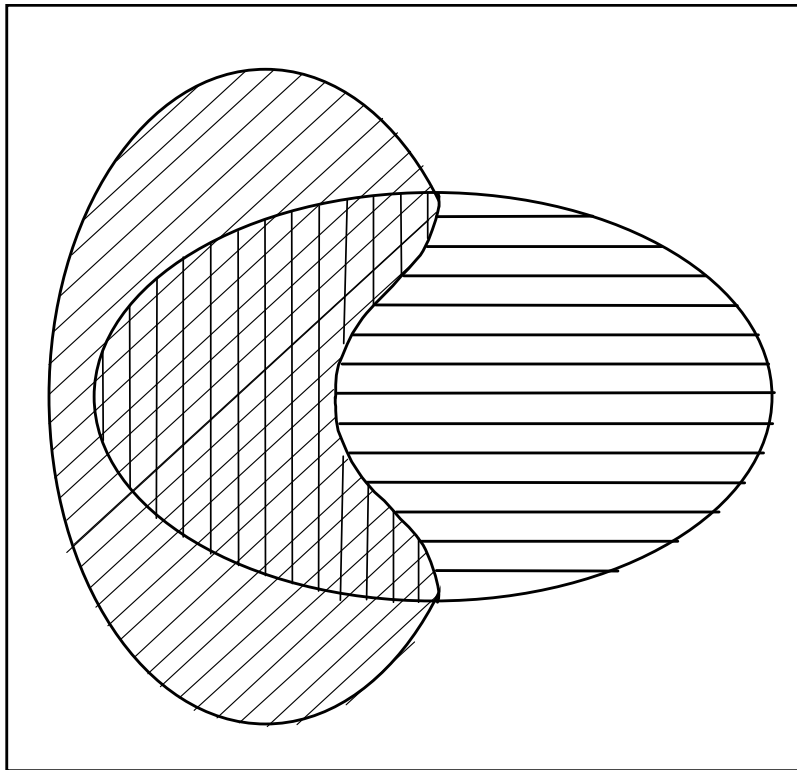
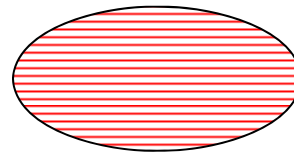


Výpočetní teorie strojového učení a pravděpodobně skoro správné (PAC) učení

Cíl induktivního strojového učení



Na základě omezeného **vzorku příkladů** $E = E^+ \cup E^-$, charakterizovat (popsat) zamýšlenou skupinu objektů (**koncept**) tak, aby **navržený popis** (hypotéza)



- pro prvky E z co nejlépe odpovídal právě jen prvkům reprezentujícím koncept, tj. prvkům z E^+
- a byl použitelný pro určení příslušnosti ke konceptu i pro objekty mimo E

Koncept a hypotéza

- **Koncept** = reálně existující třída objektů vymezená na známém definičním oboru (např. 3D vektory reálných čísel). Objekty klasifikujeme podle toho, zda ke konceptu patří nebo nepatří. Obvykle známe jen *konečnou množinu příkladů* prvků daného konceptu (*pozitivní příklady*), případně i množinu prvků, které ke konceptu nepatří (*negativní příklady*), t.j. *konečnou trénovací množinu konceptu*.
- **Trénovací množina** konceptu *nenabízí* definici konceptu, ale *popisuje koncept implicitně* (t.j. pomocí příkladů).
- **Hypotéza** = pokus o formální charakterizaci konceptu, např. formule ve výrokové logice (pravidla) nebo množina aritmetických výrazů, která má být splněna. *Hypotéza je pokus o explicitní popis konceptu*.

Které hypotézy hledáme?

Hypotéza je

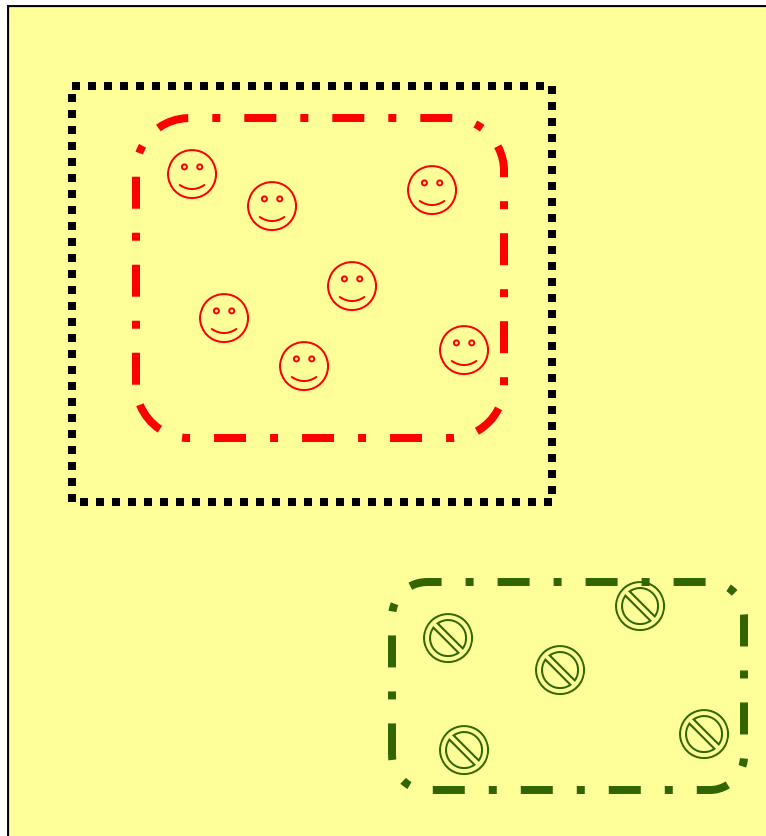
- *korektní*, když nepokrývá žádný negativní příklad.
- *úplná*, když pokrývá všechny pozitivní příklady.

Hledáme **konsistentní (= korektní + úplnou) hypotézu** pro daná trénovací data.

Trénovací data by měla mít stejnou distribuci jako všechna data o uvažované úloze. Tento požadavek má být zárukou toho, že hypotéza se bude chovat nad daty, která nebyla v trénovací množině, podobně jako nad trénovací množinou, tj. **bude „většinou správně“ přiřazovat klasifikaci** .

Korektní hypotéza

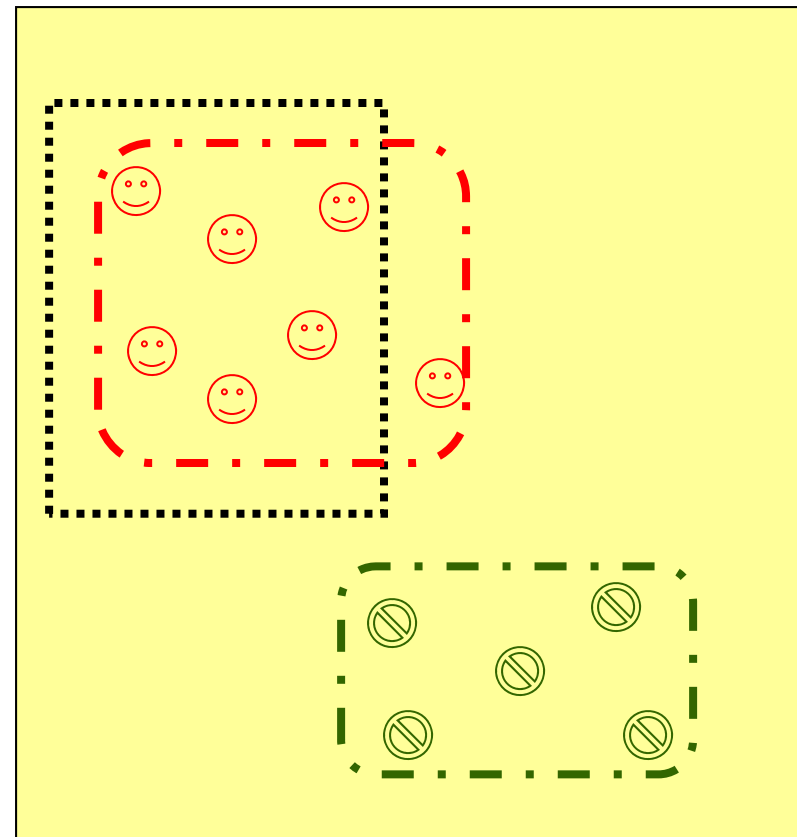
úplná (complete)



pro

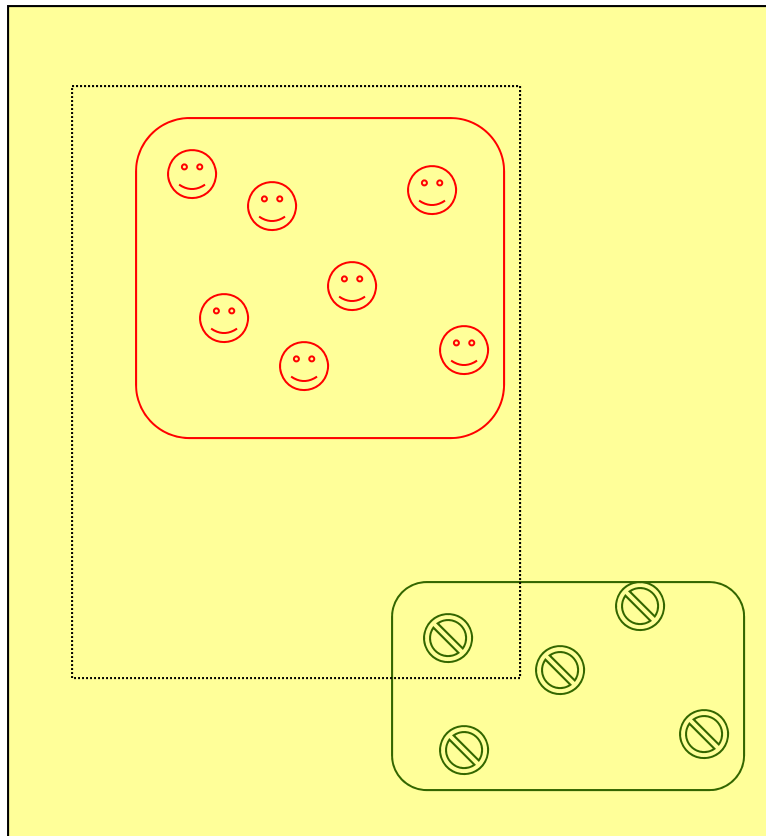
Trénovací množinu

neúplná (incomplete)

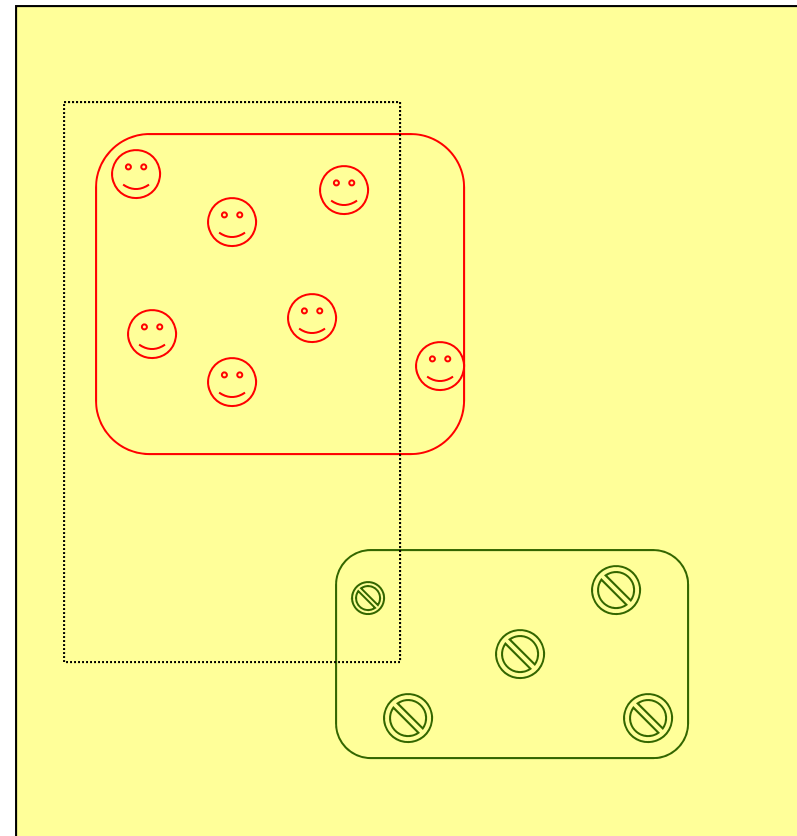


Nekorektní hypotéza

úplná (complete)



neúplná (incomplete)



Výpočetní teorie strojového učení

Věta o ošklivém kačátku. Necht' E je klasifikovaná trénovací množina pro koncept K , který tvoří podmnožinu *konečného* definičního oboru D všech myslitelných instancí objektů, které lze popsat v uvažovaném jazyce pro popis trénovacích dat.

Pokud $E \subseteq D$ a $E \neq D$, pak pro každý prvek $g \in D \setminus E$ platí, že

- počet korektních a úplných hypotéz pro E , pro které „ g patří ke konceptu K “
- počet korektních a úplných hypotéz pro E , pro které platí opak, tedy „ g nepatří ke konceptu K “

jsou úplně stejné. Tedy libovolný objekt z množiny $D \setminus E$ má stejnou pravděpodobnost, že do konceptu K patří (i že do něj nepatří).

Situace se zásadně změní, jakmile jako hypotéza nemůže sloužit libovolná podmnožina definičního oboru! Necht' například pro objekty v rovině uvažujeme jen hypotézy odpovídající konvexním tvarům. Pak existuje řada objektů z $D \setminus E$ pro které si můžeme být jisti jejich klasifikací!

Výpočetní teorie strojového učení

Přijetí nějakých omezení na typ hledaných hypotéz tedy zvyšuje šanci na to, že budeme schopni správně rozhodovat o neznámých objektech.

V takovém případě je přirozené, že přestaneme striktně trvat na tom, že hledáme **korektní a úplné hypotézy!**

Cílem strojového učení pak už není hledání *přesně správné hypotézy*, ale hledání **skoro správné hypotézy** (approximately correct), která splňuje *doplňkové požadavky*, např. je dostatečně jednoduchá, má předem stanovený tvar (bias) , ...

Hledáme vhodný kompromis mezi paměťovými nároky pro reprezentaci konceptu a mezi pravděpodobnostmi chybné klasifikace.

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?

D definiční obor s distribucí δ

X \subseteq **D** množina všech možných příkladů (s distribucí δ)

f skutečný popis konceptu

H množina všech možných hypotéz, **h** \in **H** je aktuální hypotéza

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

= $P(\{ \mathbf{x}: \mathbf{x} \in \mathbf{X} \text{ s distribucí } \delta \text{ a platí } \mathbf{f}(\mathbf{x}) \neq \mathbf{h}(\mathbf{x}) \})$

tuto hodnotu studují „křivky učení“

er(h) zkrácený zápis **er_D(h)**

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

Základní otázka PAC

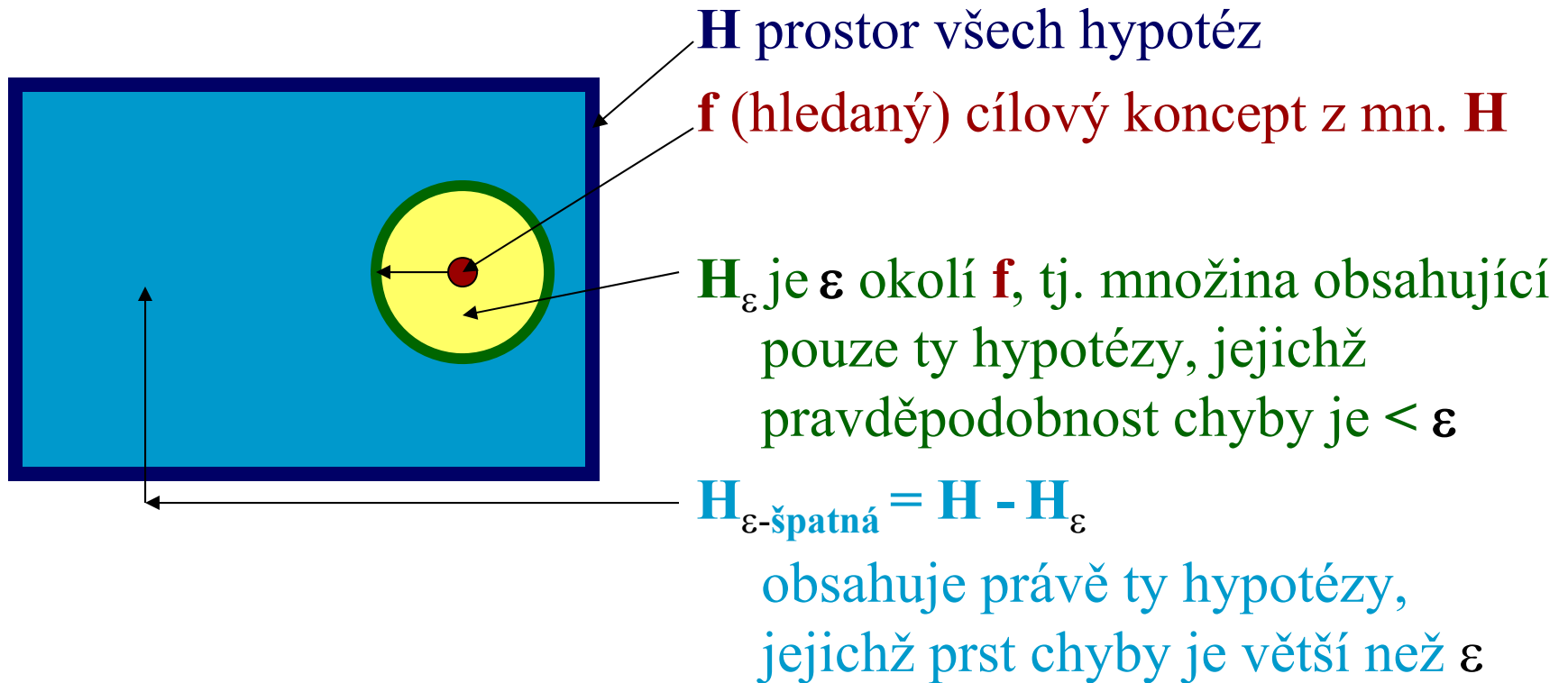
Bud' m mohutnost trénovací množiny

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



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

Algoritmus **M** učící se pojem „středně velký objekt“, jehož definice má tvar $\text{Interval}_1 \times \text{Interval}_2$

Pro všechny vstupující **pozitivní příklady** sledujte hodnoty MAX_i a MIN_i ve všech attributech i pro dosud přečtené vzorky. Hypotéza necht' je pak interval $\langle \text{MIN}_1, \text{MAX}_1 \rangle \times \langle \text{MIN}_2, \text{MAX}_2 \rangle$

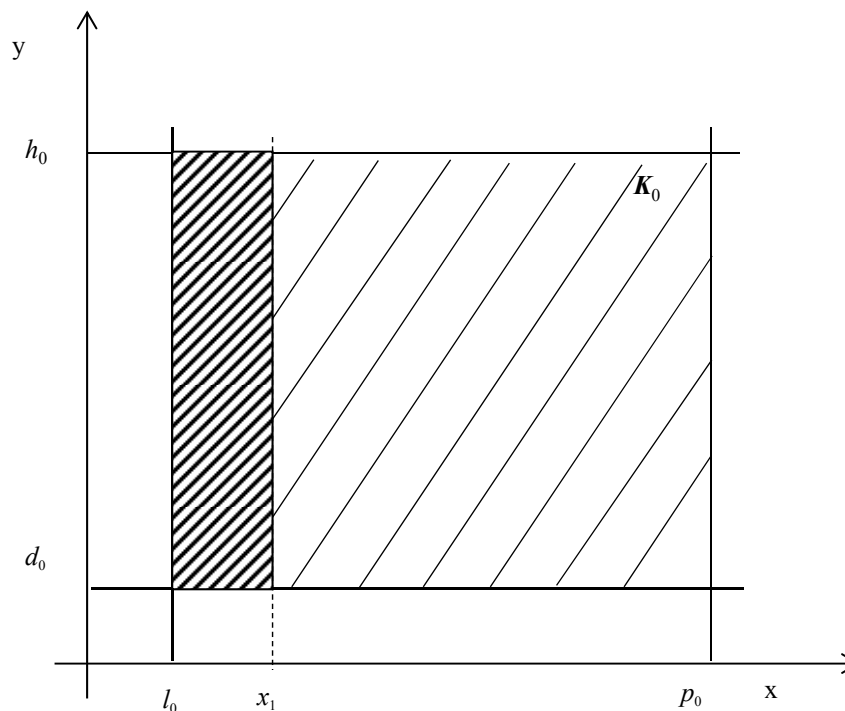
Tvrzení:

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

Důkaz:

Pro navržený algoritmus **M** platí „všechny generované hypotézy jsou konzistentní, tj. pro libovolnou navrženou hypotézu h platí, $h \subseteq k$ “.

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



Důsledek *Inkluze*:

Pro libovolnou hypotézu h navrženou algoritmem M platí, že pravděpodobnost jevu „nový objekt bude splňovat h (tj. spadne do odpovídajícího intervalu)“ je menší než tato pravděpodobnost pro cílový koncept $k = \langle l, p \rangle \times \langle d, h \rangle$, ozn. $P(k)$.

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

Jistě platí, že $P(x: h(x) \neq k(x)) < P(k)$.

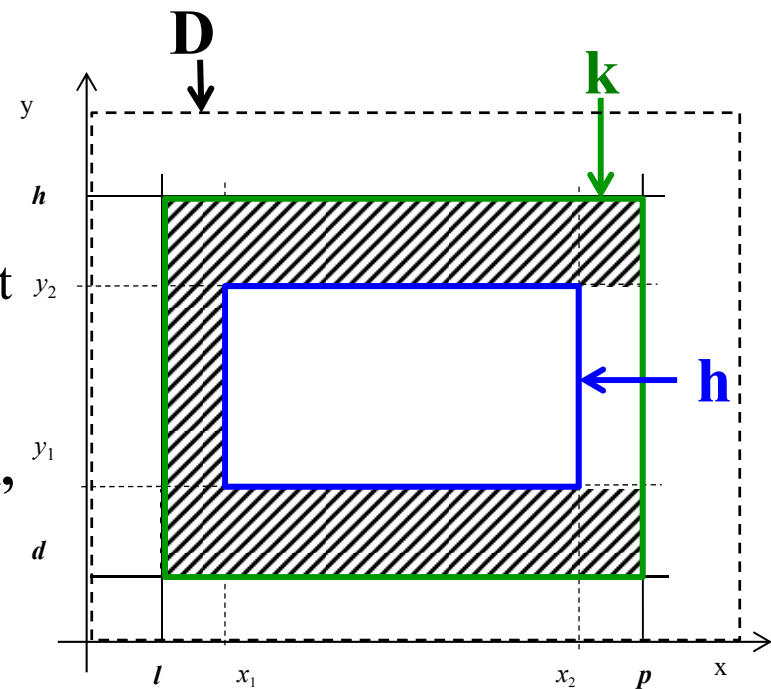
Neboť $P(x: h(x) \neq k(x)) = P(x: \neg h(x) \ \& \ k(x)) =$
 $= P(x: k(x) \ \& \ h(x) \neq k(x)) < P(k)$.

- Necht' $P(k) < \varepsilon$, pak jistě i pravděpodobnost chyby je menší než ε .
- Necht' $P(k) > \varepsilon$. Zvolme x_0 tak, aby se vzorek nedostal do bezprostřední blízkosti l , tj. padne mimo levý sloupec

$$x_0 = \inf\{x: P(\langle l, x \rangle \times \langle d, h \rangle) > \varepsilon/4\}$$

Jistě platí $P(\langle l, x_0 \rangle \times \langle d, h \rangle) > \varepsilon/4$,

a tedy pravděpodobnost, že vzorek padne mimo levý sloupec je menší než $(1 - \varepsilon/4)$



Odhad počtu trénovacích příkladů pro koncept vyjádřený intervalem (např. „středně velký objekt“)

Pro m různých vzorků platí:

Pravděpodobnost toho, že žádný z m vzorků se nedostane do bezprostřední blízkosti levého rohu (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

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

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

$$\mathbf{m > 4 / \varepsilon \ln(4 / \delta)}$$

Odhad potřebného počtu příkladů – obecný případ *

Tvrzení: Necht' h je hypotéza konzistentní se všemi trénovacími příklady.

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

Zdůvodnění:

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

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

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

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

$h \in H_{\varepsilon}$ 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 \mathbf{h} konzistentní s trén. daty platilo $P(\mathbf{h} \in \mathbf{H}_{\varepsilon\text{-špatná}}) < \delta$?

*

Necht' \mathbf{b} je libovolná opravdu špatná hypotéza. Jde tedy o h., tž. $\text{er}(\mathbf{b}) = \text{pravděpodobnost jevu „}\mathbf{x} \in \mathbf{X} \text{ a platí } \mathbf{f}(\mathbf{x}) \neq \mathbf{b}(\mathbf{x})\text{“} > \varepsilon$, tj. $\mathbf{b} \in \mathbf{H}_{\varepsilon\text{-špatná}}$. Platí:

- Pravděpodobnost, že \mathbf{b} *správně* klasifikuje 1 zvolený příklad: $P(\mathbf{b} \text{ správně klasifikuje 1 zvolený příklad}) \leq (1 - \varepsilon)$
- Pravděpodobnost, že \mathbf{b} *správně* klasifikuje m zvol. příkladů: $P(\mathbf{b} \text{ správně klasifikuje } m \text{ zvolených příkladů}) \leq (1 - \varepsilon)^m$

Pravděpodobnost, že *existuje* prvek z $\mathbf{H}_{\varepsilon\text{-špatná}} = \{\mathbf{b}_1, \dots, \mathbf{b}_k\}$, 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.

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

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

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

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

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

Proč dáváme přednost jednoduchým hypotézám?

Argument : Jednoduchých hypotéz je výrazně méně než složitých. Proto, pokud data odpovídají některé z jednoduchých h. , pak asi nejde o „náhodný jev“

Occamova břitva :

Nejlepší hypotéza je ta nejjednodušší, která odpovídá datům.

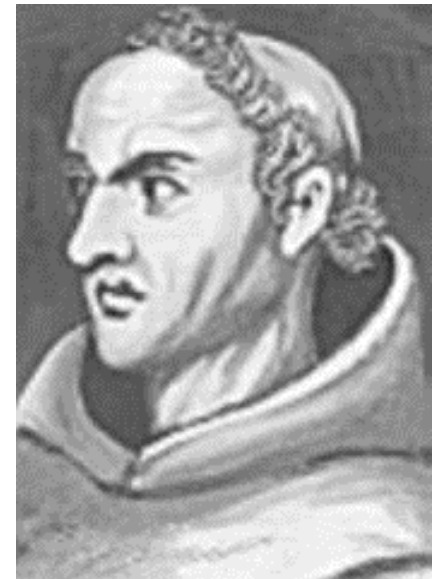
Související problémy:

- proč zrovna **tato** malá množina?
- pozor na použitý jazyk!

William of Ockham, born in the village of Ockham in Surrey (England) about 1285, was the most influential philosopher of the 14th century and a controversial theologian.

PAC učení

witten & eibe



Praktické použití odhadu počtu příkladů

$$m \geq \frac{1}{\epsilon} * (\ln |H| + \ln (1/\delta))$$

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

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

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

Postačující podmínka pro počet příkladů $m_B(n)$, které potřebujeme k tomu, abychom se skoro správně naučili koncept popsany booleovskou funkcí o n attributech, je

$$m_B(n) \geq \frac{c}{\epsilon} * (2^n + \lg (1/\delta)),$$

*Tedy: máme-li se skoro správně naučit koncept popsany **obecnou bool. funkcí**, pak potřebujeme více než 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 $(1 - \delta)$) naučit hypotézu, jejíž chyba je menší než ε , pokud máme m trénovacích příkladů a platí $m \geq 1/\varepsilon * (\ln |H| + \ln (1/\delta))$. (i)

Pozorování: m je funkcí $|H|$

Podarí-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 $\mathfrak{R} = [t_1:c_1, \dots, t_m:c_m]$, 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. $t_1(o)=0, \dots, t_{i-1}(o)=0, t_i(o)=1$). 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.

$(s_1 \& s_3 \& \text{not } s_3) \vee (\text{not } s_1 \& s_3 \& \text{not } s_6)$ odpovídá

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

PAC učení

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)

$|1\text{-DL}(n)| <$ počet permutací z n (tedy $n!$) krát 5^n .

Z pevně zvolené permutace lze totiž vytvořit 5^n různých rozhodovacích seznamů (test je buď „nezařazen“ nebo je použit s nebo negace a s výsledkem 1, 0).

$$|1\text{-DL}(n)| < n! * 5^n$$

Protože $\ln(n!) < n * \ln n$, platí

$$\ln |1\text{-DL}(n)| < \ln(n! * 5^n) < O(n * \ln n) + n \ln 5$$

Skoro správného učení lze dosáhnout při počtu trénovacích příkladů

$m \geq \frac{1}{\epsilon} * (\ln |H| + \ln(1/\delta))$, což je pro 1-DL(n) číslo srovnatelné s $(n * \lg n)$.

Tato 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

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

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

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

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 „ \times “ chápeme tak, že daná konjunkce v rozhodovacím seznamu není. Zřejmě tedy

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

Odhady mohutnosti $k\text{-DL}(n)$ a $|k\text{-DL}(n)|$ *

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

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

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

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

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

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

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

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

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

Věta o PAC učení rozhodovacího stromu

Nechť objekty jsou charakterizovány pomocí n binárních atributů a necht' připouštíme jen hypotézy ve tvaru rozhodovacího stromu s maximální délkou větve k . Dále necht' 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 $(1-\delta)$.

Některá zajímavá pozorování vyplývající z PAC učení

- **Je možné** naučit se koncept z omezené množiny trénovacích příkladů skoro správně i pokud možných hypotéz je nekonečně !
- Při učení se hypotéz z konečné množiny možných h . bývá počet potřebných příkladů funkcí n^k , kde n je počet atributů použitých v popisu příkladů a k charakterizuje složitost hypotézy (např. pro rozhodovací strom je k jeho maximální hloubka) - *lze dosáhnout snížení tohoto čísla?*
- *Změna reprezentace trénovacích příkladů. Používané metody:*
 - Snížení počtu atributů prostřednictvím nových **odvozených atributů**, které jsou funkcí nějaké skupiny původních atributů (NN, statistické řešení – Support Vector Machine)
 - Eliminace zbytečného zavádění nových atributů (např. při převodu multirelační úlohy na atributovou)
 - zjišťování **relevance**