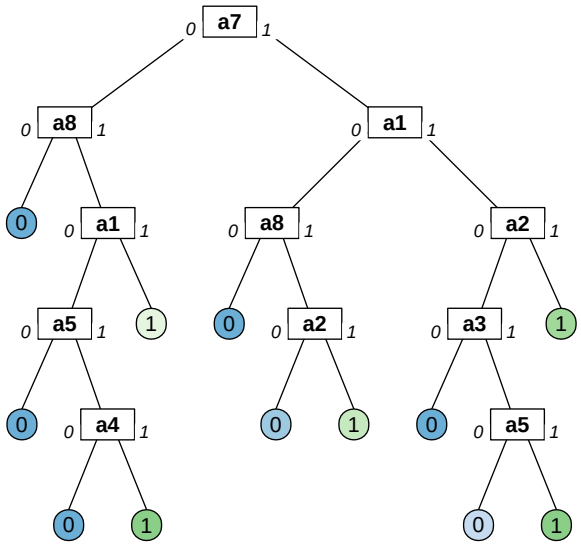
Uczenie maszynowe:
wykład 3

Paweł Cichosz

Przykład: binarne drzewa decyzyjne

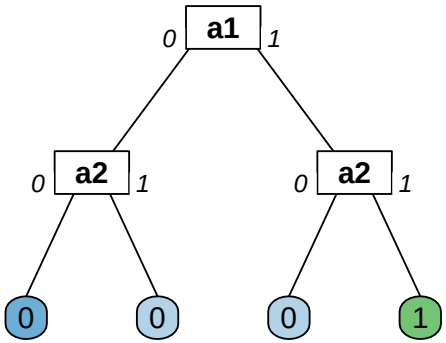


- Bardziej złożona reprezentacja modeli dla dziedziny $\{0,1\}^n$ i zbioru klas $C = \{0,1\}$:

  węzeł: binarny podział według wartości jednego atrybutu,

  liść: klasa.

Przykład: binarne drzewa decyzyjne



- Bez ograniczenia rozmiaru: przestrzeń modeli równoważna przestrzeni dowolnych funkcji boolowskich:

$$|\mathbb{H}| = 2^{2^n}$$

- Przy ograniczeniu do dwóch poziomów węzłów (węzeł w korzeniu, jego dwa węzły potomne, poniżej cztery liście):

$$|\mathbb{H}| = n(n-1)^2 2^4$$

Agnostyczne uczenie się

- Brak pewności, czy $c \in \mathbb{H}$.
- Nie można zagwarantować dowolnie małego błędu, ale można zagwarantować dowolnie małą różnicę między błędem rzeczywistym a błędem na zbiorze trenującym.
- Ryzyko zbyt dużego błędu dla pewnego ustalonego modelu h (na podstawie nierówności Hoeffdinga):

$$P(e_{\Omega,c}(h) > e_{T,c}(h) + \epsilon) \leq e^{-2m\epsilon^2}$$

- Ryzyko zbyt dużego błędu dla któregokolwiek modelu $h \in \mathbb{H}$:

$$P(e_{\Omega,c}(h) > e_{T,c}(h) + \epsilon) \leq |\mathbb{H}|e^{-2m\epsilon^2}$$

Agnostyczne uczenie się

- Ograniczamy przez δ:

$$|\mathbb{H}|e^{-2m\epsilon^2} \leq \delta$$

$$m \geq \frac{1}{2\epsilon^2}(\ln|\mathbb{H}| + \ln\frac{1}{\delta})$$

$$\epsilon \geq \sqrt{\frac{1}{2m}(\ln|\mathbb{H}| + \ln\frac{1}{\delta})}$$

- Stąd z prawdopodobieństwem $1-\delta$:

$$e_{\Omega,c}(h) \leq e_{T,c}(h) + \sqrt{\frac{1}{2m}(\ln|\mathbb{H}| + \ln\frac{1}{\delta})}$$

- Ograniczenie dotyczy to wszystkich modeli.
- Nie ma gwarancji, że algorytm znajdzie najlepszy model.

Wymiar VC

- Dla k przykładów i $C = \{0, 1\}$ istnieje 2^k możliwych etykietowań.
- $\mathrm{VC}(\mathbb{H})$ – maksymalna wartość k taka, że istnieje k przykładów z X, dla których każde spośród 2^k możliwych etykietowań jest realizowane przez pewien model z $\mathbb{H}$.
- $\mathrm{VC}(\mathbb{H}) = \infty$ jeśli dla dowolnego k istnieje k przykładów z X, dla których każde spośród 2^k możliwych etykietowań jest realizowane przez pewien model z $\mathbb{H}$.
- Miara złożoności (pojemności, siły wyrazu) przestrzeni modeli lepsza niż jej rozmiar.
- Maksymalna liczba przykładów, którą na pewno można dokładnie klasyfikować dla dowolnego pojęcia c.
- Jeśli $\mathrm{VC}(\mathbb{H}) = k$, to $2^k \leq |\mathbb{H}|$, czyli $\mathrm{VC}(\mathbb{H}) \leq \log_2 |\mathbb{H}|$.

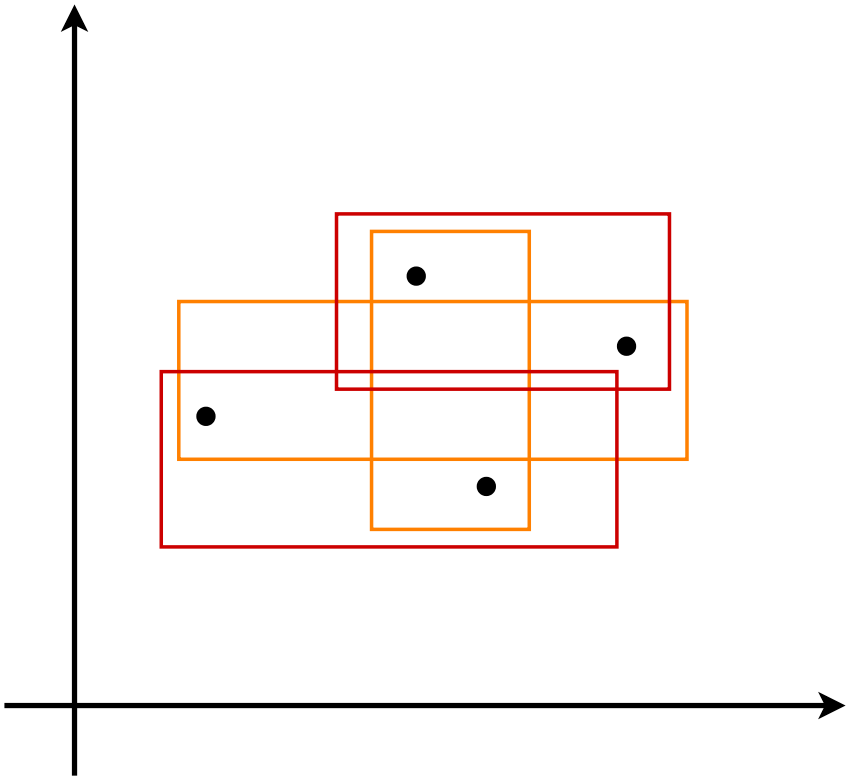
Przykład: prostokąty

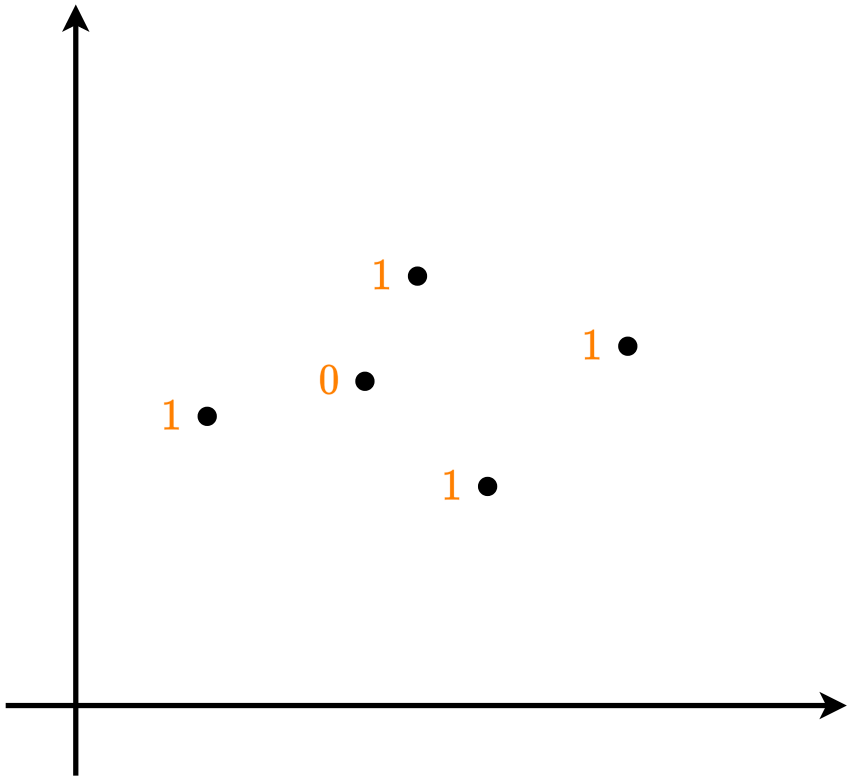
$n = 1$ (przedziały na prostej): $\mathrm{VC}(\mathbb{H}) = 2$.

$n = 2$ (prostokąty na płaszczyźnie): $\mathrm{VC}(\mathbb{H}) = 4$ (łatwo wskazać 4 punkty dla których są możliwe wszystkie etykietowania, ale nie jest to możliwe dla żadnych 5 punktów – nie można nadać innej etykiety czterem punktom leżącym na najmniejszym prostokącie obejmującym wszystkie przykłady a innej piątemu punktowi).

$n > 2$ (hiperprostopadłościany): $\mathrm{VC}(\mathbb{H}) = 2n$?

Przykład: prostokąty





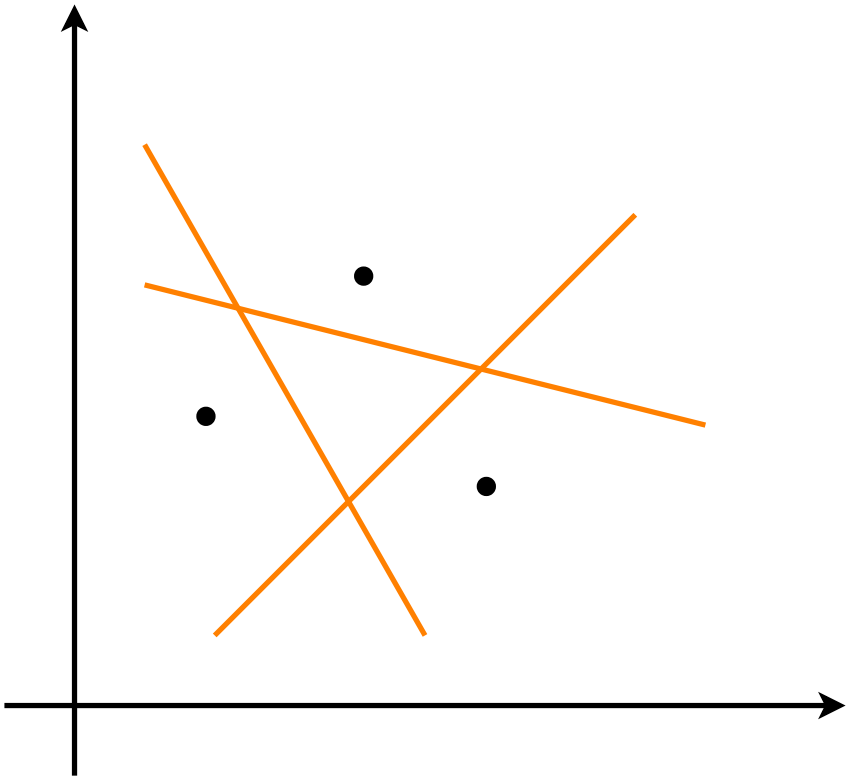
Przykład: proste

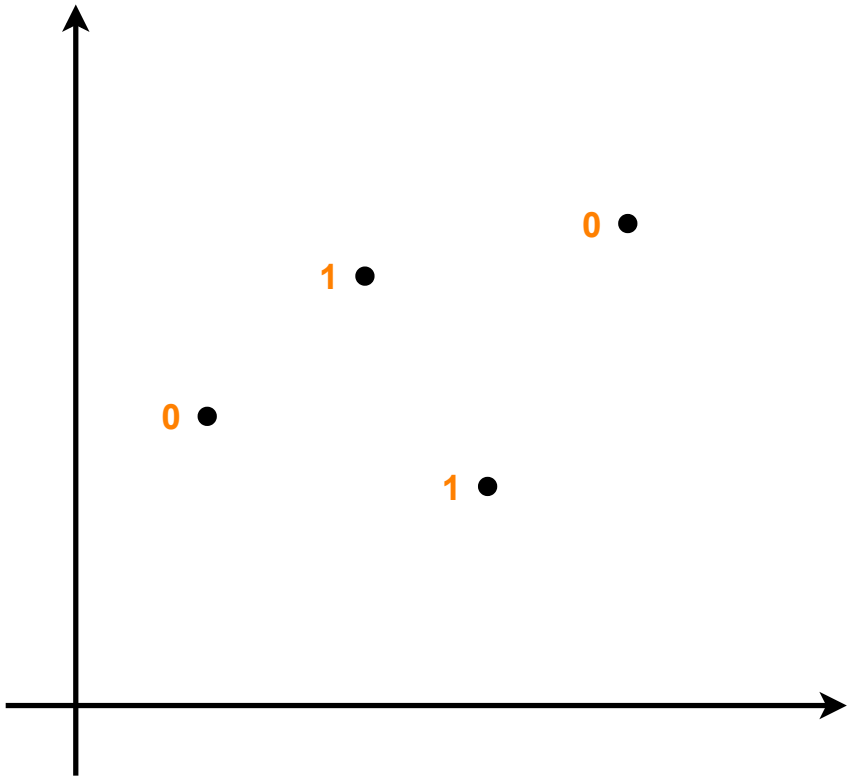
$n = 1$ (podział prostej punktem): $\mathrm{VC}(\mathbb{H}) = 2$.

$n = 2$ (podział płaszczyzny prostą): $\mathrm{VC}(\mathbb{H}) = 3$ (łatwo wskazać 3 punkty dla których są możliwe wszystkie etykietowania, ale nie jest to możliwe dla żadnych 4 punktów – nie można dać innej etykiety punktom leżącym na dwóch różnych przekątnych czworokąta).

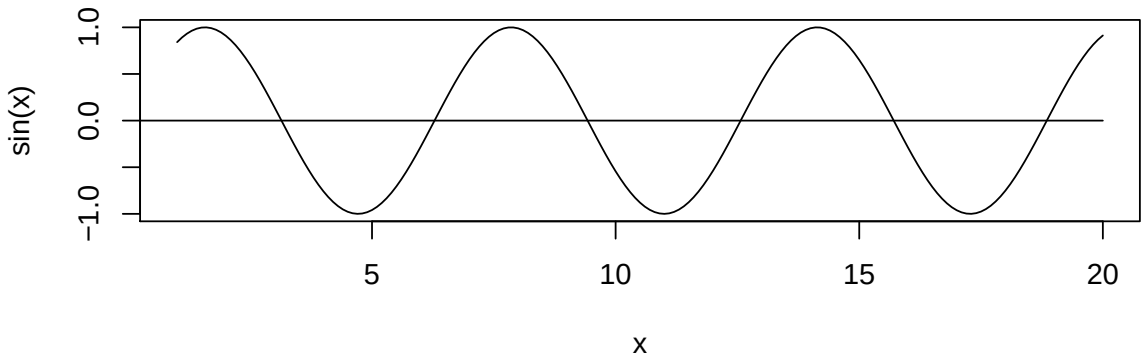
$n > 2$ (podział przestrzeni hiperpłaszczyzną): $\mathrm{VC}(\mathbb{H}) = n + 1$?

Przykład: proste





Przykład: sinus



Dziedzina: $X = \mathcal{R}$, atrybut $a_1(x) = x$.

Przestrzeń modeli: $\mathbb{H}$ zawiera wszystkie modele reprezentowane przez funkcję *sinus*:

$$h(x) = \begin{cases} 1 & \text{jeśli } \sin(\alpha x) \geq 0 \\ 0 & \text{w przeciwnym przypadku} \end{cases}$$

dla dowolnie ustalonego parametru α.

Wymiar VC: dla dowolnego k przykłady $\frac{1}{2}, \frac{1}{4}, \ldots, \frac{1}{2^k}$ mogą być dowolnie etykietowane modelami z $\mathbb{H}$ – a więc $\mathrm{VC}(\mathbb{H}) = \infty$.

Przykład: koniunkcje

a_1	a_2	a_3
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

$n = 3$: $\mathrm{VC}(\mathbb{H}) \geq 3$ (wszystkie etykietowania możliwe dla 100, 010, 001).

$n > 3$: $\mathrm{VC}(\mathbb{H}) = n$?

Zastosowanie wymiaru VC

Spójne uczenie się: do uzyskania przez spójny algorytm uczenia się z prawdopodobieństwem co najmniej $1 - \delta$ modelu o błędzie rzeczywistym nieprzekraczającym ϵ wystarczy:

$$m \geq \frac{1}{\epsilon}\left(4 \log_2 \frac{2}{\delta} + 8\mathrm{VC}(\mathbb{H}) \log_2 \frac{13}{\epsilon}\right)$$

przykładów trenujących.

Agnostyczne uczenie się: z prawdopodobieństwem $1 - \delta$:

$$e_{\Omega,c}(h) \leq e_{T,c}(h) + \sqrt{\frac{1}{m}\left(\mathrm{VC}(\mathbb{H})\left(\ln \frac{2m}{\mathrm{VC}(\mathbb{H})} + 1\right) + \ln \frac{4}{\delta}\right)}$$

Podsumowanie wniosków z teorii

- Zbyt bogata przestrzeń modeli – duże $|\mathbb{H}|$ lub duże $\mathrm{VC}(\mathbb{H})$ – zwiększa ryzyko nadmiernego dopasowania (potrzeba więcej przykładów trenujących, aby uzyskać z wystarczającą pewnością wystarczająco mały błąd).
- Zbyt uboga przestrzeń modeli zwiększa ryzyko niedopasowania – zmniejsza szansę, że istnieje model o wystarczająco małym błędzie).
- Konieczna równowaga.
- W praktyce spójne uczenie się zwykle nie jest pożądane, gdyż nie ma dostępu do wystarczająco wielu przykładów trenujących, a często nie jest także możliwe, gdyż używana przestrzeń modeli nie zawiera modelu o zerowym błędzie na zbiorze trenującym (np. dlatego, że zestaw dostępnych atrybutów nie wystarcza do odróżniania przykładów różnych klas).
- Praktyczne algorytmy mogą używać pojemnych przestrzeni modeli, lecz zmniejszać efektywny wymiar VC przez dodatkowe mechanizmy ograniczające nadmierne dopasowanie.
- Brzytwa Ockhama – preferencja dla prostych modeli.