Prohledávání prostoru hypotéz

abyste si představili balík modelů

PAC učení

(věta; PAC = pravděpodobně přibližně správně)

možná i BIC kritérium, VC dimenze

Prostor hypotéz

- **Hypotézy** formulujeme v určitém vyjadřovacím jazyce v našem případě konjunkce testů vstupních atritutů, které charakterizují hodnotu cílového atributu `Yes`
  - hypotézy jsou formátu $\langle ?, Cold, High, ?, ?, ? \rangle$, kde
  - znak na pozici odpovídá podmínce na odpovídající vstupní (ne–cílový) atribut
  - znakem je buď konkrétní hodnota atributu, znak ? nekladoucí žádnou podmínku na daný atribut, znak $\emptyset$ odpovídající nesplňitelné podmínce
- Pro binární atributy máme $4^{|\text{počet atributů}|}$ hypotéz, hypotézy obsahující $\emptyset$ jsou ekvivalentní, tj. máme $3^{|\text{počet atributů}|} + 1$.
- Budeme prohledávat systematicky.

- Prostor hypotéz je **částečně uspořádaný inkluzí** $h_1 >_g h_2$ hypotéza h_1 je **obecnější** než h_2 (píšeme $h_1 >_g h_2$), pokud každý příklad splňující h_2 splňuje i h_1. V tom případě se h_2 nazývá **specifičtější** než h_1.
- Např. $\langle ?, ?, \ldots, ? \rangle$ je obecnější než $\langle Sunny, ?, \ldots, Same \rangle$.
- **Nejobecnější hypotéza** je $\langle ?, ?, \ldots, ? \rangle$, tu splňují všechna data
- **maximálně specifická hypotéza** je $\langle \emptyset, \emptyset, \ldots, \emptyset \rangle$, kterou nesplňuje žádný záznam.
- Prostor všech hypotéz tvoří svaz, viz. obrázek na tabuli.

Ohodnocovací funkce

- **ohodnocovací funkce** určuje, nakolik hypotéza odpovídá datům.
- Hledáme takovou hypotézu, kterou by splňovaly všechny pozitivní příklady a nesplňoval žádný negativní příklad.
- tj. aby byla implikace $hypoteza \Rightarrow (EnjoySport = Yes)$ pro všechna data pravdivá
- (to lze, pokud máme data bez náhody a šumu).

Nalezení maximálně specifické hypotézy odpovídající datům

Algoritmus FIND–S

1. $h \leftarrow \langle \emptyset, \ldots, \emptyset \rangle$ max. specifická hypotéza
2. pro každý positivní příklad x v datech
   pro každou podmínku na atribut $A_i = a_i$ v h
   Pokud příklad x nesplňuje $A_i = a_i$
   nahraď podmínku nejbližší obecnější podmínkou,
   kterou x splňuje
   jinak nech h beze změny
3. vydej hypotézu h

Ale:

- Je hypotéza nalezená FIND–S jediná konzistentní s daty?
- Proč tedy volit ji, ne nějakou maximálně obecnou či něco mezi?
- V jiném prostoru hypotéz nemusí být ani maximálně specifická hypotéza jednoznačná.
- **Budeme hledat všechny hypotézy konzistentní s daty.**
- Pokud nejsou trénovací data konzistentní, máme problém. Řešení je jiný typ hypotéz a jiná ohodnocovací funkce.

Prostor verzí

Prostor verzí vzhledem k prostoru hypotéz H a trénovacích dat D je podmnožina hypotéz z H konzistentní s trénovacími daty D,

$$VS_{H,D} = \{h \in H | Consistent(h, D)\}$$

- Tento prostor může být charakterizován obecnou a specifickou hranicí; každá hypotéza mezi těmito hranicemi spadá do prostoru verzí.
- **Obecná hranice G** vzhledem k H a D je množina maximálně obecných hypotéz z H konzistentních s daty, tj.

$$\begin{aligned} G = \quad \{ \quad & g \in H | Consistent(g, D) \& \\ & (\neg \exists g^{|} \in H)[(g^{|} >_g g) \& Consistent(g^{|}, G)]\} \end{aligned}$$

- **Specifická hranice S** vzhledem k H a D je množina maximálně specifických hypotéz z H konzistentních s daty, tj.

$$S = \{ \; s \in H | Consistent(g, D) \& \\ (\neg \exists s^{|} \in H)[(s >_g s^{|}) \& Consistent(s^{|}, G)]\}$$

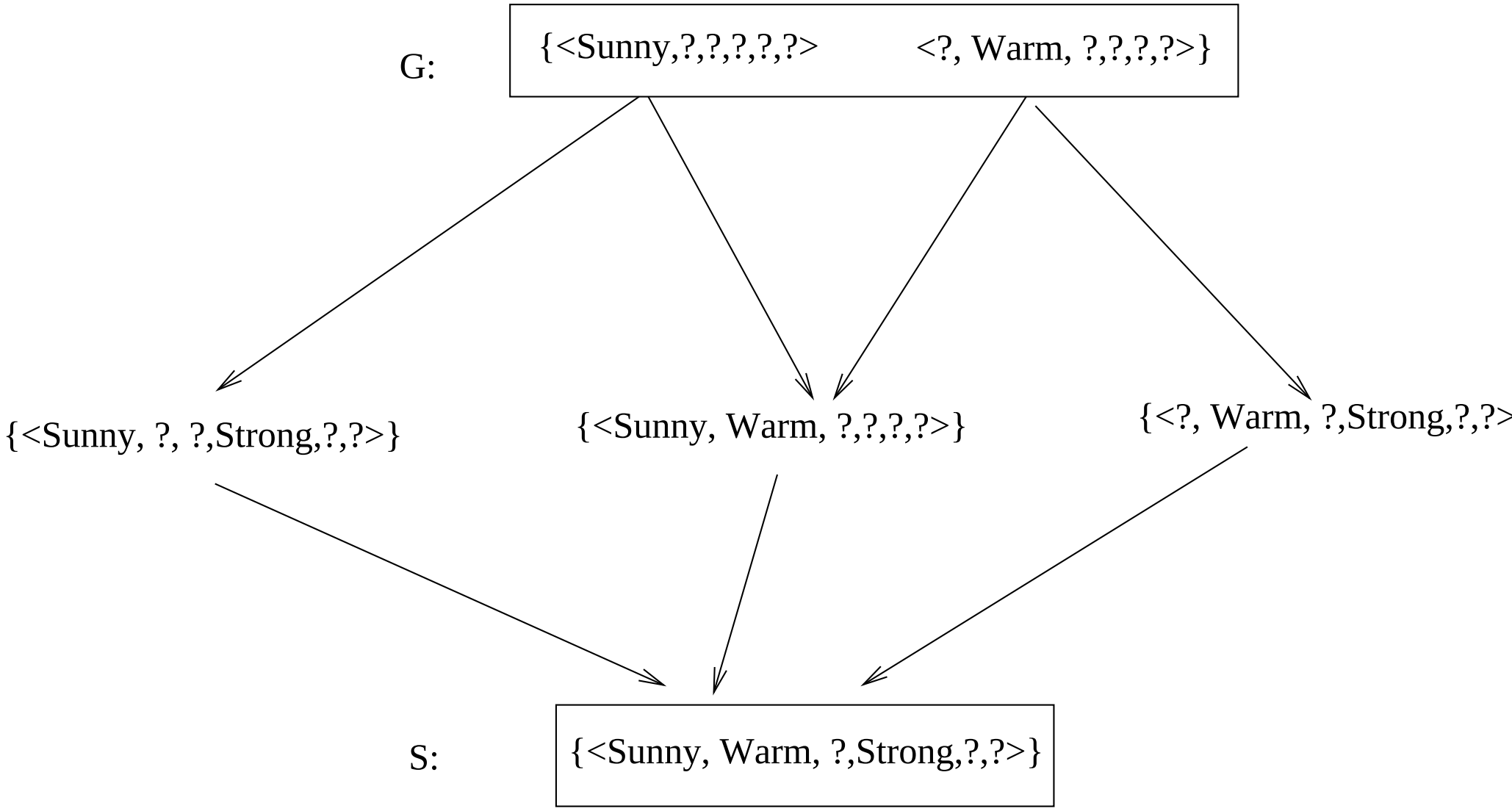

Figure 1: Prostor verzí s částečným uspořádáním inkluzí.

Algoritmus **Candidate–Elimination**

$G \leftarrow$ maximálně obecné hypotézy v H

$S \leftarrow$ maximálně specifické hypotézy v H

pokračuje

Pro každý trénovací příklad d, **do**

If d je positivní příklad

Odstaň z G všechny hypotézy nekonzistentní s d

For each $s \in S$, s nekonzistentní s d

Odstraň s z S

Přidej do S všechna h; minimální zobecnění s taková, že

h je konzistentní s d a zároveň $\exists g \in G; g >_g h$

Odstraň z S hypotézy, které nejsou maximálně specifické v S

If d je negativní příklad

Odstaň z **S** všechny hypotézy nekonzistentní s d

For each $g \in G$, g nekonzistentní s d

Odstraň g z G

Přidej do G všechna h; minimálně specifičtější než g taková, že

h je konzistentní s d a zároveň $\exists s \in S; h >_g s$

Odstraň z G hypotézy, které nejsou maximálně obecné v G

PAC – učení

(pravděpodobně přibližně správně)

- Za jakých podmínek je možné se úspěšně naučit?
- Jak u konkrétního algoritmu zaručíme, že se naučil dobře?
- chceme teoretické odpovědi; tj. jinak než z testovacích dat

Definice problému

- $\mathcal{X}$: množina datových příkladů, každý o p atributech
- cílový koncept $f(\langle x \rangle)$, pro každý příklad buď pravda, nebo nepravda
- $N = |\mathcal{X}|$ počet prvků v trénovací množině
- $\mathcal{D}$ pravděpodobnostní rozložení p atributů datových příkladů
- $\mathcal{H}$ prostor možných hypotéz
- Předpokládáme, že $f(x) \in \mathcal{H}$

[pravdivá chyba] Definujeme pravdivou chybu (true error)

$$trueError(h) = P(h(x) \neq f(x)|x \text{ náhodně zvoleno dle } \mathcal{D})$$

[přibližně správně] Hypotéza h je přibližně správná, pokud

$$trueError(h) \leq \epsilon$$

kde ϵ je malá konstanta.

Odhad počtu špatných hypotéz

Špatné hypotézy $h_b \in \mathcal{H}_{bad}$ jsou pro nás ty, které jsou konzistentní se všemi m příklady a zároveň $trueError(h_b) > \epsilon$.

- pravděpodobnost konzistence špatné hypotézy s jedním příkladem je $\leq (1 - \epsilon)$
- pravděpodobnost konzistence h_b se všemi m příklady je $\leq (1 - \epsilon)^m$
- tedy:

$$P(|\mathcal{H}_{bad}| \neq 0) \leq |\mathcal{H}_{bad}| \cdot (1 - \epsilon)^m \leq |\mathcal{H}| \cdot (1 - \epsilon)^m$$

- Protože $0 \leq \epsilon \leq 1$, tedy $(1 - \epsilon) \leq e^{-\epsilon}$, je

$$P(|\mathcal{H}_{bad}| \neq 0) \leq |\mathcal{H}| \cdot (1 - \epsilon)^m \leq |\mathcal{H}| \cdot e^{-\epsilon \cdot m}$$

- chceme, aby pravděpodobnost existence špatné hypotézy byla menší než δ, což určitě bude, pokud:

$$\begin{aligned} |\mathcal{H}| \cdot e^{-\epsilon \cdot m} &\leq \delta \\ e^{-\epsilon \cdot m} &\leq \frac{\delta}{|\mathcal{H}|} \\ ln(e^{-\epsilon \cdot m}) &\leq ln\frac{\delta}{|\mathcal{H}|} \\ m \cdot (-\epsilon) &\leq ln\frac{\delta}{|\mathcal{H}|} \\ m &\geq \frac{1}{\epsilon} ln\frac{|\mathcal{H}|}{\delta} \end{aligned}$$

$$N \geq \frac{1}{\epsilon} ln \frac{|\mathcal{H}|}{\delta}$$

proto N volíme

$$N \geq \frac{1}{\epsilon} (ln \frac{1}{\delta} + ln|\mathcal{H}|)$$

Příklad: Konjunkce literálů

- máme p dvouhodnotových atributů, pak každý literál může být v konjunkci ve třech verzích:
  - positivně
  - negativně
  - vůbec

  tj. $|\mathcal{H}| = 3^p$.
- dosazením dostaneme $N \geq \frac{1}{\epsilon}(p \cdot ln3 + ln(\frac{1}{\delta}))$
- pro $p = 10$, $\epsilon = 0.1$, $\delta = 5\%$ dostaneme $N \geq \frac{1}{0.1}(10ln3 + ln\frac{1}{0.05}) = 140$.

Vapnik–Chernovenkis dimenze (VC–dimenze)

[VC–dimenze] VC–dimenze třídy $\{f(x,\alpha)\}$ je definovaná jako největší možný počet bodů (v obecné konfiguraci), které je možné separovat prvky $\{f(x,\alpha)\}$. Například:

- přímky v rovině mají VC–dimenzi 3.
- třída $\{\sin(\alpha x)\}$ má nekonečnou VC–dimenzi (pro body z -1,1).

[VC–dimenze] Pro spojité funkce $\{g(x,\alpha)\}$ je VC–dimenze definovaná jako VC–dimenze třídy indikátorů$\{I(g(x,\alpha)-\beta)>0\}$, kde β jde přes hodnoty oboru hodnot g.

Počet příkladů a VC dimenze

- kolik příkladů na $1 - \epsilon$ dobrou hypotézu s pravděpodobností $1 - \delta$?

$$m \leq (4log(\frac{2}{\delta}) + 8VC(H)log(\frac{13}{\epsilon}))$$

VC–dimenze

Odhad testovací chyby Err z chyby na trénovacích datech $e\hat{r}r$, N příkladů, VC–dimenze h s pravděpodobností $1 - \eta$:

$$Err \leq e\hat{r}r + \frac{\epsilon}{2}(1 + \sqrt{1 + \frac{4 \cdot e\hat{r}r}{\epsilon}})$$

kde

$$\epsilon = a_1 \cdot \frac{h[log(a_2 N/h) + 1] - log(\eta/4)}{N}$$

kde a_1, a_2 jsou parametry bez doporučení jaké volit, $a_1 = 4$ a $a_2 = 2$ odpovídá nejhoršímu případu.

BIC kriterium

- mám–li nekonečný prostor modelů (parametrů), mám problém
- chci kritérium, který vybrat
- teoretické; testovací chybu chci zkombinovat se složitostí modelu
- kriteria AIC (Akaike), BIC (Bayesian IC), MDL (minimal description length)

BIC kriterium

- Mám modely (hypotézy) h_m, $m = 1, \ldots, M$
- v každém musím doladit parametry θ_m, např. spočítat průměr, regresní β, parametry BN atd.
- chci nejpravděpodobnější model při znalosti pozorovaných dat $DATA$, tj.

$$argmax_{h_m} P(h_m|DATA)$$

$$
\begin{aligned}
P(h_m|DATA) &= \frac{1}{P(DATA)} \cdot P(h_m) \cdot P(DATA|h_m) \\
&= \frac{P(h_m)}{P(DATA)} \int P(DATA|\theta_m, h_m) P(\theta_m|h_m) d\theta_m
\end{aligned}
$$

Integrál aproximujeme (Laplace aprox.) a trochu zjednodušíme a dostaneme:

$$
logP(DATA|h_m) = logP(DATA|\hat{\theta}_m, h_m) - \frac{d_m}{2} \cdot logN + O(1)
$$

- $\hat{\theta}_m$ maximálně věrohodný odhad θ_m,
  tj. maximalizující $P(DATA|\theta_m, h_m)$ pro pevná $DATA$ a h_m
- d_m počet parametrů modelu h_m

BIC kriterium

- místo maximalizace $P(DATA|h_m)$ budeme minimalizovat $-2 \cdot logP(DATA|h_m)$, tj.

$$BIC = -2 \cdot logP(DATA|\hat{\theta}_m, h_m) + d \cdot logN$$

- značíme $loglik = \sum_{i=1}^{N} logP_{\hat{\theta}}(y_i)$, dostaneme:
- $BIC = -2 \cdot loglik + d \cdot logN$
- $AIC = -2 \cdot loglik + d \cdot 2$
- MDL odpovídá BIC, jinak odvozeno

Kódování – úvod k MDL principu

- Chceme postat data z příjemci. Chceme kódovat tak, aby to zabralo co nejméně bitů.
- Možné zprávy jsou $z_1, \ldots, z_m$, my sdělujeme, která z alternativ nastala.
- Kódujeme binárně kódy délky $A = 2$.
- Jedno možné kódování:

Zpráva	z_1	z_2	z_3	z_4
Kód	0	10	110	111

žádný kód není prefixem jiného kódu, tj. příjemce ví, kdy je celý kód odeslán (omezíme se na takovéto kódy, instaneous prefix codes).

- kódy bychom mohli permutovat

- hodí se časté zprávy kódovat krátkými kódy – pak je průměrná délka zprávy kratší.
- **Shannon**: máme používat kódy délky $l_i = -log_2 P(z_i)$
- očekávaná délka zprávy pak bude

$$E(delka) \geq -\sum P(z_i) \cdot log_2(P(z_i)).$$

- Pravá strana nerovnice se nazývá Shannonova **entropie**.
- Pokud máme pravděpodobnosti $p_i = A^{-l_i}$, pak kód dosáhne dolní hranice.
- V našem případě pro $P(z_i) = \langle \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8} \rangle$.
- Obecně dolní hranici nedosáhneme, ale můžeme blízko (např. Huffmannovy kódy).
- Závěr: **Pro přenos hodnoty z s pravděpodobnostním**

rozložením $P(z)$ potřebujeme zhruba $-log_2 P(z)$ bitů informace.

Minimální délka zápisu

Minimal description length (MDL)

- Naučíme model M s parametry θ z dat $\mathbf{Z} = (\mathbf{X}, \mathbf{y})$.
- Pravděpodobnost předpovědi y modelem je $P(\mathbf{y}|\theta, M, \mathbf{X})$.
- Předpokládáme, že příjemce zná všechna X nového příkladu. Na poslání y potřebujeme:

$$\begin{aligned} delka &= -logP(\mathbf{y}|\theta, M, \mathbf{X}) - logP(\theta|M) \\ &= \sum_{i=1}^{|Z|} -logP(y|\theta, M, X) - logP(\theta|M) \end{aligned}$$

- v druhé části kódujeme parametry modelu, v první konkrétní y

kódované vzhledem k pravděpodobnostem předpovězeným modelem.

Pokud vezmeme jednoduchou třídu modelů, kde množina parametrů X bude prázdná a jen budeme učit číslo y, kde y budou náhodně rozložena dle normálního rozložení okolo neznámé střední hodnoty θ se známým rozptylem σ^2 a modely (tj. pravděpodobnosti střední hodnoty θ budou náhodně rozloženy dle $N(0,1)$, pak je délka jednoho příkladu:

$$delka = -log(\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y-\theta)^2}{2\sigma^2}}) - log(\frac{1}{\sqrt{2\pi}}e^{-\frac{(\theta)^2}{2}})$$

$$delka = const + log\sigma + \frac{(y-\theta)^2}{2\sigma^2} + \frac{\theta^2}{2}$$

Pozn. MDL princip se hodí, i na ohodnocení užitečnosti klastrování, kde jiné metody skoro nejsou (nejlépe ohodnotit použitelností v

praxi, samozřejmě).