



# Rozhodovací stromy a jejich konstrukce z dat

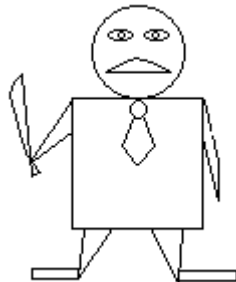
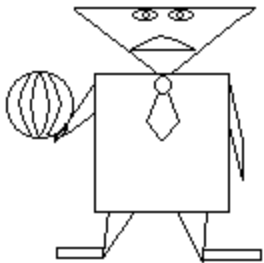
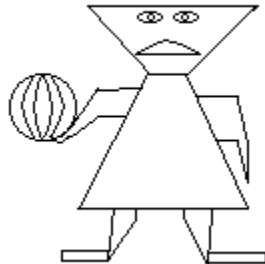
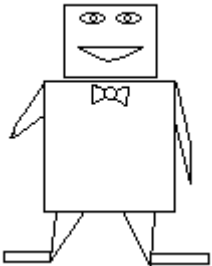


Nature Inspired  
Technologies Group

# Příklad 1 „počítačová hra“.



Můžeme se naučit rozlišit přátelské a nepřátelské roboty?



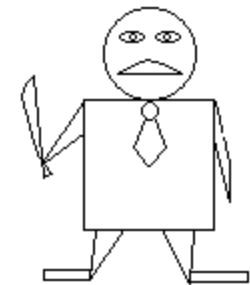
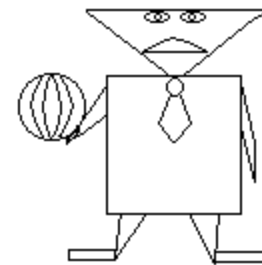
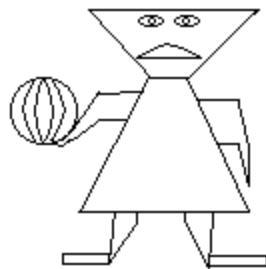
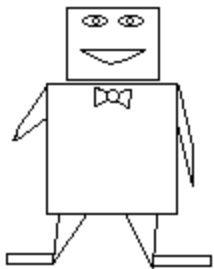
## Učení s učitelem:

u některých už víme, jakou mají povahu (**klasifikace**)

## Neparametrická úloha:

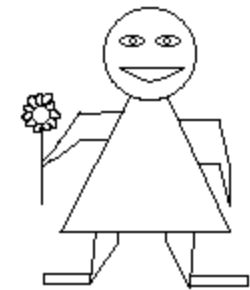
Nic nevíme o pravděpodobnostní distribuci jednotlivých objektů

# Příklad 1 „počítačová hra“. Můžeme se naučit roboty rozlišit na základě krátké zkušenosti?



přátelští

nepřátelští



# Příklad 1: Roboti a atributový popis

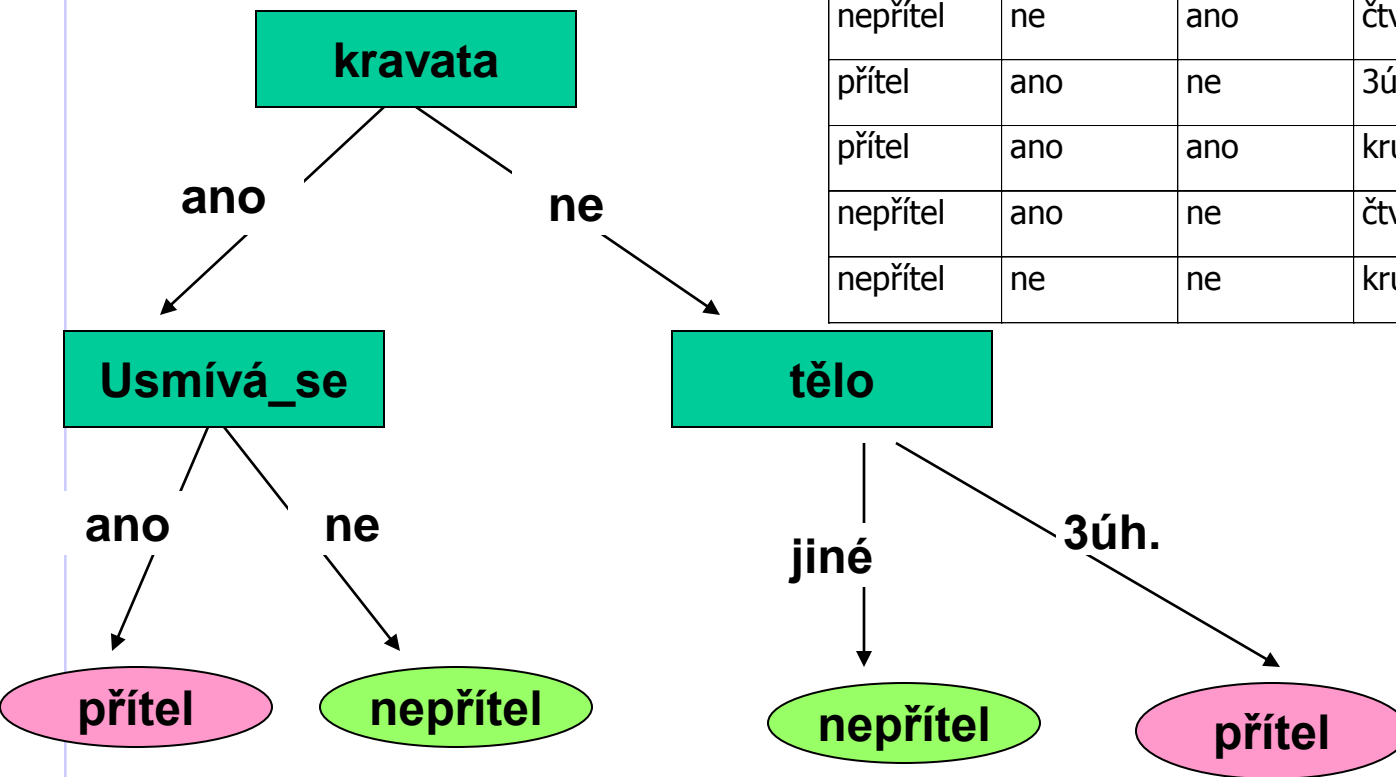


| <i>tvar hlavy</i> | <i>úsměv</i> | <i>ozdoba krku</i> | <i>tvar těla</i> | <i>předmět v ruce</i> | <i>přátelský</i> |
|-------------------|--------------|--------------------|------------------|-----------------------|------------------|
| Kruh              | ne           | kravata            | čtverec          | šavle                 | ne               |
| Čtverec           | ano          | motýlek            | čtverec          | nic                   | ano              |
| Kruh              | ne           | motýlek            | Kruh             | šavle                 | ano              |
| Trojúhelník       | ne           | kravata            | čtverec          | balón                 | ne               |
| Kruh              | ano          | nic                | trojúhelník      | květina               | ne               |
| Trojúhelník       | ne           | nic                | trojúhelník      | balon                 | ano              |
| Trojúhelník       | ano          | kravata            | Kruh             | nic                   | ne               |
| Kruh              | ano          | kravata            | Kruh             | nic                   | ano              |

# Rozhodovací strom 1 pro danou množinu příkladů



| Klasifikace | Usmívá_se | kravata | tělo     | hlava    | v_ruce |
|-------------|-----------|---------|----------|----------|--------|
| přítel      | ano       | ano     | kruh     | 3úhelník | nic    |
| přítel      | ne        | ne      | 3úhelník | 3úhelník | balon  |
| nepřítel    | ne        | ano     | čtverec  | 3úhelník | balón  |
| nepřítel    | ne        | ano     | čtverec  | kruh     | meč    |
| přítel      | ano       | ne      | 3úhelník | kruh     | květ   |
| přítel      | ano       | ano     | kruh     | kruh     | nic    |
| nepřítel    | ano       | ne      | čtverec  | čtverec  | nic    |
| nepřítel    | ne        | ne      | kruh     | kruh     | meč    |

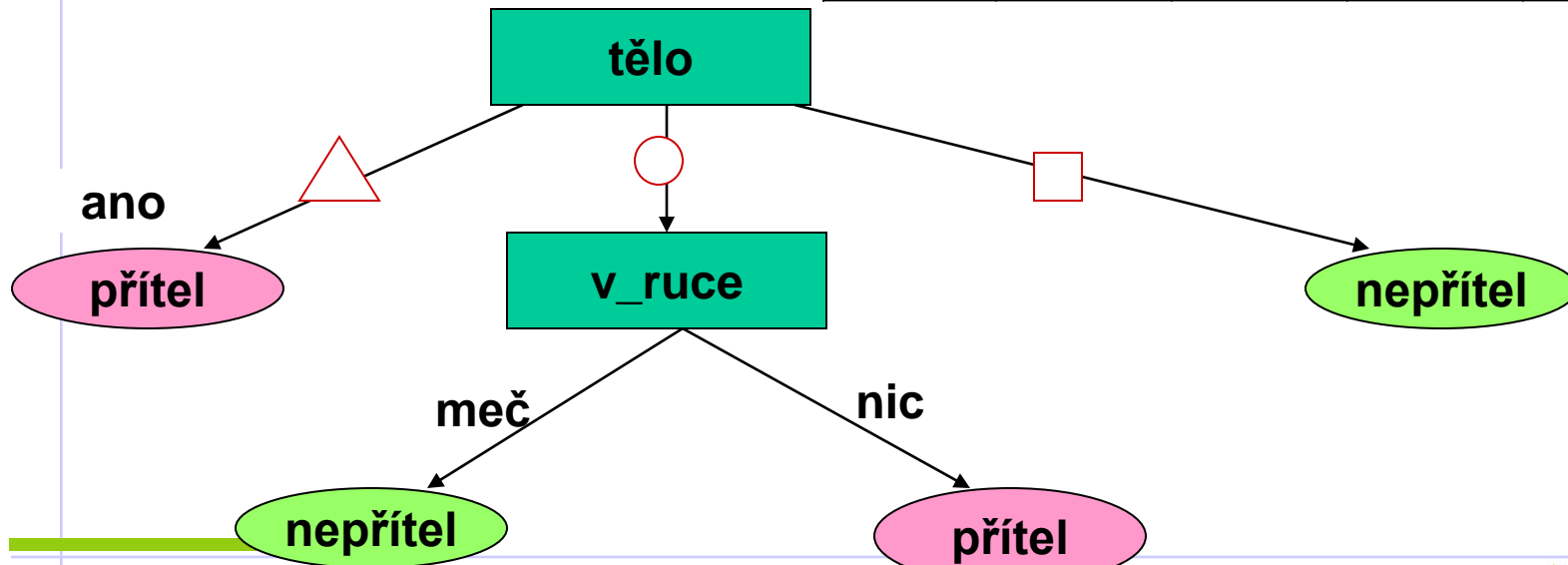


# Rozhodovací strom 2 pro tutéž množinu příkladů



| Klasifikace | Usmívá_se | kravata | tělo     | hlava    | v_ruce |
|-------------|-----------|---------|----------|----------|--------|
| přítel      | ano       | ano     | kruh     | 3úhelník | nic    |
| přítel      | ne        | ne      | 3úhelník | 3úhelník | balon  |
| nepřítel    | ne        | ano     | čtverec  | 3úhelník | balón  |
| nepřítel    | ne        | ano     | čtverec  | kruh     | meč    |
| přítel      | ano       | ne      | 3úhelník | kruh     | květ   |
| přítel      | ano       | ano     | kruh     | kruh     | nic    |
| nepřítel    | ano       | ne      | čtverec  | čtverec  | nic    |
| nepřítel    | ne        | ne      | kruh     | kruh     | meč    |

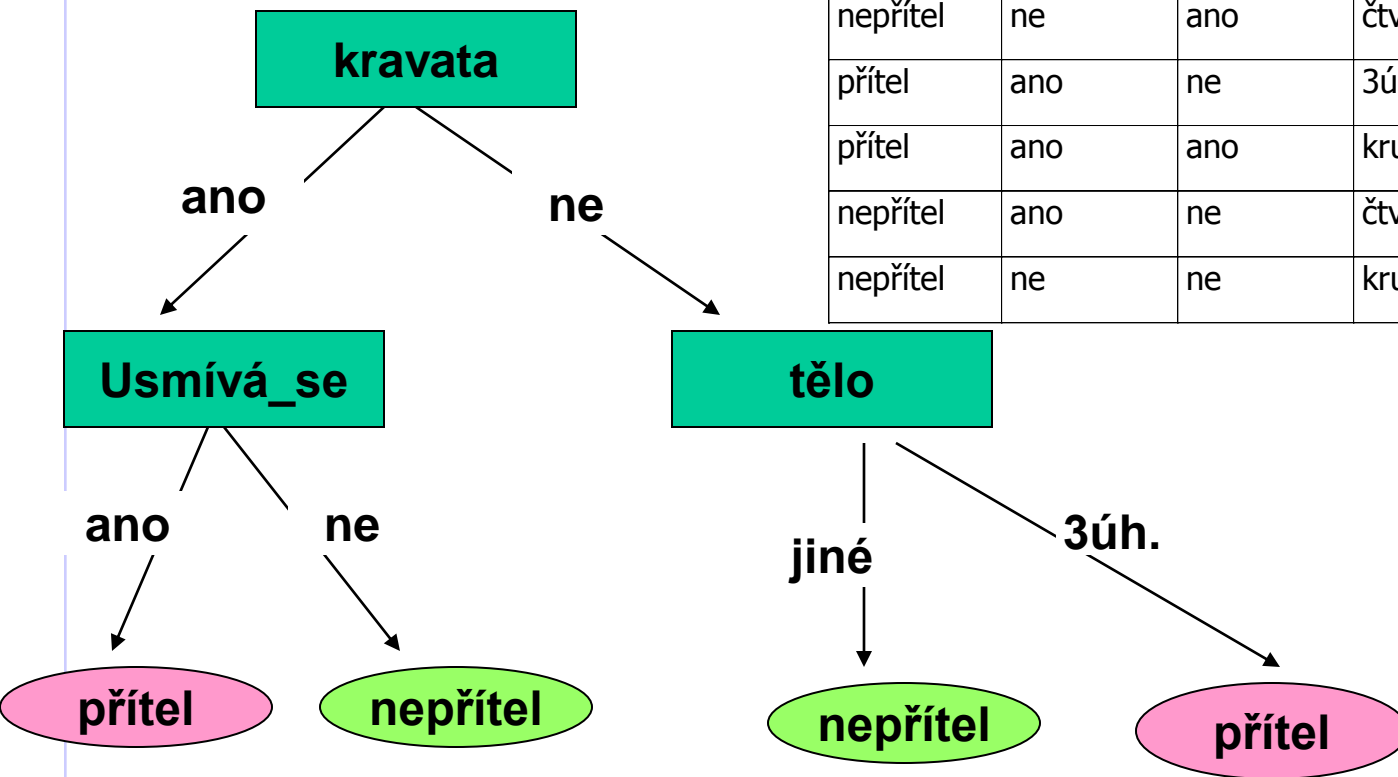
Který strom je lepší?



# Rozhodovací strom jako logický výraz



| Klasifikace | Usmívá_se | kravata | tělo     | hlava    | v_ruce |
|-------------|-----------|---------|----------|----------|--------|
| přítel      | ano       | ano     | kruh     | 3úhelník | nic    |
| přítel      | ne        | ne      | 3úhelník | 3úhelník | balon  |
| nepřítel    | ne        | ano     | čtverec  | 3úhelník | balón  |
| nepřítel    | ne        | ano     | čtverec  | kruh     | meč    |
| přítel      | ano       | ne      | 3úhelník | kruh     | květ   |
| přítel      | ano       | ano     | kruh     | kruh     | nic    |
| nepřítel    | ano       | ne      | čtverec  | čtverec  | nic    |
| nepřítel    | ne        | ne      | kruh     | kruh     | meč    |



**(Kravata=ano & usmívá\_se=ano) V (Kravata=ne & tělo=3úh.) -> přítel**



|                    |                        |                        |                        |                       |                        |
|--------------------|------------------------|------------------------|------------------------|-----------------------|------------------------|
| <b>Klasifikace</b> | Usmívá_se              | kravata                | tělo                   | hlava                 | v_ruce                 |
| přítel             | ano                    | ano                    | kruh                   | 3úhelník              | nic                    |
| přítel             | ne                     | ne                     | 3úhelník               | 3úhelník              | balon                  |
| nepřítel           | ne                     | ano                    | čtverec                | 3úhelník              | balón                  |
| nepřítel           | ne                     | ano                    | čtverec                | kruh                  | meč                    |
| přítel             | ano                    | ne                     | 3úhelník               | kruh                  | květ                   |
| přítel             | ano                    | ano                    | kruh                   | kruh                  | nic                    |
| nepřítel           | ano                    | ne                     | čtverec                | čtverec               | nic                    |
| nepřítel           | ne                     | ne                     | kruh                   | kruh                  | meč                    |
|                    | Usmívá_se              | Kravata                | tělo=3úhel.            | hlava=3úh.            | v_r.=nic               |
| <b>Klasifikace</b> | Ano:3P,1N<br>Ne: 1P,3N | Ano:2P,2N<br>Ne: 2P,2N | Ano:2P,0N<br>Ne: 2P,4N | Ano:2P,1N<br>Ne:2P,3N | Ano:2P,1N<br>Ne: 2P,3N |



# Indukce rozhodovacího stromu z trénovací množiny



dáno:  $S$  ... trénovací množina (množina klasifikovaných příkladů)

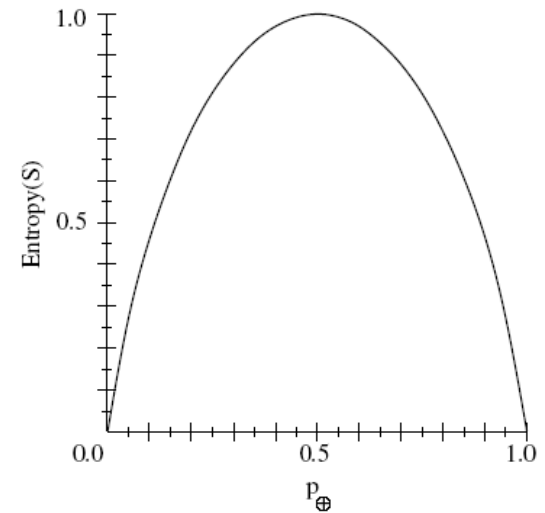
1. Nalezni "**nejlepší**" atribut  $at_0$  (t.j. atribut, jehož hodnoty nejlépe diskriminují mezi pozitivní a neg. příklady) pro  $S$  a tím ohodnot' kořen vytvářeného stromu.
2. Rozděl množinu  $S$  na podmnožiny  $S_1, S_2, \dots, S_n$  podle hodnot atributu  $at_0$  a pro každou množinu příkladů  $S_i$  vytvoř nový uzel jako následníka právě zpracovávaného uzlu (kořenu)
3. Pro každý nově vzniklý uzel s přiřazenou podmnožinou  $S_i$  proved':
  - **Jestliže** všechny příklady v  $S_i$  mají tutéž klasifikaci (všechny jsou pozitivní nebo všechny jsou negativní),
  - **pak** uzel ohodnocený  $S_i$  je prohlášen za list vytvářeného rozhodovacího stromu (a tedy se už dále nevětví),
  - **jinak** jdi na bod 1 s tím, že  $S := S_i$ .

# Entropie množiny $\mathcal{S}$

vzhledem k dané klasifikaci



- ❖ Posuzuje „různorodost“ klasifikace prvků z množiny  $\mathcal{S}$
- ❖ Necht' klasifikaci představuje atribut  $y$ , který má jen 2 hodnoty  $\{0,1\}$ . Pak označme  $\mathcal{S}^0 = \{\mathbf{z} \in \mathcal{S} : \mathbf{z}_y = 0\}$  a  $\mathcal{S}^1 = \{\mathbf{z} \in \mathcal{S} : \mathbf{z}_y = 1\}$



$$\text{Entropy}(\mathcal{S}) = - |\mathcal{S}^0|/|\mathcal{S}| * \log_2 |\mathcal{S}^0|/|\mathcal{S}| - |\mathcal{S}^1|/|\mathcal{S}| * \log_2 |\mathcal{S}^1|/|\mathcal{S}|,$$

Kde  $|\mathcal{A}|$  označuje mohutnost množiny  $\mathcal{A}$

- ❖ Je-li  $\mathcal{S}^0 = \emptyset$ , pak Entropy ( $\mathcal{S}$ )=0 ...
- ❖ Je-li  $|\mathcal{S}^0| = |\mathcal{S}^1|$ , pak Entropy ( $\mathcal{S}$ )=1



## Kriterium minimální entropie rozkladu (KMER)

- ❖ Necht'  $at$  je pevně zvolený atribut, který může nabývat hodnot  $v_1$  až  $v_n$ .
- ❖ Označme  $S_i = \{z \in S : z_{at} = v_i\}$  podmnožinu  $S$ , která obsahuje právě ty objekty, které v atributu  $at$  mají hodnotu  $v_i$ .
- ❖ Vážená entropie  $E(S, at)$  rozkladu  $S$  podle hodnot atributu  $at$  charakterizuje „čistotu“ klasifikace v jednotlivých složkách rozkladu  $S$  a je definována  $E(S, at) = \sum_{i=1}^n |S_i|/|S| * E(S_i)$

**KMER** vypočte  $E(S, at)$  pro všechny atributy  $at$  a jako nejlepší atribut  $at^0$  zvolí ten z nich, pro který je hodnota  $E(S, at^0)$  nejmenší

# ✦ Základní algoritmus ID3



❖ Realizuje prohledávání prostoru všech stromů, které lze zkonstruovat v jazyku trénovacích dat :

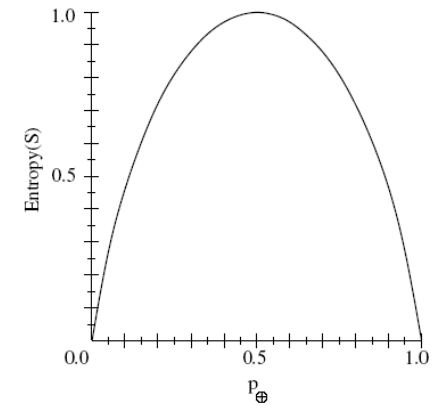
- ◆ shora dolů
- ◆ s použitím hladové strategie

❖ Volba atributu pro větvení na základě charakterizace „(ne)homogenity nově

vzniklého pokrytí“ (používají se různé míry), např.:

- ◆ **Kriterium minimalní entropie rozkladu**
- ◆ **Informační zisk** (gain) odhaduje předpokládané snížení entropie pro pokrytí vzniklé použitím hodnot odpovídajícího atributu

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$





# Nebezpečí použití kriteria zisku (*gain*)

Co se stane, pokud některý atribut má hodně „skoro“ unikátních hodnot, které takřka jednoznačně charakterizují každý trénovací příklad? Například pro rodné číslo je  $|S_i| = 1$  a tedy

$$E(S_i) = 0 \text{ a } E(S, \text{rodne\_cislo}) = 0$$

Tento argument je kriteriem KMER vybrán jako nejlepší.

*Je takový atribut opravdu užitečný pro testovací data?*

Ne, má malou generalizační schopnost! Nebylo by vhodné takovou situaci nějak „penalizovat“? JAK? Zde (pro hodnotu 0) nepomůže multiplikační koeficient! Raději pracujeme s hodnotou doplňkovou, kterým je **kriterium zisku**:

$$\text{Gain } E(S,at) = E(S) - E(S,at) = E(S) - \sum_{i=1}^n |S_i|/|S| * E(S_i)$$



**Jak charakterizovat rozklad množiny  $S$  na  $c$  disjunktních podmnožin  $S_i$  podle všech hodnot uvažovaného atributu  $A$  ?**

$$SplitInformation(S, A) \equiv - \sum_{i=1}^c \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

*SplitInformation* odpovídá entropii rozdělení  $S$  podle všech hodnot atributu  $A$ . Např. je-li  $|S_i|/|S|=1$  pro všechna  $i < (c+1)$ , je jeho hodnota rovna  $(\log_2 c)$ . Používá se pro výpočet *GainRatio*, který penalizuje atributy s příliš mnoha hodnotami:

$$GainRatio(S, A) \equiv \frac{Gain(S, A)}{SplitInformation(S, A)}$$

# Volba atributů a další „speciální situace“



- ❖ Reálné hodnoty.

  - používá se **diskretizace**

- ❖ Různé ceny pro získání hodnoty atributu

- ❖ Atribut má příliš rozsáhlý definiční obor

  - používá se **změna kriteria pro volbu atributů**



❖ Reálné hodnoty → používá se **diskretizace**

❖ **JAK SE VOLÍ VHODNÉ MEZNÍ HODNOTY?**

**Vhodné řešení (Fayyad 91):** Uspořádejte příklady podle velikosti zpracovávaného atributu a zvolte jako kandidátní mezní hodnoty ty, které leží v intervalu, kde se mění klasifikace. Hodnota, která maximalizuje InfoGain, je nutně jednou z nich.

|                     |    |    |     |     |     |    |
|---------------------|----|----|-----|-----|-----|----|
| <i>Temperature:</i> | 40 | 48 | 60  | 72  | 80  | 90 |
| <i>PlayTennis:</i>  | No | No | Yes | Yes | Yes | No |



# Volba atributů a „speciální situace“ 2



- ❖ Různé ceny pro získání hodnoty atributu.
- ❖ Určíme-li cenu  $Cost(A)$  v intervalu  $\langle 0,1 \rangle$ , pak použijeme změněné kritérium, např.

- Tan and Schlimmer (1990)

$$\frac{Gain^2(S, A)}{Cost(A)}$$

- Nunez (1988)

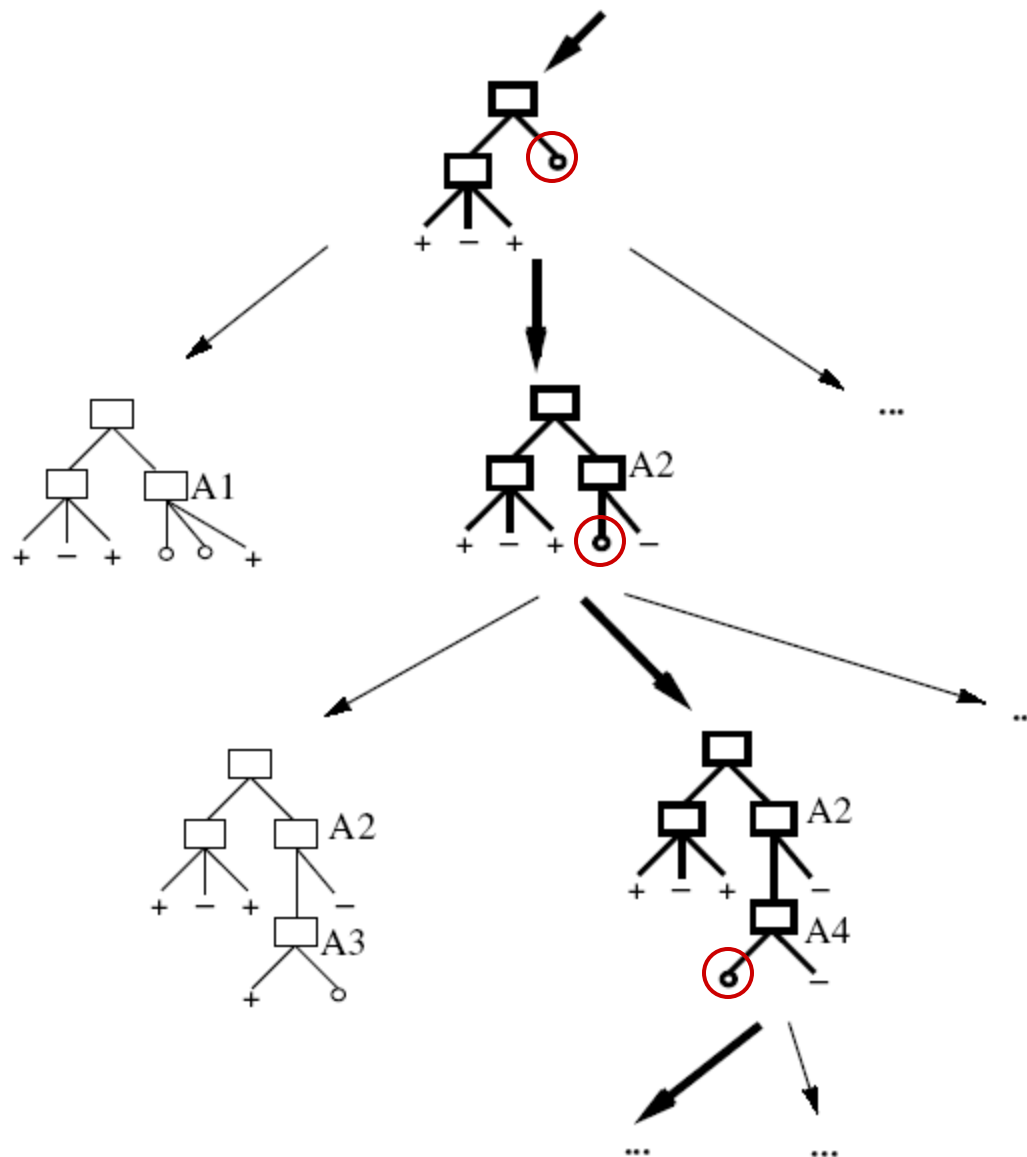
$$\frac{2^{Gain(S,A)} - 1}{(Cost(A) + 1)^w}$$

# Kdy je vhodné použít algoritmy pro konstrukci rozhodovacího stromu?



- ❖ Cílová funkce má diskrétní hodnoty (jedná se o **klasifikační problém**)
- ❖ Instance trénovacích dat mají jednotný formát popisující hodnoty atributů
- ❖ Trénovací data mohou
  - ◆ být zašuměná
  - ◆ obsahovat chybějící hodnoty
- ❖ Je potřeba reprezentovat disjunkci podmínek (pravidla)

# Postup prohlédávání



# Vlastnosti ID3: důsledky postupu prohledávání



- ❖ Pro klasifikační úlohu s diskrétními atributy je prohledávaný **prostor hypotéz úplný** (tj. je schopný reprezentovat libovolnou možnou cílovou funkci), *na rozdíl např. od prostoru verzí* --> **existuje mnoho hypotéz konzistentních s daty!**
- ❖ Aktuální **množina hypotéz je vždy jednoprvková** (hladová volba následníka), nelze jej tedy použít pro odpověď na dotaz „kolik existuje alternativních stromů konzistentních s daty?“
- ❖ Nepoužívá zpětný chod --> **možnost uvíznutí v lokálním optimu**
- ❖ **Rozhoduje se na základě všech příkladů** (nikoliv inkrementálně) --> metoda není příliš ovlivněna šumem

# Proč dáváme přednost jednoduchým hypotézám?

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

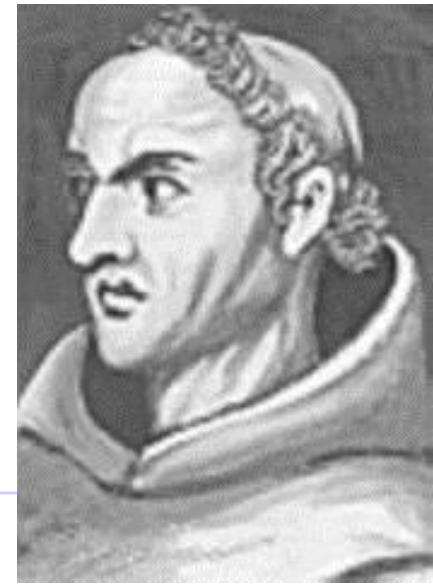
***Occamova břitva*** :

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.

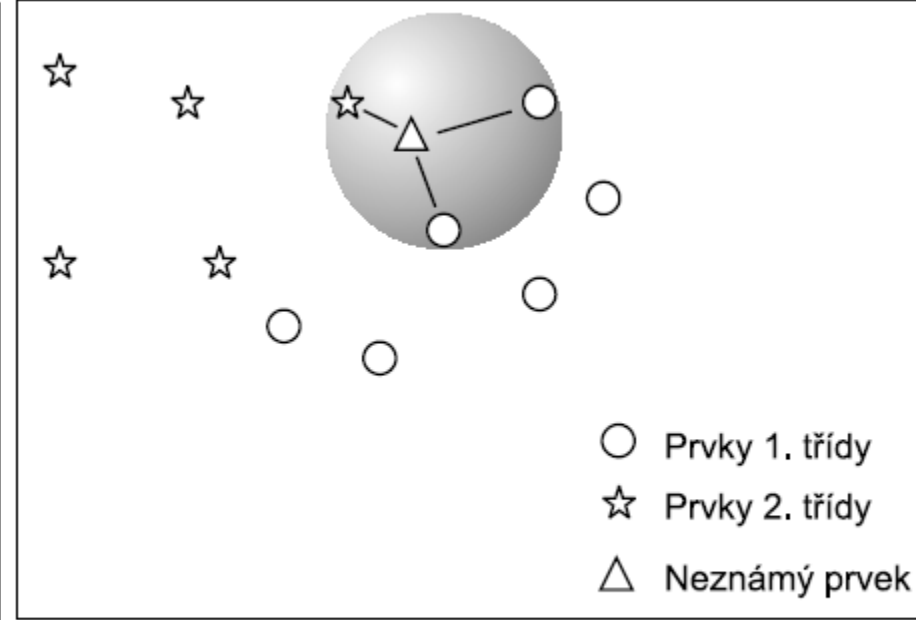
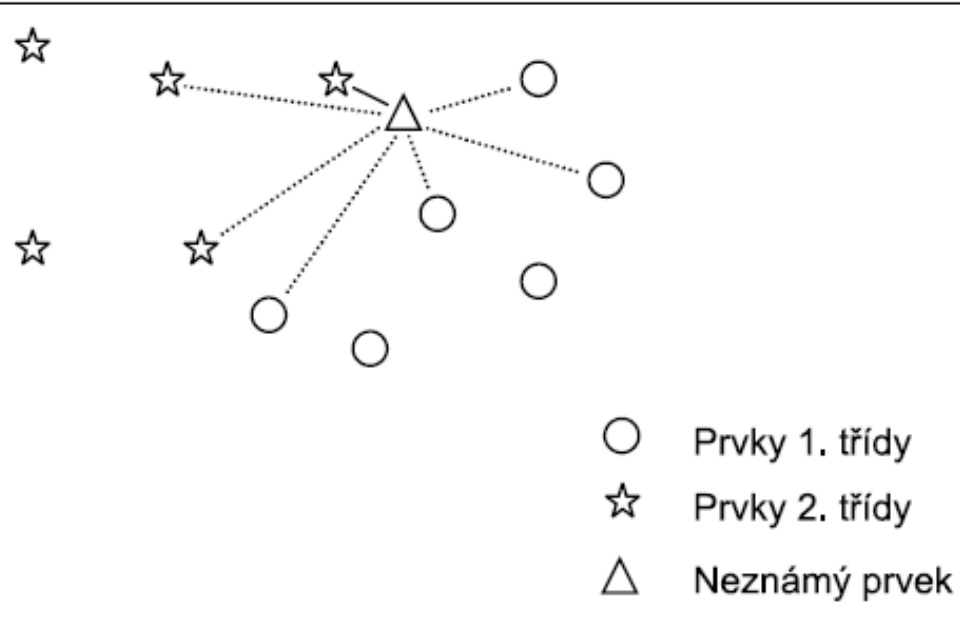


# † Otázky související s ID3



- ❖ Jak velké stromy konstruovat? Až do pokrytí všech příkladů? Co s přeučení?
- ❖ Spojitý definiční obor atributů
- ❖ Metody volby nejvhodnějšího atributu
- ❖ Atributy o různých cenách
- ❖ Chybějící hodnoty
- ❖ ... ???

# Metoda $n$ nejbližších sousedů



Obrázek 1: Popis klasifikace 1-NN

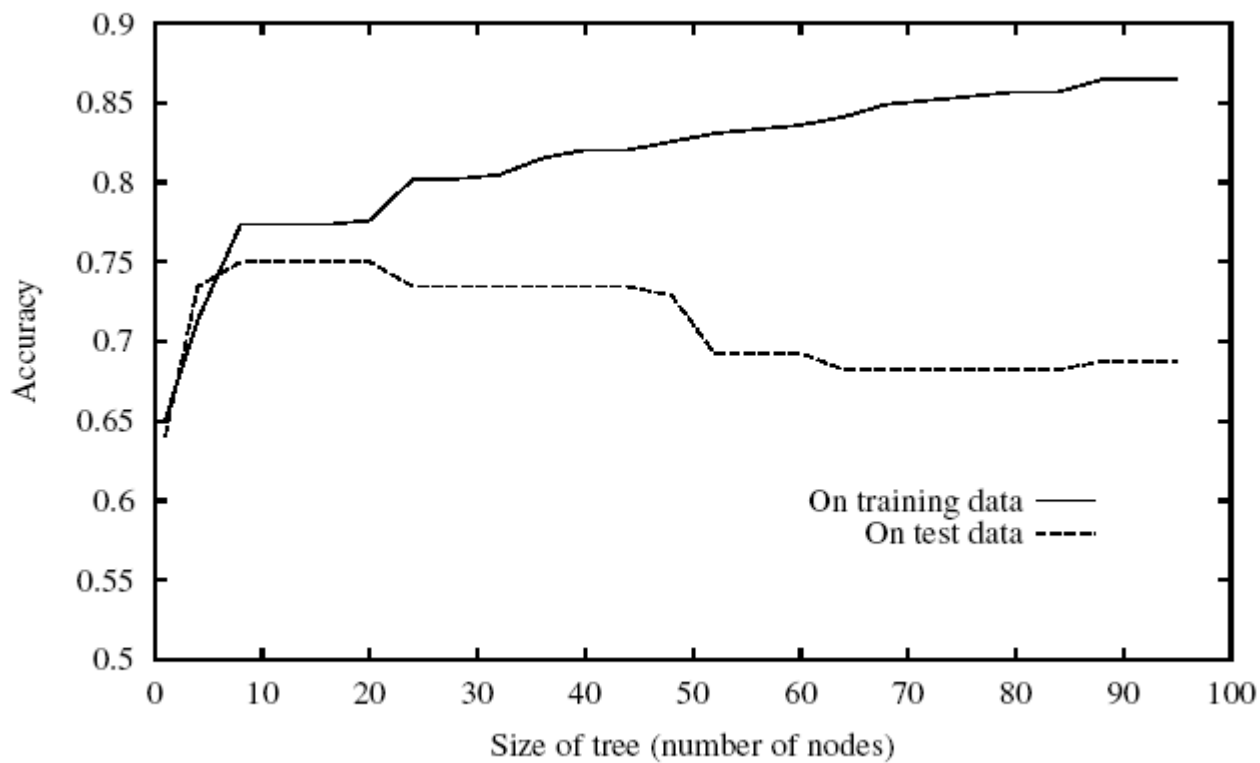
Obrázek 2: Popis klasifikace 3-NN

- ❖ Pro nový objekt je vypočtena vzdálenost od všech objektů v trén. příkl.
- ❖ Je nalezeno všech  $n$  trén. příkladů (= množina  $T^n$ ), které jsou k novému objektu nejbliž. Nový objekt získá klasifikaci, která je v  $T^n$  nejčastější.
- ❖ Možné zobecnění: hledá se nejlepší vážící koeficient pro jednotlivé atributy.

# Přeučení



❖ Necht'  $\mathbf{H}$  je prostor hypotéz. Hypotéza  $\mathbf{h} \in \mathbf{H}$  je přeučená, pokud existuje jiná hypotéza  $\mathbf{h1} \in \mathbf{H}$  taková, že chyba  $\mathbf{h}$  na trénovacích datech je menší než chyba  $\mathbf{h1}$ , avšak na celém prostoru instancí uvažovaných objektů je chyba  $\mathbf{h1}$  menší než chyba  $\mathbf{h}$



Často  
pozorovaná  
vlastnost  
zkonstruovaných  
stromů

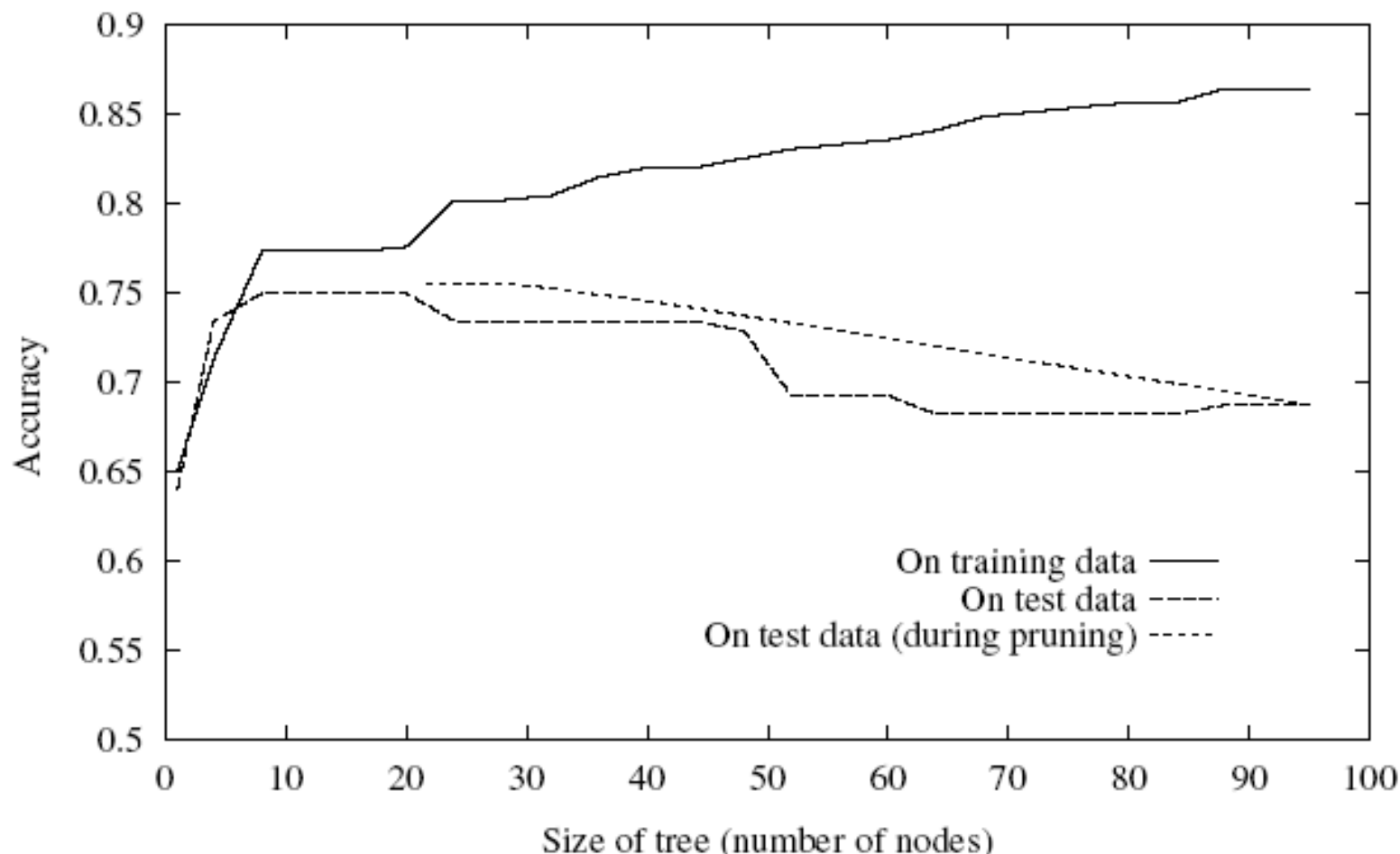
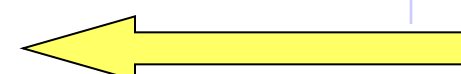


# Jak se vyhnout přeučení?



- ❖ Jak zvolit správnou velikost stromu?
- ❖ **Jak takový strom získat?**
  1. Zastavit růst stromu dřív než jsou vyčerpána všechna trénovací data
  2. Prořezávání hotového stromu – ukazuje se jako zvlášť užitečné! Volba vhodného prořezání pomocí **validační množiny dat** (vybraná nezávisle, tedy bez náhodných vlivů případně přítomných v trénovacích datech).
- ❖ **Algoritmus prořezávání „redukce chyby“:**
  - ◆ Vyberte uzel, odstraňte podstrom, v něm začínající a přiřadte většinovou klasifikaci.
  - ◆ Pokud se chyba na validačních datech zmenšila proveďte uvedené proříznutí (ze všech možností vyberte tu s největším zlepšením).

# Hodnocení přesnosti klasifikace v závislosti na složitosti použitého stromu



Výsledky na „hladké linii“ odpovídají stromům získaným prořezáním tak, jak bylo otestováno na validačních datech (jiná než testovací!!!)

# Závěrečné prořezávání pravidel

(rule post-pruning) použité v **C4.5**

1. Vytvořte přeúčtený rozhodovací strom
2. Zapište výsledný strom ve tvaru disjunkce pravidel (každá větev = jedno pravidlo)
3. Každé jednotlivé pravidlo co nejvíc prořežte (odstraní se postupně ty podmínky, které nezhorší klasifikační přesnost)
4. Uspořádejte výsledná pravidla podle jejich odhadnuté přesnosti a dále je používejte jako rozhodovací seznam

## Odhad přesnosti pravidla

- ◆ Validační množina
- ◆ Na trénovacích datech jako „pesimistický odhad na trénovacích datech za předpokladu binomického rozdělení“

# Příklad: Létání na simulátoru F16



**Úkol:** sestavit řídicí systém pro ovládání leteckého simulátoru F16 tak, aby splnil předem definovaný plán letu daný takto:

1. vzlet a výstup do výšky 2000 stop
2. let v dané výšce směrem N do vzdálenosti 32000 stop od místa startu
3. zahnout vpravo v kurzu  $330^\circ$
4. ve vzdálenosti 42000 stop od místa startu (ve směru S-N) provést obrát vlevo a zamířit zpět do místa startu (obrat je ukončen při kurzu mezi  $140^\circ$  a  $180^\circ$ )
5. vyrovnat směr letu s přistávací dráhou, tolerance  $5^\circ$  pro kurz a  $10^\circ$  pro výchylku křídel oproti horizontu
6. klesat směrem k počátku přistávací dráhy
7. přistát

**Trénovací data:** 3x30 letů (od 3 pilotů). Každý let popsán pomocí 1000 záznamů (poloha a stav letounu, pilotem provedený řídicí zásah)

# Záznam: Poloha a stav



|                      |  |
|----------------------|--|
| <b>on_gound</b>      | boolean: je letadlo na zemi?   |
| <b>g_limit</b>       | boolean: je překročen g limit letadla?   |
| <b>wing_stall</b>    | boolean: je letadlo stabilní?  |
| <b>twist</b>         | integer: $0^{\circ}$ - $360^{\circ}$ , výchylka křídel vůči obzoru                     |
| <b>elevation</b>     | integer: $0^{\circ}$ - $360^{\circ}$ , výchylka trupu vůči obzoru                      |
| <b>azimuth</b>       | integer: $0^{\circ}$ - $360^{\circ}$ , směr letu                                       |
| <b>roll_speed</b>    | integer: $0^{\circ}$ - $360^{\circ}$ , rychlost změny výchylky křídel [ $^{\circ}/s$ ] |
| <b>elev_speed</b>    | integer: $0^{\circ}$ - $360^{\circ}$ , rychlost změny výchylky trupu [ $^{\circ}/s$ ]  |
| <b>azimuth_speed</b> | integer: $0^{\circ}$ - $360^{\circ}$ , rychlost změny kurzu [ $^{\circ}/s$ ]           |
| <b>airspeed</b>      | integer: rychlost letadla v uzlech   |
| <b>climbspeed</b>    | integer: rychlost změny výšky [stop/s]   |

# Záznam: Poloha a stav + Řízení



|              |  |
|--------------|--|
| E/W distance | real: vzdálenost ve směru východ-západ od místa startu |
| N/S distance | real: vzdálenost ve směru sever-jih od místa startu    |
| fuel         | integer: váha paliva v librách                         |

## Řízení:

|          |   |
|----------|---|
| rollers  | real: nastavení ovladače horizontálního vychýlení       |
| elevator | real: nastavení ovladače vertikálního vychýlení         |
| thrust   | integer: 0-100%, plyn                                   |
| flaps    | integer: 0°, 10° nebo 20°, nastavení křídlových lopatek |

Každá ze 7 fází letu vyžaduje vlastní typ řízení (jiné zásahy pilota):  
trénovací příklady rozděleny do 7 odpovídajících skupin. V každé skupině je zkonstruován zvlášť rozhodovací strom pro každý typ řídicího zásahu (rollers, elevator, thrust, flaps), t.j. *7 x 4 stromů*