itakadémia

Univerzita Pavla Jozefa Šafárika v Košiciach, Prírodovedecká fakulta

Strojové učenie
Konzistentné algoritmy

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

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

Cieľ:

- Potenciálna naučiteľnosť hypotézového priestoru
- Ľubovoľný konečný hypotézový priestor je potenciálne naučiteľný
- Nech H je potenciálne naučiteľný a L je učiaci algoritmus pre H. Môže byť L PAC?
- Rozhodovacie zoznamy (RZ)
- Algoritmus pre RZ, ktorý je konzistentný
- VCdim a potenciálna naučiteľnosť

Algoritmus pre učenie polpriamok (lúčov)

Pripomeňme algoritmus pre učenie v hyp. priestore reál. čísel. Pre každé reálne číslo Θ lúč r_Θ (polpriamka $\langle\Theta, +\infty)$) je koncept definovaný na príkladovom priestore $\Re$ (reálne čísla) funkciou

$$r_\Theta(y) \iff y \geq \Theta$$

Hypotézový priestor $H = \{r_\Theta | \Theta \in \Re\}$

Algoritmus učenia: za aktuálnu hypotézu vezmeme "najmenší" lúč obsahujúci všetky pozitívne príklady v tréningovej vzorke. Ak neexistujú kladné príklady, vtedy hovoríme o prázdnom lúči, označenie r_∞.

Definícia: (**PAC - probably approximately correct**)

Algoritmus L je **PAC algoritmus**, *pre hypotézový priestor H, ak*

- *k ľubovoľnému reálnemu číslu* ϵ, $0 \le \epsilon \le 1$
- *k ľubovoľnému reál. číslu* δ, $0 \le \delta \le 1$
- *existuje kladné celé číslo* $m_0 = m_0(\delta, \epsilon)$ *také ,že*
- *pre ľub. cieľový koncept* $t \in H$, *pre ľub. pravdepodobnostné rozdelenie* μ *na* X,
- *pre všetky* $m \ge m_0$ *platí*

$$\mu^m\{\bar{s} \in S(m,t) \quad | \quad err_\mu(L(\bar{s}),t) < \epsilon\} > 1 - \delta$$

Veta:

Algoritmus L pre učenie lúčov je PAC.

Z dôkazu vyplýva hranica na počet príkladov tréningovej vzorky

$$m \geq m_0 = \frac{1}{\epsilon} * ln\frac{1}{\delta}$$

Exaktné učenie

Keď **príkladový priestor** X **je konečný**, pojem "PAC - učenie" má ďalšie dodatočné obmedzenia.

Záverom úvah je, že učenie na konečnom priestore príkladovom je **"PEC - probably exactly correct"**.

Príklad: Štandartný učiaci algoritmus pre monočleny (Valiant) na $\{0,1\}^n$ pre pevné n.

Potenciálna naučiteľnosť

Učiaci algoritmus je PAC – vlastnosť algoritmu

Ak je algoritmus daný, môžeme

- sa pokúšať dokázať, že je PAC, ale tiež
- uvažovať o vlastnostiach H, na ktorom algoritmus pracuje.

Podstatná otázka: Je hypotézový priestor H naučiteľný?

Uvedieme **vlastnosť hypotézového priestoru H**, ktorá zaručí, že konzistentný algoritmus pre učenie, ak H je konceptový aj hypotézový priestor, bude **PAC**.

Ukážeme, že mnoho priestorov má túto vlastnosť.

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

Definícia: (Konzistentný učiaci algoritmus)

Nech H je hypotézový priestor, X je príkladový priestor. Učiaci algoritmus L pre H je **konzistentný**, *ak pre ľubovoľnú tréningovú vzorku $\overline{s}$ a cieľový koncept $t \in H$, výstupná hypotéza $h = L(\overline{s}) \in H$ súhlasí s t na príkladoch v $\overline{s}$, t. j. platí $h(x_i) = t(x_i), \quad 1 \leq i \leq m$.*

Pre dané $\overline{s} \in S(m, t)$ označíme

$$H[\overline{s}] = \{h \in H| \quad h(x_i) = t(x_i), \quad 1 \leq i \leq m\},$$

$H[\overline{s}]$ je množina všetkých hypotéz konzistentných s $\overline{s}$.

Je každý konzistentný algoritmus PAC?

Algoritmus L je konzistentný $\Leftrightarrow L(\bar{s}) \in H[\bar{s}]$ pre každú tréningovú vzorku $\bar{s}$.

Dôležité sú množiny $H[\bar{s}]$ a ich vlastnosti.

Z toho vyplýva, že na to, aby konzistentný učiaci algoritmus bol PAC, postačí zadávať podmienky na množiny $H[\bar{s}]$.

Postačujúca podmienka na $H[\bar{s}]$.

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

Predpokladajme, že je dané pravdepodobnostné rozdelenie μ na X.
Na moment zafixujme cieľový koncept $t \in H$.
K danému $\epsilon \in (0, 1)$ položíme

$$B_\epsilon = \{h \in H| \quad er_\mu(h, t) \geq \epsilon\},$$

čo je množina ϵ-zlých hypotéz pre t.

Konzistentný algoritmus pre H dáva výstup, ktorý je v $H[\overline{s}]$
a PAC vlastnosť vyžaduje, aby výstup, ktorý je dôveryhodný, nebol ϵ-zlý;
inak povedané – **je nepravdepodobné, že zlá hypotéza je korektná na tréningovej vzorke**.

Definícia: (Potenciálne naučiteľný hypotézový priestor)

Hovoríme, že hypotézový priestor H je **potenciálne naučiteľný**, *ak k daným reálnym číslam $\delta, \epsilon, 0 < \delta, \epsilon < 1$ existuje kladné celé číslo $m_0 = m_0(\delta, \epsilon)$ také, že pre všetky $m \geq m_0$, platí*

$$\mu^m\{\bar{s} \in S(m,t) | \quad H[\bar{s}] \cap B_\epsilon = \emptyset\} > 1 - \delta$$

pre ľubovoľné pravdepodobnostné rozdelenie μ na X a ľub. $t \in H$.

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

Veta:

Ak H je potenciálne naučiteľný a L je konzistentný učiaci algoritmus pre H, potom L je PAC.

Dôkaz: L je konzistentný, teda $L(\bar{s}) \in H[\bar{s}]$.

$$H[\bar{s}] = \{h \in H |\ h(x_i) = t(x_i),\ 1 \leq i \leq m\},\ B_\epsilon = \{h \in H |\ er_\mu(h,t) \geq \epsilon\}$$

$L(\bar{s}) \in H[\bar{s}]$ a zároveň $H(\bar{s}) \cap B_\epsilon = \emptyset$, teda chyba $L(\bar{s})$ je menšia ako ϵ.

$$\forall_\delta \forall_\epsilon \exists_{m_0} \forall_{m>m_0} \forall_\mu \forall_{t \in H} \mu^m \{\bar{s} \in S(m,t) | \quad H[\bar{s}] \cap B_\epsilon = \emptyset\} > 1 - \delta$$

$$\forall_\delta \forall_\epsilon\ \exists_{m_0}\ \forall_{m>m_0}\ \forall_\mu\ \forall_{t \in H}\ \mu^m \{\bar{s} \in S(m,t) | \quad err_\mu(L(\bar{s})) < \epsilon\} > 1 - \delta$$

$$err_\mu(L(\bar{s})) = \mu\{x \in X | \quad t(x) \neq L(\bar{s})\} = 0 < \epsilon$$

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

Konečný prípad

Definícia potenciálnej naučiteľnosti je celkom zložitá (viac-menej je podobná PAC učeniu).
Potvrdíme potrebu definície - ukážeme jej významné aplikácie.

Veta:
Ľub. konečný hypotézový priestor H je potenciálne naučiteľný.

Dôkaz:
Nech H je konečný hypotézový priestor a δ, ϵ, t a μ sú dané.

Dokážeme, že pravdepodobnosť udalosti $H[\bar{s}] \cap B_\epsilon \neq \emptyset$ môže byť zvolená menšia než δ vybratím dostatočne veľkej dĺžky vzorky $\bar{s}$.

Potenciálna naučiteľnosť

$$B_\epsilon = \{h \in H| \quad er_\mu(h,t) \geq \epsilon\}$$

$$B_\epsilon = \{h \in H| \quad \mu\{x \in X, h(x) \neq t(x)\} \geq \epsilon\}$$

Pretože B_ϵ obsahuje ϵ-zlé hypotézy, pre ľub. $h \in B_\epsilon$, platí

$$\mu\{x \in X| \quad h(x) = t(x)\} = 1 - er_\mu(h,t) \leq 1 - \epsilon$$

Teda pre celú vzorku dĺžky m máme

$$\mu^m\{\overline{s}| \quad h(x_i) = t(x_i),\ 1 \leq i \leq m\} \leq (1-\epsilon)^m$$

Toto je pravdepodobnosť, že hypotéza h je ϵ-zlá pre celú vzorku.

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

Teda je to pravdepodobnosť, že nejaká ϵ-zlá hypotéza je v $H[\bar{s}]$.

Pravdepodobnosť, že nejaká ϵ-zlá hypotéza je v $H[\bar{s}]$, je vyjadriteľná

$$\mu^m\{\bar{s} \mid \quad H[\bar{s}] \cap B_\epsilon \neq \emptyset\}$$

a je preto menšia než $|H| \cdot (1-\epsilon)^m$.

Toto bude menej ako δ (sledujeme definíciu) za predpokladu, že položíme $m \geq m_0 = \left\lceil \frac{1}{\epsilon} \cdot \ln \frac{|H|}{\delta} \right\rceil$

Potenciálna naučitelnosť

V tomto prípade

$$|H| \cdot (1-\epsilon)^m \leq |H| \cdot (1-\epsilon)^{m_0} < |H| \cdot \exp(-\epsilon \cdot m_0) \leq |H| \cdot e^{\ln \frac{\delta}{|H|}} = \delta$$

Dokázali sme, že pre ľub. $\delta,\ \epsilon,\ t,\ \mu$ existuje m_0

$$\forall_{m \geq m_0} \forall_t \quad \mu^m \{\bar{s} \in S(m,t) | \quad H[\bar{s}] \cap B_\epsilon \neq \emptyset\} < \delta$$

Ak vezmeme komplementárnu udalosť, dostaneme správny záver.

Potenciálna naučitelnosť

Veta pokrýva všetky bool. prípady, kde príkladový priestor je $\{0,1\}^n$ (alebo podmnožina) s pevným n.

V ľubovoľnej takej situácii konzistentný algoritmus je automaticky PAC. Napríklad algoritmus pre učenie monočlenov a disjunkcií malých monočlenov sú PAC.

Dôkaz nám ďalej hovorí, koľko príkladov postačí na dosiahnutie predpísaných úrovní dôvery a presnosti.

$$m \geq m_0 = \left\lceil \frac{1}{\epsilon} \cdot \ln \frac{|H|}{\delta} \right\rceil$$

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

Pre algoritmus monočlenov vieme, že veľkosť $|M_n|$ hypotézového priestoru je 3^n. Preto

$$m_0 = \left\lceil \frac{1}{\epsilon} \ln \frac{|M_n|}{\delta} \right\rceil = \left\lceil \frac{1}{\epsilon} \left(n \cdot \ln 3 + \ln \frac{1}{\delta} \right) \right\rceil$$

je postačujúci počet príkladov na zabezpečenie, že s pravdepodobnosťou väčšou než $1 - \delta$ výstup algoritmu má chybu menšiu než ϵ.

Navyše pre ľub. konečný hypotézový priestor existuje konzistentný učiaci algoritmus: metóda učenia očíslovaním hypotéz, ktorú sme popísali.

Teda bezprostredným dôsledkom vety je, že k ľub. konečnému hypotézovému priestoru H existuje učiaci algoritmus, ktorý je PAC.

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

Potenciálna naučiteľnosť

Úvahy:

- Vytvorili sme komplikovanú podmienku, aby sme dokázali, že je vždy splnená v konečnom prípade, ktorý je ale najdôležitejší v praxi.
- Avšak praktické predpoklady zavádzajú dodatočné ohraničenie, a síce že počet príkladov by mal byť "zvládnuteľný", a teda nie je nutné použiť metódu učenia sa očíslovaním.
- Ak napr. hypotézový priestor je množina B_n všetkých booleovských funkcií n premenných, $|B_n| = 2^{2^n}$, potom hranica pre veľkosť vzorky je $m_0 = \left\lceil \frac{2^n}{\epsilon} \ln \frac{2}{\delta} \right\rceil$. Čo ak $n = 50$?

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

- Jeden spôsob popisovania zložitých konceptov je ich vytváranie z menších jednotiek. Viď vytváranie $D_{n,k}$.
- Metóda konštrukcie, ktorá môže byť aplikovaná na ľub. danú množinu vytvárajúcich blokov.

Definícia: (Rozhodovací zoznam)

Nech K je ľubovoľná množina bool. funkcií na $\{0,1\}^n$, n je pevné. Booleovská funkcia f s tým istým def. oborom ako funkcie v K sa nazýva **rozhodovacím zoznamom** *(podľa Rivesta) vytvoreným nad K, ak môže byť vyhodnotená nasledovne:*

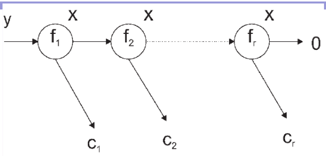
Nech je daný príklad y. Najprv vyhodnotíme $f_1(y)$ pre nejaké pevné $f_1 \in K$.

Ak $f_1(y) = 1$, priradíme do $f(y)$ pevnú hodnotu c_1.

Ak $f_1(y) \neq 1$, vyhodnotíme $f_2(y)$ pre nejaké pevné $f_2 \in K$.

Ak $f_2(y) = 1$, priradíme do $f(y)$ pevnú hodnotu c_2, inak vyhodnotíme $f_3(y)$, atď.

Rozhodovacie zoznamy



Obr.: Vyhodnotenie funkcie pomocou rozhodovacieho zoznamu

V inom zápise

if $f_1(y) = 1$ then set $f(y) = c_1$
else if $f_2(y) = 1$ then set $f(y) = c_2$
...
else if $f_r(y) = 1$ then set $f(y) = c_r$
else set $f(y) = 0$

Definícia: (DL(K))

Hypotézový priestor rozhodovacích zoznamov na množine K, $DL(K)$, je množina konečných postupností $f = (f_1, c_1), (f_2, c_2), \ldots, (f_r, c_r)$ takých, že $f_i \in K$, $c_i \in \{0, 1\}, 1 \leq i \leq r$. Hodnoty f sú definované

$$f(y) = \begin{cases} c_j, & ak\ j = min\{i| \quad f_i(y) = 1\}\ existuje \\ 0, & inak \end{cases}$$

Vyžadujeme, aby všetky členy v DL(K) boli rôzne; opakovanie danej funkcie $g \in K$ môže byť odstránené a nemá účinok na vyhodnotenie.

Teda dĺžka DL(K) vytvoreného nad konečnou množinou K je najviac $|K|$ a $|DL(K)|$ je ohraničená zhora konštantou $|K|$.

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

Predpokladajme, že $K = M_{3,2}$ je priestor monočlenov dĺžky najviac 2 z 3 premenných. Rozhodovací zoznam

$$(\langle u_2 \rangle, 1), (\langle u_1 \overline{u_3} \rangle, 0), (\langle \overline{u_1} \rangle, 1)$$

môže byť spracovávaný nasledujúcim spôsobom na príkladovom priestore $\{0,1\}^3$.

Najprv sú vybrané tie príklady, pre ktoré $\langle u_2 \rangle$ dáva hodnotu 1. Sú to 010, 011, 110, 111;

ďalej zostanú len tie, ktoré pre $u_1 \overline{u_3}$ dajú 0, je to len 10;

ďalej tie, ktoré pre $\langle \overline{u_1} \rangle$ dávajú hodnotu 1, sú to 000, 001.

Zostáva jediný príklad 101, ktorému je priradená hodnota 0.

Rozhodovacie zoznamy

Je dôležité poznamenať, že disjunkcia 2 funkcií v K je špeciálny prípad rozhodovacieho zoznamu založeného na K. Explicitne $f \vee g$ je reprezentovaná rozhodovacím zoznamom $(f, 1), (g, 1)$.

To znamená, že rozhodovací zoznam je zovšeobecnenie disjunkcie.

Napríklad, priestor $D_{n,k}$ je obsiahnutý v $DL(M_{n,k})$; v skutočnosti je vlastnou podmnožinou.

Iný dôsledok je, že pre dané n, ľubovoľná bool. funkcia n premenných je v nejakom $DL(M_{n,k})$ pre k dostatočne veľké. (Bezprostredne, toto je pravda pre $k = n$ podľa existencie disjunktnej normálnej formy.)

Konzistentný učiaci algoritmus pre DL(K)

Učiaci algoritmus pre $DL(K)$, ktorý pracuje, keď K je ľub. konečná množina.
Algoritmus je konzistentný, ale nie je bezpamäťový on-line algoritmus.
Samozrejme, učiaci algoritmus očíslovaním má podobné vlastnosti avšak ukážeme, že tento algoritmus je podstatným vylepšením.

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

Konzistentný učiaci algoritmus pre rozhodovacie DL(K)

Nech $\bar{s}$ je tréningová vzorka ohodnotených príkladov (x_i, b_i), $1 \leq i \leq m$. V každom kroku konštrukcie rozhodovacieho zoznamu niektoré príklady budú vymazané, iné zostanú.

Procedúra prebehne cez K hľadajúc funkciu $g \in K$ a hodnotu c takú, že pre ešte nezaradené príklady x_i platí, ak $g(x_i) = 1$, tak b_i je konštantná booleovská hodnota c.

Dvojica (g, c) je potom vybratá ako ďalší člen postupnosti definujúcej rozhodovací zoznam, a všetky príklady spĺňajúce g sú vymazané.

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

Rozhodovacie zoznamy

Procedúra je opakovaná, pokým všetky príklady z $\bar{s}$ neboli vymazané.
Nech $\{g_1, g_2, \ldots g_p\}$ je očíslovanie funkcií v K.

Algoritmus:

```
set I= {1,2, ..., m };
j:=1;
repeat
   if forall i in I, g[j](x[i])=1 implikuje b[i]=c then
     begin
       select (g[j],c);
       delete from I all i for which g[j](x[i])=1;
       j:=1;
     end
   else j:=j+1;
until I is empty set;
```

Príklad

$K = M_{5,2}$ a predpokladajme, že K je zaevidované v "slovníkovom poradí" nad literálmi $u_1\, u_2\, u_3\, u_4\, u_5\, \overline{u_1}\, \overline{u_2}\, \overline{u_3}\, \overline{u_4}\, \overline{u_5}$.
Prvých niekoľko funkcií v slovníku je: monočleny dĺžky 1, atď.

$$\langle\rangle \;\langle u_1\rangle \;\langle u_1 u_2\rangle \;\langle u_1 u_3\rangle \;\langle \overline{u_1} u_4\rangle \;\langle \overline{u_2} u_3\rangle$$

Predpokladajme, že trénovacia vzorka je:

$$\begin{array}{llll} x_1 = 10000 & b_1 = 0, & x_4 = 10101 & b_4 = 1 \\ x_2 = 01110 & b_2 = 0, & x_5 = 01100 & b_5 = 1 \\ x_3 = 11000 & b_3 = 0, & x_6 = 10111 & b_6 = 1 \end{array}$$

Vyberieme prvú položku zo slovníka, ktorá splňuje požiadavky podmienky.

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

Príklad

- $\langle\rangle$ *nepoužijeme,*
- $\langle u_1 \rangle$ *vylúčime príklady* x_1 *a* x_4 *majú* $b_1 \neq b_4$
- $\langle u_1 u_2 \rangle$ *je splnené len pre* x_3 *a* $b_3 = 0$*, teda vyberieme ako prvý term* $(\langle u_1 u_2 \rangle, 0)$ *do rozhodovacieho zoznamu a vymažeme príklad* x_3.

Príklad

Ďalšie kroky:

- $\langle u_1 u_3 \rangle$ *je splnené pre* x_4 *a* x_6 *a* $b_4 = b_6 = 1$, *teda vyberieme* $(\langle u_1 u_3 \rangle, 1)$; *vymažeme* x_4 *a* x_6
- $\langle u_1 \rangle$ *je splnené pre* x_1, $b_1 = 0$, *vyberieme* $(\langle u_1 \rangle, 0)$
- $(\langle \overline{u_1} u_4 \rangle, 0)$ *vymaže* x_2
- $(\langle \rangle, 1)$ *vymaže* x_5

Vytvorený rozhodovací zoznam je:

$$(\langle u_1 u_2 \rangle, 0), (\langle u_1 u_3 \rangle, 1), (\langle u_1 \rangle, 0), (\langle \overline{u_1} u_4 \rangle, 0), (\langle \rangle, 1)$$

Rôzne usporiadania $M_{5,2}$ dávajú rôzne odpovede.

Nie je bezprostredne zrejmé, prečo hľadanie g a c je vždy úspešné.

Aby sme dokázali, že algoritmus pracuje správne, je nutné ukázať, že vždy keď je daná tréningová vzorka $\bar{s}$ pre cieľový koncept v $DL(K)$, potom vždy bude nejaká dvojica (g, c), ktorá vyhovuje niekoľkým príkladom.

O tréningovej vzorke danej vyššie nebolo na začiatku známe, či je kompatibilná s konceptom v $DL(M_{5,2})$, napriek tomu úspešné dokončenie algoritmu ukazuje, že je v skutočnosti v tomto tvare.

Konzistentný učiaci algoritmus pre DL(K)

Veta:
Predpokladajme, že K je hypotézový priestor obsahujúci identicky 1-kovú funkciu. Nech t je funkcia v $DL(K)$ a nech $\bar{s}$ je konečná množina príkladov. Potom existuje $g \in K$ a $c \in \{0,1\}$ také, že:

1. *množina $S^g = \{x \in \bar{s} | \quad g(x) = 1\}$ je neprázdna;*
2. *pre všetky $x \in S^g, \quad t(x) = c$.*

Dôkaz: Je dané $t \in DL(K)$, reprezentácia $t = (f_1, c_1), (f_2, c_2), \dots, (f_r, c_r)$

Ak $f_i(x) = 0$ pre všetky $x \in \overline{x}$ a všetky $i \in \{1, 2, \dots, r\}$, potom všetky príklady v $\bar{s}$ sú negatívne príklady pre t. V tomto prípade g bude identicky 1-ková funkcia a $c = 0$.

Naopak, ak existuje nejaké i také, že množina $\{x_{i_1}, x_{i_2}, \ldots\}$, $x_{i_j} \in \bar{s}$, pre ktoré $f_i(x_{i_j}) = 1$ nie je prázdna, potom nech g bude najmenším takým indexom, t.j. $g = min\{i_1, i_2, \ldots\}$.

Z definície rozhodovacieho zoznamu vyplýva, že $t(x) = c_g$ pre všetky x také, že $f_i(x) = 1$. V tomto prípade môžeme vybrať $g = f_i$ a $c = c_g$.

Z tohto vyplýva, že k ľub. tréningovej vzorke pre funkciu v $DL(K)$ existuje vhodný výber dvojice (g, c) pre "prvý člen" rozhodovacieho zoznamu. Aplikáciou tohto výsledku rekurzívne vidíme, že algoritmus popísaný vyššie bude vždy úspešný.

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

Vzťah VCdim a potenciálnej naučiteľnosti

Jednoduchý príklad priestoru H s nekonečnou VCdim:

Pre každú podmnožinu $A \subseteq \mathcal{R}$ definujeme charakteristickú funkciu χ_A

$$\chi_A(y) = 1, \text{ ak } y \in A; = 0 \text{ inak}.$$

Nech U je množina všetkých podmnožín $\mathcal{R}$, ktoré môžu byť vyjadrené ako konečné zjednotenie uzavretých intervalov. Nech

$$J = \{\chi_A | A \in U\}$$

je priestor zjednotených intervalov. Tento priestor má nekonečnú VCdim. Ukážeme.

VC dim a potenciálna naučiteľnosť

Nech $\overline{x} = (x_1, x_2, \ldots, x_m)$ je vzorka rôznych bodov v $\mathcal{R}$. Nech E_x označuje odpovedajúcu množinu príkladov.
K ľub. $S \subseteq E_x$ vieme skonštruovať množinu $A \in U$ takú, že $S \subseteq A$ a $(E_x \setminus S) \cap A = \emptyset$ takto:

Pre každé $x_i \in S$ nech A_i bude uzavretý interval, ktorý obsahuje x_i, ale nie iné prvky z E_x a nech A je zjednotenie všetkých takých A_i. Množina A je konečné zjednotenie uzavretých intervalov a χ_A je rovná 1 na S a 0 na $E_x \setminus S$. Inak povedané J rozbíja $\overline{x}$.

Pomocou rôznych S dosiahneme všetky možné ohodnotenia.

Vzťah VCdim a potenciálnej naučiteľnosti

Záver: Pretože to platí pre ľub. konečnú vzorku, ľub. dĺžky, z toho vyplýva, že VCdim(J) je nekonečná.

Veta:
Ak má hypotézový priestor nekonečnú VC dimenziu, tak nie je potenciálne naučiteľný.

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

Veta: (Fundamentálna.)

Ak má hypotézový priestor konečnú VC dimenziu, tak je potenciálne naučiteľný.

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. Dokážte, že algoritmus pre DL(K) je konzistentný.
2. Ukážte, že pre hypotézový priestor $D_{n,k}$ $(n \geq k > 1)$ je postačujúce vziať hodnotu $m_0(\delta, \epsilon) = \left[\frac{k}{\epsilon} \cdot \ln 2n + \frac{1}{\epsilon} \cdot \ln \frac{1}{\delta}\right]$ v def. potenciálnej naučiteľnosti.
3. Dokážte, že pre všetky $n \geq k \geq 1$, $D_{n,k} \subset DL(M_{n,k})$. Odvoďte, že ľubovoľná bool. funkcia môže byť reprezentovaná rozhodovacím zoznamom.
4. Skonštruujte bool. funkciu 3 premenných, ktorá nie je $D_{3,2}$, ale je v $DL(M_{3,2})$.
5. Skonštruujte bool. funkciu 3 premenných, ktorá nie je v priestore $DL(M_{3,2})$.

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

6. Dokážte, že pre ľubovoľnú množinu K booleovských funkcií, $3^{|K|} \cdot |K|!$ je horná hranica na $|DL(K)|$.

7. Komplement bool. funkcie h je bool. funkcia $\overline{h}$ taká, že $\overline{h}(x) = 1 \Leftrightarrow h(x) = 0$. Dokážte, že pre ľub. množinu K bool. funkcií obsahujúcu identickú funkciu 1, platí

$$h \in DL(K) \Leftrightarrow \overline{h} \in DL(K).$$

Toto znamená, že $DL(K)$ je uzavretá vzhľadom na komplement.

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