Výpočetní teorie strojového učení

Koncept = pojem

- Pojmy často **vysvětlujeme** prostřednictvím příkladů: dobrý politik, krásný člověk, nudná přednáška, ..
- *Definice potřebujeme kvůli vzájemné komunikaci*
- Mějme definiční obor **X** popsaný jako množinu všech možných instancí objektů, kde objekt je charakterizován hodnotami svých **atributů.**
- **Koncept** na množině objektů **X** je definován tak, že každý prvek **X** je označen
  - buď jako příklad daného pojmu,
  - nebo o něm víme, že uvažovanému pojmu **neodpovídá** (nereprezentuje jej).

  Koncept vlastně představuje nějakou podmnožinu **X**, její charakteristická funkce **c** se někdy označuje jako **cílová funkce (target f.)**.

  Obecně může být **c** libovolná bool.funkce nad X, tj. **c: X → {0,1}**

Hypotéza pro klasifikované příklady

- Naučit se koncept z příkladů znamená vlastně naučit se booleovskou funkci **c** (na celém definičním oboru **X**) na základě příkladů jejího chování na nějaké podmnožině **X,** tedy na základě znalostí o chování na množině **Y** ⊂ **X**.
- Nechť **D** označuje aktuální množinu **trénovacích příkladů** chování ve tvaru **<x, c(x) >,** což znamená, že **D** je rozdělena na pozitivní a negativní příklady
- **Hypotéza** = pokus o popis cílového konceptu.
- **H** nechť označuje množinu všech možných *přípustných hypotéz*. Obecně hypotéza **h** může být rovněž boolovská funkce nad **X**, tj. **h: X → {0,1}** . Jinými slovy hypotéza také reprezentuje podmnožinu **X.**

Cíl učení: najít hypotézu, která je **korektní** pro všechny příklady z **X**, tj. platí **h(x) = c(x)** pro všechna **x** ∈ **X.**

Induktivní bias

Věta o ošklivém kačátku pro konečný def.obor**.** Nechť ***D*** je klasifikovaná trénovací množina pro koncept ***K***, který tvoří podmnožinu *konečného* definičního oboru **X** všech myslitelných možností, které lze vyjádřit v uvažovaném jazyce pro popis trénovacích dat.

Pokud ***D*** ⊆ **X** a ***D*** ≠ **X**, pak pro každý prvek ***y*** ∈ **X** \ ***D*** platí, že pravděpodobnost tvrzení „***y*** patří ke konceptu ***K***“ je rovna **0,5.** Tedy všechny objekty z množiny ***D*** \ ***E*** mají stejnou pravděpodobnost, že do konceptu ***K*** patří (i že do něj nepatří).

Cílem strojového učení je najít **korektní popis konceptu** na základě omezeného počtu příkladů (množina výrazně menší než **X**). ***Lze toho vůbec dosáhnout***? *Za jakých podmínek a jak postupovat*?

- Využití doménové znalosti pro volbu **tvaru hypotéz**, která zaručí, že algorimus učení nalezne hypotézu, která je korektní = **induktivní bias** (předpojatost).

Charakteristika prostoru (množiny) hypotéz

- **Příklad:** Pojem „*střední postava*“.

  Zde hypotézou může být libovolný ***obdélník rovnoběžný s osami***, tj. podinterval intervalu „*výška* X *váha*“, tj. **<150, 200>** X **< 50, 120 >.**

  Taková hypotéza nemůže popsat libovolnou podmnožinu def.oboru. Musíme mít tedy dobrý důvod pro volbu takového omezení!
- Říkáme, že **množina hypotéz *H* plně pokrývá** (shatters) **množinu dat obsahující *M* bodů**, pokud pro libovolnou z **2^M** možných klasifikačních úloh na uvažované množině dat existuje korektní hypotéza ***h*** z ***H***, která tuto úlohu řeší (dokonale odděluje poz. a neg. příklady v uvažované úloze).
- **Kapacitu** dané množiny hypotéz ***H*** charakterizuje maximální počet bodů, které mohou být touto množinou hypotéz ***H*** pokryty. Tato hodnota se také označuje jako **Vapnik-Chervonenkis** (VC) **dimense** množiny hypotéz ***H.***
- Např. *obdélník rovnoběžný s osami* v rovině může oddělit 4 body, které nejsou v přímce. VC takové množiny hypotéz je 4. To je velmi málo. Je takový prostor hypotéz vůbec užitečný?

Co se lze naučit, je-li def.obor nekonečný?

Platí obdoba „Věty o ošklivém káčátku“ i v případě definičního oboru, který **není** konečný?

Co když uvažujeme *nekonečný* definiční obor, např. N (přirozená čísla)?

V takovém případě dokonce musí existovat koncepty, které nelze popsat pomocí předem zvoleného jazyka pro popis hypotéz!

Důvod? Množina **C** všech konceptů (= množina všech podmnožin N) je ***výrazně mohutnější*** než množina **H** hypotéz (= mn. formulí ve zvoleném jazyce)!

Přesněji: neexistuje žádné vzájemně jednoznačné zobrazení mezi **C** a **H**

Důsledky věty o ošklivém kačátku

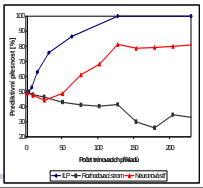
Upřesnění cíle strojového učení:

Nemá smysl pátrat po *přesně správné hypotéze*, ale hledáme

skoro správnou hypotézu (approximately correct), která splňuje *doplňkové požadavky* – Occamova břitva, bias, …, vhodný kompromis mezi paměťovými nároky pro reprezentaci konceptu a mezi pravděpodobností chybné klasifikace.

Lze nějak charakterizovat chování známé hypotézy pro cílový koncept, jehož popis pro celý definiční obor není k dispozici?

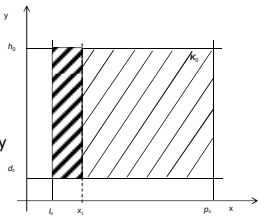
„Křivka učení“ a testovací data.



Algoritmus A, který se skoro přesně učí pojem „středně velký objekt“:

Pro všechny vstupující pozitivní příklady sledujte hodnoty MAX_i a MIN_i ve všech atributech *i* pro dosud přečtené vzorky. Hypotéza nechť je pak interval $\langle MIN_1, MAX_1 \rangle \times \langle MIN_2, MAX_2 \rangle$

Tvrzení: K tomu, aby se tento algoritmus naučil hypotézu skoro přesně, stačí mu prohlédnout $4/\varepsilon * \ln(4/\delta)$ příkladů.



Důkaz:

Pro navržený algoritmus **A** platí „všechny generované hypotézy jsou konzistentní, tj. pro libovolnou navrženou hypotézu **h** platí, **h** ⊆ **k**“, kde **k** je cíl.koncept.

Tuto vlastnost budeme dále označovat jako ***Inkluze***.

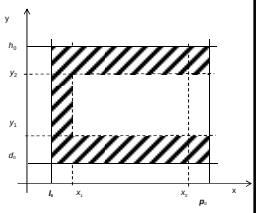
Důsledek *Inkluze:*

Pro libovolnou hypotézu **h** navrženou algoritmem **A** platí, že pravděpodobnost jevu **„nový objekt bude splňovat h (tj. spadne do odpovídajícího intervalu)“** je menší než tato pravděpodobnost pro cílový koncept **k = <l,p> x < d,h>, ozn. P(k).**

Jaká je pravděpodobnost toho, že nový objekt **x** bude špatně klasifikován?

Jistě platí, že **P(x: h(x) ≠ k(x)) < P(k).**

- Nechť **P(k) < ε**, pak jistě i pravděpodobnost chyby je menší než ε.
- Nechť **P(k) > ε** .



Zvolme x_0 tak, že $x_0 = \inf\{x: P(< l, x > x < d, h >) > \varepsilon/4\}$
Jistě platí $P(< l, x_0 > x < d, h >) > \varepsilon/4$,
a tedy pravděpodobnost, že vzorek padne mimo levý sloupec je menší než **(1- ε)/4**

Odhad počtu trénovacích příkladů pro koncept vyjádřený intervalem

Pro **m** různých vzorků platí:

pravděpodobnost toho, že každý vzorek padne mimo levý sloupec, je menší než

$$(1- \varepsilon /4)^m$$

Tentýž postup lze použít i pro ostatní rohy:

Pravděpodobnost jevu „žádný prvek z trénovací množiny obsahující **m** vzorků nepadnul do rohu “ je menší než

$$4\,(1- \varepsilon /4)^m < \delta$$

$$4\,(1- \varepsilon /4)^m \cong 4\, e^{-m * \varepsilon /4} < \delta$$

$$m > 4/\varepsilon \ln(4/\delta)$$

Východiska výpočetní teorie PAC stroj.učení

Výchozí předpoklad - **stacionarita**: Trénovací i testovací množina jsou vybírány z téže populace za použití totožné distribuce pravděpodobnosti

(Probably Aproximately Correct) **PAC učení**: Kolik trénovacích příkladů je třeba, aby se podařilo eliminovat všechny velmi špatné hypotézy?

X množina všech možných příkladů (s distribucí **D**)

f skutečný popis konceptu

H množina všech možných hypotéz, **h ∈ H** je aktuální hypotéza

er(h) = pravděpodobnost jevu „ **x ∈ X** a platí **f(x) ≠ h(x)**“
= P({**x**: **x ∈ X** s distribucí **D** a platí **f(x) ≠ h(x)** })
tuto hodnotu studují „křivky učení“

Hypotéza **h** je **ε-skoro správná**, pokud **er(h) < ε**

Základní otázka PAC

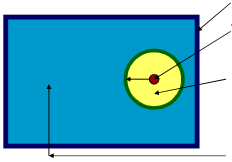
Buď **m** mohutnost trénovací množiny

Můžeme určit **m** *tak, aby pouhá konsistence hypotézy s trénovací množinou byla dostatečnou zárukou toho, že jsme našli skoro správnou hypotézu?*

Takový odhad může sloužit např. jako vodítko při shromažďování či posuzování trénovacích dat

V dalším předpokládáme, že hledaný popis konceptu **f** je prvkem uvažované množiny hypotéz **H** (tedy, že existuje korektní, tj.konsistentní a úplná hypotéza pro daný koncept)

Základní pojmy PAC



H konečný prostor všech hypotéz

f (hledaný) cílový koncept z mn. **H**

H_ε je **ε** okolí **f**, tj. množina obsahující pouze ty hypotézy, jejichž pravděpodobnost chyby je **< ε**

$H_{\varepsilon\text{-špatná}} = H - H_\varepsilon$

obsahuje právě ty hypotézy, jejichž pravděpodobnost chyby je větší než **ε**

Pokusme se odhadnout, za jakých okolností platí, že

P(„nalezena hypotéza **konzistentní** se všemi trén. příklady a současně z $H_{\varepsilon\text{-špatná}}$ **“) < δ**

Odhad potřebného počtu příkladů

Tvrzení: Nechť **h** je hypotéza konzistentní se všemi trénovacími příklady.

Pokud platí $P(\mathbf{h} \in H_{\varepsilon\text{-špatná}}) < \delta$, pak pravděpodobnost toho, že „**h** je **ε-skoro správná**“ je větší než **(1- δ).**

Zdůvodnění:

Předpokládáme, že v **H** *existuje nějaká hypotéza konzistentní se všemi trénovacími příklady.* Jistě platí

$$P(\mathbf{h} \in \mathbf{H} - H_{\varepsilon\text{-špatná}}) + P(\mathbf{h} \in H_{\varepsilon\text{-špatná}}) = 1$$

Víme, že $H_\varepsilon = H - H_{\varepsilon\text{-špatná}}$, a proto

$$P(\mathbf{h} \in H_\varepsilon) \geq (1-\delta).$$

$\mathbf{h} \in H_\varepsilon$ znamená, že **h** klasifikuje prvky z **X** s chybou menší než **ε**, tj. **h** je **ε-skoro správná**

Kdy pro h konzistentní s trén. daty platí $P(h \in H_{\varepsilon\text{-špatná}}) < \delta$?

Nechť **b** je libovolná opravdu špatná hypotéza, tedy h., jejíž pravděpodobnost chyby **er(b)** = pravděpodobnost jevu „ **x ∈ X** a platí **f(x) ≠ h(x)**“ **> ε**. V tomto případě je $\mathbf{b} \in H_{\varepsilon\text{-špatná}}$ *a platí:*

- Pravděpodobnost opačného jevu „**b** *správně* klasifikuje jeden náhodně zvolený příklad“ je P(**b** správně klasifikuje 1 zvolený příklad) ≤ (1- **ε**)
- Pravděpodobnost, že **b** správně klasifikuje **m** náhodně zvol. příkladů: P(**b** správně klasifikuje **m** zvolených příkladů) ≤ $(1-\varepsilon)^m$

Pravděpodobnost, že *existuje* prvek z $H_{\varepsilon\text{-špatná}} = \{b_1,\ldots, b_k\}$, který správně klasifikuje **m** zvol. příkladů, je rovna pravděpodobnosti, že „b_1 správně klasifikuje **m** zvolených příkladů“ **nebo** …**nebo** „b_k správně klasifikuje **m** zvolených příkladů“, tj.

$$P(\mathbf{h} \in H_{\varepsilon\text{-špatná}}) \leq \sum_{i \leq k} (1-\varepsilon)^m = |H_{\varepsilon\text{-špatná}}| * (1-\varepsilon)^m \leq |H| * (1-\varepsilon)^m$$

Pokud $|H| * (1-\varepsilon)^m < \delta$, pak jistě $P(\mathbf{h} \in H_{\varepsilon\text{-špatná}}) < \delta$

Jak lze zajistit, aby $P(h \in H_\varepsilon) \geq 1-\delta$?

Stačí, aby platilo $P(h \in H_{\varepsilon\text{-špatná}}) < \delta$

To dosáhneme, když $|H| * (1-\varepsilon)^m < \delta$

Fakt: Pro $|\varepsilon| < 1$, platí, že $(1-\varepsilon)^m \cong e^{-\varepsilon * m}$

Podmínku lze tedy přepsat do tvaru

$$\ln(|H| * e^{-\varepsilon * m}) < \ln \delta \quad \textit{čili} \quad \ln|H| - \varepsilon * m < \ln \delta$$

Postačující podmínka pro počet příkladů **m** je tedy

$$\mathbf{m} \geq (\ln|H| - \ln \delta) / \varepsilon = 1/\varepsilon * (\ln|H| + \ln(1/\delta))$$

Máme-li k dispozici alespoň $1/\varepsilon * (\ln|H| + \ln(1/\delta))$ příkladů a učící algoritmus navrhuje hypotézu **h**, která je se všemi těmito příklady konzistentní, pak pravděpodobnost toho, že „chyba **h** je menší než **ε** (**h** je **ε-skoro správná**)“ je větší než (**1- δ**).

Praktické použití odhadu počtu příkladů

$m \geq 1/\varepsilon * (\ln|H| + \ln(1/\delta))$

Nechť $\mathbf{H_B(n)}$ je *množina všech bool. funkcí pro* **n** bool. atributů, tj. zobrazení z n-tic skládajících se z 0 a1 do { 0, 1}.

- ***Velikost definičního oboru***: $\mathbf{2^n}$
- Počet funkcí z množiny mohutnosti **a** do { 0, 1}: $\mathbf{2^a}$.

Tedy mohutnost $\mathbf{H_B(n)}$ je 2^{2^n}

Postačující podmínka pro počet příkladů $\mathbf{m_B(n)}$, které potřebujeme k tomu, abychom se skoro správně naučili koncept popsaný booleovskou funkcí o **n** atributech, je

$$\mathbf{m_B(n)} \geq 1/\varepsilon * (2^n + \lg(1/\delta)),$$

Důsledek:

Máme-li se skoro správně naučit koncept popsaný **obecnou bool**. funkcí, pak potřebujeme více než $\mathbf{2^n}$ příkladů. Jinými slovy: **musíme znát celý definiční obor**. *Věta o ošklivém kačátku.*

Důsledek odhadu

Je-li H množina možných hypotéz, pak se lze skoro správně (s pravděpodobností větší než (1- δ)) naučit hypotézu, jejíž chyba je menší než **ε**, pokud máme **m** trénovacích příkladů a platí

$\mathbf{m} \geq 1/\varepsilon * (\ln|H| + \ln(1/\delta))$. **(i)**

Pozorování : **m** je funkcí **|H|**

Podaří-li se nám získat nějakou doplňkovou informaci (omezení na tvar přípustných hypotéz), která omezuje rozsah **H**, pak vystačíme s menším počtem trénovacích příkladů !!! Zde hraje významnou roli **doménová znalost**.

Pokusme se

- provést odhad mohutnosti množiny hypotéz pro některé běžné typy hypotéz (rozhodovací stromy,…)
- a zjistit vliv tohoto odhadu na požadovaný počet trénovacích příkladů

Věta o PAC učení rozhodovacího stromu

Nechť objekty jsou charakterizovány pomocí n binárních atributů a nechť připouštíme jen hypotézy ve tvaru rozhodovacího stromu s maximální délkou větve k. Dále nechť δ, ε jsou malá pevně zvolená kladná čísla blízká 0. Pokud algoritmus strojového učení vygeneruje hypotézu φ, která je konzistentní se všemi m příklady trénovací množiny a platí

$$m \geq m_{k\text{-DT}}(n) \geq c\,(n^k + \ln(1/\delta))/\varepsilon$$

pak φ je ε-skoro správná hypotéza s pravděpodobností větší než $(1-\delta)$, t.j. chyba hypotézy φ na celém definičním oboru konceptu je menší než ε s pravděpodobností větší než $(1-\delta)$.

Doporučená literatura:

Mařík et al.: ***UI(3)*** – kapitola 7 Teorie složitosti a úlohy UI (Demlová, Štěpánková)

Odhad mohutnosti množiny hypotéz pro rozhodovací seznam (lin. reprezentace stromu)

L_n jazyk obsahující přesně **n** binárních atributů

Ω def. obor uvažovaného konceptu má tedy 2^n různých prvků

Rozhodovací seznam (decision list) v jazyce L_n je uspořádaný seznam $\mathfrak{R}=[\,t_1{:}c_1, \ldots, t_m{:}c_m]$, kde

- t_i je test vyjádřený ve tvaru konjunkce literálů z L_n
- $c_i \in \{0,1\}$ je přiřazená klasifikace.

Nechť $o \in \Omega$, pak $\mathfrak{R}(o) = c_i$, kde t_i je první test, který objekt o splňuje (tj. $t_1(o)=0, \ldots, t_{i-1}(o)=0, t_i(o)=1$).

Pokud $t_k(o)=0$ pro všechna $k \leq m$, pak $\mathfrak{R}(o) = 0$

Každý strom hloubky **n** *(= délka nejdelší větve) nebo formuli v disjunktivní normální formě lze napsat jako rozhodovací seznam v jazyce L_n: např.*

$(s_1 \,\&\, s_3 \,\&\, \text{not } s_3)$ **v** $(\text{not } s_1 \,\&\, s_3 \,\&\, \text{not } s_6)$ *odpovídá*

$[(s_1 \,\&\, s_3 \,\&\, \text{not } s_3){:}1, (\text{not } s_1 \,\&\, s_3 \,\&\, \text{not } s_6){:}1]$

Odhad mohutnosti *k*-DL(*n*)

Nechť ***k*-DL(*n*)** je množina všech rozhodovacích seznamů, jejichž testy mají přípustnou délku omezenou pevně zvoleným číslem $k < n$. *Jak ovlivní volba* ***k*** *mohutnost množiny hypotéz?*

***Odhad mohutnosti prostoru hypotéz pro rozhodovací seznamy* 1-DL(*n*)**

$|1\text{-DL}(n)| <$ počet permutací z n (tedy $n!$) krát 3^n.

Zdůvodnění: Z 1 pevně zvolené permutace n prvků lze totiž vytvořit 3^n různých rozhodovacích seznamů (test je použit s výsledkem **1**, **0** nebo *„nezařazen"*).

$$|1\text{-DL}(n)| < n! \cdot 3^n$$

Protože $\ln(n!) < n \cdot \ln n$, *platí*

$$\ln|1\text{-DL}(n)| < \ln(n! \cdot 3^n) < O(n \cdot \ln n) + n \ln 3$$

Skoro správného učení lze dosáhnout při rozumném počtu trén. příkladů

$m \geq 1/\varepsilon \cdot (\ln|H| + \ln(1/\delta))$: pro 1-**DL(*n*)** jde o číslo srovnatelné s $(n \cdot \lg n)$, což je výrazně méně než mohutnost 2^n celého uvažovaného definičního oboru

Odhad mohutnosti k-DL(*n*)

Conj(n,k): počet různých konjunkcí *nejvýše* ***k*** literálů sestrojených z ***n*** atributů.

Conj0(n,j): počet všech konjunkcí *přesně* ***j*** literálů sestrojených z ***n*** atributů (pomocná veličina pro odhad ***Conj(n,k)***).

Postupujeme takto

$Conj0(n,j) < 2^j \cdot n^j = (2n)^j$ člen vyjadřující „znaménko" atomu

$Conj(n,k) < \sum_{i=k+1} Conj0(n,i)$

$$< \sum_{i=k+1} (2n)^i = 2n\,(2^k n^k - 1)/(2n - 1) \approx O(n^k) \qquad \text{(ii)}$$

Horní odhad pro počet prvků ***k*-DL(*n*)**: Rozhodovací seznam je vlastně uspořádaná posloupnost neopakujících se prvků z ***Conj(n,k)***, z nichž každý je klasifikován jednou z hodnot $\{0,1,\times\}$, kde „×" chápeme tak, že daná konjunkce v rozhodovacím seznamu není. Zřejmě tedy

$$\text{k-DL(n)} < 3^{|\text{Conj(j,k)}|} \cdot |\text{Conj(j,k)}|!$$

Odhady mohutnosti k-DL(*n*) a |k-DL(*n*)|

Víme, že $|\text{k-DL(n)}| < 3^{|\text{Conj(j,k)}|} \cdot |\text{Conj(j,k)}|!$ *Z toho plyne, že*

$$\ln|\text{k-DL(n)}| < |\text{Conj(j,k)}| \cdot \ln 3 + \ln(\text{Conj(j,k)})!$$

Použitím vztahu $\lg n! < n \cdot \lg n$ *dostáváme*

$$\ln|\text{k-DL(n)}| < |\text{Conj(j,k)}| \cdot (\ln 3 + \ln|\text{Conj(j,k)}|)$$
$$< O(n^k) \cdot (\ln 3 + \ln O(n^k)) \approx O(n^k \ln(n^k))$$

Po dosazení do vzorce $m \geq 1/\varepsilon \cdot (\ln|H| + \ln(1/\delta))$ *dostáváme odhad pro hypotézy ve tvaru rozhodovacího listu*

$$m_{k\text{-DL}}(n) \geq c/\varepsilon \cdot (O(n^k \ln n^k) + (1/\delta))$$

Pro rozhodovací stromy s omezenou hloubkou je odhad ještě poněkud nižší, protože pro mohutnost prostoru hypotéz platí

$$|\text{k-DT(n)}| < 3^{|\text{Conj(j,k)}|}$$

Odpovídající počet trén. příkladů je $m_{k\text{-DT}}(n) \geq c/\varepsilon \cdot (n^k + (1/\delta))$