Uczenie maszynowe: *wykład 2*

Paweł Cichosz

Estymacja błędu rzeczywistego

- Błąd na zbiorze $e_{S,c}(h)$ jest wiarygodnym estymatorem błędu rzeczywistego $e_{\Omega,c}(h)$, jeśli:
  - S jest wystarczająco duży,
  - S jest wybrany z dziedziny niezależnie od h,
  - S jest wybrany z dziedziny zgodnie z rozkładem Ω.
- W praktyce: $S \subset D$, $T \subset D$, gdzie $D \subset X$ jest zbiorem przykładów o znanych klasach, $S \cap T = \emptyset$.
- Można użyć zastępczo estymatora $e_{S,c}(h)$ zamiast nieznanego $e_{\Omega,c}(h)$ do wykrywania nadmiernego dopasowania.
- Uproszczone praktyczne kryterium ryzyka nadmiernego dopasowania: $e_{S,c}(h) \gg e_{T,c}(h)$.

Estymacja przedziałowa błędu

Przedział ufności: Przedział wokół $e_{S,c}(h)$ który prawdopodobieństwem $1-\delta$ zawiera $e_{\Omega,c}(h)$ (gdzie δ – poziom ufności).

Przedział Walda: (tylko jeśli błąd $e_{S,c}(h)$ nie jest zbyt bliski 0 ani 1)

$$|e_{\Omega,c}(h) - e_{S,c}(h)| < u_\delta \sqrt{\frac{e_{S,c}(h)(1-e_{S,c}(h))}{|S|}}$$

z prawdopodobieństwem $1-\delta$, gdzie dla $U \sim N(0,1)$:

$$P(|U| < u_\delta) = 1-\delta$$

Przedział Wilsona: (także jeśli błąd $e_{S,c}(h)$ bliski 0 lub 1)

$$\left| e_{\Omega,c}(h) - \frac{e_{S,c}(h) + \frac{u_\delta^2}{2|S|}}{1+\frac{u_\delta^2}{|S|}} \right| < \frac{u_\delta}{1+\frac{u_\delta^2}{|S|}} \sqrt{\frac{e_{S,c}(h)(1-e_{S,c}(h))}{|S|} + \frac{u_\delta^2}{4|S|^2}}$$

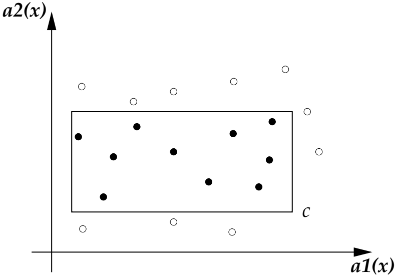
z prawdopodobieństwem $1-\delta$, gdzie u_δ – jak wyżej.

Probably Approximately Correct

Klasa pojęć $\mathbb{C}$ dla dziedzinie X jest PAC-nauczalna za pomocą przestrzeni modeli $\mathbb{H}$, jeśli

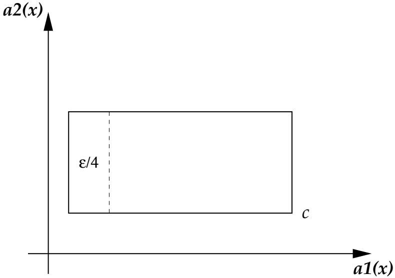
- istnieje algorytm uczenia się używający $\mathbb{H}$,
- którego uruchomienie z dostępem do źródła przykładów $\mathrm{EX}(\Omega, c)$ oraz z parametrami ϵ i δ,
- daje w wyniku z prawdopodobieństwem co najmniej $1 - \delta$ model $h \in \mathbb{H}$, dla którego $e_{\Omega,c}(h) \leq \epsilon$,
- dla dowolnego pojęcia $c \in \mathbb{C}$, dowolnego rozkładu prawdopodobieństwa Ω na X oraz dowolnych $0 < \epsilon < 1$ i $0 < \delta < 1$.

Przykład: prostokąty



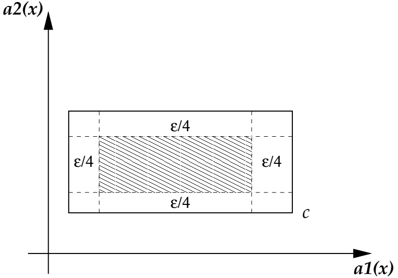
- Algorytm najciaśniejszego dopasowania (model – najmniejszy prostokąt zawierający wszystkie przykłady pozytywne).
- R_c, R_h – prostokąty reprezentujące pojęcie c, model h (oczywiście $R_h \subseteq R_c$).
- Wystarczy rozważyć przypadek c takiego, że $P(R_c) > \epsilon$ (w przeciwnym przypadku błąd nie może przekroczyć ϵ).

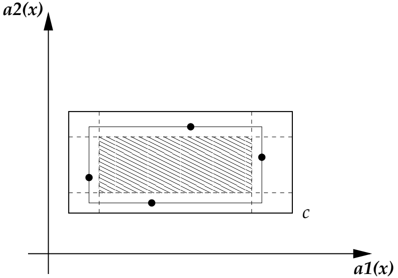
Przykład: prostokąty



- Odcinamy z boku R_c margines o prawdopodobieństwie $\frac{\epsilon}{4}$.

Przykład: prostokąty





- Powtarzając to dla każdego boku uzyskujemy „ramkę” o prawdopodobieństwie poniżej ϵ.
- Model ma błąd poniżej ϵ, jeśli w każdym marginesie $\frac{\epsilon}{4}$ znajduje się przykład trenujący (warunek wystarczający).

Przykład: prostokąty

1. Model ma błąd poniżej ϵ, jeśli w każdym marginesie $\frac{\epsilon}{4}$ znajduje się przykład trenujący (warunek wystarczający).
2. Prawdopodobieństwo sytuacji przeciwnej można ograniczyć przez:

$$4(1-\frac{\epsilon}{4})^m \leq 4e^{-m\frac{\epsilon}{4}} \leq \delta$$

   gdzie m jest liczbą przykładów trenujących (w ostatnim kroku wykorzystana nierówność $1+\alpha \leq e^\alpha$ dla dowolnego α).
3. Wystarczy odpowiednio wiele przykładów:

$$m \geq \frac{4}{\epsilon}\left(\ln 4 + \ln \frac{1}{\delta}\right)$$

4. Można również określić, jaki poziom błędu rzeczywistego da się zagwarantować przy ustalonej liczbie przykładów:

$$\epsilon \geq \frac{4}{m}\left(\ln 4 + \ln \frac{1}{\delta}\right)$$

Spójne uczenie się

Spójny model: zerowy błąd na zbiorze trenującym.

Spójny algorytm uczenia się: zwraca spójny model albo zawodzi, jeśli takiego modelu nie ma w przestrzeni modeli $\mathbb{H}$.

Niezawodność spójnego uczenia się: można zagwarantować tylko, jeśli $\mathbb{C} \subseteq \mathbb{H}$.

Przestrzeń wersji: zbiór wszystkich spójnych modeli:

$$\mathrm{VS}_{\mathbb{H},T}(c) = \{h \in \mathbb{H} \mid e_{T,c}(h) = 0\}$$

Przykładowe algorytmy spójne

Prostokąty: algorytm znajdujący najmniejszy prostokąt zawierający wszystkie przykłady pozytywne.

Koniunkcje boolowskie: algorytm usuwający z początkowej koniunkcji $a_1 \wedge \neg a_1 \wedge a_2 \wedge \neg a_2 \wedge \cdots \wedge a_n \wedge \neg a_n$ wszystkie literały, które nie są spełnione dla któregokolwiek przykładu pozytywnego.

Błąd rzeczywisty spójnych modeli

1. Ponieważ algorytm spójny może zwrócić dowolny model spójny, więc potrzebujemy ograniczenia błędu rzeczywistego dowolnego takiego modelu.
2. Przestrzeń wersji jest ϵ-wyczerpana jeśli błąd rzeczywisty wszystkich należących do niej modeli nie przekracza ϵ.
3. Prawdopodobieństwo, że pewien model o błędzie rzeczywistym powyżej ϵ należy do przestrzeni wersji nie przekracza:

$$(1-\epsilon)^m \leq e^{-m\epsilon}$$

gdzie m jest liczbą przykładów trenujących.
4. Dla skończonej przestrzeni modeli prawdopodobieństwo, że którykolwiek model z tej przestrzeni o błędzie rzeczywistym powyżej ϵ należy do przestrzeni wersji, nie przekracza:

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

Błąd rzeczywisty spójnych modeli

1. Zatem dla dowolnego spójnego modelu h:
$$P(e_{\Omega,c}(h) > \epsilon) \leq |\mathbb{H}|e^{-m\epsilon}$$
2. Ograniczamy przez δ:
$$|\mathbb{H}|e^{-\epsilon m} \leq \delta$$
$$m \geq \frac{1}{\epsilon}(\ln \| \mathbb{H}| + \ln \frac{1}{\delta}) \qquad \epsilon \geq \frac{1}{m}(\ln \| \mathbb{H}| + \ln \frac{1}{\delta})$$
3. Stąd z prawdopodobieństwem $1 - \delta$:
$$e_{\Omega,c}(h) \leq \frac{1}{m}(\ln \| \mathbb{H}| + \ln \frac{1}{\delta})$$
4. Algorytmy spójne używające skończonej przestrzeni modeli mogą osiągnąć dowolnie mały błąd, mimo podatności na nadmierne dopasowanie (wystarczy odpowiednio wiele przykładów).
5. Dla ustalonej liczby przykładów można określić, jaki poziom błędu rzeczywistego może być probabilistycznie zagwarantowany.

Przykłady wyznaczania $|\mathbb{H}|$

Koniunkcje boolowskie:

- dla każdego atrybutu koniunkcja może zawierać literał pozytywny, negatywny lub nie zawierać go wcale,
- dodatkowo należy uwzględnić koniunkcję „zerową" (stale niespełnioną):

$$|\mathbb{H}| = 3^n + 1$$

Dowolne funkcje boolowskie: wszystkie możliwe sposoby etykietowania wszystkich możliwych 2^n przykładów z dziedziny:

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