Pravděpodobně skoro správné (PAC) učení

Výpočetní teorie strojového učení

Věta o ošklivém kačátku. Nechť E je klasifikovaná trénovací množina pro koncept K, který tvoří podmnožinu *konečného* definičního oboru $\mathbf{D}$ všech myslitelných možností, které lze vyjádřit v uvažovaném jazyce pro popis trénovacích dat.

Pokud $E \subseteq \mathbf{D}$ a $E \neq \mathbf{D}$, pak pro každý prvek $g \in \mathbf{D} \setminus E$ platí, že pravděpodobnost tvrzení „g patří ke konceptu K“ je rovna 0,5. Tedy všechny objekty z množiny $\mathbf{D} \setminus E$ mají stejnou pravděpodobnost, že do konceptu K patří (i že do něj nepatří).

Cílem strojového učení není hledání *přesně správné hypotézy*, ale hledání **skoro správné hypotézy** (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.

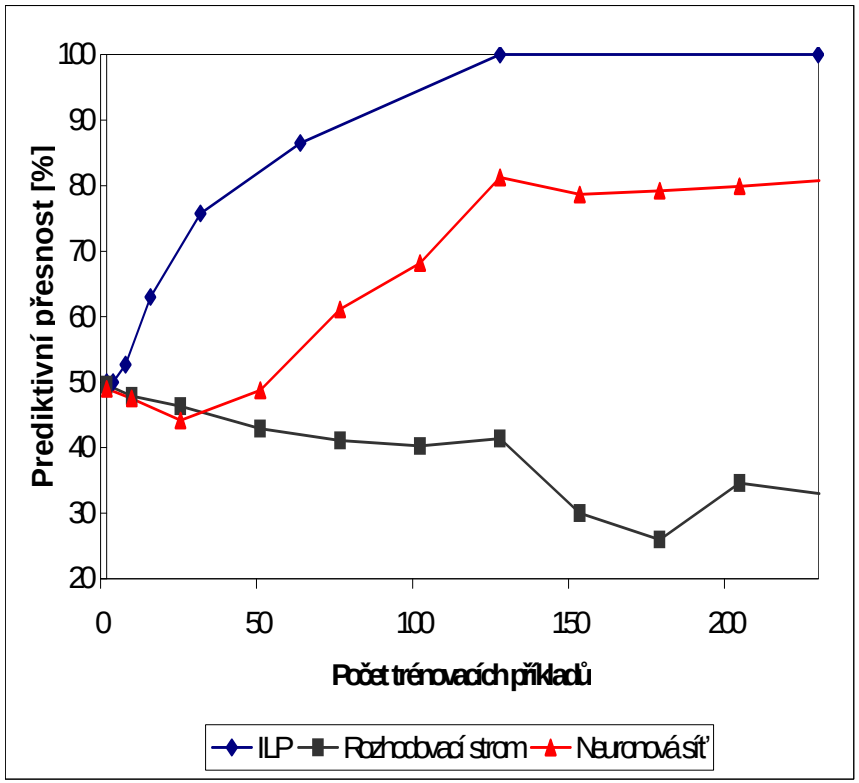
Křivka učení

Jak se křivka učení generuje?

- množina všech instancí se náhodně rozdělí postupně na 0%, 10%, 20% … 90% **trénovacích** a zbytek **testovacích** příkladů,
- křivka znázorňuje prediktivní přesnost – tj. ***výkonnost nad testovacími*** (neznámými daty), výsledky se průměrují přes více náhodných běhů.

Umožňuje hodnotit ***schopnost algoritmu naučit se*** *daný koncept*

Křivka učení pro úlohu „*parita*“ v prostoru $X = \{0,1\}^8$ (256 instancí)



Východiska výpočetní teorie stroj.učení PAC

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** $\in$ **H** je aktuální hypotéza

er(h) = pravděpodobnost jevu „ **x** $\in$ **X** a platí **f(x)** $\neq$ **h(**x**)**“

= P({**x**: **x** $\in$ **X** s distribucí **D** a platí **f(x)** $\neq$ **h(x)** })

tuto hodnotu studují „křivky učení“

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

Základní otázka PAC

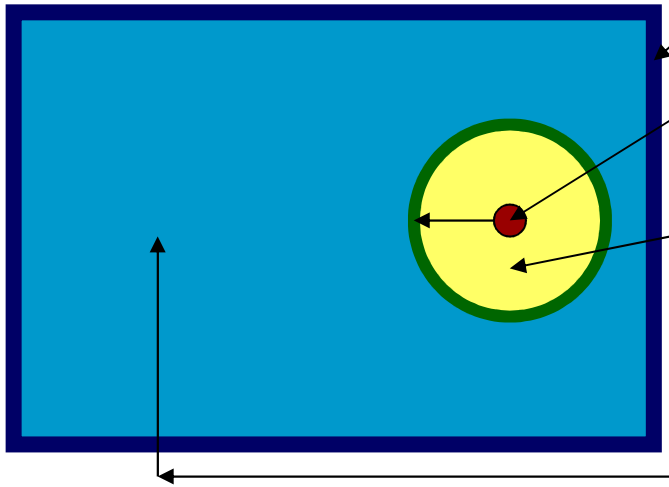
Buď **m** mohutnost trénovací množiny

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 v uvažované množině hypotéz **H** (tedy existuje korektní a úplná hypotéza pro daný koncept)

Základní pojmy PAC



H prostor všech hypotéz

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

$\mathbf{H}_{\varepsilon}$ je **ε** okolí **f**, tj. množina obsahující pouze ty hypotézy, jejichž pravděpodobnost chyby je < **ε**

$\mathbf{H}_{\varepsilon\text{-špatná}} = \mathbf{H} - \mathbf{H}_{\varepsilon}$

obsahuje právě ty hypotézy, jejichž prst chyby je větší než ε

Pokusme se odhadnout, za jakých okolností platí, že
P(**hypotéza** konzistentní se všemi učicími příklady je z $\mathbf{H}_{\varepsilon\text{-špatná}}$) < 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

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

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**

Za jakých okolností můžeme zajistit, aby pro **h** konzistentní s trén. daty platilo $P(\mathbf{h} \in H_{\varepsilon\text{-špatná}}) < \delta$? *

Nechť **b** je libovolná opravdu špatná hypotéza. Jde tedy o h., tž. **er(b)** = pravděpodobnost jevu „ **x** ∈ **X** a platí **f(x) ≠ h(**x**)**“ > **ε**, tj. **b** ∈ $H_{\varepsilon\text{-špatná}}$. *Platí:*

- Pravděpodobnost, že **b** *správně* klasifikuje 1 zvolený příklad: P(**b** správně klasifikuje 1 zvolený příklad) ≤ (1- **ε**)
- Pravděpodobnost, že **b** správně klasifikuje **m** 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 pravděpod., že „$\mathbf{b_1}$ správně klasifikuje m zvolených příkladů“ nebo …nebo „$\mathbf{b_k}$ správně klasifikuje m zvolených příkladů“, tj.

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

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

Za jakých okolností můžeme zajistit, aby $P(\mathbf{h} \in \mathbf{H}_{\varepsilon\text{-špatná}}) < \delta$? *

Stačí, aby platilo $|\mathbf{H}|*(1-\varepsilon)^m < \delta$.

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

Podmínku lze tedy přepsat do tvaru

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

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

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

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

Praktické použití odhadu počtu příkladů
$\mathbf{m \geq = 1/\varepsilon * (\ln |H| + \ln (1/\delta))}$

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 $\mathbf{2^{2^n}}$

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

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

Tedy: 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 **m ≥ 1/ ε *(**ln **|H|** + ln (1/δ)). **(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

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 & not s_3) v (not s_1 & s_3 & not s_6) *odpovídá*

[(s_1 & s_3 & not s_3):1, (not s_1 & s_3 & 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 H, *je-li prostor hypotéz* 1-DL(*n*)

|1-DL(*n*)| < počet permutací z ***n*** (tedy ***n!***) krát 3^n.

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

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

Protože ln (**n!**) < ***n*** ∗ ln ***n*****,** *platí*

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

Skoro správného učení lze dosáhnout při počtu trénovacích příkladů

m ≥ = 1/ ε *(ln |H| + ln(1/δ)), což je pro 1-DL(*n*) číslo srovnatelné s (*n* ∗ lg *n*), to je výrazně méně něž 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

$Conj0(n,j) < 2^j * n^j = (2n)^j$ člen vyjadřující „znaménko“ atomu

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

$< \sum_{i<k+1} (2n)^i = 2n(2^{k-1} n^{k-1} - 1)/ (2n - 1) \approx O(n^k)$ (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, × }, kde „×“ chápeme tak, že daná konjunkce v rozhodovacím seznamu není. Zřejmě tedy

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

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

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

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

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

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

Po dosazení do vzorce **m ≥ 1/ ε *(**ln **|H|** + ln (1/δ)) *dostáváme odhad pro hypotézy ve tvaru rozhodovacího listu*

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

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

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

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

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 $(1-\delta)$.