Czego mogą się nauczyć komputery?

Andrzej Skowron, Hung Son Nguyen

son@mimuw.edu.pl; skowron@mimuw.edu.pl

Instytut Matematyki,

Wydział MIM, UW

Spis treści

- Wprowadzenie do uczenia maszynowego;
- Co to znaczy, że komputer się wyuczy jakichś pojęć?
  - Model PAC
  - Efektywne uczenie się
- Jakich klas pojęć jest się w stanie wyuczyć komputer?
  - Wymiar Vapnika-Chervonenkisa
  - Ciekawe oszacowania

Kto się uczy?

Ograniczymy się do programów komputerowych zwanych ”*algorytmami uczącymi się*”. Tak jak zwykli uczniowie, są na nie nałożone pewne rygorystyczne wymagania: np.

- *muszą być skuteczne*,
- *mogą korzystać z ograniczonej pamięci*,
- *mogą używać pewnej specyficznej reprezentacji wiedzy*.

Dziedzina:“czego” się uczy?

Przykłady uczenia się:

- *pojęć*: uczeń próbuje znaleźć reguł odróżniających pozytywne przykłady od negatywnych. (np. pojęcie “krzesło”).
- *nieznanych urządzeń* – np. używanie VCR
- *nieznanych środowisk* – (np. nowe miasto)
- *procesów* (np. pieczenie ciasta)
- *nieznanych rodzin podobnych wzorców* (np. rozpoznawanie mowy, twarzy lub pisma)
- *funkcji*: (np. funkcje boolowskie)

Model uczenia

Każdy “model uczenia” powinien uwzględnić proces uczenia się różnych przedmiotów.

Np. jeśli uczymy się funkcji, to ważne jest aby “algorytm uczenia się” nie ograniczał się do jednej konkretnej funkcji. Żądamy aby “modele uczenia” działały skutecznie na klasach funkcji.

Źródło informacji:

Uczeń może pozyskać informacje o dziedzinie poprzez:

1. **Przykłady:** Uczeń dostaje pozytywne i/lub negatywne przykłady. Przykłady mogą być wybrane
   (a) losowo według pewnego znanego lub nieznanego rozkładu;
   (b) arbitralnie;
   (c) złośliwie (np. przez kontrolera, który chciałby poznać najgorsze zachowanie algorytmu uczenia się);
   (d) specjalnie przez życzliwego nauczyciela (aby ułatwiać proces uczenia się)
2. **Zapytania:**
3. **Eksperymentowanie:** (aktywne uczenie się)

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 $(wzrost(c^m), waga(Kg))$ 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: Znaleźć 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.

Inne przykłady

- **Uczenie półosi (lub dyskretyzacji)**:
  $X = \Re; \mathbb{C} = \mathbb{H} = \{[\lambda, \infty) : \alpha \in \Re\}$
- **Uczenie hiperpłaszczyzny**:

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

  gdzie
  $f_{w_0,...,w_n}(x_1, ..., x_n) = sgn(w_0 + w_1x_1 + ... + w_nx_n)$.
- **Uczenie jednomianów Boolowskich**:
  $X = \{0,1\}^n$; $c : \{0,1\}^n \to \{0,1\}$;
  $\mathbb{H} = M_n$ = zbiór jednomianów Boolowskich o n zmiennych, t.j. zbiór koniunkcji literałów będących albo zmiennymi bądź negacjami 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):

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

Z prawdopodobieństwem $(1 - \varepsilon)$ możemy oszacować er^c_Ω:

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

The No Free Lunch Theorem

Algorytm L dobrze się uczy pojęcia c jeśli er_{Ω}^{c} jest mały.
Niech $\mathbb{C}(X) = \{c : X \to \{0,1\}\}$. Czy można powiedzieć, że L_1 uczy się wszystkich pojęć z $\mathbb{C}(X)$ lepiej od L_2?
”No Free Lunch theorem” (Wolpert, Schaffer) głosi, że:

- Żaden algorytm nie może być najlepszy w uczeniu wszystkich pojęć.
- Każdy algorytm jest najlepszy dla takiej samej liczby pojęć
- Ale interesują nam tylko konkretne problemy czyli klasy pojęć $\mathbb{C} \subset \mathbb{C}(X)$
- Wniosek: Dopasuj algorytm do problemu.

Model uczenia się PAC

Dane: dziedzina X, klasa pojęć $\mathbb{C}$ i przestrzeń hipotez ucznia $\mathbb{H}$.

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

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

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

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

- Celem do wyuczenia przez ucznia jest znalezienie hipotezy minimalizującej błąd rzeczywisty względem c dla rozkładu Ω, czyli er^c_Ω.

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

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

- Celem do wyuczenia przez 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

PAC uczenie się

Definicja: Mówimy, że *prawdopodobnie algorytm uczenia się* L *jest aproksymacyjnie poprawny* $\Leftrightarrow$ 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, jeśli $m > m_0$

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

Wówczas mówimy w skrócie, że L **jest PAC**.
(*Probably Approximately Correct*)

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

Przykład problemu dyskretyzacji

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

Przykład problemu dyskretyzacji

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

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 $D = \{\langle x_1, f_{\lambda_0}(x_1)\rangle, ..., \langle x_m, f_{\lambda_0}(x_m)\rangle\}$

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 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. $L(D) := f_{\lambda^*}$;

Twierdzenie: Powyższy algorytm jest PAC

Przykład (c.d.)

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

Przykład (c.d.)

- $er^c_\Omega = \mu([\lambda_0, \lambda^*))$
- 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 $[\lambda_0, \beta_0]$;

Przykład (c.d.)

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

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

Przykład (c.d.)

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

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

Przykład (c.d.)

- $er^c_\Omega = \mu([\lambda_0, \lambda^*))$
- 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 $x_1, ..., x_n \in X$ – takich, że $\mu_1 + ... + \mu_n = 1$. Niech $\varepsilon_{\min} = \min_i \mu_i$.
- Jeśli 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 L jest dokładnym algorytmem (*jest PEC – probably exactly correct*)

Potencjalna wyuczalność

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

Definicja: 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

Twierdzenie: Jeśli

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

Wówczas L jest PAC

Twierdzenie: (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 $m \geqslant m_0 = \left\lceil \frac{1}{\varepsilon} \ln \frac{|\mathbb{H}|}{\delta} \right\rceil$

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 $\Pi_{\mathbb{H}}(m) = \max\limits_{\mathbf{x} \in X^m} \Pi_{\mathbb{H}}(\mathbf{x})$

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 $\mathbb{H}$

Definicja: 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

- Niech $\mathbb{H}_1$ będzie przestrzenią półosi. Wówczas $VCdim(\mathbb{H}_1) = 1$
- Niech $X = \Re^2$ i $\mathbb{H}$ = zbiór wszystkich półpłaszczyzn. Wówczas $VCdim(\mathbb{H}) = 3$
- Niech $\mathbb{H}$ = zbiór wszystkich półprzestrzeni wyznaczonych przez hiperpłaszczyzny w przestrzeni $\Re^m$. Wówczas $VCdim(\mathbb{H}) = m + 1$
- **Twierdzenie** Jeśli $\mathbb{H}$ jest skończoną przestrzenią, to $VCdim(\mathbb{H}) \leqslant \log |\mathbb{H}|$
- Niech M_n bedzie przestrzenią wszystkich jednomianów Boolowskich o n zmiennych. Ponieważ $|M_n| = 3^n$, mamy

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

Lemat Sauer'a

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

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

- Wnioski

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

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

VC vs. potencjalna wyuczalność

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

Twierdzenie Niech $\mathbb{H}$ będzie przestrzenią hipotez określonych na X. Dla dowolnych c, μ, ε (ale ustalonych) mamy

$$\mu^m\{D : \mathbb{H}[D] \cap \mathbb{B}_\varepsilon \neq \emptyset\} < 2\Pi_{\mathbb{H}}(2m)2^{-\varepsilon m/2}$$

o ile $m \geqslant 8/\varepsilon$.

Twierdzenie (**Fundamentalne twierdzenie**) Jeśli przestrzeń hipotez ma skończony wymiar VC, to jest ona potencjalnie wyuczalna.

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 L jest PAC oraz wymagana liczba przykładów $m_L(\mathbb{H}, \delta, \varepsilon)$ dla 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 L jest PAC oraz wymagana liczba przykładów dla 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 $m_L(\mathbb{H}, \delta, \varepsilon) > \frac{d-1}{32\varepsilon}$
  - $m_L(\mathbb{H}, \delta, \varepsilon) > \frac{1-\varepsilon}{\varepsilon} \ln \frac{1}{\delta}$