Univerzita Pavla Jozefa Šafárika v Košiciach,
Prírodovedecká fakulta

Strojové učenie

Booleovské formuly a reprezentácie

Gabriela Andrejková, Ľubomír Antoni

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Ako chápeme pojem učenie sa (learning)

Učenie sa je proces, v ktorom sa doplňajú už naučené znalosti a získavajú nové znalosti.
Zaradenie predchádzajúcich znalostí ovplyvňuje proces učenia a je nevyhnutné pre úspech postupov (algoritmov) učenia sa. Zhruba povedané:

- čím silnejšie sú predchádzajúce vedomosti (alebo predchádzajúce predpoklady), s ktorými sa začína proces učenia, tým ľahšie sa učí z ďalších príkladov,
- čím silnejšie sú predchádzajúce predpoklady, tým je učenie menej flexibilné - je a priori viazané k týmto predpokladom.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Opakovanie

- Konceptový priestor, C - domáce zvieratá (DZ)
- Abeceda Σ - popis prvkov patriacich do konceptu, $\{0, 1\}$
- Atribúty konceptu - 4 nohy, rohy, nos, uši, oči, ústa, chvost, žerie, pije; počet atribútov n;
- Príkladový priestor - koza=(1,0,1,1,1,1,0,1,1), pes=((1,0,1,1,0,1,0,1,1);
- Tréningová vzorka $\bar{s}$ - ((1,0,1,1,1,1,0,1,1),1), ((1,0,1,1,1,1,0,0,0),0), ...; dĺžka tréningovej vzorky m;
- Hypotézový priestor, H - napríklad Booleovské funkcie

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Opakovanie

- Učiaci algoritmus pre (C, H) - akceptuje tréningové vzorky pre koncepty v C a výstupmi sú hypotézy v H.
- Hypotéza $h \in H$ je konzistentná s tréningovou vzorkou $\overline{s}$ alebo súhlasí s $\overline{s}$, ak $h(x_i) = b_i$, pre $1 \leq i \leq m$.
- Učiaci algoritmus L je konzistentný - ak každá hypotéza $L(\overline{s})$ je konzistentná s tréningovou vzorkou $\overline{s}$
- Učenie monočlenov (monomials)

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Učenie disjunkcií malých monočlenov, Valiant, 1984

Konjunktívna a disjunktívna normálna forma - KNF a DNF
Disjunktívna normálna forma (DNF)

$$\mu_1 \vee \mu_2 \vee \cdots \vee \mu_n$$

kde μ_i je monočlen, $1 \leq i \leq n$.
Konjuktívna normálna forma (KNF)

$$\gamma_1 \wedge \gamma_2 \wedge \cdots \wedge \gamma_n$$

kde γ_i je, $1 \leq i \leq n$ je klauzula, t.j. disjunkcia literálov.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Označenie:

- M_n-množina monočlenov nad $\{0,1\}^n$,
- $M_{n,k}$-množina monočlenov nad $\{0,1\}^n$, z ktorých každý má najviac k literálov,
- $D_{n,k}$-množina disjunkcií členov z množiny $M_{n,k}$.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Algoritmus

begin
  h:=disjunkcia všetkých monočlenov dlžky najviac k;
  for i:=1 to m do
    if $(b_i = 0) \wedge (h(x_i) = 1)$ then
      vymazať monočleny μ, pre ktoré $\mu(x_i) = 1$;
  $L(\overline{s}) := h$;
end;

Toto učenie má zjednodušené predpoklady, a síce že konceptový priestor je ten istý ako hypotézový. V skutočnosti sme uvažovali o tom, že cieľové koncepty majú nejaký popis pomocou formúl. Hoci tento predpoklad sa môže zdať reštriktívny, je prirodzený z matematického hľadiska.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Príklad:

Pracujeme s $D_{3,2}$
Množina relevantných monočlenov je:

$$\{u_1, u_2, u_3, \overline{u}_1, \overline{u}_2, \overline{u}_3, u_1 u_2, u_1 u_3, u_2 u_3, \overline{u}_1 u_2, \overline{u}_1 u_3,$$
$$\overline{u}_2 u_3, u_1 \overline{u}_2, u_1 \overline{u}_3, u_2 \overline{u}_3, \overline{u}_1 \overline{u}_2, \overline{u}_1 \overline{u}_3, \overline{u}_2 \overline{u}_3\}$$

Máme 5 kladných príkladov a 3 negatívne: 101, 111 a 100.
Cieľový koncept je $\langle \overline{u}_1 \vee u_2 \overline{u}_3 \rangle$

Aplikovaním algoritmu dostaneme hypotézu

$$h = \langle \overline{u}_1 \vee \overline{u}_1 u_2 \vee \overline{u}_1 u_3 \vee u_2 \overline{u}_3 \vee \overline{u}_1 \overline{u}_2 \vee \overline{u}_1 \overline{u}_3 \rangle$$

Po úpravách: $h = \langle \overline{u}_1 \vee u_2 \overline{u}_3 \rangle$, teda algoritmus dal iné vyjadrenie tejto funkcie.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Učenie pomocou konštrukcie:

X je príkladový priestor, X^+, $X^+ \subseteq X$ množina kladných príkladov, t cieľový koncept, m je dĺžka tréningovej vzorky,

Metóda učenia hypotézy h: skonštruovať množinu X^+ explicitne. Môžme začať s prázdnou množinou kladných príkladov a potom prechodom cez tréningovú vzorku pridať každý pozitívny príklad. Formálne to môžeme vyjadriť (príklady sú vhodne uložené),

set h(x) = 0 for all x in X;
for i: = 1 to m do if $b_i = 1$ **then set h(**x_i**) = 1;**
$L(\overline{s}) = h$;

Hypotéza sa naučí všetky príklady tréningovej vzorky správne.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Učenie očíslovaním:

Predpokladáme, že hypotézový priestor H je spočítateľný a má explicitné očíslovanie hypotéz, $H = \{h^{(1)}, h^{(2)}, \ldots\}$

Predpokladajme, že $\overline{s}$ je tréningová vzorka dĺžky m pre cieľový koncept t.

Metóda: Vyhodnotiť hypotézu pre každý príklad v $\overline{s}$, odmietnuť hypotézu vždy, keď jej ohodnotenie niektorého príkladu nesúhlasí s jeho daným ohodnotením v tréningovej vzorke.
Po odmietnutí hypotézy je testovaná ďalšia hypotéza tým istým spôsobom. Proces sa zastaví, keď je nájdená hypotéza, ktorá vyhovuje všetkým príkladom tréningovej vzorky.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Učenie očíslovaním

r - poradové číslo hypotézy, i - poradové číslo príkladu vo vzorke

```
begin
   r: = 1, i: = 1;
   repeat
      if h^(r)(x_i) <> b_i then
         r: = r+ 1; i : = 1
      else i: = i + 1;
   until i = m + 1;
   L (s̄) : = h^(r);
end;
```

Množina hypotéz H môže byť konečná, a teda môže sa stať, že sa vhodná hypotéza nenájde. Algoritmus vieme ľahko modifikovať.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Poznámky

V praxi sa musíme vyhnúť používaniu neprimerane veľkých hypotézových priestorov.
Napríklad, počet všetkých hypotéz (Booleovských funkcií) $h : \{0,1\}^n \to \{0,1\}$ je 2^{2^n}. Ak $n = 10, 2^{2^n} = 2^{1024} = 8^{256} = 16^{128}$

Na to, aby sa táto metóda stala vhodnou metódou učenia, je potrebné urobiť určité obmedzenia na hypotézový priestor H a jeho vzťah k priestoru konceptov C.
Toto vedie k pojmu **"induktive bias (nastavenie)"**.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Poznámky

Je to predpoklad, že učiaci sa má nejakú vopred danú ideu o tom, akú metódu učenia (klasifikácie) učiteľ používa, t.j. učiaci vie, alebo má nejaké informácie o konceptovom priestore.

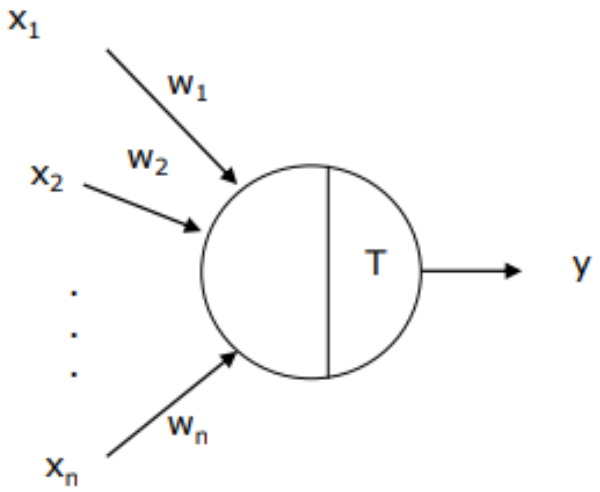
Najjednoduchší spôsob modelovať taký predpoklad je stanoviť $H = C$. Napríklad, pripraviť dáta pre Booleovské funkcie.

V tomto prípade hovoríme o učiacom algoritme pre H, čo znamená (H, H). Niekoľko preberaných algoritmov v ďalšom bude tohto typu.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Lineárne prahové jednotky (LPJ)

Algoritmus, ktorý predpokladá lineárnu závislosť od vstupov.



$$y = \begin{cases} 1 & \text{ak} \\ 0 & \text{inak} \end{cases} \qquad \mathbf{w}^T\mathbf{x} = \sum_{i=1}^{n} w_i x_i \geq T$$

kde

$\mathbf{x}=(x_1,x_2,\ldots,x_n)^T$	sú vstupné hodnoty
$\mathbf{w}=(w_1,w_2,\ldots,w_n)$	hodnoty váh
$\boldsymbol{T}$	je prahová hodnota.

Obr.: Tvar a funkcia lineárnej prahovej jednotky

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

Počet realizovateľných BF pomocou LPJ

n	BF_n pomocou LPJ	Počet BF celkom
1	4	4
2	14	16
3	104	256
4	1882	65536
5	94572	$4.3 * 10^9$
6	15028134	$1.8 * 10^{19}$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Úlohy:

1. Napíšte *DNF* formuly pre koncepty *parita* a *palindrom* na príkladových priestoroch $\{0,1\}^5$.
2. Uveďte príklad bool. funkcie 3 premenných, ktorá nie je $D_{3,2}$.
3. Diskusia o prahových jednotkách. Nájdite prahovú jednotku pre tréningovú vzorku: (11,1), (10,0), (01,0), (00,0) a pre tréningovú vzorku (11,1), (10,0), (01,0), (00,1)
4. Nech $C_n = C_n^1$ je priestor booleovských funkcií na $\{0,1\}^n$, ktoré môžu byť reprezentované jedinou klauzulou. Sformulujte konzistentný učiaci algoritmus pre C_n, ktorý je duálny ku štandardnému algoritmu pre monočleny a vyhodnoťte jeho časovú zložitosť.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Čas behu učiacich algoritmov

Príkladový priestor je $\{0,1\}^n$ pre nejaké pevné n a hypotézový priestor je množina funkcií definovaných na príkladovom priestore.

Pre každý z týchto algoritmov parameter n je ľubovoľný v zmysle, že algoritmus je definovaný pre ľubovoľné n a navyše operuje v podstate tým istým spôsobom pre každú hodnotu n.

Napríklad, štandartný učiaci algoritmus pre priestor M_n monočlenov je definovaný celkom všeobecne, hoci výpočet vždy prebieha pre danú konkrétnu hodnotu n.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Čas behu učiacich algoritmov

Chceme kvantifikovať chovanie sa učiacich algoritmov vzhľadom na n, a je obvykle používať nasledujúce definície.

Definícia:

Hovoríme, že zjednotenie hypotézových priestorov $H = \bigcup H_n$ je odstupňované (graded) príkladmi veľkosti n, ak H_n označuje hypotézový priestor definovaný len na príkladoch z X^n.

Definícia:

Nech je daný učiaci algoritmus pre $H = \bigcup H_n$, t. j. $L : X^* \to H$, taký, že ak $\overline{s}$ je tréningová vzorka pre $h \in H_n$, tak $L(\overline{s})) \in H_n$. Hovoríme, že L podporuje odstupňovanie (grading) priestoru H.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Čas behu učiacich algoritmov

Uvažujme učiaci algoritmus L pre booleovský hypotézový priestor $H = \bigcup H_n$, odstupňovaný veľkosťou príkladov.

Vstup do L je tréningová vzorka, ktorá pozostáva z m n-bitových vektorov spolu s m 1-bitovými označeniami.
Celkový počet bitov na vstupe je $m(n+1)$ a bolo by možné použiť toto jediné číslo ako mieru veľkosti vstupu. Avšak je výhodné sledovať aj m aj n oddelene.

Použijeme $R_L(m, n)$ na označenie **najhoršieho času behu algoritmu L na tréningovej vzorke m n-bitových príkladov.**

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Príklad:

Nech L je učiaci algoritmus pre monočleny popísaný vyššie. Hypotézový priestor je zjednotenie $\bigcup M_n$. *Hlavný krok algoritmu vyžaduje kontrolu každého bitu u každého pozitívneho príkladu a možno vymazanie niektorých literálov.*

V najhoršom prípade, každý príklad v tréningovej vzorke môže byť pozitívny príklad, a tak by sme mali ošetriť tento krok m-krát, každý krok obsahuje kontrolu n bitov. Iné časti výpočtu vyžadujú porovnateľne toľko operácií, takže môžme hovoriť, že čas behu $R_L(m, n)$ *je v tomto prípade* $O(m * n)$.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Konzistencia tréningovej vzorky

Nech $H = \bigcup H_n$ je hypotézový priestor booleovských funkcií odstupňovaný príkladmi veľkosti n. **Problém konzistencie tréningovej vzorky pre H** môže byť vyjadrený nasledovne:

H - KONZISTENCIA:

- **Inštancia:** Tréningová vzorka $\overline{s}$ vyjadrená n-bitovými ohodnotenými príkladmi.
- **Otázka:** Existuje hypotéza v H_n konzistentná s $\overline{s}$?

Ukážeme, že v niektorých netriviálnych prípadoch (podmnožinách H) je tento problém NP-ťažký.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Konzistencia tréningovej vzorky

Označme C_n^k konjunkciu k klauzúl (disjunkcií) vytvorených výberom literálov z n premenných.

Ukážeme, že pre pevné $k \geq 3$, problém konzistencie pre $C^k = \bigcup C_n^k$ je NP-ťažký. Teda nie je pravda, že existuje polynomiálny učiaci algoritmus pre C^k, ktorý produkuje konzistentnú hypotézu.

Dôkaz spočíva v prevedení problému farbenia grafu s n vrcholmi pomocou k farieb na tento problém, čo je NP-úplný problém pre $k \geq 3$.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

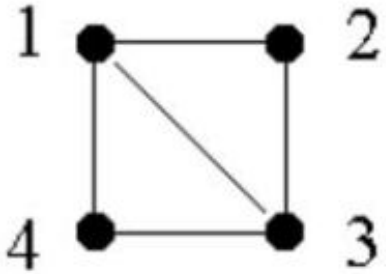
Konzistencia tréningovej vzorky

Problém farbenia grafu: $G = (V, E)$, k-farbenie je funkcia $\chi : V \to \{1, 2, \ldots, k\}$ s vlastnosťou: ak $\langle v_i, v_j \rangle \in E$, potom $\chi(v_i) \neq \chi(v_j)$.

Predpokladajme, že máme $G = (V, E)$, $V = \{1, 2, \ldots k\}$.
Skonštruujeme tréningovú vzorku $s(G)$ nasledovne:
Pre každý vrchol $i \in V$ vytvoríme záporný príklad vektor v_i, ktorý má 1 v pozícii i-tej súradnice a 0 inde.
Pre každú hranu $\langle i, j \rangle \in E$ vezmeme ako pozitívny príklad vektor $v_i + v_j$.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Príklad:



	v_i
1000	0
0100	0
0010	0
0001	0
1100	1
1010	1
1001	1
0110	1
0011	1

Obr.: *Graf spolu s pripravenou tréningovou vzorkou s(G)*

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Veta:

Existuje funkcia $h \in C_n^k$, ktorá je konzistentná so vzorkou $s(G) \iff$ graf G je k-zafarbiteľný.

$\Rightarrow$

Predpokladajme, že $h \in C_n^k$ a je konzistentná s tréningovou vzorkou. Podľa definície h je konjunkcia $h = h_1 \wedge h_2 \wedge \ldots h_k$ klauzúl. Pre každý vrchol $i \in V$, $h(v_i) = 0$ a teda musí ex. aspoň 1 klauzula h_f, pre ktorú $h_f(v_i) = 0$. Funkcia $\chi : V \to \{1, 2, \ldots k\}$ taká, že:

$$\chi(i) = min\{f \quad | \quad h_f(v_i) = 0\}$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Konzistencia tréningovej vzorky

Zostáva ukázať, že χ je farbenie grafu G; inak povedané, ak i a j sú dva vrcholy, pre ktoré $\chi(i) = \chi(j)$, potom $\langle i, j \rangle \notin E$.

Predp., že $\chi(i) = \chi(j) = f$ a teda $h_f(v_i) = h_f(v_j) = 0$. Pretože h_f je klauzula, každý literál, ktorý sa v nej vyskytuje musí byť 0 pri aplikácii na v_i a aj na v_j. Teraz v_i je príklad, ktorý má 1 len v i-tej pozícii, a teda $h_f(v_i) = 0$ implikuje, že len negovaný literál $\overline{u}_i$ sa môže vyskytnúť v h_f. Pretože to isté platí pre $\overline{u}_j$, dostávame, že h_f obsahuje len niektoré literály u_z, pre ktoré $z \neq i, j$. Teda $h_f(v_i + v_j) = 0$ a $h(v_i + v_j) = 0$. Ak by $\langle i, j \rangle$ bola hranou v G, potom by platilo $h(v_i + v_j) = 1$, pretože sme predpokladali, že h je konzistentná s $s(G)$. Teda $\langle i, j \rangle$ nie je hranou v G a χ je farbenie.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

$\Leftarrow$

Predpokladajme, že je dané farbenie $\chi : V \to \{1, 2, \ldots, k\}$. Pre $1 \leq f \leq k$ definujme h_f ako klauzulu $\langle \vee u_{i_{\chi(i) \neq f}} \rangle$ a definujme $h = h_1 \wedge h_2 \wedge \cdots \wedge h_k$. Tvrdíme, že h je konzistentná s $s(G)$.

Najprv, predpokladajme, že pre vrchol i $\chi(i) = g$. Klauzula h_g je definovaná tak, že obsahuje len tie (nie negované) literály, ktoré zodpovedajú vrcholom nezafarbeným farbou g, a teda u_i sa nenachádza v h_g. Teda $h_g(v_i) = 0$ a $h(v_i) = 0$.
Nech $\langle i, j \rangle$ je hrana v G. Pre každú farbu f aspoň jeden v_i alebo v_j nemá farbu f; označme vhodný výber $i(f)$. Potom h_f obsahuje literál $u_{i(f)}$, ktorý je 1 na $v_i + v_j$. Teda klauzula h_f je 1 na $v_i + v_j$ a $h(v_i + v_j) = 1$, ako sme požadovali.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Poznámky:

- Graf s n vrcholmi má najviac $n * (n-1)/2$ hrán, a teda dĺžka tréningovej vzorky je $n + n * (n-1)/2$
- Ak je dĺžka tréningovej vzorky polynomiálne ohraničená v n, potom sa jedná o **ohraničený** problém.
- Ukázali sme, že ohraničený tvar problému – C^k-KONSISTENCIA (k-FARBENIA GRAFU), v ktorom m je $O(n^2)$, je NP–ťažký.
- Ak je ohraničený tvar – C^k-KONSISTENCIA NP-ťažký problém, potom aj H-KONSISTENCIA je tiež NP-ťažká.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Úlohy:

1. Nech $G = (V, G)$ je graph s vrholmi $V = \{1, 2, 3, 4, 5\}$ a hranami $E = \{12, 13, 15, 23, 25, 34, 45\}$. Vytvorte tréningovú vzorku $s(G)$. Nájdite najmenšiu hodnotu k, pre ktorú existuje funkcia v C_5^k konzistentná s $s(G)$. Vyjadrite túto funkciu.
2. Nech $C_n = C_n^1$ je priestor boleovských funkcií na $\{0, 1\}^n$, ktoré môžu byť reprezentované jedinou klauzulou. Sformulujte konzistentný učiaci algoritmus pre C_n, ktorý je duálny ku štandardnému algoritmu pre monočleny a určte jeho časovú zložitosť.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk