Teoria systemów uczących się i wymiar Vapnika-Chervonenkisa

Hung Son Nguyen

Wydział Matematyki, Informatyki i Mechaniki
Uniwersytet Warszawski
email: son@mimuw.edu.pl

Grudzień 2009

Plan wykładu

Podejście indukcyjne

- Np. Pokazać, że dla każdego $n \in \mathbb{N}$ zachodzi

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

Podejście indukcyjne

- Np. Pokazać, że dla każdego $n \in \mathbb{N}$ zachodzi

$$\Psi(n): \boxed{1^2 + 2^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6}}$$

- Indukcja pełna:

$$\Psi(1) \quad \text{oraz} \quad \forall_{n \geq 1}[\Psi(n) \implies \Psi(n+1)]$$

Podejście indukcyjne

- Np. Pokazać, że dla każdego $n \in \mathbb{N}$ zachodzi

$$\Psi(n): \boxed{1^2 + 2^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6}}$$

- Indukcja pełna:

$$\Psi(1) \quad \text{oraz} \quad \forall_{n \geq 1}[\Psi(n) \implies \Psi(n+1)]$$

- Indukcja niepełna: czy wystarczy sprawdzić, np.

$$\Psi(1), \Psi(2), \Psi(3), \Psi(4)?$$

Komputerowe uczenie się pojęć

Podejście indukcyjne:

Wnioskowanie na podstawie skończonego zbioru obserwacji

Komputerowe uczenie się pojęć

Podejście indukcyjne:

Wnioskowanie na podstawie skończonego zbioru obserwacji

Jakie prawa rządzą procesem indukcyjnego uczenia się pojęć?

Komputerowe uczenie się pojęć

Podejście indukcyjne:

Wnioskowanie na podstawie skończonego zbioru obserwacji

Jakie prawa rządzą procesem indukcyjnego uczenia się pojęć?

Szukamy teorii obejmującej zagadnienia:

- **Szansy na skuteczne wyuczanie się pojęć;**

Komputerowe uczenie się pojęć

Podejście indukcyjne:

Wnioskowanie na podstawie skończonego zbioru obserwacji

Jakie prawa rządzą procesem indukcyjnego uczenia się pojęć?

Szukamy teorii obejmującej zagadnienia:

- **Szansy na skuteczne wyuczanie się pojęć;**
- **Niezbędnej liczby przykładów treningowych;**

Komputerowe uczenie się pojęć

Podejście indukcyjne:

Wnioskowanie na podstawie skończonego zbioru obserwacji

Jakie prawa rządzą procesem indukcyjnego uczenia się pojęć?

Szukamy teorii obejmującej zagadnienia:

- **Szansy na skuteczne wyuczanie się pojęć;**
- **Niezbędnej liczby przykładów treningowych;**
- **Złożoności przestrzeni hipotez;**

Komputerowe uczenie się pojęć

Podejście indukcyjne:

Wnioskowanie na podstawie skończonego zbioru obserwacji

Jakie prawa rządzą procesem indukcyjnego uczenia się pojęć?

Szukamy teorii obejmującej zagadnienia:

- **Szansy na skuteczne wyuczanie się pojęć;**
- **Niezbędnej liczby przykładów treningowych;**
- **Złożoności przestrzeni hipotez;**
- **Jakości aproksymacji;**

Komputerowe uczenie się pojęć

Podejście indukcyjne:

Wnioskowanie na podstawie skończonego zbioru obserwacji

Jakie prawa rządzą procesem indukcyjnego uczenia się pojęć?

Szukamy teorii obejmującej zagadnienia:

- **Szansy na skuteczne wyuczanie się pojęć;**
- **Niezbędnej liczby przykładów treningowych;**
- **Złożoności przestrzeni hipotez;**
- **Jakości aproksymacji;**
- **Metod reprezentacji danych treningowych;**

Ogólny model uczenia indukcyjnego

- Niech

Ogólny model uczenia indukcyjnego

- Niech
  - $\mathcal{X}$ – (skończony lub nieskończony) zbiór obiektów;

Ogólny model uczenia indukcyjnego

- Niech
  - $\mathcal{X}$ – (skończony lub nieskończony) zbiór obiektów;
  - $\mathbb{C}$ – klasa pojęć w $\mathcal{X}$, tj. $\mathbb{C} = \{f : \mathcal{X} \to \{0, 1\}\}$

Ogólny model uczenia indukcyjnego

- Niech
  - $\mathcal{X}$ – (skończony lub nieskończony) zbiór obiektów;
  - $\mathbb{C}$ – klasa pojęć w $\mathcal{X}$, tj. $\mathbb{C} = \{f : \mathcal{X} \to \{0, 1\}\}$
  - $c \in \mathbb{C}$ – pojęcie docelowe lub funkcja celu;

Ogólny model uczenia indukcyjnego

- Niech
  - $\mathcal{X}$ – (skończony lub nieskończony) zbiór obiektów;
  - $\mathbb{C}$ – klasa pojęć w $\mathcal{X}$, tj. $\mathbb{C} = \{f : \mathcal{X} \to \{0, 1\}\}$
  - $c \in \mathbb{C}$ – pojęcie docelowe lub funkcja celu;
- **Dane są**

Ogólny model uczenia indukcyjnego

- Niech
  - $\mathcal{X}$ – (skończony lub nieskończony) zbiór obiektów;
  - $\mathbb{C}$ – klasa pojęć w $\mathcal{X}$, tj. $\mathbb{C} = \{f : \mathcal{X} \to \{0, 1\}\}$
  - $c \in \mathbb{C}$ – pojęcie docelowe lub funkcja celu;
- **Dane są**
  - skończona próbka etykietowanych obiektów:

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

gdzie $x_1, ..., x_m \in \mathcal{X}$.

Ogólny model uczenia indukcyjnego

- Niech
  - $\mathcal{X}$ – (skończony lub nieskończony) zbiór obiektów;
  - $\mathbb{C}$ – klasa pojęć w $\mathcal{X}$, tj. $\mathbb{C} = \{f : \mathcal{X} \to \{0, 1\}\}$
  - $c \in \mathbb{C}$ – pojęcie docelowe lub funkcja celu;
- **Dane są**
  - skończona próbka etykietowanych obiektów:

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

    gdzie $x_1, ..., x_m \in \mathcal{X}$.
  - przestrzeń hipotez $\mathbb{H} = \{h : \mathcal{X} \to \{0, 1\}\}$;

Ogólny model uczenia indukcyjnego

- Niech
  - $\mathcal{X}$ – (skończony lub nieskończony) zbiór obiektów;
  - $\mathbb{C}$ – klasa pojęć w $\mathcal{X}$, tj. $\mathbb{C} = \{f : \mathcal{X} \to \{0, 1\}\}$
  - $c \in \mathbb{C}$ – pojęcie docelowe lub funkcja celu;
- **Dane są**
  - skończona próbka etykietowanych obiektów:

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

    gdzie $x_1, ..., x_m \in \mathcal{X}$.
  - przestrzeń hipotez $\mathbb{H} = \{h : \mathcal{X} \to \{0, 1\}\}$;
- **Szukana**

Ogólny model uczenia indukcyjnego

- Niech
  - $\mathcal{X}$ – (skończony lub nieskończony) zbiór obiektów;
  - $\mathbb{C}$ – klasa pojęć w $\mathcal{X}$, tj. $\mathbb{C} = \{f : \mathcal{X} \to \{0, 1\}\}$
  - $c \in \mathbb{C}$ – pojęcie docelowe lub funkcja celu;
- **Dane są**
  - skończona próbka etykietowanych obiektów:

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

    gdzie $x_1, ..., x_m \in \mathcal{X}$.
  - przestrzeń hipotez $\mathbb{H} = \{h : \mathcal{X} \to \{0, 1\}\}$;
- **Szukana**
  - hipoteza $h \in \mathbb{H}$ będąca dobrą aproksymacją pojęcia c.

Ogólny model uczenia indukcyjnego

- Niech
  - $\mathcal{X}$ – (skończony lub nieskończony) zbiór obiektów;
  - $\mathbb{C}$ – klasa pojęć w $\mathcal{X}$, tj. $\mathbb{C} = \{f : \mathcal{X} \to \{0, 1\}\}$
  - $c \in \mathbb{C}$ – pojęcie docelowe lub funkcja celu;
- **Dane są**
  - skończona próbka etykietowanych obiektów:

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

    gdzie $x_1, ..., x_m \in \mathcal{X}$.
  - przestrzeń hipotez $\mathbb{H} = \{h : \mathcal{X} \to \{0, 1\}\}$;
- **Szukana**
  - hipoteza $h \in \mathbb{H}$ będąca dobrą aproksymacją pojęcia c.
- **Wymagane**

Ogólny model uczenia indukcyjnego

- Niech
  - $\mathcal{X}$ – (skończony lub nieskończony) zbiór obiektów;
  - $\mathbb{C}$ – klasa pojęć w $\mathcal{X}$, tj. $\mathbb{C} = \{f : \mathcal{X} \to \{0, 1\}\}$
  - $c \in \mathbb{C}$ – pojęcie docelowe lub funkcja celu;
- **Dane są**
  - skończona próbka etykietowanych obiektów:

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

    gdzie $x_1, ..., x_m \in \mathcal{X}$.
  - przestrzeń hipotez $\mathbb{H} = \{h : \mathcal{X} \to \{0, 1\}\}$;
- **Szukana**
  - hipoteza $h \in \mathbb{H}$ będąca dobrą aproksymacją pojęcia c.
- **Wymagane**
  - dobra jakość aproksymacji

Ogólny model uczenia indukcyjnego

- Niech
  - $\mathcal{X}$ – (skończony lub nieskończony) zbiór obiektów;
  - $\mathbb{C}$ – klasa pojęć w $\mathcal{X}$, tj. $\mathbb{C} = \{f : \mathcal{X} \to \{0, 1\}\}$
  - $c \in \mathbb{C}$ – pojęcie docelowe lub funkcja celu;
- **Dane są**
  - skończona próbka etykietowanych obiektów:

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

    gdzie $x_1, ..., x_m \in \mathcal{X}$.
  - przestrzeń hipotez $\mathbb{H} = \{h : \mathcal{X} \to \{0, 1\}\}$;
- **Szukana**
  - hipoteza $h \in \mathbb{H}$ będąca dobrą aproksymacją pojęcia c.
- **Wymagane**
  - dobra jakość aproksymacji
  - szybki czas wyuczania.

Przykład

Przykład

Przykład

- Pojęcie: "człowieka o średniej budowie ciała".

Przykład

- Pojęcie: "człowieka o średniej budowie ciała".
- Dane – czyli osoby – są reprezentowane przez ich *wagę*(kg) i *wzrost*(cm) i są etykietowane przez $+$ i $-$.

Przykład

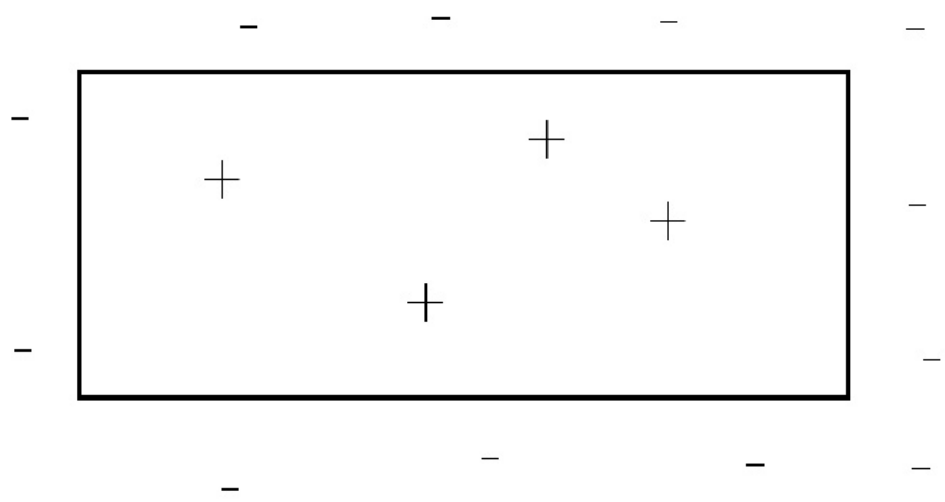
- Pojęcie: "człowieka o średniej budowie ciała".
- Dane – czyli osoby – są reprezentowane przez ich *wagę*(kg) i *wzrost*(cm) i są etykietowane przez $+$ i $-$.
- Dodatkowa wiedza: szukane pojęcie można wyrazić za pomocą PROSTOKĄTA

Przykład

- Pojęcie: "człowieka o średniej budowie ciała".
- Dane – czyli osoby – są reprezentowane przez ich *wagę*(kg) i *wzrost*(cm) i są etykietowane przez $+$ i $-$.
- Dodatkowa wiedza: szukane pojęcie można wyrazić za pomocą PROSTOKĄTA

Przykład

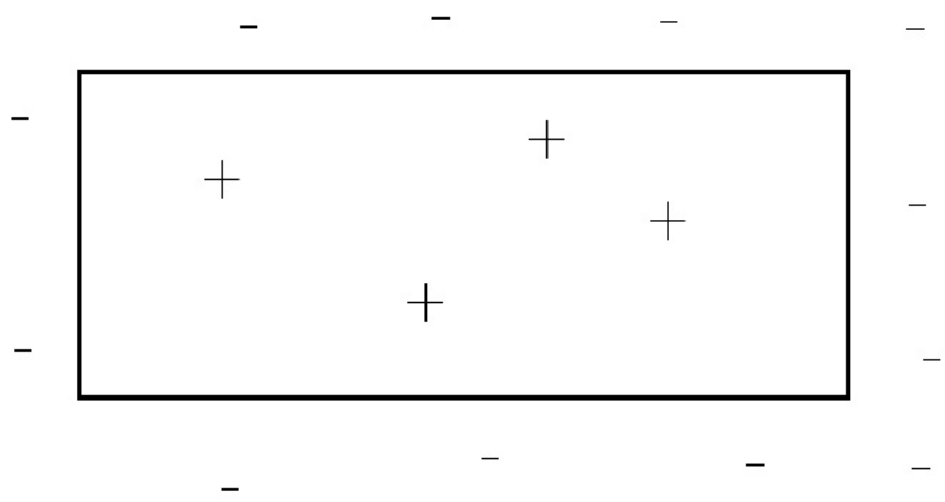
- Pojęcie: "człowieka o średniej budowie ciała".
- Dane – czyli osoby – są reprezentowane przez ich *wagę*(kg) i *wzrost*(cm) i są etykietowane przez $+$ i $-$.
- Dodatkowa wiedza: szukane pojęcie można wyrazić za pomocą PROSTOKĄTA



Uczenie prostokąta

Przykład

- Pojęcie: "człowieka o średniej budowie ciała".
- Dane – czyli osoby – są reprezentowane przez ich *wagę*(kg) i *wzrost*(cm) i są etykietowane przez $+$ i $-$.
- Dodatkowa wiedza: szukane pojęcie można wyrazić za pomocą PROSTOKĄTA

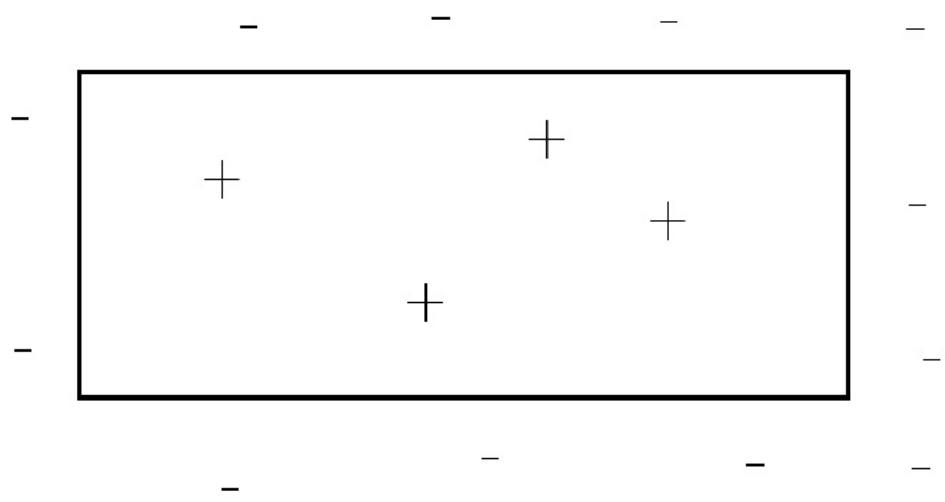


Uczenie prostokąta

- $\mathcal{X} = \Re^2$;

Przykład

- Pojęcie: "człowieka o średniej budowie ciała".
- Dane – czyli osoby – są reprezentowane przez ich *wagę*(kg) i *wzrost*(cm) i są etykietowane przez $+$ i $-$.
- Dodatkowa wiedza: szukane pojęcie można wyrazić za pomocą PROSTOKĄTA

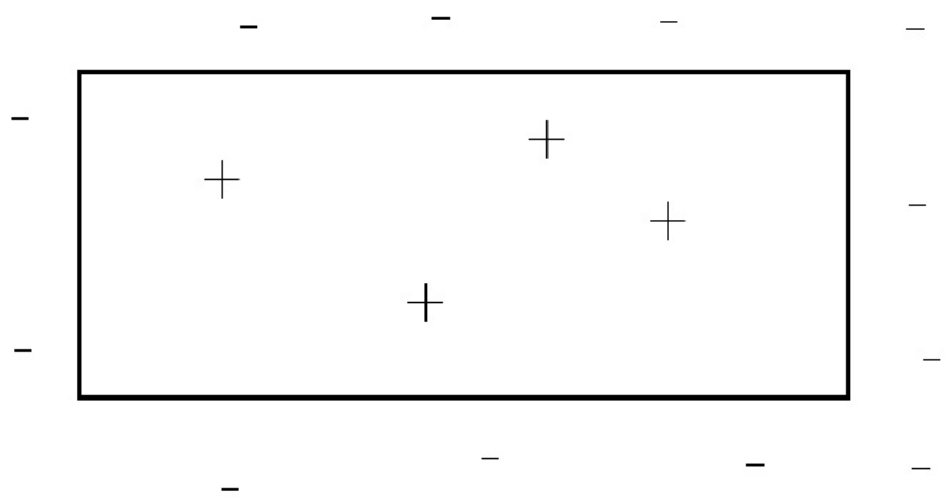


Uczenie prostokąta

- $\mathcal{X} = \Re^2$;
- $\mathbb{C} = \mathbb{H} =$ zbiór prostokątów;

Przykład

- Pojęcie: "człowieka o średniej budowie ciała".
- Dane – czyli osoby – są reprezentowane przez ich *wagę*(kg) i *wzrost*(cm) i są etykietowane przez $+$ i $-$.
- Dodatkowa wiedza: szukane pojęcie można wyrazić za pomocą PROSTOKĄTA

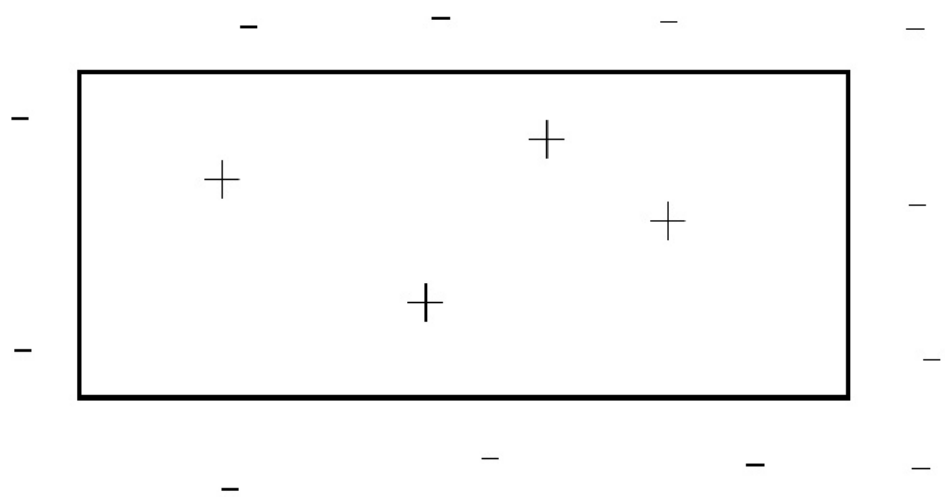


Uczenie prostokąta

- $\mathcal{X} = \Re^2$;
- $\mathbb{C} = \mathbb{H} =$ zbiór prostokątów;
- Przykład zbioru treningowego
  $((84, 184), +)$,
  $((70, 170), +)$,
  $((75, 163), -)$,
  $((80, 180), +)$,
  $((81, 195), -)$,
  $((63, 191), -)$,
  $((77, 187), -)$,
  $((68, 168), +)$

Przykład

- Pojęcie: "człowieka o średniej budowie ciała".
- Dane – czyli osoby – są reprezentowane przez ich *wagę*(kg) i *wzrost*(cm) i są etykietowane przez $+$ i $-$.
- Dodatkowa wiedza: szukane pojęcie można wyrazić za pomocą PROSTOKĄTA



Uczenie prostokąta

- $\mathcal{X} = \Re^2$;
- $\mathbb{C} = \mathbb{H} =$ zbiór prostokątów;
- Przykład zbioru treningowego $((84, 184), +)$, $((70, 170), +)$, $((75, 163), -)$, $((80, 180), +)$, $((81, 195), -)$, $((63, 191), -)$, $((77, 187), -)$, $((68, 168), +)$
- $((79, 183, ?)$

Inne przykłady

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

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

Inne przykłady

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

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

- **Uczenie hiperpłaszczyzny**:

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

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

Inne przykłady

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

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

- **Uczenie hiperpłaszczyzny**:

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

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

- **Uczenie jednomianów Boolowskich**:

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

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

Błąd hipotezy

Błąd rzeczywisty

Błąd hipotezy

Błąd rzeczywisty

- $\Omega = (\mathcal{X}, \mu)$ – przestrzeń probabilistyczna na $\mathcal{X}$;

Błąd hipotezy

Błąd rzeczywisty

- $\Omega = (\mathcal{X}, \mu)$ – przestrzeń probabilistyczna na $\mathcal{X}$;
- Błąd hipotezy $h \in \mathbb{H}$ względem funkcji celu c:

$$\boxed{er_\Omega(h, c) = er^c_\Omega(h) = \mu(\mathcal{X}_{h \neq c})}$$

gdzie $\mathcal{X}_{h \neq c} = \{x \in \mathcal{X} : h(x) \neq c(x)\}$.

Błąd hipotezy

Błąd rzeczywisty

- $\Omega = (\mathcal{X}, \mu)$ – przestrzeń probabilistyczna na $\mathcal{X}$;
- Błąd hipotezy $h \in \mathbb{H}$ względem funkcji celu c:

$$\boxed{er_\Omega(h, c) = er^c_\Omega(h) = \mu(\mathcal{X}_{h \neq c})}$$

gdzie $\mathcal{X}_{h \neq c} = \{x \in \mathcal{X} : h(x) \neq c(x)\}$.

Statystyka: Jeśli przykłady z D są wybrane zgodnie z miarą prawdopodobieństwa μ w sposób niezależny oraz $|D| \geqslant 30$, to

Błąd hipotezy

Błąd rzeczywisty

- $\Omega = (\mathcal{X}, \mu)$ – przestrzeń probabilistyczna na $\mathcal{X}$;
- Błąd hipotezy $h \in \mathbb{H}$ względem funkcji celu c:

$$\boxed{er_\Omega(h, c) = er^c_\Omega(h) = \mu(\mathcal{X}_{h \neq c})}$$

gdzie $\mathcal{X}_{h \neq c} = \{x \in \mathcal{X} : h(x) \neq c(x)\}$.

Statystyka: Jeśli przykłady z D są wybrane zgodnie z miarą prawdopodobieństwa μ w sposób niezależny oraz $|D| \geqslant 30$, to

- $er^c_\Omega(h) \approx er^c_D(h) = \frac{|D \cap \mathcal{X}_{h \neq c}|}{|D|}$,

Błąd hipotezy

Błąd rzeczywisty

- $\Omega = (\mathcal{X}, \mu)$ – przestrzeń probabilistyczna na $\mathcal{X}$;
- Błąd hipotezy $h \in \mathbb{H}$ względem funkcji celu c:

$$\boxed{er_\Omega(h, c) = er^c_\Omega(h) = \mu(\mathcal{X}_{h \neq c})}$$

gdzie $\mathcal{X}_{h \neq c} = \{x \in \mathcal{X} : h(x) \neq c(x)\}$.

Statystyka: Jeśli przykłady z D są wybrane zgodnie z miarą prawdopodobieństwa μ w sposób niezależny oraz $|D| \geqslant 30$, to

- $er^c_\Omega(h) \approx er^c_D(h) = \frac{|D \cap \mathcal{X}_{h \neq c}|}{|D|}$,
- z prawdopodobieństwem $(1 - \varepsilon)$

$$|er^c_\Omega - er^c_D| \leqslant s_{\frac{\varepsilon}{2}} \cdot \sqrt{\frac{er^c_D(1 - er^c_D)}{|D|}}$$

Plan wykładu

Model uczenia się PAC

Idea modelu PAC (Probably Approximately Correct):

Określenie warunków, przy których uczeń (algorytm uczenia się) z „dużym prawdopodobieństwem" znajdzie „dobrą hipotezę" na podstawie danych D.

Model uczenia się PAC

Idea modelu PAC (Probably Approximately Correct):

Określenie warunków, przy których uczeń (algorytm uczenia się) z „dużym prawdopodobieństwem" znajdzie „dobrą hipotezę" na podstawie danych D.

PAC-owy uczeń

Niech $\mathfrak{L}$ będzie algorytmem uczenia się, jeśli

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

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

Wówczas mówimy w skrócie, że $\mathfrak{L}$ *jest* **PAC** *dla klasy* $\mathbb{C}$ (*"prawdopodobnie aproksymacyjnie poprawny"*).

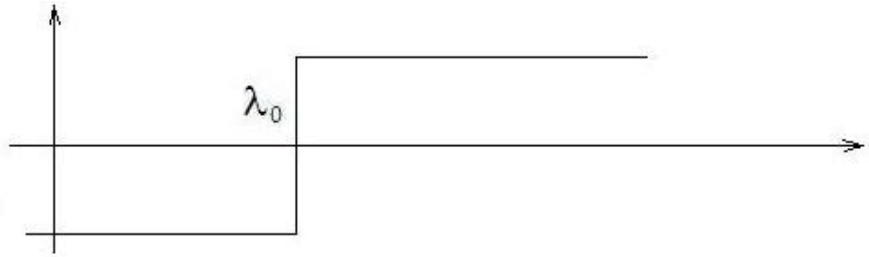
ε = dopuszczalny poziom błędu; $(1 - \delta)$ = poziom zaufania.

Przykład problemu dyskretyzacji

- $\mathbb{H} = \mathbb{C} = \{f_\lambda : \Re \to \{0,1\} : f_\lambda(x) = 1 \Leftrightarrow x \geqslant \lambda\}$

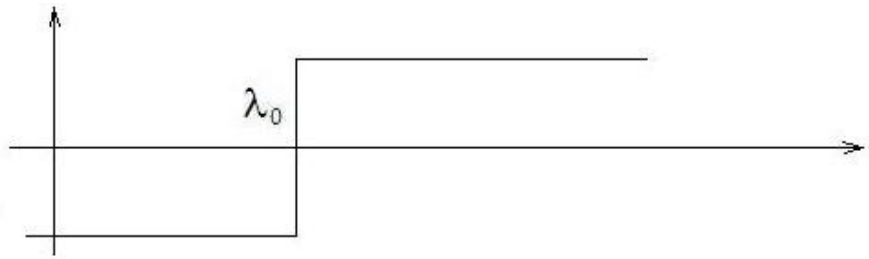
Przykład problemu dyskretyzacji

- $\mathbb{H} = \mathbb{C} = \{f_\lambda : \Re \to \{0,1\} : f_\lambda(x) = 1 \Leftrightarrow x \geqslant \lambda\}$
- $c = f_{\lambda_0}$



Przykład problemu dyskretyzacji

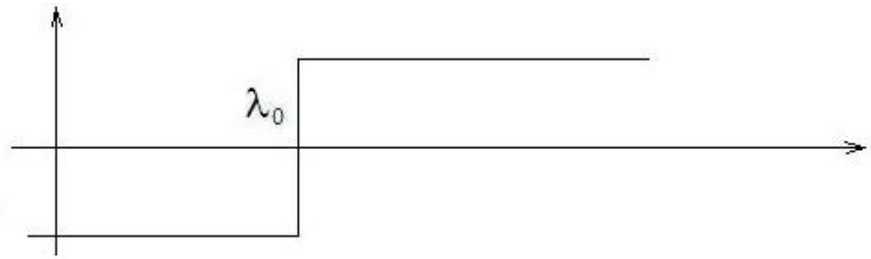
- $\mathbb{H} = \mathbb{C} = \{f_\lambda : \Re \to \{0,1\} : f_\lambda(x) = 1 \Leftrightarrow x \geqslant \lambda\}$
- $c = f_{\lambda_0}$



- znaleźć λ_0 na podstawie losowo wygenerowanych przykładów $D = \{\langle x_1, f_{\lambda_0}(x_1)\rangle, ..., \langle x_m, f_{\lambda_0}(x_m)\rangle\}$

Przykład problemu dyskretyzacji

- $\mathbb{H} = \mathbb{C} = \{f_\lambda : \Re \to \{0,1\} : f_\lambda(x) = 1 \Leftrightarrow x \geqslant \lambda\}$
- $c = f_{\lambda_0}$

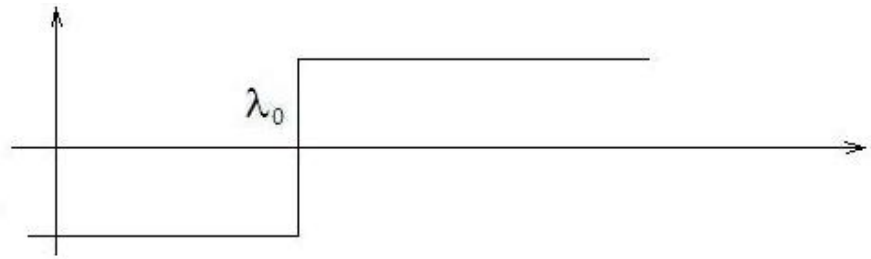


- znaleźć λ_0 na podstawie losowo wygenerowanych przykładów $D = \{\langle x_1, f_{\lambda_0}(x_1)\rangle, ..., \langle x_m, f_{\lambda_0}(x_m)\rangle\}$

Algorytm:

Przykład problemu dyskretyzacji

- $\mathbb{H} = \mathbb{C} = \{f_\lambda : \Re \to \{0,1\} : f_\lambda(x) = 1 \Leftrightarrow x \geqslant \lambda\}$
- $c = f_{\lambda_0}$



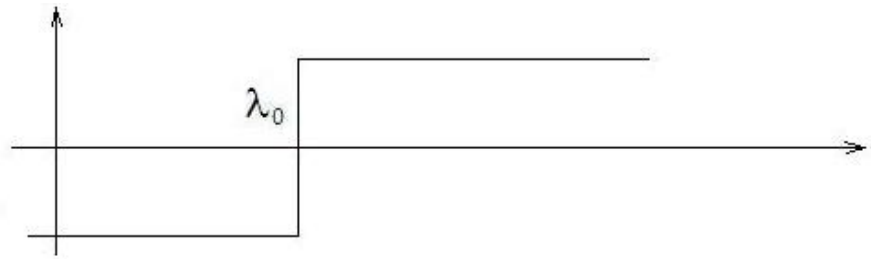
- znaleźć λ_0 na podstawie losowo wygenerowanych przykładów $D = \{\langle x_1, f_{\lambda_0}(x_1)\rangle, \ldots, \langle x_m, f_{\lambda_0}(x_m)\rangle\}$

Algorytm:

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

Przykład problemu dyskretyzacji

- $\mathbb{H} = \mathbb{C} = \{f_\lambda : \Re \to \{0,1\} : f_\lambda(x) = 1 \Leftrightarrow x \geqslant \lambda\}$
- $c = f_{\lambda_0}$



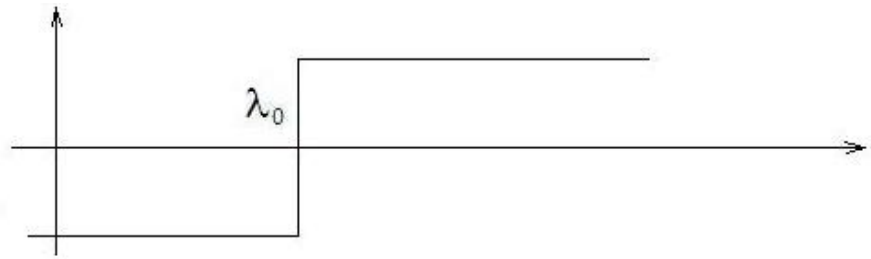
- znaleźć λ_0 na podstawie losowo wygenerowanych przykładów $D = \{\langle x_1, f_{\lambda_0}(x_1)\rangle, \ldots, \langle x_m, f_{\lambda_0}(x_m)\rangle\}$

Algorytm:

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

Przykład problemu dyskretyzacji

- $\mathbb{H} = \mathbb{C} = \{f_\lambda : \Re \to \{0,1\} : f_\lambda(x) = 1 \Leftrightarrow x \geqslant \lambda\}$
- $c = f_{\lambda_0}$



- znaleźć λ_0 na podstawie losowo wygenerowanych przykładów $D = \{\langle x_1, f_{\lambda_0}(x_1)\rangle, \ldots, \langle x_m, f_{\lambda_0}(x_m)\rangle\}$

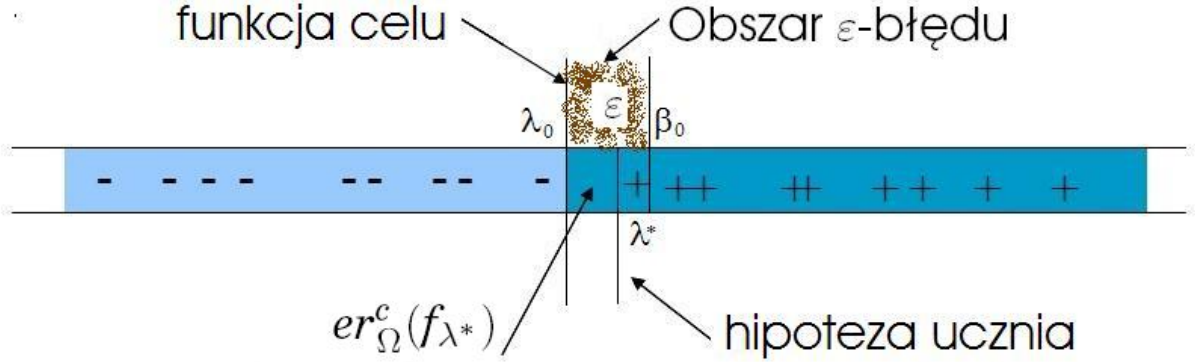
Algorytm:

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

Twierdzenie:

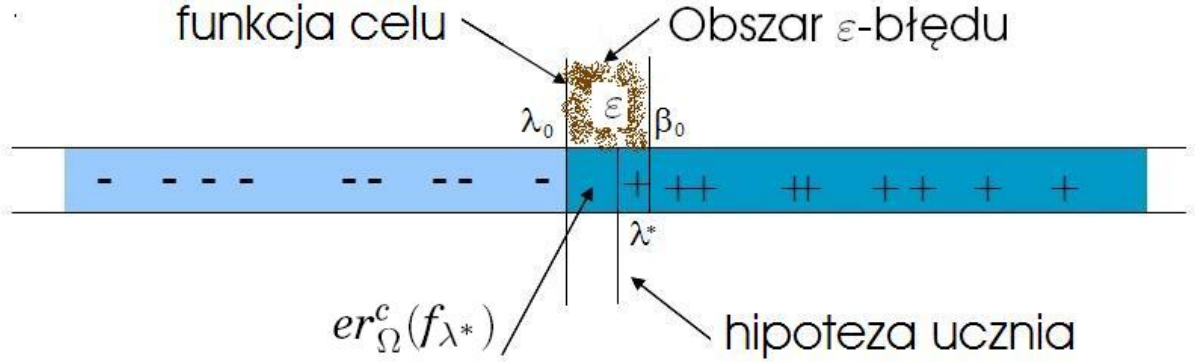
Powyższy algorytm jest PAC

Dowód



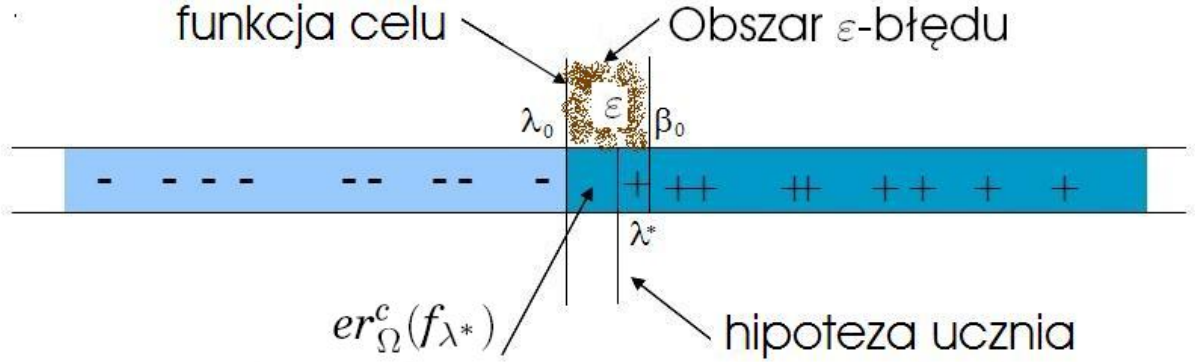
- $er^c_\Omega(f_{\lambda^*}) = \mu([\lambda_0, \lambda^*))$.

Dowód



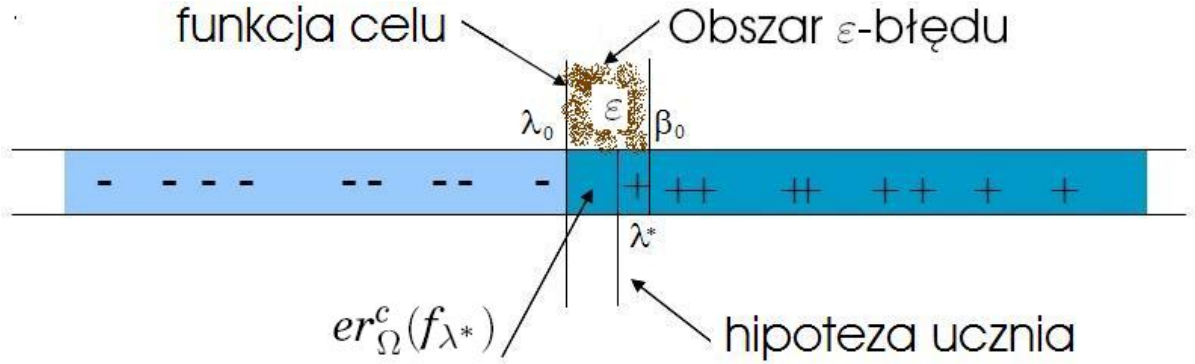
- $er^c_\Omega(f_{\lambda^*}) = \mu([\lambda_0, \lambda^*))$.
- Niech $\beta_0 = \mathbf{sup}\{\beta : \mu([\lambda_0, \beta)) < \varepsilon\}$.

Dowód



- $er^c_\Omega(f_{\lambda^*}) = \mu([\lambda_0, \lambda^*))$.
- Niech $\beta_0 = \mathbf{sup}\{\beta : \mu([\lambda_0, \beta)) < \varepsilon\}$.
- Wówczas $er^c_\Omega(f_{\lambda^*}) \geqslant \varepsilon \Leftrightarrow \forall_{x_i \in D} : x_i \notin [\lambda_0, \beta_0]$;

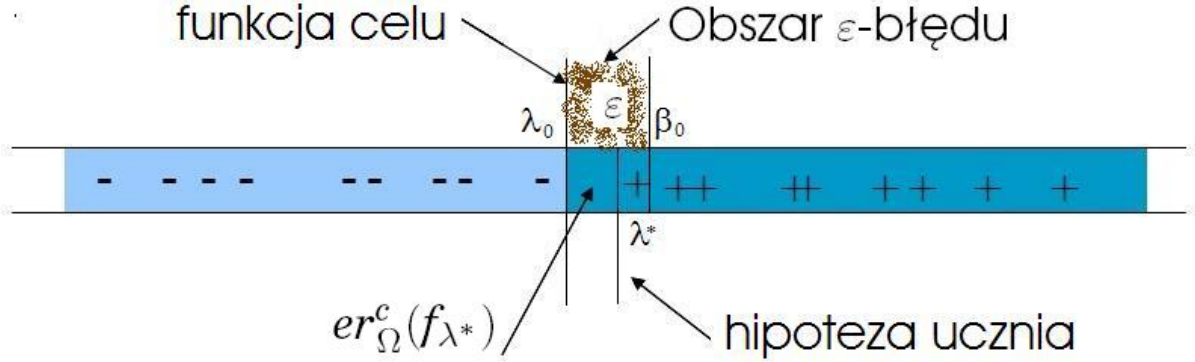
Dowód



- $er^c_\Omega(f_{\lambda^*}) = \mu([\lambda_0, \lambda^*))$.
- Niech $\beta_0 = \mathbf{sup}\{\beta : \mu([\lambda_0, \beta)) < \varepsilon\}$.
- Wówczas $er^c_\Omega(f_{\lambda^*}) \geqslant \varepsilon \Leftrightarrow \forall_{x_i \in D} : x_i \notin [\lambda_0, \beta_0]$;
- Stąd

$$\mu^m\{(x_1, ..., x_m) : \forall_{x_i \in D} : x_i \notin [\lambda_0, \beta_0]\} \leqslant (1-\varepsilon)^m$$
$$\mu^m\{D \in \mathcal{S}(m, f_{\lambda_0}) : er_\Omega(f_{\lambda^*}) \leqslant \varepsilon\} \geqslant 1 - (1-\varepsilon)^m$$

Dowód



- $er^c_\Omega(f_{\lambda^*}) = \mu([\lambda_0, \lambda^*))$.
- Niech $\beta_0 = \mathbf{sup}\{\beta : \mu([\lambda_0, \beta)) < \varepsilon\}$.
- Wówczas $er^c_\Omega(f_{\lambda^*}) \geqslant \varepsilon \Leftrightarrow \forall_{x_i \in D} : x_i \notin [\lambda_0, \beta_0]$;
- Stąd

$$\mu^m\{(x_1, ..., x_m) : \forall_{x_i \in D} : x_i \notin [\lambda_0, \beta_0]\} \leqslant (1-\varepsilon)^m$$
$$\mu^m\{D \in \mathcal{S}(m, f_{\lambda_0}) : er_\Omega(f_{\lambda^*}) \leqslant \varepsilon\} \geqslant 1 - (1-\varepsilon)^m$$

- Aby to prawdopodobieństwo było $> 1 - \delta$, wystarczy przyjąć $m \geqslant m_0 = \left\lceil \frac{1}{\varepsilon} \ln \frac{1}{\delta} \right\rceil$

Dokładne uczenie się

- Niech Ω będzie rozkładem dyskretnym zdefiniowanym przez $\mu_1 = \mu(x_1)$, ..., $\mu_n = \mu(x_n)$ – dla pewnych $x_1, ..., x_n \in X$ – takich, że $\mu_1 + ... + \mu_n = 1$. Niech $\varepsilon_{\min} = \min_i \mu_i$.

Dokładne uczenie się

- Niech Ω będzie rozkładem dyskretnym zdefiniowanym przez $\mu_1 = \mu(x_1)$, ..., $\mu_n = \mu(x_n)$ – dla pewnych $x_1, ..., x_n \in X$ – takich, że $\mu_1 + ... + \mu_n = 1$. 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

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

Dokładne uczenie się

- Niech Ω będzie rozkładem dyskretnym zdefiniowanym przez $\mu_1 = \mu(x_1)$, ..., $\mu_n = \mu(x_n)$ – dla pewnych $x_1, ..., x_n \in X$ – takich, że $\mu_1 + ... + \mu_n = 1$. 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*)

Plan wykładu

Potencjalna wyuczalność

- Niech $D = \{\langle x_1, c(x_1)\rangle, ..., \langle x_m, c(x_m)\rangle\}$ i niech

$$\boxed{\mathbb{H}^c(D) = \{h \in \mathbb{H} : h|D = c|D\}}$$

zbiór hipotez zgodnych z c na próbce D.

Potencjalna wyuczalność

- Niech $D = \{\langle x_1, c(x_1)\rangle, ..., \langle x_m, c(x_m)\rangle\}$ i niech

$$\boxed{\mathbb{H}^c(D) = \{h \in \mathbb{H} : h|D = c|D\}}$$

zbiór hipotez zgodnych z c na próbce D.

- $\boxed{\mathbb{B}^c_\varepsilon = \{h \in \mathbb{H} : er_\Omega(h) \geqslant \varepsilon\}}$ – zbiór ε-złych hipotez

Potencjalna wyuczalność

- Niech $D = \{\langle x_1, c(x_1)\rangle, ..., \langle x_m, c(x_m)\rangle\}$ i niech

$$\boxed{\mathbb{H}^c(D) = \{h \in \mathbb{H} : h|D = c|D\}}$$

zbiór hipotez zgodnych z c na próbce D.

- $\boxed{\mathbb{B}^c_\varepsilon = \{h \in \mathbb{H} : er_\Omega(h) \geqslant \varepsilon\}}$ – zbiór ε-złych hipotez

Potencjalna wyuczalność

- Niech $D = \{\langle x_1, c(x_1)\rangle, ..., \langle x_m, c(x_m)\rangle\}$ i niech

$$\boxed{\mathbb{H}^c(D) = \{h \in \mathbb{H} : h|D = c|D\}}$$

zbiór hipotez zgodnych z c na próbce D.

- $\boxed{\mathbb{B}^c_\varepsilon = \{h \in \mathbb{H} : er_\Omega(h) \geqslant \varepsilon\}}$ – zbiór ε-złych hipotez

Definicja: Potencjalna wyuczalność

Mówimy, że $\mathbb{C}$ jest potencjalnie wyuczalna za pomocą $\mathbb{H}$, jeśli dla każdego rozkładu Ω na $\mathcal{X}$ 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 $\mathfrak{L}(D) \in \mathbb{H}^c(D)$ dla każdego zbioru D.

Twierdzenie

W przestrzeni potencjalnie wyuczalnej, każdy wzorowy uczeń (niesprzeczny algorytm) jest PAC-owy.

Potencjalna wyuczalność

Algorytm $\mathfrak{L}$ nazywamy niesprzecznym jeśli $\mathfrak{L}(D) \in \mathbb{H}^c(D)$ dla każdego zbioru D.

Twierdzenie

W przestrzeni potencjalnie wyuczalnej, każdy wzorowy uczeń (niesprzeczny algorytm) jest PAC-owy.

Twierdzenie (Haussler, 1988)

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

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

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

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

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

Plan wykładu

Wymiar Vapnika-Chervonenkisa

- Niech $\vec{\mathbf{x}} = \langle x_1, ..., x_m \rangle \in \mathcal{X}^m$. Niech

$$\Pi_{\mathbb{H}}(\vec{\mathbf{x}}) = |\{\langle h(x_1), ..., h(x_m) \rangle \in \{0,1\}^m : h \in \mathbb{H}\}|$$

Wymiar Vapnika-Chervonenkisa

- Niech $\overrightarrow{\mathbf{x}} = \langle x_1, ..., x_m \rangle \in \mathcal{X}^m$. Niech

$$\Pi_{\mathbb{H}}(\overrightarrow{\mathbf{x}}) = |\{\langle h(x_1), ..., h(x_m)\rangle \in \{0,1\}^m : h \in \mathbb{H}\}|$$

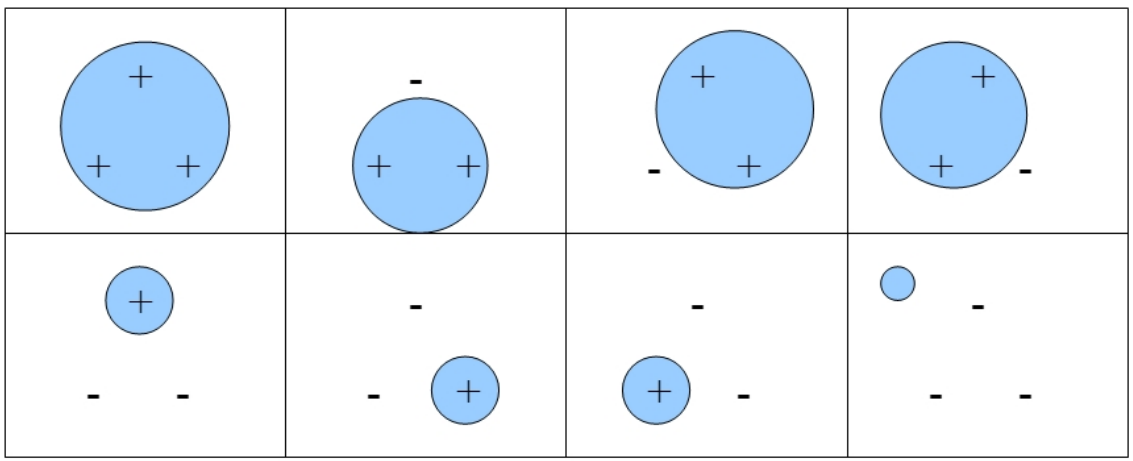
- $\Pi_{\mathbb{H}}(\overrightarrow{\mathbf{x}})$ jest liczbą podziałów zbioru elementów $\overrightarrow{\mathbf{x}}$ wyznaczonych przez $\mathbb{H}$. Mamy $\Pi_{\mathbb{H}}(\overrightarrow{\mathbf{x}}) \leqslant 2^m$.

Wymiar Vapnika-Chervonenkisa

- Niech $\vec{\mathbf{x}} = \langle x_1, ..., x_m \rangle \in \mathcal{X}^m$. Niech

$$\Pi_{\mathbb{H}}(\vec{\mathbf{x}}) = |\{\langle h(x_1), ..., h(x_m) \rangle \in \{0,1\}^m : h \in \mathbb{H}\}|$$

- $\Pi_{\mathbb{H}}(\vec{\mathbf{x}})$ jest liczbą podziałów zbioru elementów $\vec{\mathbf{x}}$ wyznaczonych przez $\mathbb{H}$. Mamy $\Pi_{\mathbb{H}}(\vec{\mathbf{x}}) \leqslant 2^m$.
- Gdy $\Pi_{\mathbb{H}}(\vec{\mathbf{x}}) = 2^m$, mówimy, że $\mathbb{H}$ rozbija $\mathbf{x}$.

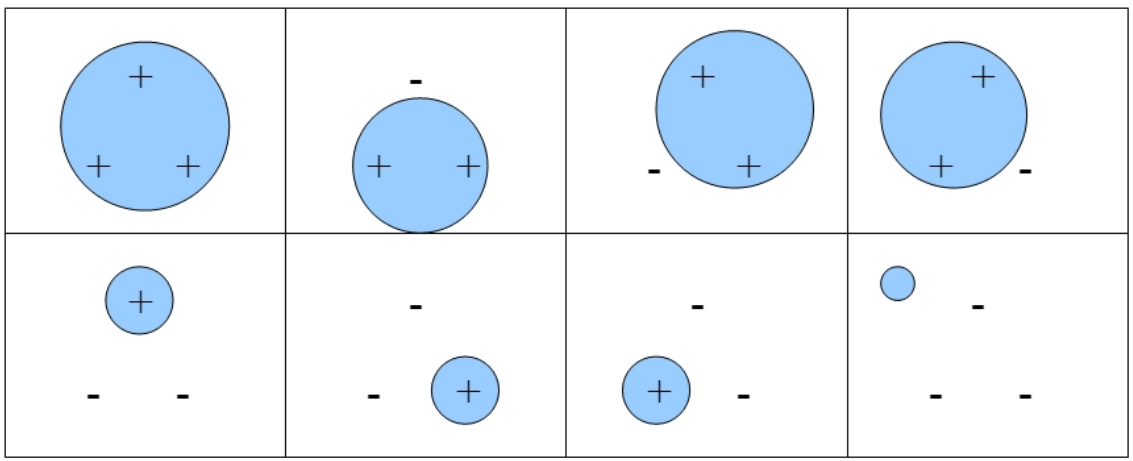


Wymiar Vapnika-Chervonenkisa

- Niech $\vec{\mathbf{x}} = \langle x_1, ..., x_m \rangle \in \mathcal{X}^m$. Niech

$$\Pi_{\mathbb{H}}(\vec{\mathbf{x}}) = |\{\langle h(x_1), ..., h(x_m) \rangle \in \{0,1\}^m : h \in \mathbb{H}\}|$$

- $\Pi_{\mathbb{H}}(\vec{\mathbf{x}})$ jest liczbą podziałów zbioru elementów $\vec{\mathbf{x}}$ wyznaczonych przez $\mathbb{H}$. Mamy $\Pi_{\mathbb{H}}(\vec{\mathbf{x}}) \leqslant 2^m$.
- Gdy $\Pi_{\mathbb{H}}(\vec{\mathbf{x}}) = 2^m$, mówimy, że $\mathbb{H}$ rozbija $\mathbf{x}$.



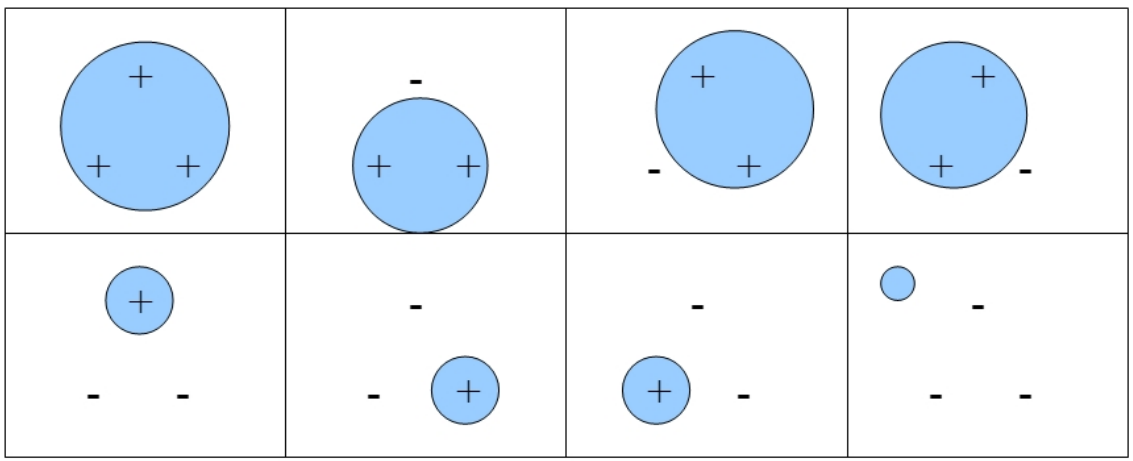
- Niech $\Pi_{\mathbb{H}}(m) = \max_{\vec{\mathbf{x}} \in \mathcal{X}^m} \Pi_{\mathbb{H}}(\vec{\mathbf{x}})$

Wymiar Vapnika-Chervonenkisa

- Niech $\vec{\mathbf{x}} = \langle x_1, ..., x_m \rangle \in \mathcal{X}^m$. Niech

$$\Pi_{\mathbb{H}}(\vec{\mathbf{x}}) = |\{\langle h(x_1), ..., h(x_m) \rangle \in \{0,1\}^m : h \in \mathbb{H}\}|$$

- $\Pi_{\mathbb{H}}(\vec{\mathbf{x}})$ jest liczbą podziałów zbioru elementów $\vec{\mathbf{x}}$ wyznaczonych przez $\mathbb{H}$. Mamy $\Pi_{\mathbb{H}}(\vec{\mathbf{x}}) \leqslant 2^m$.
- Gdy $\Pi_{\mathbb{H}}(\vec{\mathbf{x}}) = 2^m$, mówimy, że $\mathbb{H}$ rozbija $\mathbf{x}$.



- Niech $\Pi_{\mathbb{H}}(m) = \max_{\vec{\mathbf{x}} \in \mathcal{X}^m} \Pi_{\mathbb{H}}(\vec{\mathbf{x}})$
- **Na przykład:** W przypadku klasy pojęć "półosi"postaci $[\alpha, \infty)$ mamy $\Pi_{\mathbb{H}}(m) = m + 1$.

Uwagi:

- Jeśli $\Pi_{\mathbb{H}}(m) = 2^m$, to istnieje pewien zbiór o mocy m taki, że $\mathbb{H}$ może definiować każdy jego podzbiór ($\mathbb{H}$ rozbija ten zbiór).

Uwagi:

- Jeśli $\Pi_{\mathbb{H}}(m) = 2^m$, to istnieje pewien zbiór o mocy m taki, że $\mathbb{H}$ może definiować każdy jego podzbiór ($\mathbb{H}$ rozbija ten zbiór).
- Maksymalna wartość m, dla której $\Pi_{\mathbb{H}}(m) = 2^m$ można uważać za siłę wyrażalności przestrzeni $\mathbb{H}$

Uwagi:

- Jeśli $\Pi_{\mathbb{H}}(m) = 2^m$, to istnieje pewien zbiór o mocy m taki, że $\mathbb{H}$ może definiować każdy jego podzbiór ($\mathbb{H}$ rozbija ten zbiór).
- Maksymalna wartość m, dla której $\Pi_{\mathbb{H}}(m) = 2^m$ można uważać za siłę wyrażalności przestrzeni $\mathbb{H}$

Definicja: wymiar $VCdim$

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 wymiarów $VCdim$

Przykłady wymiarów $VCdim$

Przykłady wymiarów $VCdim$

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

Przykłady wymiarów $VCdim$

- $H = \{\text{okręgi ... }\} \implies VC(H) = 3$
- $H = \{\text{prostokąty ... }\} \implies VC(H) = 4$

Przykłady wymiarów $VCdim$

- $H = \{\text{okręgi ... }\} \implies VC(H) = 3$
- $H = \{\text{prostokąty ... }\} \implies VC(H) = 4$
- $H = \{\text{funkcje progowe ... }\} \implies$

  $VC(H) = 1$ jeśli "+" są zawsze po prawej stronie;

  $VC(H) = 2$ jeśli "+" mogą być po obu stronach

Przykłady wymiarów $VCdim$

- $H = \{\text{okręgi ... }\} \implies VC(H) = 3$
- $H = \{\text{prostokąty ... }\} \implies VC(H) = 4$
- $H = \{\text{funkcje progowe ... }\} \implies$

  $VC(H) = 1$ jeśli "+" są zawsze po prawej stronie;

  $VC(H) = 2$ jeśli "+" mogą być po obu stronach
- $H = \{\text{przedziały ... }\} \implies$

  VC(H) = 2 jeśli "+" są zawsze w środku

  VC(H) = 3 jeśli w środku mogą być zarówno "+" i "-"

Przykłady wymiarów $VCdim$

- $H = \{\text{okręgi ... }\} \implies VC(H) = 3$
- $H = \{\text{prostokąty ... }\} \implies VC(H) = 4$
- $H = \{\text{funkcje progowe ... }\} \implies$

  $VC(H) = 1$ jeśli "+" są zawsze po prawej stronie;

  $VC(H) = 2$ jeśli "+" mogą być po obu stronach
- $H = \{\text{przedziały ... }\} \implies$

  VC(H) = 2 jeśli "+" są zawsze w środku

  VC(H) = 3 jeśli w środku mogą być zarówno "+" i "-"
- $H = \{$ półpłaszczyzny w $\Re^2$ ... $\}$
  $\implies VC(H) = 3$

Przykłady wymiarów $VCdim$

- $H = \{$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 $VC(H) = \infty$?

Przykłady wymiarów $VCdim$

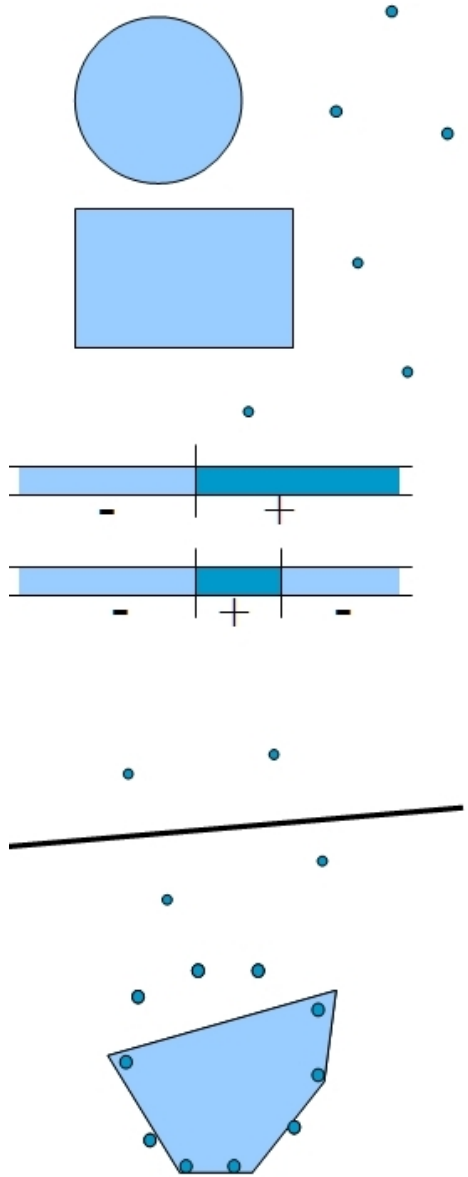
- $H = \{\text{okręgi ... }\} \implies VC(H) = 3$
- $H = \{\text{prostokąty ... }\} \implies VC(H) = 4$
- $H = \{\text{funkcje progowe ... }\} \implies$

  $VC(H) = 1$ jeśli "+" są zawsze po prawej stronie;

  $VC(H) = 2$ jeśli "+" mogą być po obu stronach
- $H = \{\text{przedziały ... }\} \implies$

  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 $VC(H) = \infty$?



Wymiar $VCdim$ dla perceptronu

Twierdzenie

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

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

Dowód:

Wymiar $VCdim$ 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

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

Wymiar $VCdim$ 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.

Wymiar $VCdim$

Twierdzenie

Wymiar $VCdim$

Twierdzenie

1. Jeśli $|\mathbb{H}| < \infty$ to $\boxed{VCdim(\mathbb{H}) \leqslant \log|\mathbb{H}|}$.

Wymiar $VCdim$

Twierdzenie

1. Jeśli $|\mathbb{H}| < \infty$ to $\boxed{VCdim(\mathbb{H}) \leqslant \log|\mathbb{H}|}$.
2. (**Lemat Sauer'a**) Jeśli $VCdim(\mathbb{H}) = d \geq 0$ i $m \geq 1$, to

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

Wymiar $VCdim$

Twierdzenie

1. Jeśli $|\mathbb{H}| < \infty$ to $\boxed{VCdim(\mathbb{H}) \leqslant \log|\mathbb{H}|}$.
2. (**Lemat Sauer'a**) Jeśli $VCdim(\mathbb{H}) = d \geq 0$ i $m \geq 1$, to

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

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

Wymiar $VCdim$

Twierdzenie

1. Jeśli $|\mathbb{H}| < \infty$ to $\boxed{VCdim(\mathbb{H}) \leqslant \log|\mathbb{H}|}$.
2. (**Lemat Sauer'a**) Jeśli $VCdim(\mathbb{H}) = d \geq 0$ i $m \geq 1$, to

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

3. Wniosek: $\Phi(d,m) \leqslant \left(\frac{em}{d}\right)^d \Rightarrow \Pi_{\mathbb{H}}(m) \leqslant \left(\frac{em}{d}\right)^d$
4. Jeśli $|\mathcal{X}| < \infty$, $\mathbb{H} \subset 2^{\mathcal{X}}$ oraz $\mathbb{H} > 1$

$$|\mathcal{X}| < \infty \implies \boxed{VCdim(\mathbb{H}) > \frac{ln|\mathbb{H}|}{1 + ln|\mathcal{X}|}}$$

Plan wykładu

Fundamentalne twierdzenia

Twierdzenie: (Warunek konieczny)

Jeśli przestrzeń hipotez ma *nieskończony* wymiar $VCdim$ to *nie jest potencjalnie wyuczalna*.

Fundamentalne twierdzenia

Twierdzenie: (Warunek konieczny)

Jeśli przestrzeń hipotez ma *nieskończony* wymiar $VCdim$ to *nie jest potencjalnie wyuczalna*.

Twierdzenie: (fundamentalne)

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

Szkic dowodu fundamentalnego twierdzenia

1. Definiujemy

$$Q_m^\varepsilon = \{D \in \mathcal{S}(m,c) : H^c[D] \cap B_\varepsilon \neq \emptyset\}$$

Szkic dowodu fundamentalnego twierdzenia

1. Definiujemy

$$Q_m^\varepsilon = \{D \in \mathcal{S}(m,c) : H^c[D] \cap B_\varepsilon \neq \emptyset\}$$

2. Szukamy górnego ograniczenia $f(m,\varepsilon)$ dla $\mu^m(Q_m^\varepsilon)$, które powinno
   - być niezależne od $c \in \mathbb{C}$ i μ (rozkład).
   - dążyć do 0 przy $m \to \infty$

Szkic dowodu fundamentalnego twierdzenia

1. Definiujemy

$$Q_m^\varepsilon = \{D \in \mathcal{S}(m,c) : H^c[D] \cap B_\varepsilon \neq \emptyset\}$$

2. Szukamy górnego ograniczenia $f(m,\varepsilon)$ dla $\mu^m(Q_m^\varepsilon)$, które powinno
   - być niezależne od $c \in \mathbb{C}$ i μ (rozkład).
   - dążyć do 0 przy $m \to \infty$
3. **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 $m \geqslant 8/\varepsilon$.

Szkic dowodu fundamentalnego twierdzenia

1. Definiujemy

$$Q_m^\varepsilon = \{D \in \mathcal{S}(m,c) : H^c[D] \cap B_\varepsilon \neq \emptyset\}$$

2. Szukamy górnego ograniczenia $f(m,\varepsilon)$ dla $\mu^m(Q_m^\varepsilon)$, które powinno
   - być niezależne od $c \in \mathbb{C}$ i μ (rozkład).
   - dążyć do 0 przy $m \to \infty$
3. **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 $m \geqslant 8/\varepsilon$.
4. Korzystamy z lematu Sauer'a, aby pokazać, że $\mu^m(Q_m^\varepsilon) < \delta$ dla dostatecznie dużych m.

Złożoność zbioru treningowego (c.d.)

- 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.)

- Dla skończonych przestrzeni hipotez $\mathbb{H}$ mamy

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

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

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

Złożoność zbioru treningowego (c.d.)

- Dla skończonych przestrzeni hipotez $\mathbb{H}$ mamy

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

- **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:

Złożoność zbioru treningowego (c.d.)

- Dla skończonych przestrzeni hipotez $\mathbb{H}$ mamy

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

- **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:
  - $m_L(\mathbb{H}, \delta, \varepsilon) \geqslant d(1 - \varepsilon)$

Złożoność zbioru treningowego (c.d.)

- Dla skończonych przestrzeni hipotez $\mathbb{H}$ mamy

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

- **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 $m_L(\mathbb{H}, \delta, \varepsilon) > \frac{d-1}{32\varepsilon}$

Złożoność zbioru treningowego (c.d.)

- Dla skończonych przestrzeni hipotez $\mathbb{H}$ mamy

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

- **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 $m_L(\mathbb{H}, \delta, \varepsilon) > \frac{d-1}{32\varepsilon}$
  - $m_L(\mathbb{H}, \delta, \varepsilon) > \frac{1-\varepsilon}{\varepsilon} \ln \frac{1}{\delta}$

Konkluzje

1. Wyuczalność

Kiedy każdy „wzorowy uczeń” będzie PAC-owy?

Konkluzje

1. Wyuczalność

Kiedy każdy „wzorowy uczeń” będzie PAC-owy?

2. Liczba przykładów

Ile przykładów musi mieć uczeń, by się nauczyć?

Konkluzje

1. Wyuczalność

Kiedy każdy „wzorowy uczeń” będzie PAC-owy?

2. Liczba przykładów

Ile przykładów musi mieć uczeń, by się nauczyć?

Skończoność wymiaru $VCdim()$

1. $VCdim(\mathbb{C}) = d < \infty \Leftrightarrow \mathbb{C}$ jest wyuczalna;
2. Wówczas $L(\frac{1}{\varepsilon}, \frac{1}{\delta}, d) < m(\varepsilon, \delta) < U(\frac{1}{\varepsilon}, \frac{1}{\delta}, d)$

3. Ocena ucznia

$$R(\alpha) = \min_{\alpha \in A} \int Q^c_\Omega(h_\alpha) d\mu$$

na podstawie N losowych przykładów

$$R(\alpha_N) = \min_{\alpha_i \in D} \frac{1}{N} \sum_{i=1}^{N} Q^c(h_{\alpha_i})$$

Kiedy i jak szybko $R(\alpha_N) \to R(\alpha)$?

3. Ocena ucznia

$$R(\alpha) = \min_{\alpha \in A} \int Q^c_\Omega(h_\alpha) d\mu$$

na podstawie N losowych przykładów

$$R(\alpha_N) = \min_{\alpha_i \in D} \frac{1}{N} \sum_{i=1}^{N} Q^c(h_{\alpha_i})$$

Kiedy i jak szybko $R(\alpha_N) \to R(\alpha)$?

Skończoność wymiaru $VCdim()$

3 Dla algorytmów typu ERM, $R(\alpha_N) \to R(\alpha)$ szybko.

$$Prob\{R(\alpha) - R(\alpha_N) > \varepsilon\} < e^{-2c\varepsilon^2 N}$$

Plan wykładu

O co chodzi w NFL?

- Znaleźć optimum nieznanej funkcji $f : S \to W$ ($f \in \mathcal{F}$), gdzie S, W są skończonymi zbiorami.

O co chodzi w NFL?

- Znaleźć optimum nieznanej funkcji $f : S \to W$ ($f \in \mathcal{F}$), gdzie S, W są skończonymi zbiorami.
- Działanie algorytmu przeszukiwania $\mathcal{A}$ dla funkcji f jest identyfikowany z wektorem:

$$V_{\mathcal{A}}(f,t) = \langle (s_1, f(s_1)), (s_2, f(s_2)), ..., (s_t, f(s_t)) \rangle$$

O co chodzi w NFL?

- Znaleźć optimum nieznanej funkcji $f : S \to W$ ($f \in \mathcal{F}$), gdzie S, W są skończonymi zbiorami.
- Działanie algorytmu przeszukiwania $\mathcal{A}$ dla funkcji f jest identyfikowany z wektorem:

$$V_{\mathcal{A}}(f,t) = \langle (s_1, f(s_1)), (s_2, f(s_2)), ..., (s_t, f(s_t)) \rangle$$

- Ocena algorytmu: $M : \{V_{\mathcal{A}}(f,t) | \mathcal{A}, f, t\} \to \mathbb{R}$;
  Np. $M(V_{\mathcal{A}}(f,t)) = \min\{i | f(s_i) = f_{\max}\}$

- **Warunek NFL:** Dla dowolnej funkcji M, i dla dowolnych algorytmów $\mathcal{A}, \mathcal{A}'$

$$\sum_{f \in \mathcal{F}} M(V_{\mathcal{A}}(f, |S|)) = \sum_{f \in \mathcal{F}} M(V_{\mathcal{A}'}(f, |S|))$$

- **Warunek NFL:** Dla dowolnej funkcji M, i dla dowolnych algorytmów $\mathcal{A}, \mathcal{A}'$

$$\sum_{f \in \mathcal{F}} M(V_{\mathcal{A}}(f, |S|)) = \sum_{f \in \mathcal{F}} M(V_{\mathcal{A}'}(f, |S|))$$

- $\mathcal{F}$ **jest zamknięta wzg. permutacji:** dla dowolnej funkcji $f \in \mathcal{F}$ i dowolnej permutacji $\sigma \in Perm(S)$ mamy $\sigma f \in \mathcal{F}$

- **Warunek NFL:** Dla dowolnej funkcji M, i dla dowolnych algorytmów $\mathcal{A}, \mathcal{A}'$

$$\sum_{f \in \mathcal{F}} M(V_{\mathcal{A}}(f, |S|)) = \sum_{f \in \mathcal{F}} M(V_{\mathcal{A}'}(f, |S|))$$

- $\mathcal{F}$ **jest zamknięta wzg. permutacji:** dla dowolnej funkcji $f \in \mathcal{F}$ i dowolnej permutacji $\sigma \in Perm(S)$ mamy $\sigma f \in \mathcal{F}$

Twierdzenie o NFL

- **Warunek NFL:** Dla dowolnej funkcji M, i dla dowolnych algorytmów $\mathcal{A}, \mathcal{A}'$

$$\sum_{f \in \mathcal{F}} M(V_{\mathcal{A}}(f, |S|)) = \sum_{f \in \mathcal{F}} M(V_{\mathcal{A}'}(f, |S|))$$

- $\mathcal{F}$ **jest zamknięta wzg. permutacji:** dla dowolnej funkcji $f \in \mathcal{F}$ i dowolnej permutacji $\sigma \in Perm(S)$ mamy $\sigma f \in \mathcal{F}$

Twierdzenie o NFL

- zachodzi równoważność

$$NFL \Leftrightarrow \mathcal{F} \text{ jest zamknięta wzg. permutacji}$$

- **Warunek NFL:** Dla dowolnej funkcji M, i dla dowolnych algorytmów $\mathcal{A}, \mathcal{A}'$

$$\sum_{f \in \mathcal{F}} M(V_{\mathcal{A}}(f, |S|)) = \sum_{f \in \mathcal{F}} M(V_{\mathcal{A}'}(f, |S|))$$

- $\mathcal{F}$ **jest zamknięta wzg. permutacji:** dla dowolnej funkcji $f \in \mathcal{F}$ i dowolnej permutacji $\sigma \in Perm(S)$ mamy $\sigma f \in \mathcal{F}$

Twierdzenie o NFL

- zachodzi równoważność

$$NFL \Leftrightarrow \mathcal{F} \text{ jest zamknięta wzg. permutacji}$$

- Prawdopodobieństwo wylosowania niepustej klasy funkcji zamkniętej wzg. permutacji wynosi:

$$\frac{2^{\binom{|S|+|W|-1}{|S|}} - 1}{2^{|S|^{|W|}} - 1}$$

The No Free Lunch Theorem for learning

- Algorytm $\mathfrak{L}$ dobrze się uczy pojęcia c jeśli er_{Ω}^{c} jest mały.

The No Free Lunch Theorem for learning

- Algorytm $\mathfrak{L}$ dobrze się uczy pojęcia c jeśli er_{Ω}^{c} 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 $\mathfrak{L}_2$?

The No Free Lunch Theorem for learning

- Algorytm $\mathfrak{L}$ dobrze się uczy pojęcia c jeśli er_{Ω}^{c} 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:

The No Free Lunch Theorem for learning

- Algorytm $\mathfrak{L}$ dobrze się uczy pojęcia c jeśli er_{Ω}^{c} 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.

The No Free Lunch Theorem for learning

- Algorytm $\mathfrak{L}$ dobrze się uczy pojęcia c jeśli er_{Ω}^{c} 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

The No Free Lunch Theorem for learning

- Algorytm $\mathfrak{L}$ dobrze się uczy pojęcia c jeśli er_{Ω}^{c} 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 $\mathbb{C} \subset \mathbb{P}(X)$

The No Free Lunch Theorem for learning

- Algorytm $\mathfrak{L}$ dobrze się uczy pojęcia c jeśli er_{Ω}^{c} 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.