Univerzita Pavla Jozefa Šafárika v Košiciach
Prírodovedecká fakulta, Ústav informatiky

Strojové učenie
Motivácia, základné pojmy

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

Úvod

Umelá inteligencia (AI) vyvíja a stáva sa sofistikovanejšou, jej rôzne aplikácie a prípady použitia sú čoraz zreteľnejšie.

Existuje šesť rôznych podskupín AI, ktoré sa oplatí sledovať:

- Strojové učenie (Machine learning)
- Hlboké učenie (Deep learning)
- Robotika (Robotics)
- Neurónové siete (Neural networks)
- Spracovanie prirodzeného jazyka (NLP)
- Genetické algoritmy (Genetic algorithms)

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

Úvod

Strojové učenie – počítačové programy, ktoré automaticky rozpoznávajú zložité vzory a robia rozhodnutia na základe dát, adaptujú svoje postupy na základe dát.

Dva pohľady na strojové učenie:

1. **Používateľský** - hotové naprogramované procedúry a funkcie (knižnice); potrebné rozumieť, kedy ktorú procedúru (metódu) použiť
2. **Konštruktérsky** - tvorba metód a algoritmov, na základe ktorých sú procedúry a funkcie vytvárané; vyžaduje dôsledný teoretický prístup, definíciu pojmov

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

Používateľský prístup - 10 TOP algoritmov v roku 2023, Podľa: United States Data Science Institute

1. Lineárna regresia
2. Logistická regresia
3. Testovanie hypotéz
4. Klastrovanie
5. ANOVA
6. Rozhodovacie stromy
7. KNN klasifikačný algoritmus
8. Naivný Bayesov algoritmus
9. Neurónové siete
10. Spoločná analýza

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

Dve školy myslenia

- deduktívne uvažovanie a
- induktívne učenie

Rozdiely súvisia s tým, ako sa zaobchádza s hypotézami spojenými so závermi.

Hypotéza: predbežné vysvetlenie pozorovania, javu alebo vedeckého problému, ktoré možno otestovať ďalším skúmaním.

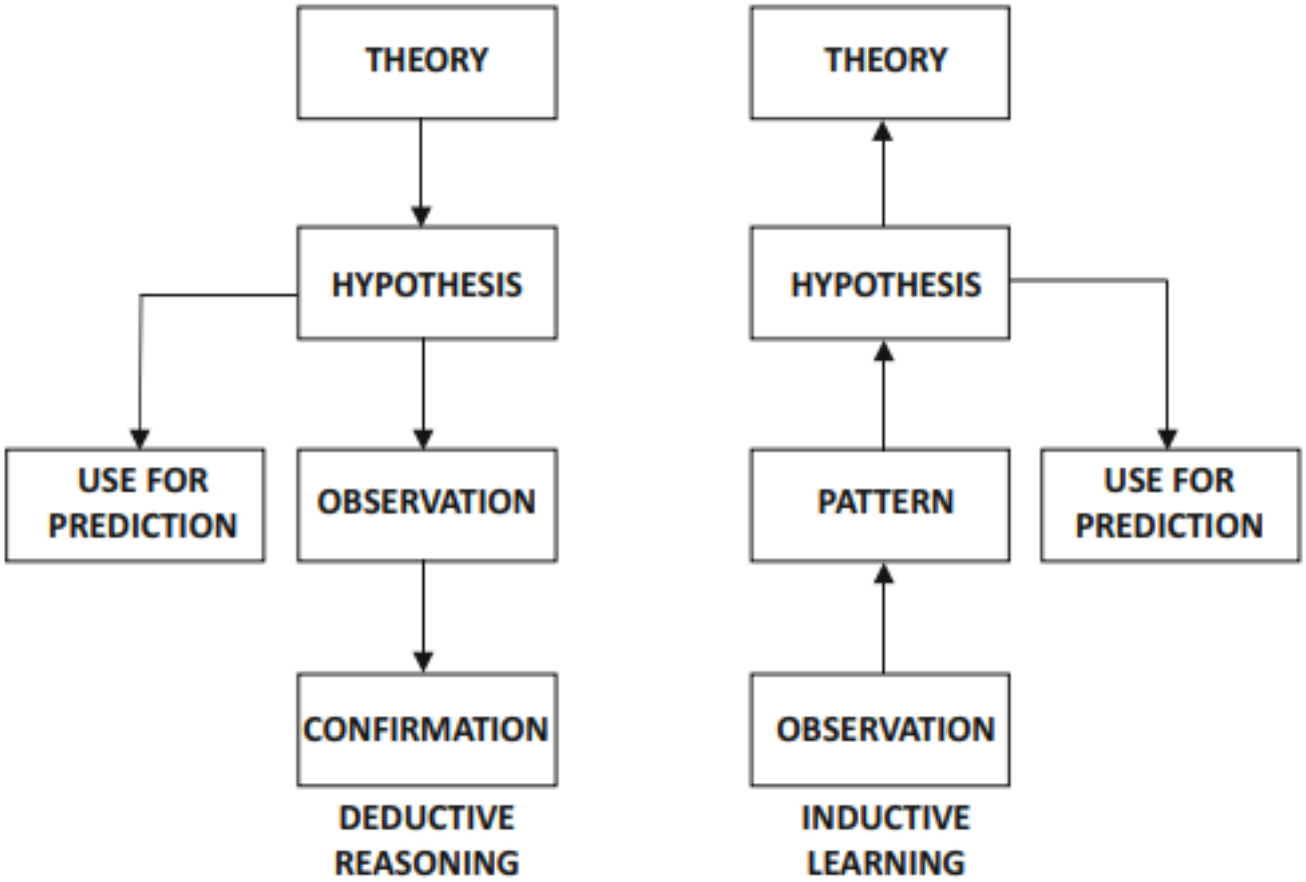
EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja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

Dve školy myslenia

1. dedukcia versus indukcia
   - dedukcia - systematický posun od všeobecného ku špecifickému; deduktívny prístup vždy poskytuje matematicky presné závery
   - indukcia - opačne; induktívny prístup je formou "chybovej" logiky založenej na zovšeobecňovaní konkrétnych skúseností
2. uvažovanie (odvodenie) versus učenie
   - metódy uvažovania, v ktorých je prepojený reťazec tvrdení
   - metódy učenia sa snažia použiť štatistické inferencie na mnohých príkladoch, aby bolo možné vyvodiť závery

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

Dve školy myslenia. Argawall. 2021



EURÓPSKA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
ĽUDSKE ZDROJE
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

Typy strojového učenia

Je mnoho typov aktivít označovaných ako **"učenie sa"**.

4 typy strojového učenia:

- pod dohľadom (supervised)
- čiastočne pod dohľadom (semi-supervised)
- bez dozoru (unsupervised)
- posilňovacie učenie (reinforcement learning)

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

V tomto predmete budeme najprv študovať **matematický model** takého procesu, ktorý pracuje na základe známych údajov.

Tento model sa zdá byť použitelný, pretože zachytáva a upresňuje základ určitých aktivít, ktoré boli popísané predtým pomocou nepresných výrazov, a zároveň umožňuje vytvoriť netriviálne matematické tvrdenia, ktoré môžu byť dokázané.

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

Ciele, ktoré budeme sledovať

1. Poskytnúť rigorózny úvod k hlavným konceptom patriacim ku strojovému učeniu:

   Čo je učenie? Ako sa môže stroj učiť? Ako určujeme zdroje potrebné na naučenie nejakého konceptu? Môžeme dopredu vedieť, či učiaci proces bude úspešný? Ako vyhodnotiť kvalitu učenia? Je učenie vždy možné?

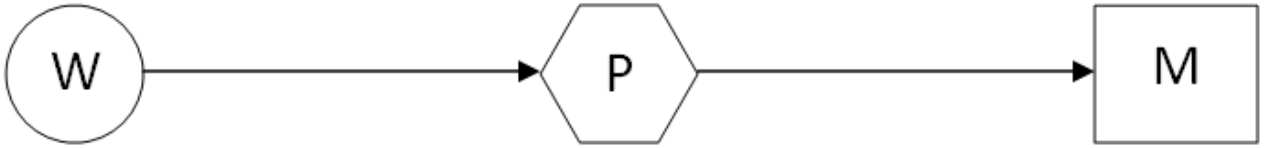
2. Prezentovať niektoré kľúčové algoritmy strojového učenia. Vybrali sme niektoré úspešne používané algoritmy v praxi, ale uvedieme tiež spektrum rôznych učiacich techník.

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

Motivačný príklad



Skutočné príklady

Reálny svet, ktorý obsahuje množinu objektov nazývané príklady

Kódované príklady

Procesor vytvára kódovaný príklad

Rozpoznané príklady

Stroj rozpoznáva určité príklady; výstupom je 0 alebo 1
1 – príklad je rozpoznaný
0 – príklad nie je rozpoznaný

Obr.: Spracovanie príkladov reálneho sveta pomocou natrénovaného stroja

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

Koncepty

Pojem koncept môže byť popísaný pomocou množiny príkladov.
Nech je Σ - abeceda na popis príkladov.
Napríklad, $\Sigma = \{0, 1\}$, $\Sigma = R$.

Množina Σ^n predstavuje množinu všetkých slov v abecede dĺžky n a množina Σ^* predstavuje množinu všetkých slov v abecede Σ .

Definícia:

Nech $X \subseteq \Sigma^*$. Koncept v abecede Σ je funkcia c, $c : X \to \{0, 1\}$. Množina X sa nazýva **priestor príkladov**. *Prvok x, $x \in X$ sa nazýva príklad. Ak pre $x \in X$ platí $c(x) = 1$, tak x je* **pozitívny príklad**, *ak platí $c(x) = 0$, tak x je* **negatívny príklad**.

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

Koncepty

Zjednotenie množiny kladných a záporných príkladov je definičný obor funkcie c. Teda za predpokladu, že definičný obor je známy, c určuje a je určované množinou svojich pozitívnych príkladov.

Príklady:

- Koncept parita
  $\Sigma = \{0, 1\}$, $p : \Sigma^* \to \{0, 1\}$, $x = x_1 \dots x_n$:

$$p(x) = \begin{cases} 1 & \text{ak v } x \text{ je párny počet jedničiek} \\ 0 & \text{ak v } x \text{ je nepárny počet jedničiek} \end{cases}$$

  1011101 záporný príklad, 1000001 kladný príklad,

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

Koncepty

- Koncept palindróm
  $\Sigma = \{0,1\}$, $p : \Sigma^* \to \{0,1\}$, $x = x_1 \dots x_n$:

$$p(x) = \begin{cases} 1 & \text{ak } x_i = x_{n-i+1} \quad i = 1, 2, \dots, \lfloor \frac{n}{2} \rfloor \\ 0 & \text{inak} \end{cases}$$

- Koncept n-rozmerná jednotková guľa
  $\Sigma = R$, $u : \Sigma^n \to \{0,1\}$, $x = x_1 \dots x_n$:

$$u(x) = \begin{cases} 1 & \text{ak } {x_1}^2 + {x_2}^2 + \dots + {x_n}^2 \leq 1 \\ 0 & \text{inak} \end{cases}$$

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

Množiny konceptov

1. **Množina konceptov odvodených z reálneho sveta**,

   ktorá je predkladaná na rozpoznanie (učenie). Táto množina môže obsahovať koncepty ako "písmeno A", "písmeno B", . . . , z ktorých každý môže byť zakódovaný.

   Každý koncept má svoje množiny kladných a záporných príkladov.

   Keď je množina konceptov určovaná týmto spôsobom, budeme pre ňu používať výraz **konceptový priestor**.

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

Množiny hypotéz

2. **Množina hypotéz povolených pre daný koncept v rámci učenia**.

   Na predchádzajúcom obrázku je to množina výstupov, ktorú stroj M je schopný určiť. Napríklad, stroj M sa môže preradiť do jedného alebo do druhého stavu a v danom stave bude klasifikovať niektoré vstupy ako kladné (výstup 1) a zvyšok ako záporné (výstup 0). Jeho výstupmi je určená hypotéza.

   Teda stroj M dáva výstup, ktorý môžeme chápať ako **hypotézu** pre danú množinu príkladov.

   Množina všetkých hypotéz, ktoré stroj M určuje, bude nazývaná **hypotézový priestor**.

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

Učiaci proces

Cieľom učiaceho procesu je nájsť hypotézu, ktorá v určitom zmysle zodpovedá konceptu z konceptového priestoru vzhľadom na vyššie uvedené úvahy. Detaily, kedy a ako toto môže byť urobené sú ústredným cieľom tejto prednášky.

Máme teda 2 množiny (východiskové a určené strojom M):

- C - konceptový priestor,
- H - hypotezový priestor,

a **problémom** je ku každému $c, c \in C$, nájsť nejaké $h, h \in H$, ktoré je dobrou aproximáciou pre c.

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

V reálnych situáciách sú koncepty tvorené na základe určitých informácií, ktoré neprinášajú explicitnú definíciu c. My budeme predpokladať, že táto informácia je daná postupnosťou kladných a záporných príkladov z X.

V praxi obyčajne nemáme dostatok zdrojov na to, aby sme mohli vybudovať veľmi veľký stroj pre nájdenie c, nemáme dostatok času na to, aby bol vytvorený a spustený program, ktorý by určil, že $h = c$, alebo že h je tak blízko c, ako si prajeme.

V praxi sú kladené obmedzenia na zdroje a my budeme preto hľadať hypotézu h, ktorá "pravdepodobne" reprezentuje c (aproximuje c) v nejakom definovanom zmysle.

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

Nech $X \subseteq \Sigma^*$ je príkladový priestor.
$\Sigma = \{0, 1\}$ alebo $\Sigma = R$.

Vzorka dĺžky m je postupnosť m príkladov, t. j. je to $m-$tica $\overline{x} = (x_1, x_2, \ldots, x_m) \in X^m$, kde x_i sú príklady.

Postupnosť môže obsahovať rovnaké príklady viackrát. Niekedy budeme predpokladať, že sú rôzne bez ujmy na všeobecnosti.

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

Tréningová vzorka $\bar{s}$ je množina príkladov $(X \times \{0,1\})^m$, t. j. $\bar{s} = ((x_1, b_1), (x_2, b_2), \ldots, (x_m, b_m))$.
Hodnota b_i vyjadruje, či príklad je kladný alebo záporný.

Budeme predpokladať, že vo vzorke nie sú žiadne sporné príklady, t.j. ak

$$x_i = x_j \quad \Rightarrow \quad b_i = b_j.$$

Teda existuje funkcia s, definovaná ako $s(x_i) = b_i \quad (1 \leq i \leq m)$.
Budeme hovoriť, že $\bar{s}$ je tréningová vzorka pre cieľový koncept t, ak platí $b_i = t(x_i)$, pre $1 \leq i \leq m$.

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

Príklady

Tréningová vzorka pre koncept "palindróm" je
$((0010, 0), (1001001001, 1), (111, 1), (010101, 0), (111101, 0))$

Cieľový koncept t: $x = (x^1 \dots x^n)$:

$$t(x) = \begin{cases} 1, & \text{ak } x^i = x^{n-i+1} \quad \text{pre } 1 \leq i \leq n \\ 0, & \text{inak} \end{cases}$$

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

Učiaci proces

Uvažujme teraz o povahe učiaceho procesu, ktorý tu chceme študovať. Majme daný konceptový priestor C a hypotézový priestor H v abecede Σ.

Definícia:

Učiaci algoritmus pre (C, H), *niekedy nazývaný (C, H)-učiaci algoritmus, je procedúra, ktorá akceptuje tréningové vzorky pre koncepty v C a výstupy zodpovedajú hypotézam v H.*

Aby táto procedúra mohla byť považovaná za algoritmus, musí byť efektívna. Ak ignorujeme problém efektívnosti, tak (C, H)-učiaci algoritmus je funkcia L, ktorá priradí ľubovoľnej tréningovej vzorke $\overline{s}$ pre cieľový koncept $t \in C$ funkciu $h \in H$. Píšeme $h = L(\overline{s})$ alebo $h = L(\overline{s}, t)$, ak chceme zdôrazniť 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

Učiaci proces

Predpokladáme, že $L(\overline{s})$ je definovaná na celom príkladovom priestore X, pričom $\overline{s}$ je konečná podmnožina $E_{\overline{s}} \subseteq C$, lebo zahŕňa len príklady vzorky $(x_1, ..., x_m)$. Teda učenie sa koná na podmnožine X.

Hypotéza $h \in H$ je konzistentná s tréningovou vzorkou $\overline{s}$ alebo súhlasí s $\overline{s}$, ak $h(x_i) = b_i$, pre $1 \leq i \leq m$.

Ak hypotéza $L(\overline{s})$ je konzistentná s každou tréningovou vzorkou $\overline{s}$ platí, hovoríme že učiaci algoritmus L je **konzistentný**. V tomto prípade je funkcia $L(\overline{s})$ ako rozširujúca funkcia pre $\overline{s}$, tzn. je definovaná na celom príkladovom priestore X.
Vo všeobecnosti nemusí platiť, že hypotéza $L(\overline{s})$ je konzistentná s každou tréningovou vzorkou $\overline{s}$.

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

Učiaci proces

Vo všeobecnosti neplatí, že každé rozšírenie tréningovej vzorky bude vhodným zovšeobecnením, pretože cieľový koncept je len parciálne definovaný príkladmi vzorky.
A tiež tréningová vzorka môže byť nereprezentatívna, alebo zavádzajúca.

Príklad: Ak vhodne zakódujeme všetky zvieratá a cieľový koncept je "mačka", tak sa môže stať, že tréningová vzorka pozostáva z bezchvostových mačiek.

V praxi musíme predpokladať, že nereprezentatívne vzorky sú nepravdepodobné a že väčšina vzoriek je dostatočne reprezentatívna, takže rozšírenia funkcií sú vyhovujúce.

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

Uvažujme koncept "kreslo". Kreslo je možné popísať pomocou zakodóvaných príkladov

(4 rohy,	chvost,	sedací priestor,	farba,	žije)	b_i
(1	0	1	0	0),	1
(1	0	1	1	0),	1
(1	0	0	1	1),	0

Hypotézový priestor by mohol byť - množina všetkých Booleovských funkcií.
Konceptový priestor by mohol byť - množina položiek nábytku (zakoódovaná).

Učiaci proces spočíva v hľadaní takej Booleovskej funkcie, ktorá dáva správne odpovede pre všetky príklady.

Napríklad: $4rohy \wedge \neg chvost \wedge sedacipriestor \wedge \neg zije$

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:
(1 1 1 0 0 0 1, 1)
(1 0 1 1 0 1 1, 1)
(1 1 0 1 0 0 0, 0)
(1 1 1 0 1 1 0, 1)
(1 1 1 1 1 1 1, 1)
(1 1 1 1 0 0 1, 0)

Nájdite učiaci algoritmus, ktorý bude hľadať hypotézu v tvare Booleovskej funkcie.
Algoritmus musí fungovať na ľubovoľnej tréningovej vzorke.

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

Booleovské funkcie, Valiant 1984

Monočlen je konjunkcia literálov (premenných alebo ich negácií). Začíname bez informácií, t.j. predpokladáme výskyt všetkých $2n$ literálov v cieľovej hypotéze

$$h_U : \quad u_1 \bar{u}_1 u_2 \bar{u}_2 \ldots u_n \bar{u}_n, \quad U = \{u_1, \bar{u}_1, u_2, \bar{u}_2, \ldots, u_n, \bar{u}_n\}$$

Každý pozitívny príklad $y = y_1 \ldots y_n$ umožňuje odstránenie tých literálov u_j, pre ktoré $y_j = 0$ a tie literály $\bar{u}_j$, pre ktoré $y_j = 1$. Predpokladajme, že $\bar{s}$ je tréningová vzorka

$$\bar{s} = ((x_1, b_1), \ldots (x_m, b_m)), \quad x_i = ((x_i)_1 (x_i)_2 \ldots (x_i)_n), \quad 1 \leq i \leq m$$

$h_U \ldots$ monočlenná funkcia obsahujúca literály v množine U.

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

Algoritmus:

begin
set $U = \{u_1, \bar{u}_1, \ldots, u_n, \bar{u}_n\}$;
for i:=1 to m do
    if b_i=1 then
        for j:=1 to n do
            if $(x_i)_j = 1$ then delete $\bar{u}_j$ from U
                else delete u_j from U;
$L(\bar{s}) = h_U$;
end;

Tento algoritmus sa nazýva štandardný učiaci algoritmus pre monočleny.

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

Redukovaná tréningová vzorka vznikne z pôvodnej tréningovej vzorky vynechaním tých položiek v tréningových príkladoch, ktoré neovplyvnia hodnotu h_U, t.j. im odpovedajúce premenné (ani ich negácie) sa v h_U nevyskytujú.

Veta:
Štandardný učiaci algoritmus pre monočleny je konzistentný s redukovanou tréningovou vzorkou, ak je bezosporná.

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

Dôkaz: To znamená, že výsledná neprázdna formula je konzistentná s redukovanou tréningovou vzorkou.
V každom kroku algoritmu je odstránených niekoľko (možno žiadny) literálov. Literály, ktoré sú vo výslednej hypotéze h_U, neboli odstránené.

Ľubovoľný negatívny príklad pre h_U je spôsobený niektorým literálom v h_U, ktorý nebol odstránený. Teda všetky negatívne príklady pre h_U (a tiež tie vo vzorke) sú korektne klasifikované pomocou $L(\overline{s})$.

Ak $h_U(x) = 1$ a $V \subset U$, tak $h_V(x) = 1$. Po každej prezentácii kladného príkladu x, mazacia procedúra zaručí, že $h_U(x) = 1$ a teda klasifikácia x je korektná.
Teda koncová hypotéza $L(\overline{s})$ korektne klasifikuje všetky pozitívne 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

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

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 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 disjunkcií

Úlohy:

1. Aký je počet kladných príkladov konceptu "palindróm", keď príkladový priestor je $\{0,1\}^n$?
2. Nech w je nasledujúci koncept: $\{0,1\}^n$, $x \in \{0,1\}^n$, $x = x_1 \dots x_n$:

$$w(x) = \begin{cases} 1 & \text{ak x obsahuje najviac 2 jedničky} \\ 0 & \text{inak} \end{cases}$$

   Ukážte, že počet kladných príkladov v tomto koncepte je kvadratickou funkciou n.
3. Napíšte postupnosť hypotéz generovaných algoritmom učenia monočlenov, keď na vstupe je prezentovaná tréningová vzorka $(11100101, 1), (00100011, 0), (11001001, 1)$. Ak cieľový koncept je $\langle u_2 \overline{u}_4 u_8 \rangle$, doplňte príklady do vzorky, ktoré sú pre to nutné.
4. Dokážte, že počet funkcií $f : \{0,1\}^n \to \{0,1\}$ je 2^{2^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