Univerzita Pavla Jozefa Šafárika v Košiciach, Prírodovedecká fakulta

Strojové učenie

Pravdepodobnostné učenie

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

Úloha: Je daná tréningová vzorka

Tr. vzorka	Hyp. z M_7	Hyp. z D_7^2
(1 1 1 0 0 0 1, 1)	1	**0**
(1 0 1 1 0 1 1, 1)	1	**0**
(1 1 1 0 1 1 0, 1)	1	**0**
(1 1 1 1 1 1 1, 1)	1	**0**
(1 1 0 1 0 0 0, 0)	0	0
(1 0 0 1 0 0 1, 0)	0	0
(1 0 0 1 1 1 1, 0)	0	0
(1 1 1 1 0 0 1, 0)	**1**	0

Učiaci algoritmus, ktorý nájde hypotézu (prediktor) v tvare Bool. funkcie.

Učenie monočlenov, výsledná hypotéza je: $u_1 u_3$

Učenie disjunkcii malých monočlenov, nech k=2; $\overline{u_1} \vee \overline{u_1}\ \overline{u_4}$

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

Chyba hypotézy, tréningovej vzorky

Predpokladajme, že máme danú tréningovú vzorku

$$\overline{s} = ((x_1, b_1), \ldots, (x_m, b_m)).$$

Príklady sú získané z nejakého pravdepodobnostného rozdelenia všetkých príkladov, ktoré nemusíme poznať.
Výstup algoritmu učenia (učiaceho sa) je hypotéza (prediktor, klasifikátor), ktorá vznikne na základe vstupných príkladov.

Miera úspechu učenia: Definujeme **chybu hypotézy** ako pravdepodobnosť ňou určených nesprávnych ohodnotení príkladov získaných z daného pravdepodobnostného rozdelenia.

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

Dostupná informácia pre algoritmus učenia:

- Nevie nič o pravdepodobnostnom rozdelení príkladov a ich rozdelení. Nepozná hypotézu.
- Dostáva len tréningovú vzorku: príklady a ich ohodnotenia.
- Po natrénovaní je možné vypočítať chybu tréningovej vzorky, čo je pomer počtu nesprávne klasifikovaných príkladov k celkovému počtu príkladov tréningovej vzorky. Hovoríme tiež o empirickej chybe alebo empirickom risku.

Má zmysel hľadať hypotézu s minimálnou chybou pre danú trén. vzorku? Minimalizácia empirického risku - Empirical Risk Minimization (ERM).

učiaci sa = algoritmus učenia

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

Poznámka k empirickej chybe

Nech $n = 4$.

Tréningová vzorka pre učenie monočlenov a monočlen je

Tr. vzorka	$\overline{u}_3 u_4$
(1 0 1 0, 0)	0
(1 0 0 1, 1)	1
(0 0 0 1, 0)	1
(0 1 0 1, 1)	1
(1 1 0 1, 0)	1

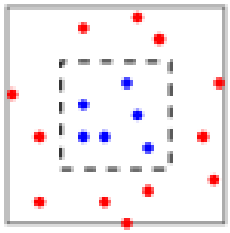
Empirická chyba tejto tréningovej vzorky: 2*100/4=50%.

Za predpokladu rovnomerného rozdelenia všetkých 16 príkladov, táto hypotéza bude správne predikovať 4*100/16=25%

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

Poznámka o overfittingu

Minimalizácia empirickej chyby môže spôsobiť problémy.
Predpokladajme, že kladné príklady sú vo veľkom štvorci. Z modrých príkladov natrénujeme malý štvorec.
Riešenie - prehľadať celý hypotézový priestor a hľadať najlepšiu hypotézu.



Našli sme prediktor, ktorého výkon na tren. vzorke je vynikajúci, no jeho výkon v skutočnom "svete" je veľmi slabý.

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

Poznámky k ERM

- Hľadať podmienky - existuje záruka, že ERM nespôsobí overfit; podmienky, za ktorých, keď má prediktor ERM dobrý výkon vzhľadom na trénovacie dáta, je tiež vysoko pravdepodobné, že bude dobre fungovať v porovnaní so základnou distribúciou dát
- Riešenie: ERM pravidlo na ohraničený vyhľadávací priestor hypotéz - vyučujúci vyberie (predtým, než videl dáta) množinu prediktorov, *triedu hypotéz* H.

$$ERM_H(\overline{s}) \in argmin_{h \in H} L(\overline{s}, h),$$

kde $L(\overline{s}, h)$ je učiaci algoritmus na tr. vzorke $\overline{s}$, $h \in H$ je funkcia mapujúca X do množiny predpokladaných výsledkov.

- *inductive bias*

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

Poznámky k ERM

- Základnou otázkou v teórii učenia je: pri ktorých triedach hypotéz nebude učenie ERM_H viesť k overfitingu. Budeme sa tým zaoberať detailnejšie.
- Príklady v tréningovej vzorke majú pravdepodobnostné rozdelenie. A teda vieme určiť pravdepodobnosť danej tréningovej vzorky.
- Vieme určiť chybu (aj pravdepodobnosť chyby) skúmanej hypotézy.
- Toto skombinujeme v pravdepodobnostnom učení.

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 pre učenie lúčov

Budeme zaoberať veľmi jednoduchým algoritmom pre učenie v reálnom hypotézovom priestore. Lúč je polpriamka na reálnej osi.

Pre každé reálne číslo Θ **lúč r_Θ (polpriamka $\langle\Theta, +\infty)$)** je koncept definovaný na príkladovom priestore $\Re$ (reálne čísla) funkciou

$$r_\Theta(y) = 1 \iff y \geq \Theta$$

Algoritmus pre učenie v hypotézovom priestore $H = \{r_\Theta | \Theta \in \Re\}$ je založený na idee, že za aktuálnu hypotézu vezmeme "najmenšie" r_θ obsahujúce všetky pozitívne príklady v tréningovej vzorke.
V prípade, že neexistujú kladné príklady, vtedy budeme hovoriť o prázdnom lúči. Bude označovaný r_∞.

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

Algoritmus pre učenie lúčov

begin

set $\lambda = \infty$;

for i:=1 to m do

if $(b_i = 1)$ and $(x_i < \lambda)$ then set $\lambda = x_i$;

L(s):=r_λ;

end

Pretože je len konečný počet príkladov v tréningovej vzorke a príkladový priestor je nespočítateľný, nemôžeme očakávať, že $\lambda = \Theta$.

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 pre učenie lúčov

Predpokladajme, že máme cieľový koncept $\Theta = 1.2856$.
Máme tréningovú vzorku

$$(8.2, 1), (4.3, 1), (2.99, 1), (-1, 0), (1.5, 1), (2.2, 1)$$

Získané λ je 1.5.
Z algoritmu je vidieť, že ak tréningová vzorka je pre cieľovú hypotézu r_Θ, potom $L(\overline{s})$ bude lúč r_λ s $\lambda = \lambda(\overline{s}) \geq \Theta$.

Tu chybu vidíme. Ale v skutočnosti máme len príklady, nemáme Θ, preto chybu potrebujeme odhadnúť.

Ak dĺžka tréningovej vzorky rastie, tak by sa mala zvyšovať pravdepodobnosť toho, že chyba medzi r_λ a r_Θ bude menšia (prípadne klesá).

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 pre učenie lúčov

Táto vlastnosť môže byť charakterizovaná nasledovne:

Predpokladajme, že spustíme algoritmus s veľkou tréningovou vzorkou a potom sa rozhodneme použiť výstupnú hypotézu r_λ pre cieľovú (neznámu) hypotézu r_Θ. Inak povedané, uspokojíme sa s tým, že **"učiaci sa"** bol adekvátne natrénovaný.

Ak λ nie je blízke k Θ, indikuje to, že pozitívne príklady, blízke Θ sa nevyskytovali v tréningovej vzorke, a teda boli málo pravdepodobné.

Z toho vyplýva, keď teraz klasifikujeme niektoré ďalšie príklady, ktoré sú prezentované podľa toho istého pravdepodobnostného rozdelenia, tak môžeme urobiť niekoľko chýb ako dôsledok použitia r_λ namiesto r_Θ.

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 pre učenie lúčov

Uvažujme model, v ktorom tréningová vzorka $\overline{s}$ pre cieľový koncept t je generovaná náhodným výberom príkladov $x_1, x_2, \ldots, x_m$ z X, ktorý má nejaké neznáme, ale pevne dané pravdepodobnostného rozdelenie.

Učiaci algoritmus L produkuje hypotézu $L(\overline{s})$, ktorá je očakávaná ako dobrá aproximácia pre cieľový koncept t.

Očakávame tiež, ak počet príkladov m v tréningovej vzorke vzrastie, tak z pravdepodobnosti vyplynie, že chyba, ktorá je výsledkom použitia $L(\overline{s})$ namiesto t, bude menšia.

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

Základné pojmy:

X - pravdepodobnostný priestor, **A** - trieda podmnožín množiny X,
μ - pravdepodobnostné rozdelenie, miera pravdepodobnosti, z **A** $\langle 0, 1\rangle$.

$$\mathbf{A} \rightarrow \quad \langle 0, 1\rangle.$$

Od triedy **A** sa vyžaduje, aby bola uzavretá vzhľadom na operácie komplementu, konečného prieniku a spočítateľného zjednotenia.
Prvok $A \in \mathbf{A}$, sa nazýva *udalosť* a $\mu(A)$ je pravdepodobnosť udalosti A.
Od μ sa vyžaduje, aby splňovala nasledujúce podmienky:

$$\mu(\emptyset) = 0, \mu(X) = 1, \mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i).$$

pre ľubovoľné po dvoch disjunktné množiny $A_1, A_2, \ldots \in \mathbf{A}$

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

Pravdepodobnostné učenie

Pre nás X je príkladový priestor, a príklady sú buď booleovské alebo reálne. V prvom prípade je X konečná alebo spočítateľná množina, a teda **A** môže byť triedou všetkých podmnožín X.

V prípade reálnych čísel môžeme zobrať za **A** ľubovoľnú dostatočne veľkú triedu obsahujúcu množiny, ktoré potrebujeme uvažovať; postačí uvažovať triedu Borelovských množín v $\Re^n$.

V oboch prípadoch budeme používať vhodnú triedu bez explicitného vyjadrovania detailov.

Borelovská množina je ľubovoľná množina v topologickom priestore, ktorú je možné získať z otvorených množín pomocou operácií spočítateľného zjednotenia, spočítateľného prieniku a relatívneho doplnku (alebo ekvivalentne z uzavretých množín). Názov majú po francúzskom matematikovi Émile Borelovi.

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

Pravdepodobnostné učenie

Budeme hovoriť **pravdepodobnostné rozdelenie μ na X**, čím mienime funkciu μ definovanú na vhodnej triede **A** a splňujúcej axiómy uvedené vyššie.
Treba zdôrazniť, že v aplikáciách, nerobíme žiadne predpoklady o μ, okrem podmienok uvedených v definícii.

Situácia, ktorú modelujeme, je taká, že svet príkladov prezentovaných učiacemu sa, sa chová (modeluje) podľa nejakého pevného, ale neznámeho rozdelenia.

Učiteľovi je povolené klasifikovať príklady ako pozitívne a negatívne, ale nemôže riadiť postupnosť, v ktorej príklady budú prezentované.

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

Pravdepodobnostné učenie

Budeme predpokladať, že cieľový koncept patrí do hypotézového priestoru H, ktorý je dostupný učiacemu sa.

K danému cieľovému konceptu $t \in H$ definujeme **chybu ľubovoľnej hypotézy $h \in H$ vzhľadom na t** a bude to pravdepodobnosť udalosti $h(x) \neq t(x)$, t. j.

$$err_\mu(h, t) = \mu\{x \in X \quad | \quad h(x) \neq t(x)\}.$$

V kučeravých zátvorkách je error set - chybová množina a predpokladáme, že existuje udalosť taká, že táto pravdepodobnosť jej môže byť priradená. Keď pôjde o t známe z konceptu, budeme tiež používať označenie $err_\mu(h)$.

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

Príklad: Nech $X = \{0,1\}^3$, predpokladajme, že cieľový koncept je $\langle u_1 \rangle$. Chybová množina pre hypotézu $\langle u_1 \overline{u_2} \rangle$ obsahuje dva príklady, 110 a 111. Tak

$$err_{<u_1 \overline{u_2}>} = \mu\{110, 111\}.$$

Napríklad, ak μ - rovnomerné rozdelenie na X: pravdepodobnosť $\frac{1}{8}$ pre každý prvok, potom

$$err_{\langle u_1 \overline{u_2} \rangle} = \frac{1}{4}.$$

Ak z nejakých dôvodov príklady, u ktorých je $y_2 = 1$ sú málo pravdepodobné, potom err_μ bude o niečo menšia.

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

Pravdepodobnostné učenie

Príklad: Nech $X = \langle -100, 1000 \rangle$, reálne čísla s 1 des. miestom; konečný; Predpokladajme rovnomerné pravdepodobnostné rozdelenie týchto čísel – príkladov do tréningovej vzorky.
Aká je pravdepodobnosť p jedného príkladu?
Počet príkladov je $N = 100 * 10 + 1 + 1000 * 10 = 11001$, takže $p = 1/N$.

Cieľový koncept v algoritme učenia lúčov bol $\Theta = 5$. Učiaci sa naučil $\lambda = 7$. Pre algoritmus sú kladné príklady $\geq \Theta$, záporné nemusíme uvažovať.

Chybová množina pre hypotézu λ sú príklady z intervalu $\langle \Theta, \lambda)$. Tak

$$err_{\langle \Theta, \lambda)} = \mu\{5.0, 5.1, 5.2, \ldots 6.9, 7\}, ||\{5.0, 5.1, 5.2, \ldots 6.9, 7\}|| = 21.$$

$$err_{\langle \Theta, \lambda)} = 21 * p.$$

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

Pravdepodobnostné učenie

Stačí jeden príklad z tohto intervalu a chyba sa zmenší. Ako?

Mohli by sme uvažovať o tom, že chyba môže byť najviac ϵ.

Pri učení je možné pripraviť nevhodné tréningové vzorky. Napríklad, ak náš cieľový koncept bude $\Theta = 5$ a tréningová vzorka bude obsahovať príklady ≥ 900, nedá sa dosiahnuť dostatočne malá chyba.

Preto je dôležité uvažovať
o pravdepodobnosti reprezentatívnosti alebo vhodnosti/nevhodnosti tréningovej vzorky.
Nevhodných tréningových vzoriek by nemalo byť veľa, δ.

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

Od X k X^m, pravdepodobnosť tréningovej vzorky

Keď je daný pravdepodobnostný priestor X, karteziánsky súčin množín – X^m preberá pravdepodobnostnú štruktúru X.

To nám umožňuje považovať komponenty m-tice $(x_1, x_2, \ldots, x_m)$ za nezávislé premenné, rozdelenie každej z nich je podľa pravdep. rozdelenia μ na X.
Odpovedajúce pravdepodobnostné rozdelenie na X^m je označované μ^m.
Pravdepodobnosť m-tice je určená súčinom.

Neformálne, pre dané $Y \subseteq X^m$ budeme interpretovať hodnotu $\mu^m(Y)$ ako "pravdepodobnosť, že náhodná vzorka m príkladov vybratých z X podľa rozdelenia μ patrí do Y".

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

Pravdepodobnostné učenie

Nech $S(m,t)$ **označuje množinu tréningových vzoriek dĺžky** m **pre daný cieľový koncept** t, kde príklady sú vyberané z príkladového priestoru X.

Ľubovoľný príklad $x_i \in X$ determinuje a je determinovaný tréningovou vzorkou $\overline{s} \in S(m,t)$:
ak $\overline{x} = (x_1, x_2, \ldots, x_m)$, potom $\overline{s} = ((x_1, t(x_1), (x_2, t(x_2), \ldots)$.

Inak povedané, existuje zobrazenie Φ

$$\Phi : X^m \to S(m,t), \quad \text{pre ktorú} \quad \Phi(\overline{x}) = \overline{s}$$

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

Pravdepodobnostné učenie

Teda môžeme vyjadriť pravdepodobnosť, že tréningová vzorka $\overline{s} \in S(m,t)$ má nejakú danú vlastnosť P, nasledujúcim spôsobom

$$\mu^m\{\overline{s} \in S(m,t) \quad | \quad \overline{s} \text{ má vlastnosť P } \},$$

to znamená

$$\mu^m\{x \in X^m \quad | \quad \Phi(x) \in S(m,t) \text{ má vlastnosť P } \}$$

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

Pravdepodobnostné učenie

Keď príkladový priestor X je vybavený pravdepodobnostným rozdelením μ, môžeme zaviesť precíznejšiu interpretáciu pre:

(i) chybu hypotézy, ktorá vznikne, keď učiaci algoritmus L pracuje s $\overline{s}$; táto veličina pracuje s $err_\mu(L(\overline{s}), t)$ a hovorí o kvalite aproximácie; pravdepodobnosť, že táto chyba je menšia než ϵ.

(ii) pravdepodobnosť kvality tréningovej vzorky; ak označíme pravdepodobnosť množiny nereprezentatívnych vzoriek δ, tak $1 - \delta$ nazveme dôveryhodnosť (confidence). Od množiny reprezentatívnych tréningových vzoriek, na ktorých je urobená dobrá aproximácia, $err_\mu(L(\overline{s}), t) < \epsilon$, budeme požadovať jej veľkú dôveryhodnosť.

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

Definícia: (**PAC - algoritmus**)

Algoritmus L je pravdepodobne aproximačne správny učiaci alg. – **probably approximately correct (PAC)**, *pre hypotézový priestor H, ak*

- *k ľubovoľnému reál. číslu* δ, *- pravdepodobnosť nereprezentatívnej vzorky,* $0 \leq \delta \leq 1$
- *k ľubovoľnému reál. číslu* ϵ, *- pravdepodobnosť chyby učenia sa,* $0 \leq \epsilon \leq 1$
- *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* $m \geq m_0$ *platí*

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

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

Pravdepodobnostné učenie

$$\forall_{(0\leq\delta,\epsilon\leq 1)}\exists_{(m_0=m_0(\epsilon,\delta))}\forall_{(t\in H)}\forall_{(\mu\ \mathbf{na}\ X)}\forall_{(m\geq m_0)}$$

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

Skutočnosť, že m_0 závisí od δ a ϵ, ale nie od t a μ odráža to, že učiaci sa môže byť schopný špecifikovať predpokladanú úroveň dôvery a presnosti, aj keď cieľový koncept a rozdelenie príkladov sú neznáme.

Dôvodom k tomu, že je možné splniť podmienku pre ľubovoľné μ je vzťah medzi dvoma veličinami, ktoré obsahujú μ: chyba err_μ a pravdepodobnosť μ^m určitej množiny príkladov.

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

Pravdepodobnostné učenie

PAC učenie je, v istom zmysle najlepšie, v čo môžeme dúfať pri tomto pravdepodobnostnom pohľade.

Nereprezentatívne tréningové vzorky, hoci nepravdepodobné, budú príležitostne prezentované učiacemu algoritmu, a tak môžeme očakávať, že to bude odhalené.

Naopak, aj keď máme reprezentatívnu tréningovú vzorku, rozšírenie tréningovej vzorky nebude vo všeobecnosti koincidovať s cieľovým konceptom, pretože výstupná hypotéza je aj tak len aproximačne správna.

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:

Algoritmus L pre učenie lúčov je PAC.

Dôkaz. Predpokladajme, že $\delta, \epsilon, r_\Theta$ a μ sú zvolené, ale sú pevné. Nech $\overline{s}$ je tréningová vzorka dĺžky m pre r_Θ a nech $L(\overline{s}) = r_\lambda$. Chybový interval je $\langle \Theta, \lambda)$. Pre danú hodnotu ϵ a dané μ definujme

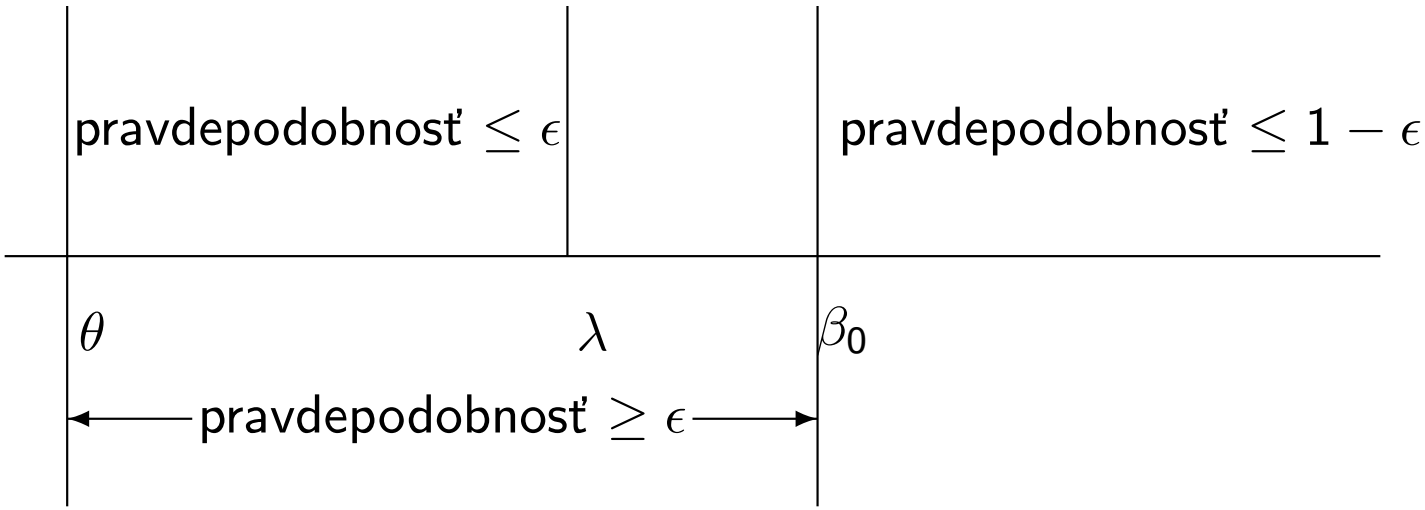
$$\beta_0 = \beta_0(\epsilon, \mu) = \sup\{\beta \quad | \quad \mu\langle \Theta, \beta) < \epsilon\}$$

Platí, že $\mu\langle \Theta, \beta_0) < \epsilon$ a $\mu\langle \Theta, \beta_0\rangle \geq \epsilon$.
Ak uvažujeme $\lambda \leq \beta_0$, tak máme

$$err_\mu(L(\overline{s})) = \mu\langle \Theta, \lambda) \leq \mu\langle \Theta, \beta_0) \leq \epsilon$$

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

Udalosť, že $\overline{s}$ má vlastnosť $\lambda \leq \beta_0$ je práve udalosť, že aspoň jeden príklad v $\overline{s}$ je v intervale $\langle \Theta, \beta_0 \rangle$.



Obr.: V tejto situácii chyba výstupu je najviac ϵ

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

Pravdepodobnostné učenie

Pretože $\mu\langle\Theta, \beta_0\rangle \geq \epsilon$, pravdepodobnosť, že jeden príklad nie je v tomto intervale, je najviac $1 - \epsilon$.
Preto pravdepodobnosť, že žiadny z m príkladov vzorky $\overline{s}$ nie je v tomto intervale je najviac $(1 - \epsilon)^m$.

Keď budeme uvažovať komplementárnu udalosť (existuje príklad, ktorý je z tohto intervalu), z toho vyplýva, že pravdepodobnosť, že $\lambda \leq \beta_0$ je aspoň $1 - (1 - \epsilon)^m$.
Ako sme už poznamenali vyššie, že udalosť $\lambda \leq \beta_0$ implikuje udalosť $err_\mu(L(\overline{s})) \leq \epsilon$ a tak

$$\mu^m\{\overline{s} \in S(m, r_\Theta) \quad | \quad er_\mu(L(\overline{s})) \leq \epsilon\} \geq 1 - (1 - \epsilon)^m$$

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

Pravdepodobnostné učenie

Položíme

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

$$(1-\epsilon)^m \leq (1-\epsilon)^{m_0} < e^{-\epsilon m_0} < e^{ln\delta} = \delta$$

tento výpočet ukazuje, že algoritmus je PAC.

Dôkaz poskytuje explicitnú formulu pre dĺžku vzorky postačujúcu na to, aby boli splnené predpísané hodnoty presnosti a dôveryhodnosti.

Predpokladajme, že $\delta = 0.001$ $\epsilon = 0.01$
$m_0 = \frac{1}{0.01} * \ln \frac{1}{0.001} = 100 * \ln 1000 = 691.$

Až 691 príkladov je treba, aby sme si boli istí na 99,9%, že najviac 1% príkladov bude klasifikovaných nesprávne, za predpokladu, že príklady sú z toho istého zdroja ako tréningová vzorka.

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

Odvodenie vzťahu:

$\delta = (1-\epsilon)^m$

$\ln\delta = m * ln(1-\epsilon) \geq m * (-\epsilon) \qquad \ln\delta \geq m * (-\epsilon)$

$\ln(1-\epsilon) \geq ln(1+0) + f'(0) * \frac{\epsilon}{1!} = 0 + \frac{1}{1-0} * (-1) * \frac{\epsilon}{1!} = -\epsilon$
v intervale $\langle 0, 1\rangle$

$-\ln\delta \leq m * \epsilon \qquad \frac{1}{\epsilon} * ln\frac{1}{\delta} \leq m$

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

Konečný príkladový priestor

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

Začneme tým, že ľubovoľné pravdepodobnostné rozdelenie na konečnej množine X je determinované hodnotami na jej 1-prvkových množinách $\{x\}$, použitím axiómy o aditivite. Budeme písať $\mu(x)$ namiesto $\mu(\{x\})$. Ak budú nejaké príklady, pre ktoré $\mu(x) = 0$, s pravdepodobnosťou 1 sa nebudú vyskytovať v konečnej náhodnej vzorke a môžu byť ignorované.

Inými slovami, môžme, ak je to nutné, predefinovať X tak, že $\mu(x) > 0$ pre všetky $x \in X$. Pretože X je konečný, veličina

$$\epsilon_\mu = min_{\{x \in X\}} \mu(x) > 0$$

je dobre definovaná.

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

Exaktné učenie

Predpokladajme, že máme algoritmus L, ktorý je PAC pre hypotézový priestor H definovaný na X. Vo význame definície PAC algoritmu máme dané δ, ϵ, μ, a t v ich obvyklom význame

$$m \geq m_0 \Rightarrow \mu^m\{\overline{s} \in S(m,t) \quad | \quad err_\mu(L(\overline{s},t)) < \epsilon\} > 1 - \delta$$

Predpokladajme, že presnosť ϵ je vybratá tak, aby nebola väčšia než ϵ_μ. Potom podmienka $err_\mu(L(\overline{s},t)) < \epsilon$ implikuje, že chybová množina pre $L(\overline{s})$ je prázdna, pretože neexistujú žiadne príklady, ktoré majú pravdepodobnosť menšiu než $\epsilon \leq \epsilon_\mu$.
Teda podmienka implikuje, že $L(\overline{s}) = t$, t.j. výstupná hypotéza je presne rovná cieľovému konceptu t.

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

Exaktné učenie

Záverom týchto úvah je, že učenie na konečnom priestore príkadov je "pec-probably exactly correct".

Ale je v tom háčik: Jednoduchá vzorka dĺžky m_0 v definícii PAC-učenia závisí od parametrov δ, ϵ ale nezávisí od μ (a t).

Argument uvedený vyššie obsahuje výber ϵ pomocou ϵ_μ, a tak hodnota m_0 vyžadovaná pre exaktné učenie bude závisieť od δ a μ.

Toto je v spore s naším originálnym cieľom dokazovania výsledkov, ktoré nie sú závislé od rozdelenia μ (možno neznáme rozdelenie 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:

Štandartný učiaci algoritmus pre monočleny na $\{0,1\}^n$ pre pevné n. Uvedieme "PEC" vlastnosť.

Kľúčovým tu je, že algoritmus poskytuje správne hypotézy, poskytované všetkými kladnými príkladmi, ktoré boli zahrnuté do tréningovej vzorky.

Dĺžka tréningovej vzorky rastie a rastie tiež pravdepodobnosť, že vzorka obsahuje všetky kladné príklady; dôsledkom toho pravdepodobnosť spôsobí, že výstup je korektný.

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

Exaktné učenie

Presnejšie, nech ϵ_μ bude najmenšia hodnota $\mu(x)$, ktorú uvažujeme nad množinou príkladov x z $\{0,1\}^n$ s nenulovou pravdepodobnosťou.

Potom pravdepodobnosť, že tréningová vzorka dĺžky m neobsahuje daný príklad je najviac $(1-\epsilon_\mu)^m$. Pravdepodobnosť, že existuje jeden z danej množiny p príkladov, ktoré nie sú v tréningovej vzorke je preto $p * (1-\epsilon_\mu)^m$.

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

Exaktné učenie

Ak X^+ je množina kladných príkladov pre daný cieľový koncept t, pravdepodobnosť, že existuje príklad v X^+, ktorý nie je vo vzorke, je najviac

$$|X^+|(1-\epsilon_\mu)^m$$

Potrebujeme vyjadriť m, teda

$$|X^+|(1-\epsilon_\mu)^m < \delta, \quad m\lg(1-\epsilon_\mu) + \lg|X^+| < \lg\delta$$

$$\lg|X^+| - \lg\delta < -m\lg(1-\epsilon_\mu) < m\epsilon_\mu$$

$$\left\lceil \frac{1}{\epsilon_\mu}\lg(|X^+|/\delta)\right\rceil < m \tag{1}$$

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

Exaktné učenie

Použijeme nejaké známe skutočnosti:

$$|X^+| \leq |X| \leq 2^n \qquad \text{a v } (0,\,1) \qquad 1-\epsilon_\mu < exp(-\epsilon_\mu)$$

$$lg|X^+| \leq n \qquad ln(1-\epsilon_\mu) < -\epsilon_\mu$$

$$m \geq \left\lceil \frac{n}{\epsilon_\mu} ln2 + \frac{1}{\epsilon_\mu} ln\frac{1}{\delta} \right\rceil$$

Poznamenajme, že dĺžka vzorky je nezávislá od t, ale závisí od rozdelenia cez parameter ϵ_μ.

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

Exaktné učenie

Skúsime analyzovať situáciu s PAC učením algoritmu učenia lúčov v konečnom priestore uvedenom na začiatku prednášky.

Nech $X = \langle -100, 1000 \rangle$, reálne čísla s 1 des. miestom; konečný;
Predpokladajme rovnomerné pravdepodobnostné rozdelenie týchto čísel – príkladov do tréningovej vzorky, pravdepodobnosť jedného príkladu máme vypočítanú $p = 1/11001, n = 11001$.
Zvolíme $\epsilon_\mu \geq p, \epsilon_\mu = 1/11001, \delta = 0.01,$.
Výsledok každej hypotézy je jedno číslo.

$$m \geq \left\lceil \frac{1}{\frac{1}{11001}} lg \frac{11001}{\delta} \right\rceil = \left\lceil 11001 * lg \frac{11001}{\delta} \right\rceil$$

Odhad hovorí, že treba použiť všetky príklady.

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. Napíšte postupnosť hypotéz generovaných algoritmom uvedeným vyššie pre učenie lúčov, ak tréningová vzorka je nasledujúca: (6.1436, 1), (1.5987, 0), (4.2381, 1), (5.7462, 1), (4.3964, 1), (4.2167, 1).
2. Aká je dĺžka tréningovej vzorky pre algoritmus učenia lúčov, aby bola dôveryhodnosť 99.5% a najviac 25% príkladov uvažovaných podľa nejakého rozdelenia do vzorky, bolo zle klasifikovaných?
3. Modifikujte algoritmus pre učenie lúčov tak, že namiesto prázdneho lúča na začiatku vezmeme nejaké veľké číslo. Je tento algoritmus konzistentný?
4. Dané sú dve reálne čísla a, b, $a \leq b$. Intervalový koncept $C_{a,b}$ je def.: $C_{a,b}(y) = 1$, ak y je v intervale $[a, b]$, inak je rovné 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

Úlohy

5. Nech H je hypotézový priestor všetkých intervalov, do ktorého patrí aj prázdny interval. Nasleduje učiaci algoritmus pre H:

empty:=true;
for i:=1 to m do
  if $b_i = 1$ then
  if empty then begin $a := x_i$; $b := x_i$; empty:= false; end
  else begin if $x_i > b$ then $b := x_i$; if $x_i < a$ then $a := x_i$ end;
if empty then $L(\overline{s})$:=prazdny interval else $L(\overline{s}) := C < a, b >$

Dokážte, že L je PAC s vhodne zvolenou dĺžkou vzorky. Akou?

Modifikujte algoritmus pre priestor intervalových konceptov, ktorý nepoužíva funkciu identicky rovnú nule.

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