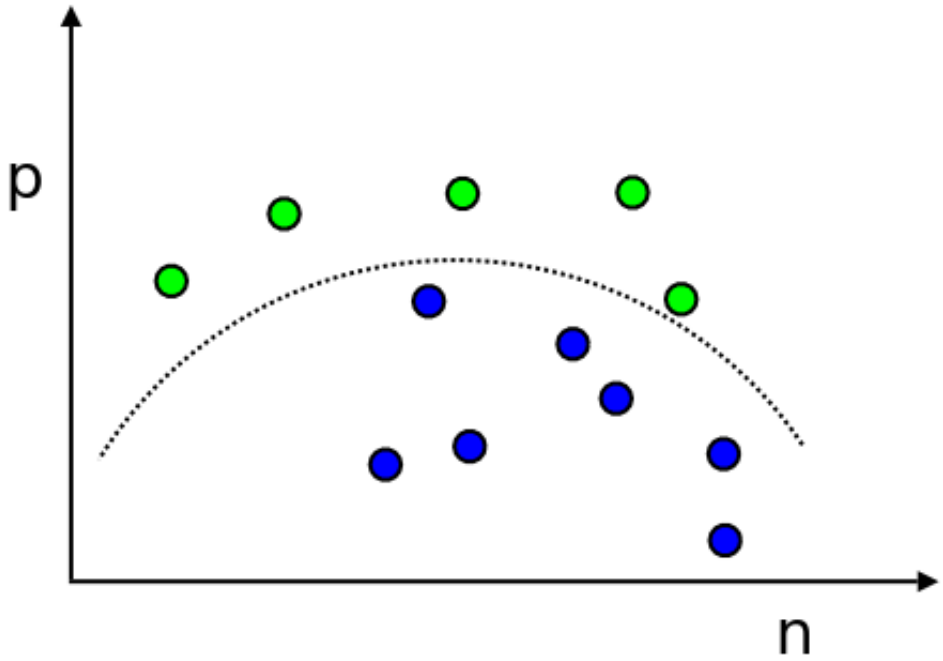
Vytěžování dat

Filip Železný

Katedra kybernetiky
skupina Inteligentní Datové Analýzy (IDA)

7. dubna 2009

Reálné příznaky



- Lineární/polynomiální modely: příznaky jsou reálná čísla

Nominální příznaky

teplota	**bolest svalů**	**diagnóza**
zvýšená	ne	nachlazení
normální	ne	hypochondr
horečka	ano	chřipka
...	...	...

Nominální příznaky

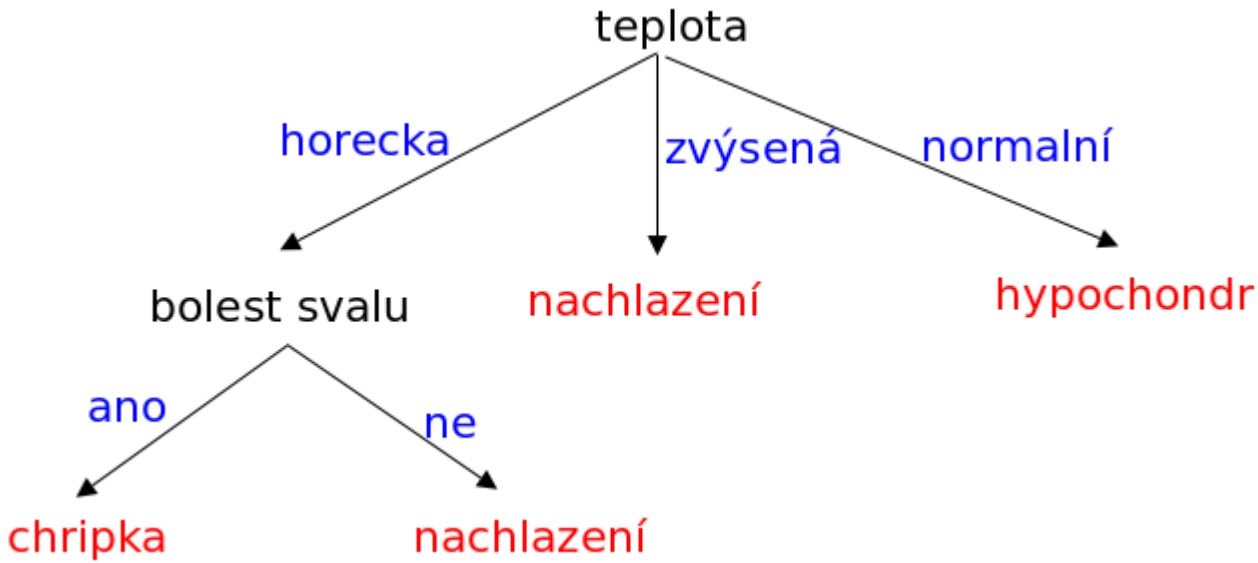
teplota	**bolest svalů**	**diagnóza**
zvýšená	ne	nachlazení
normální	ne	hypochondr
horečka	ano	chřipka
...	...	...

- Lze převést na příznaky s oborem $\{0, 1\}$: reprezentace "1 z n"

teplota			**bolest**	**diagnóza**		
normální	zvýšená	horečka	**svalů**	nachlaz.	chřipka	hypoch.
0	1	0	0	1	0	0
...	...	...	...	...	...	...

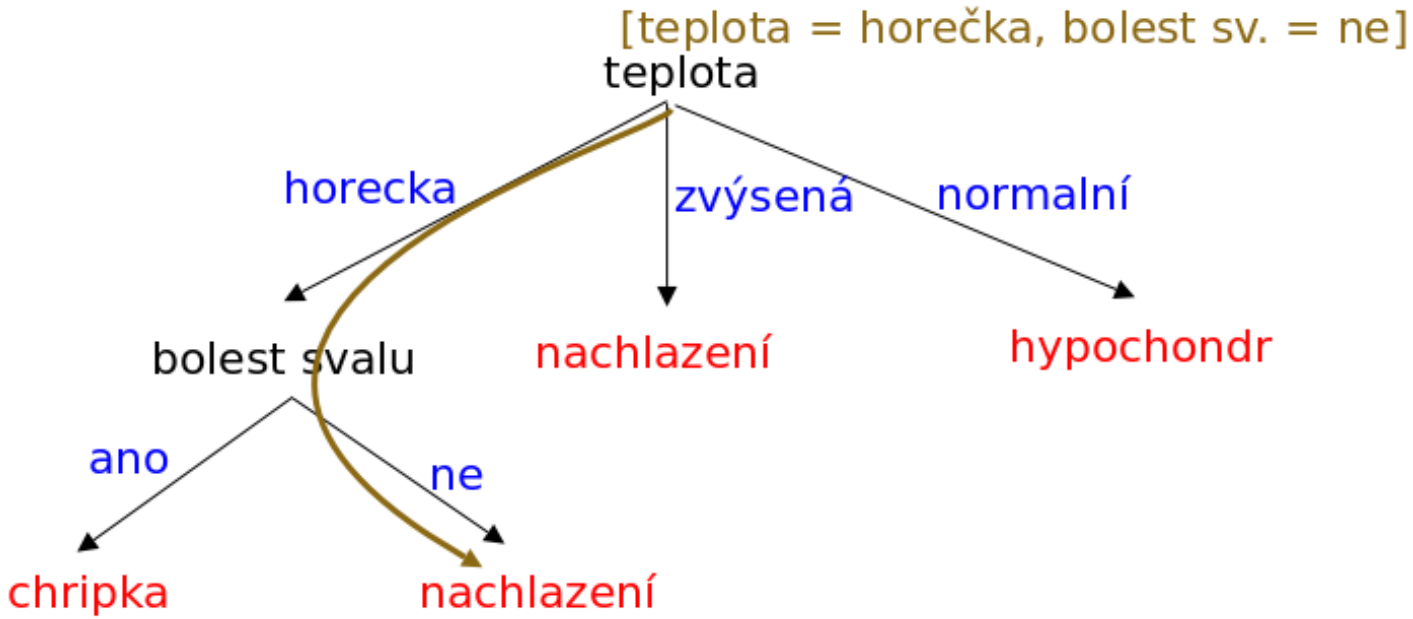
- Umožňuje využití lineárních resp. polynomiálních klasifikátorů, ale nešikovné.
- Klasifikační modely přímo pro nominální příznaky?

Rozhodovací strom



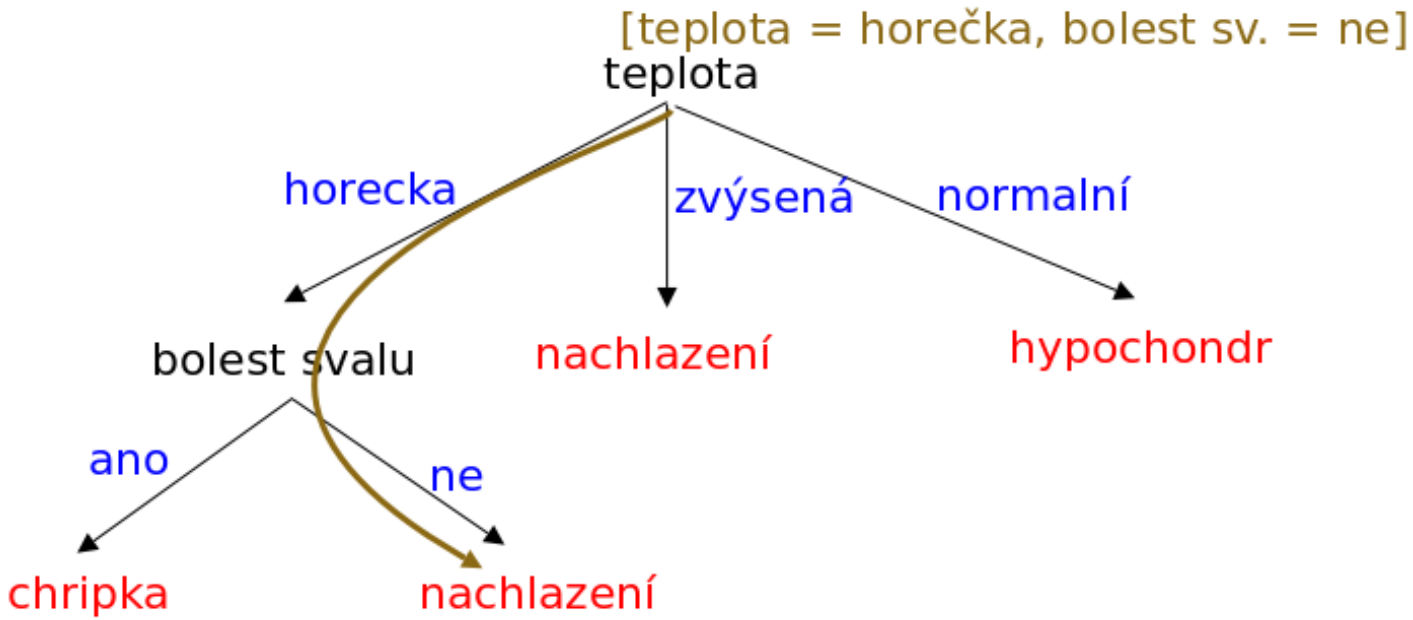
- Klasifikační model
- Uzly: testy příznaků, hrany: možné hodnoty

Rozhodovací strom



- Klasifikační model
- Uzly: testy příznaků, hrany: možné hodnoty
- Klasifikace: cesta z kořene do listu podle hodnot příznaků

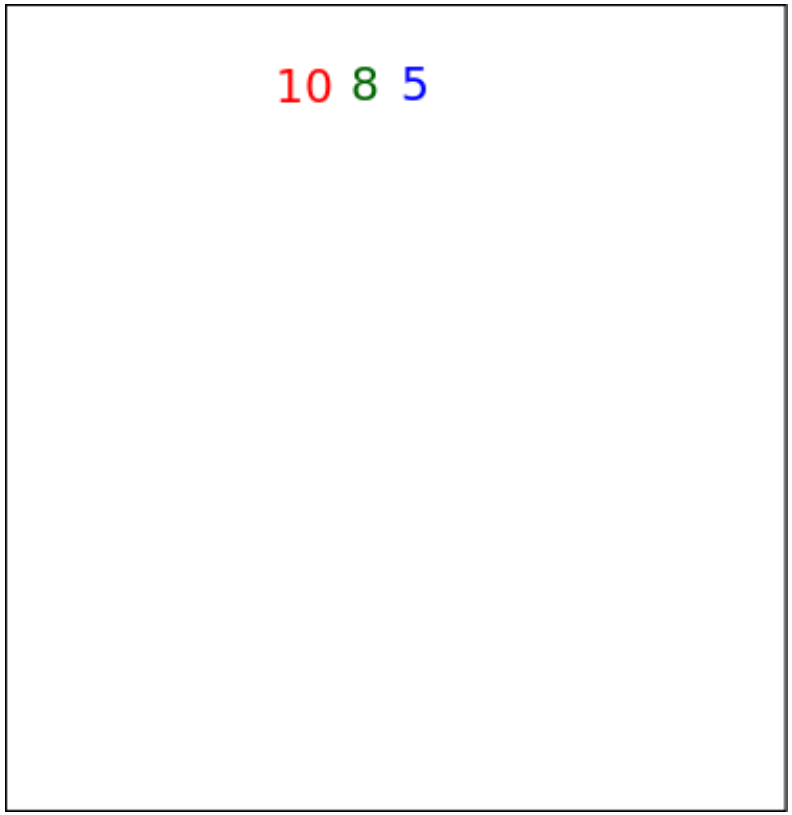
Rozhodovací strom



- Klasifikační model
- Uzly: testy příznaků, hrany: možné hodnoty
- Klasifikace: cesta z kořene do listu podle hodnot příznaků
- Jak strom zkonstruovat?

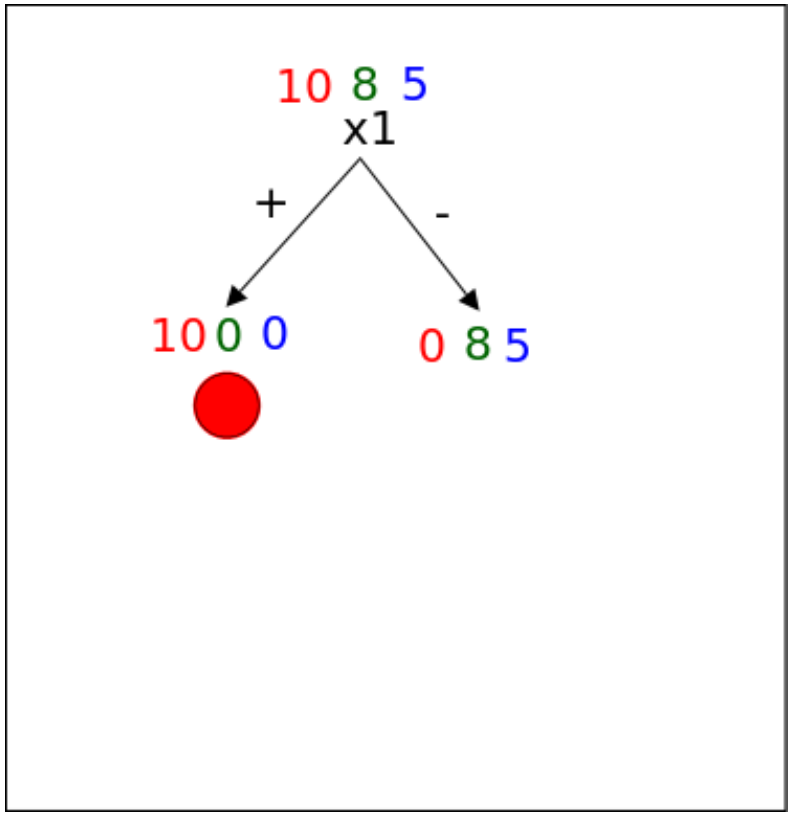
Rekurzivní rozdělování: příklad

- 2 binární příznaky
  $x_1, x_2 \in \{+, -\}$
- Instance spadají do 3 tříd:
  - 10 červených instancí
  - 8 zelených instancí
  - 5 modrých instancí
- Všech 10 s $x_1 = +$ má $y =$ ●



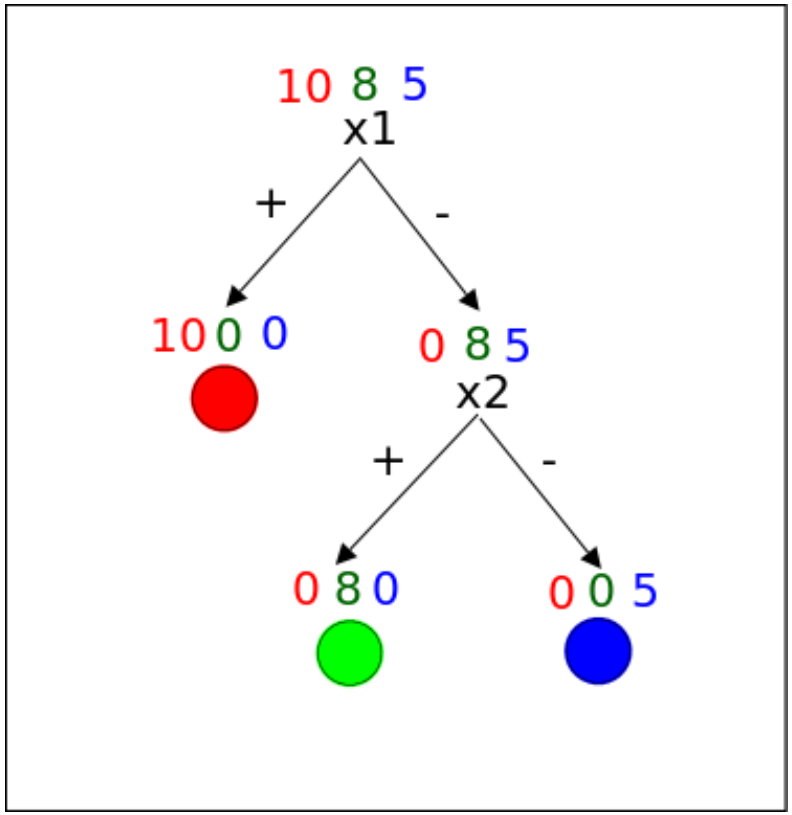
Rekurzivní rozdělování: příklad

- Všech 10 s $x_1 = +$ má $y =$ •
- Zbývá 13 instancí s $x_1 = -$



Rekurzivní rozdělování: příklad

- Všech 8 s $x_2 = +$ má $y =$ ●
- Všech 5 s $x_2 = -$ má $y =$ ●

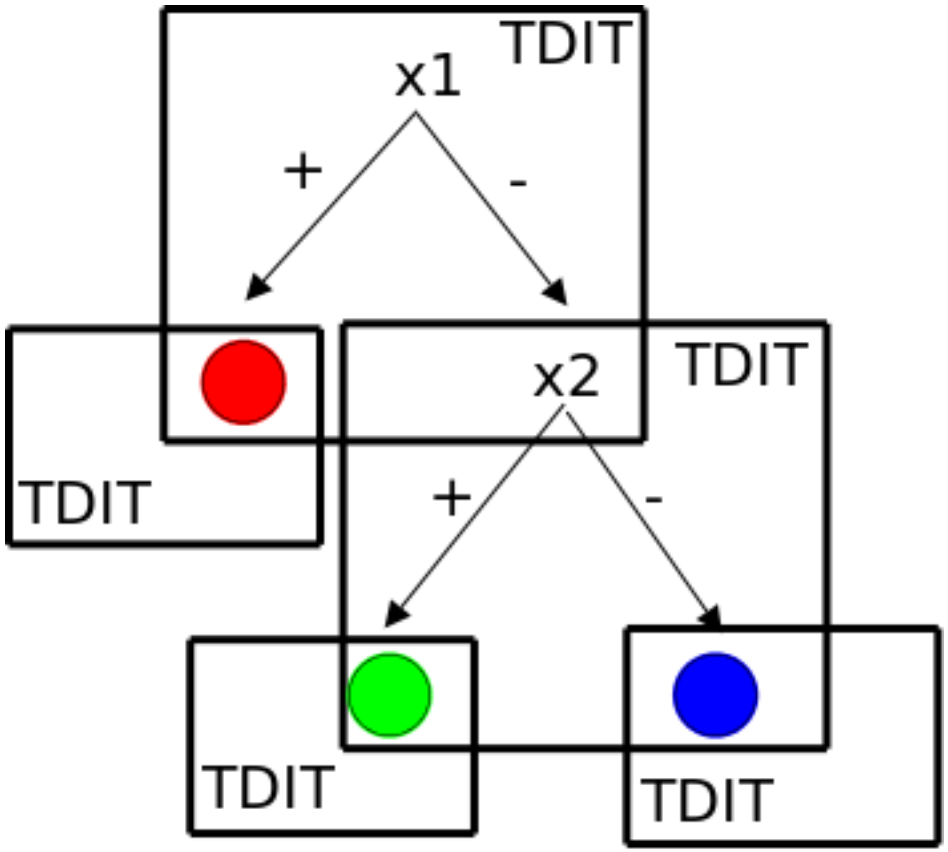


Algoritmus pro tvorbu rozhodovacího stromu

TDIT(D,I) /* *Top Down Induction of Decision Trees* */

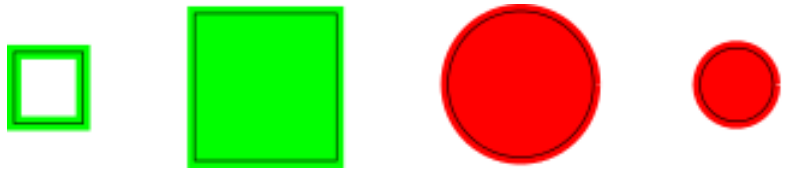
Input: D trénovací data, I indexy příznaků
if *všechny instance v* D *mají stejnou třídu* y **then return** *uzel označený* y
else
  if $I = \emptyset$ **then return** *uzel označený většinovou třídou v* D
  else
    vyber $i \in I$ a vytvoř uzel označený x_i
    for $\forall v_j \in Range(x_i)$ /* konečný obor hodnot x_i */ **do**
      $E_j =$ všechny instance z D u nichž $x_i = v_j$
      Vyveď z uzlu x_i hranu označenou v_j
      if $E_j = \emptyset$ **then** připoj list na hranu v_j označený většinovou třídou v D
      else
        připoj výsledek TDIT(E_j, $I \setminus \{i\}$) na hranu v_i
      end
    end
  end
end
return *vytvořený strom s kořenem* x_i

TDIT: rekurzivní volání



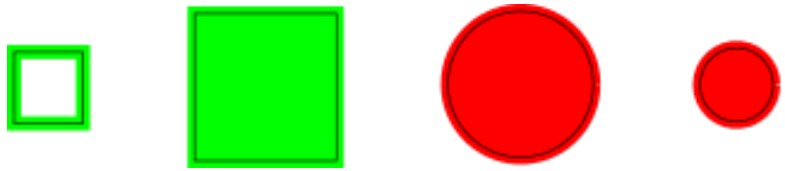
Výběr příznaku

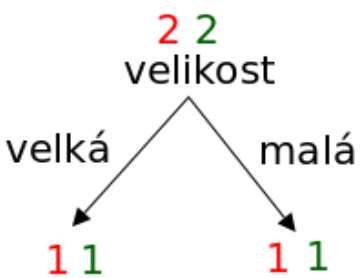
- Jak implementovat příkaz "**vyber** $i \in I$" v algoritmu TDIT?
- Příklad



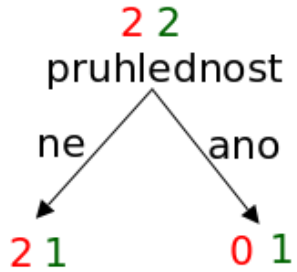
- Třída: barva
- Příznaky: tvar, velikost, průhlednost
- Začínáme konstruovat strom.
- Jaký příznak zvolit první?
- Měl by co 'nejčistěji' dělit data podle tříd

Výběr příznaku

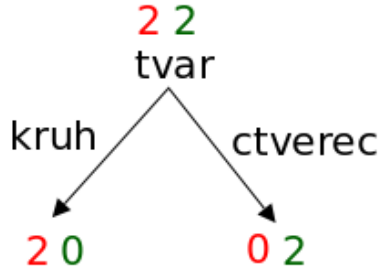




spatne



lepsi



nejlepsi

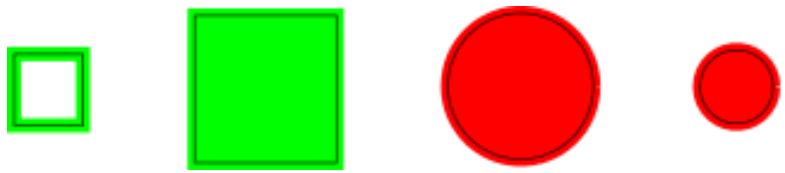
Entropie

- *Entropie* množiny instancí D s t třídami

$$H(D) = \sum_{i=1}^{t} -p_i \log_2 p_i$$

- $p_1, p_2 \ldots p_t \ \ldots$ poměrné velikosti tříd (v počtu instancí)
- Minimální $H(D) = 0$, pokud jsou všechny příklady v jedné třídě.
- Maximální $H(D) = \log_2 t$, pokud $p_1 = p_2 = \ldots = p_t$.

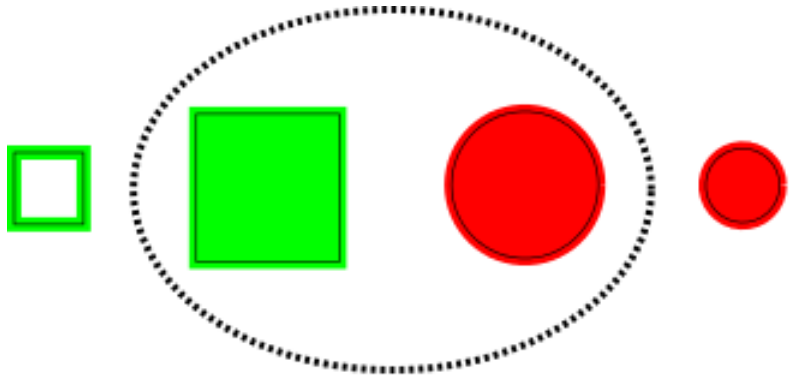
Entropie



$$p_{\text{zelená}} = p_{\text{červená}} = \frac{1}{2}$$

$$H(D) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$$

Entropie po rozdělení množiny



$E_{\text{velké}} =$ velké instance

$p_{\text{zelená}} = p_{\text{červená}} = 0.5$

$H(E_{\text{velké}}) = 1$

$E_{\text{malé}} =$ malé instance

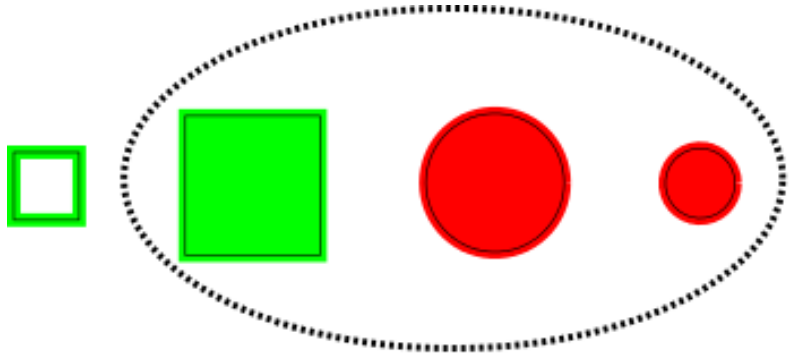
$p_{\text{zelená}} = p_{\text{červená}} = 0.5$

$H(E_{\text{malé}}) = 1$

Vážený průměr entropií

$$\sum_{j \in \{\text{velké,malé}\}} \frac{|E_j|}{|D|} \cdot H(E_j) = \frac{2}{4} \cdot 1 + \frac{2}{4} \cdot 1 = 1$$

Entropie po rozdělení množiny



$E_{\text{průhledné}} = \text{průhledné instance}$

$p_{\text{zelená}} = 1$

$p_{\text{červená}} = 0$

$H(E_{\text{průhledné}}) = 0$

$E_{\text{neprůhledné}} = \text{neprůhledné instance}$

$p_{\text{zelená}} = 1/3$

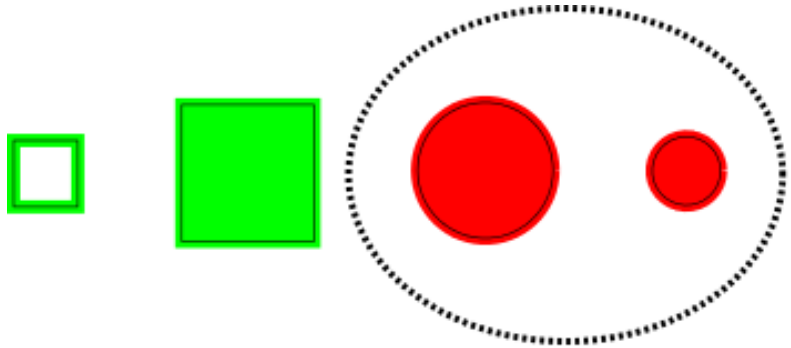
$p_{\text{červená}} = 2/3$

$H(E_{\text{neprůhledné}}) = 0.92$

Vážený průměr entropií

$$\sum_{j \in \{\text{průhledné,neprůhledné}\}} \frac{|E_j|}{|D|} \cdot H(E_j) = \frac{1}{4} \cdot 0 + \frac{3}{4} \cdot 0.92 = 0.69$$

Entropie po rozdělení množiny



$E_{\text{hranaté}} = \text{hranaté instance}$

$p_{\text{zelená}} = 1$

$p_{\text{červená}} = 0$

$H(E_{\text{hranaté}}) = 0$

$E_{\text{kulaté}} = \text{kulaté instance}$

$p_{\text{zelená}} = 0$

$p_{\text{červená}} = 1$

$H(E_{\text{kulaté}}) = 0$

Vážený průměr entropií

$$\sum_{j \in \{\text{hranaté,kulaté}\}} \frac{|E_j|}{|D|} \cdot H(E_j) = \frac{2}{4} \cdot 0 + \frac{2}{4} \cdot 0 = 0$$

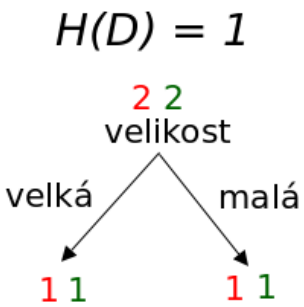
Zisk entropie

$$\Delta E(D, x_i) = H(D) - \sum_{v_j \in Range(x_i)} \frac{|E_j|}{D} H(E_j)$$

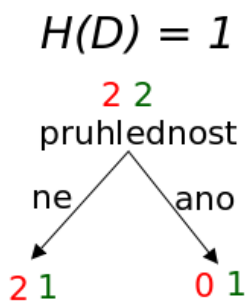
- Rozdíl entropie původní množiny D a váženého průměru entropií množiny rozdělené hodnotami příznaku x_i
- Jak implementovat příkaz "**vyber** $i \in I$" v algoritmu TDIT?
  - Vybereme i, které maximalizuje $\Delta E(D, x_i)$
- Pozn: pro výběr i není sčítanec $H(D)$ důležitý (nezávisí na i).

Zisk entropie

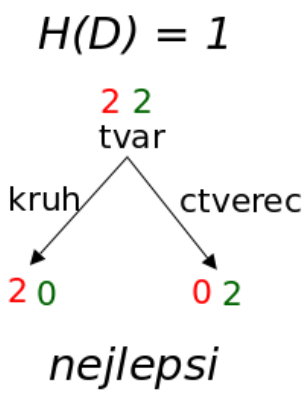
- Jak implementovat příkaz "**vyber** $i \in I$" v algoritmu TDIT?
  - Vybereme i, které maximalizuje $\Delta E(D, x_i)$



$\Delta E(D,velikost) = H(D) - 1 = 0$



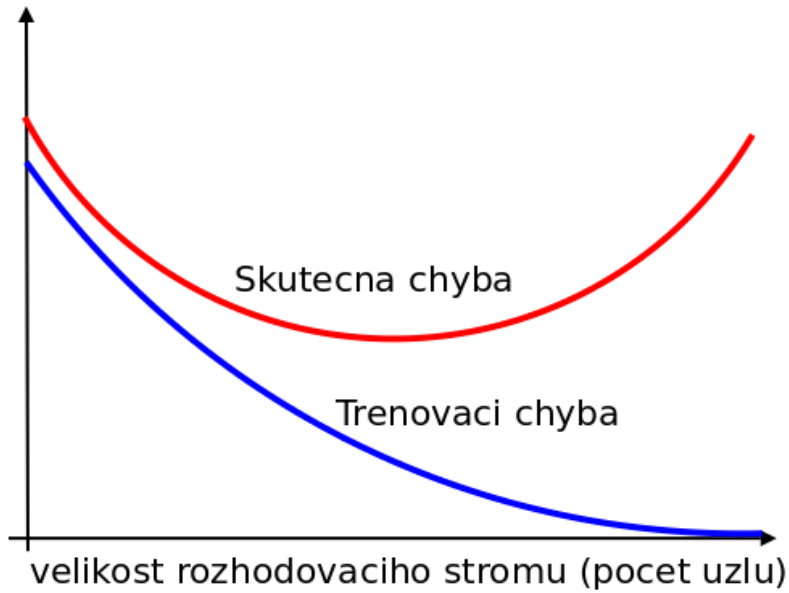
$\Delta E(D,pruhlednost) = H(D) - 0.69 = 0.31$



$\Delta E(D,tvar) = H(D) - 0 = 1$

Složitost rozhodovacího stromu

- Algoritmus TDIT se snaží minimalizovat trénovací chybu za cenu velké složitosti (košatosti) stromu
- Stále platí kompromis mezi složitostí modelu a trénovací chybou!



- Vymyslete úpravu algoritmu TDIT omezující složitost stromu
  - Upravte podmínku "**if** *všechny instance v D mají stejnou třídu y*".

Ordinální příznaky

Ordinální veličina

- Veličina, jejíž obor hodnot je uspořádán
- Např. přirozená (nebo reálná) čísla

$$1 < 2 < 3 < \dots$$

- ale i např.

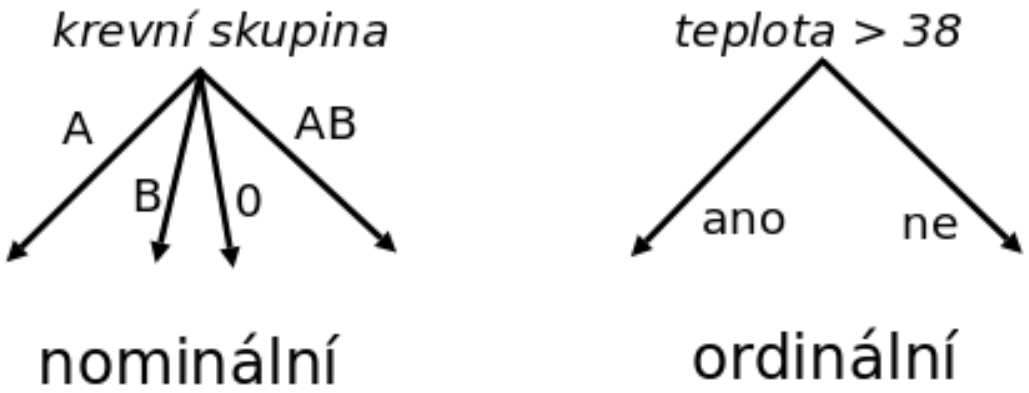
$$\text{nízký} < \text{střední} < \text{vysoký}$$

Ordinální příznaky

- Pro ordinální příznaky obvykle test

$$x > h$$

v uzlech, kde h je zvolená hraniční hodnota



Převod na nominální příznaky

- Před tvorbou stromu můžeme každý ordinální příznak, např.

  $$\text{teplota}$$

  převést na množinu nominálních příznaků

  $$\text{teplota} > h_1, \text{teplota} > h_2, \ldots, \text{teplota} > h_n$$

  z nichž každý má binární obor hodnot.
- Co jsou $h_1, h_2, \ldots h_n$?
- V nejjednodušším případě celý obor hodnot původního příznaku, je-li konečný (a malý).
- Obvykle ale jen některé z oboru hodnot. Které?

Diskretizace

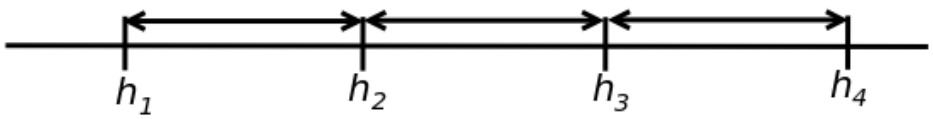
- U některých veličin se hraniční hodnoty ‘nabízejí’.

$x_i < 36.5$	podchlazení
$36.5 \leq x_i < 37$	normální teplota
$37 \leq x_i < 38$	zvýšená teplota
$38 \leq x_i < 42$	horečka
$42 \leq x_i$	smrt

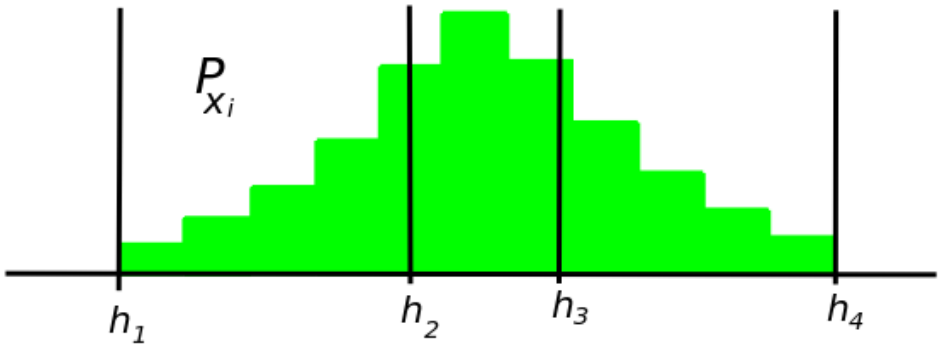
- Zde tedy uvažujeme hraniční hodnoty $\{36.5, 37, 38, 42\}$
- Pozn.: převedení *reálné* veličiny (teplota) na veličinu s konečným oborem hodnot = **diskretizace**.
- V obecném případě vhodné hraniční hodnoty předem neznáme.

Diskretizace: 3 obecné způsoby

- Intervaly stejné délky



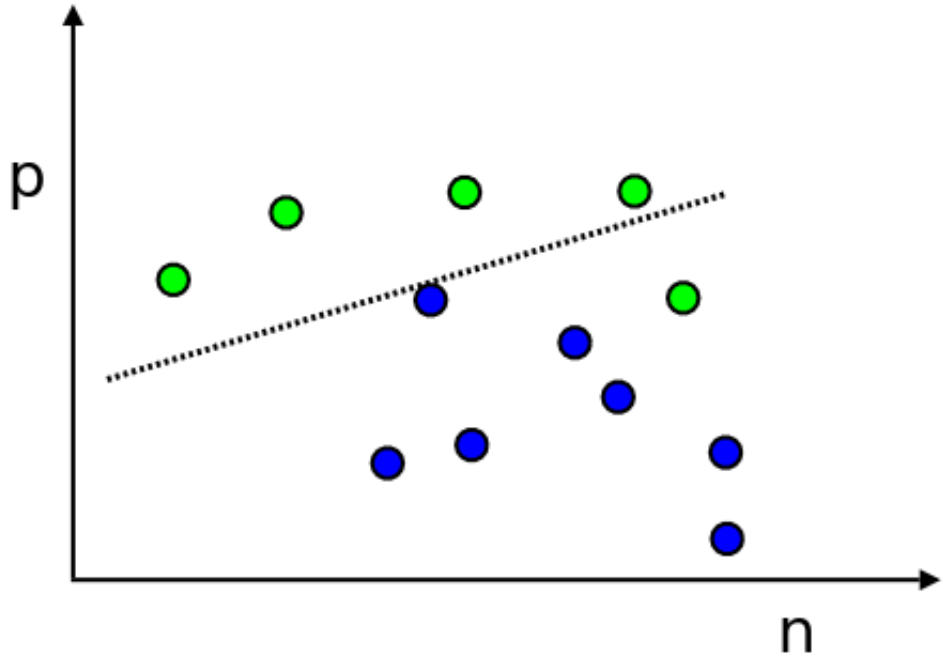
- Intervaly stejné pravděpodobnosti



- Intervaly obsahující instance stejné třídy (nejužívanější pro stromy)

Separace: srovnání

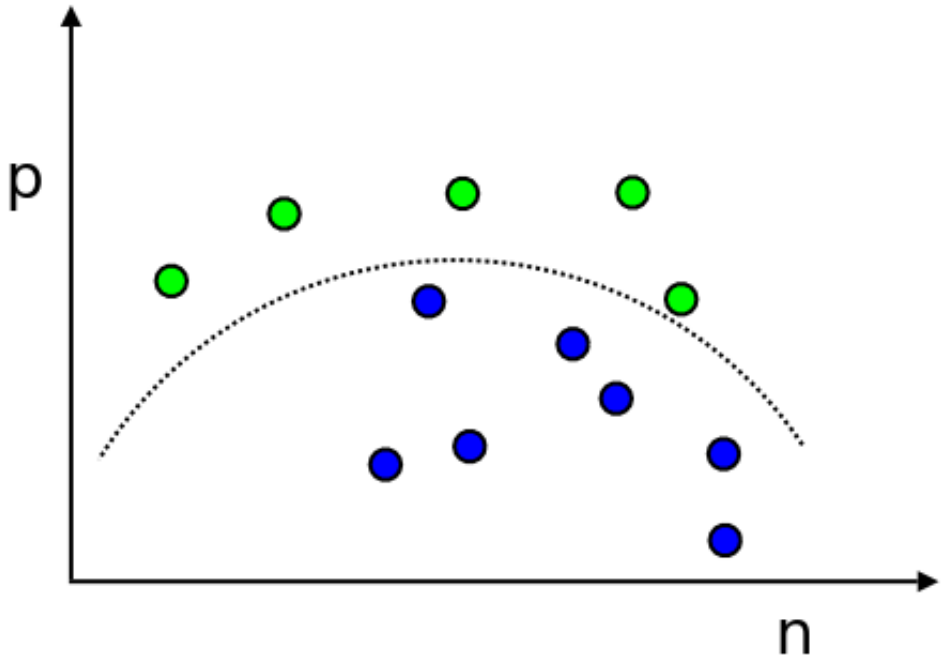
Separace v prostoru dvou reálných příznaků



Lineární klasifikátor (nelze rozdělit)

Separace: srovnání

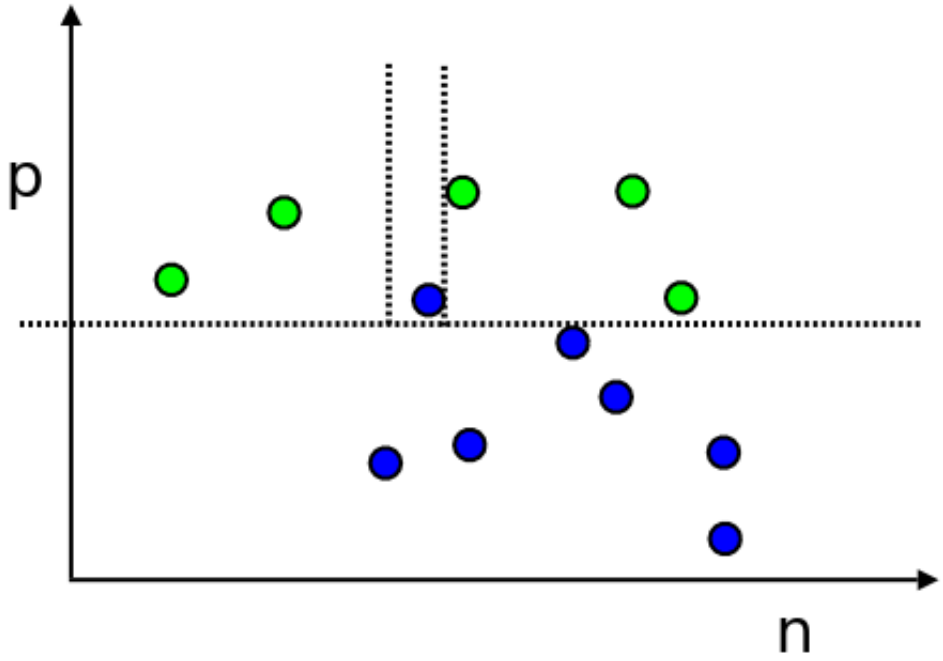
Separace v prostoru dvou reálných příznaků



Kvadratický klasifikátor

Separace: srovnání

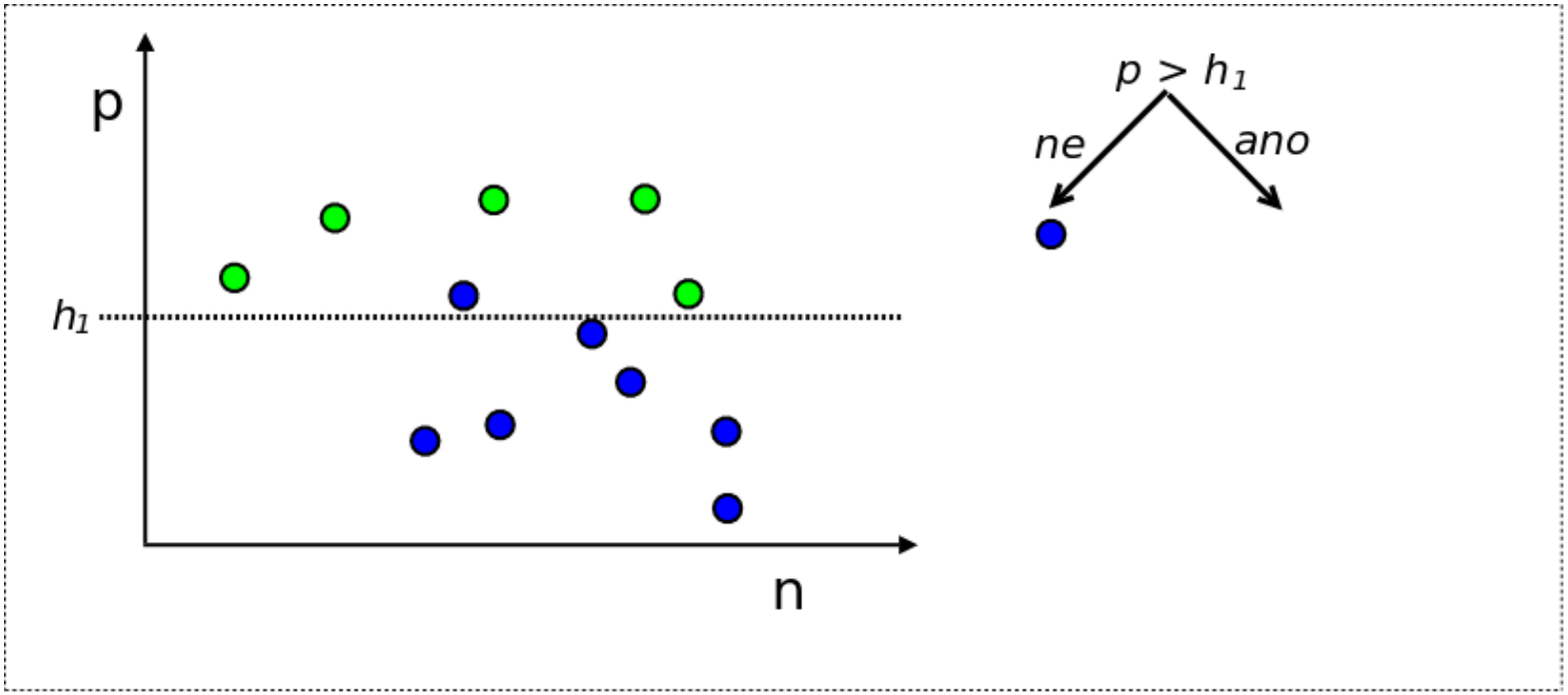
Separace v prostoru dvou reálných příznaků



Rozhodovací strom

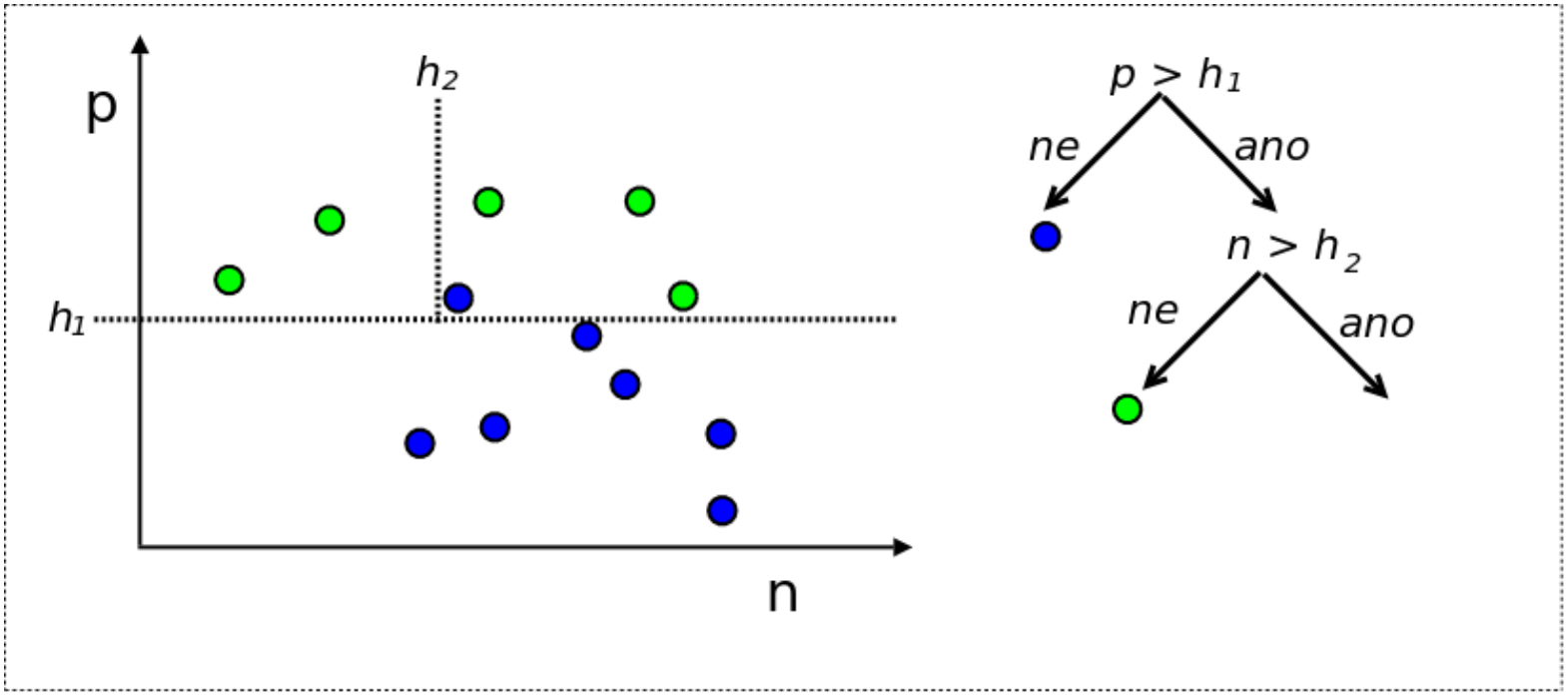
Separace rozhodovacím stromem

Separace v prostoru dvou reálných příznaků



Separace rozhodovacím stromem

Separace v prostoru dvou reálných příznaků



Separace rozhodovacím stromem

Separace v prostoru dvou reálných příznaků