziały} \ldots\} \implies$
  VC(H) = 2 jeśli "+" są zawsze w środku
  VC(H) = 3 jeśli w środku mogą być zarówno "+" i "-"
- $H = \{\text{ półpłaszczyzny w } \Re^2 \ldots\} \implies VC(H) = 3$
- czy istnieje H dla której $VC(H) = \infty$?
- **Tw.:** Jeśli $|\mathbb{H}| < \infty$ to $VCdim(\mathbb{H}) \leqslant \log |\mathbb{H}|$
- Niech M_n = zbiór jednomianów boolowskich na n zmiennych. Ponieważ, $|M_n| = 3^n$ mamy

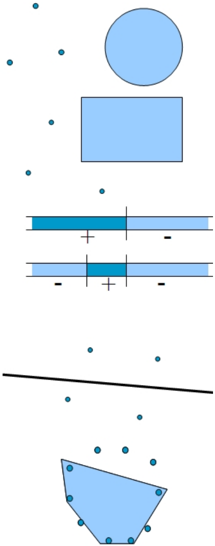
$$VCdim(M_n) \leqslant n \log 3$$



Przykłady wymiaru VC-Dim

- $H = \{$okręgi ... $\} \implies VC(H) = 3$
- $H = \{$prostokąty ... $\} \implies VC(H) = 4$
- $H = \{$funkcje progowe ... $\} \implies$
  $VC(H) = 1$ jeśli "+" są zawsze po prawej stronie;
  $VC(H) = 2$ jeśli "+" mogą być po obu stronach
- $H = \{$przedziały ... $\} \implies$
  VC(H) = 2 jeśli "+" są zawsze w środku
  VC(H) = 3 jeśli w środku mogą być zarówno "+" i "-"
- $H = \{$ półpłaszczyzny w $\Re^2$ ... $\} \implies VC(H) = 3$
- czy istnieje H dla której $VC(H) = \infty$?
- **Tw.:** Jeśli $|\mathbb{H}| < \infty$ to $VCdim(\mathbb{H}) \leqslant \log |\mathbb{H}|$
- Niech M_n = zbiór jednomianów boolowskich na n zmiennych. Ponieważ, $|M_n| = 3^n$ mamy

$$VCdim(M_n) \leqslant n \log 3$$



Wymiar VC dla perceptronu

Twierdzenie

Dla każdej liczby naturalnej n, niech P_n będzie perceptronem o n wejściach rzeczywistych. Wówczas

$$VCdim(P_n) = n + 1$$

Dowód:

- $\boxed{VCdim(P_n) \leq n + 1:}$

  Wynika z Twierdzenia Radona: Dla dowolnego zbioru E zawierającego $n+2$ punktów w przestrzeni $\mathbb{R}^n$ istnieje niepusty podzbiór $S \subset E$ taki, że

  $$conv(S) \cap conv(E \setminus S) \neq \emptyset$$

- $\boxed{VCdim(P_n) \geq n + 1 :}$ Wystarczy wybrać $\mathbf{x} = \{0, e_1, ..., e_n\}$ i pokazać, że każdy jego podzbiór jest definiowany przez jakiś perceptron.

Lemat Sauer'a

Theorem (Lemat Sauer'a)

Jeśli $VCdim(\mathbb{H}) = d \geq 0$ *i* $m \geq 1$, *to*

$$\Pi_{\mathbb{H}}(m) \leqslant \underbrace{1 + \binom{m}{1} + \binom{m}{2} + \ldots + \binom{m}{d}}_{\Phi(d,m)}$$

Wnioski

$$\Phi(d, m) \leqslant \left(\frac{em}{d}\right)^d \Rightarrow \Pi_{\mathbb{H}}(m) \leqslant \left(\frac{em}{d}\right)^d$$

$$VCdim(\mathbb{H}) > \frac{ln|\mathbb{H}|}{1 + ln|X|}$$

Spis treści

Fundamentalne twierdzenia

Theorem (Warunek konieczny)

Każda przestrzeń hipotez o nieskończonym *wymiarze VC* nie jest potencjalnie wyuczalna.

Theorem (Fundamentalne twierdzenie)

Jeśli przestrzeń hipotez ma skończony wymiar VC, to jest ona potencjalnie wyuczalna.

Fundamentalne twierdzenia

Theorem (Warunek konieczny)

Każda przestrzeń hipotez o nieskończonym *wymiarze VC* nie jest potencjalnie wyuczalna.

Theorem (Fundamentalne twierdzenie)

Jeśli przestrzeń hipotez ma skończony wymiar VC, to jest ona potencjalnie wyuczalna.

Dowód

- Definiujemy
$$Q_m^\varepsilon = \{\mathbf{x} \in X^m | H[\mathbf{x}, t] \cap B_\varepsilon \neq \emptyset\}$$
- Szukamy górnego ograniczenia $f(m, \varepsilon)$ dla $\mu^m(Q_m^\varepsilon)$, które powinno
  - być niezależne od t (target concept) i μ (rozkład prawdopodobieństwa).
  - dążyć do 0 przy $m \to \infty$
- **Twierdzenie** Niech $\mathbb{H}$ będzie przestrzenią hipotez określonych na X. Dla dowolnych c, μ, ε (ale ustalonych) mamy
$$\mu^m(Q_m^\varepsilon) < 2\Pi_{\mathbb{H}}(2m)2^{-\varepsilon m/2}$$
o ile $m \geqslant 8/\varepsilon$.
- Możemy użyć lematu Sauer'a, aby pokazać, że $\mu^m(Q_m^\varepsilon) < \delta$ dla dostatecznie dużych m.

Dowód

- Definiujemy
$$Q_m^\varepsilon = \{\mathbf{x} \in X^m | H[\mathbf{x}, t] \cap B_\varepsilon \neq \emptyset\}$$
- Szukamy górnego ograniczenia $f(m, \varepsilon)$ dla $\mu^m(Q_m^\varepsilon)$, które powinno
  - być niezależne od t (target concept) i μ (rozkład prawdopodobieństwa).
  - dążyć do 0 przy $m \to \infty$
- **Twierdzenie** Niech $\mathbb{H}$ będzie przestrzenią hipotez określonych na X. Dla dowolnych c, μ, ε (ale ustalonych) mamy
$$\mu^m(Q_m^\varepsilon) < 2\Pi_{\mathbb{H}}(2m)2^{-\varepsilon m/2}$$
o ile $m \geqslant 8/\varepsilon$.
- Możemy użyć lematu Sauer'a, aby pokazać, że $\mu^m(Q_m^\varepsilon) < \delta$ dla dostatecznie dużych m.

Dowód

- Definiujemy
$$Q_m^\varepsilon = \{\mathbf{x} \in X^m | H[\mathbf{x}, t] \cap B_\varepsilon \neq \emptyset\}$$
- Szukamy górnego ograniczenia $f(m, \varepsilon)$ dla $\mu^m(Q_m^\varepsilon)$, które powinno
  - być niezależne od t (target concept) i μ (rozkład prawdopodobieństwa).
  - dążyć do 0 przy $m \to \infty$
- **Twierdzenie** Niech $\mathbb{H}$ będzie przestrzenią hipotez określonych na X. Dla dowolnych c, μ, ε (ale ustalonych) mamy
$$\mu^m(Q_m^\varepsilon) < 2\Pi_{\mathbb{H}}(2m)2^{-\varepsilon m/2}$$
  o ile $m \geqslant 8/\varepsilon$.
- Możemy użyć lematu Sauer'a, aby pokazać, że $\mu^m(Q_m^\varepsilon) < \delta$ dla dostatecznie dużych m.

Dowód

- Definiujemy
$$Q_m^\varepsilon = \{\mathbf{x} \in X^m | H[\mathbf{x}, t] \cap B_\varepsilon \neq \emptyset\}$$
- Szukamy górnego ograniczenia $f(m, \varepsilon)$ dla $\mu^m(Q_m^\varepsilon)$, które powinno
  - być niezależne od t (target concept) i μ (rozkład prawdopodobieństwa).
  - dążyć do 0 przy $m \to \infty$
- **Twierdzenie** Niech $\mathbb{H}$ będzie przestrzenią hipotez określonych na X. Dla dowolnych c, μ, ε (ale ustalonych) mamy
$$\mu^m(Q_m^\varepsilon) < 2\Pi_{\mathbb{H}}(2m)2^{-\varepsilon m/2}$$
o ile $m \geqslant 8/\varepsilon$.
- Możemy użyć lematu Sauer'a, aby pokazać, że $\mu^m(Q_m^\varepsilon) < \delta$ dla dostatecznie dużych m.

Złożoność zbioru treningowego

- Z Fundamentalnego Twierdzenia wynika, że jeśli $VCdim(\mathbb{H}) < \infty$, to dla danych δ i ε, istnieje $m_0 = m_0(\mathbb{H}, \delta, \varepsilon)$ takie, że

$$m \geq m_0 \Rightarrow \mu^m\{D \in \mathcal{S}(m, c) : \mathbb{H}[D] \cap \mathbb{B}_\varepsilon = \emptyset\} > 1 - \delta$$

- Wówczas każdy niesprzeczny algorytm $\mathfrak{L}$ jest PAC oraz wymagana liczba przykładów $m_L(\mathbb{H}, \delta, \varepsilon)$ dla $\mathfrak{L}$ jest ograniczona z góry przez $m_0(\mathbb{H}, \delta, \varepsilon)$.
- Dla skończonych przestrzeni hipotez $\mathbb{H}$ mamy

$$m_L(\mathbb{H}, \delta, \varepsilon) \leq \left\lceil \frac{1}{\varepsilon} \ln \frac{|\mathbb{H}|}{\delta} \right\rceil = \left\lceil \frac{1}{\varepsilon} (\ln |\mathbb{H}| + \ln(1/\delta)) \right\rceil$$

Złożoność zbioru treningowego (c.d.)

- **Twierdzenie** Niech $VCdim(\mathbb{H}) = d \geq 1$. Wówczas każdy algorytm niesprzeczny $\mathfrak{L}$ jest PAC oraz wymagana liczba przykładów dla $\mathfrak{L}$ wynosi

$$m_L(\mathbb{H}, \delta, \varepsilon) \leq \left\lceil \frac{4}{\varepsilon} \left(d \log \frac{12}{\varepsilon} + \log \frac{2}{\delta} \right) \right\rceil$$

- Dolne ograniczenia:
  - $m_L(\mathbb{H}, \delta, \varepsilon) \geqslant d(1 - \varepsilon)$
  - Jeśli $\delta \leq 1/100$ i $\varepsilon \leq 1/8$, to $m_L(\mathbb{H}, \delta, \varepsilon) > \frac{d-1}{32\varepsilon}$
  - $m_L(\mathbb{H}, \delta, \varepsilon) > \frac{1-\varepsilon}{\varepsilon} \ln \frac{1}{\delta}$