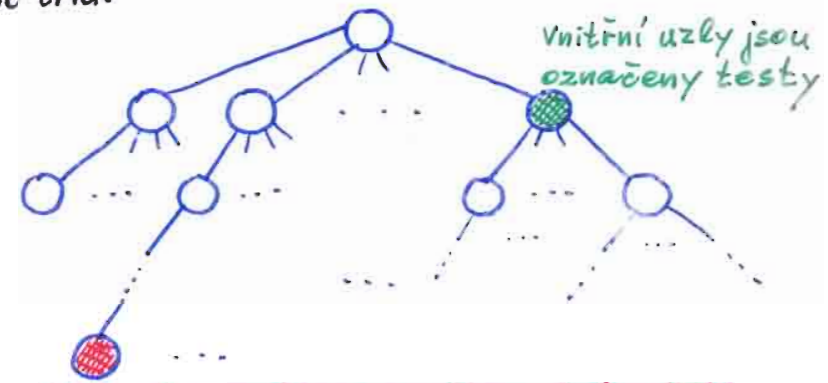
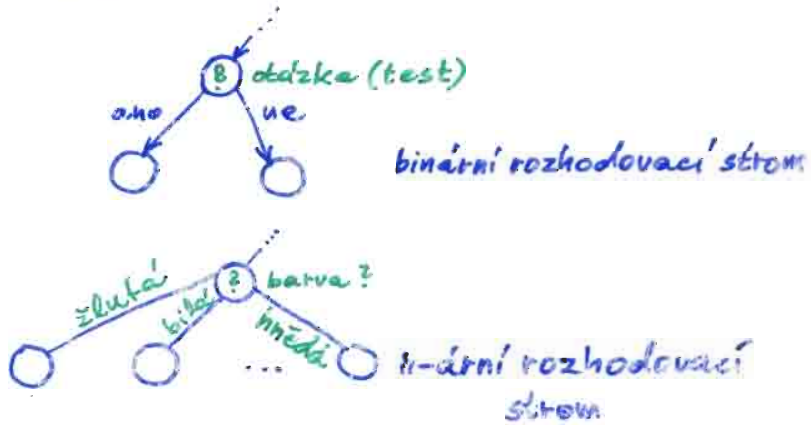
Indukování rozhodovacích stromů (ID3)

Rozhodovací stromy se používají při roztřídování dat do tříd.



Vnitřní uzel, který má asociován test s n možnými výstupy, má také n výstupních hran (jedna hrana pro každý možný výsledek testu). Je-li např. test binární (odpověď na testovací otázku je ANO/NE), pak má vnitřní uzel dvě výstupní hrany a tedy dva potomky. Jedna z hran je pak označena ANO, druhá NE:

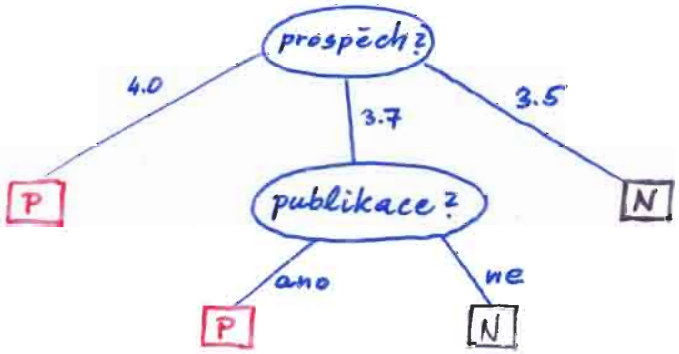


Příklad: Uvažme např. kritéria pro přijímání kandidátů postgraduálního studia na nějakou mýtickou universitu.

Každý kandidát je ohodnocen čtyřmi atributy:
- průměrný prospěch (možné hodnoty 4.0, 3.7, 3.5);
- kvalita absolvované university (top-10, top-20 [tj. 11...20], top-30 [21...30]);
- publikace (dosud publikoval / nepublikoval);
- váha (významnost) doporučení (dobré doporučení, běžné doporučení).

Kandidáti jsou klasifikovdni do dvou tříd: Přijat (P-pozitivní) a Nepřijat (N-negativní).

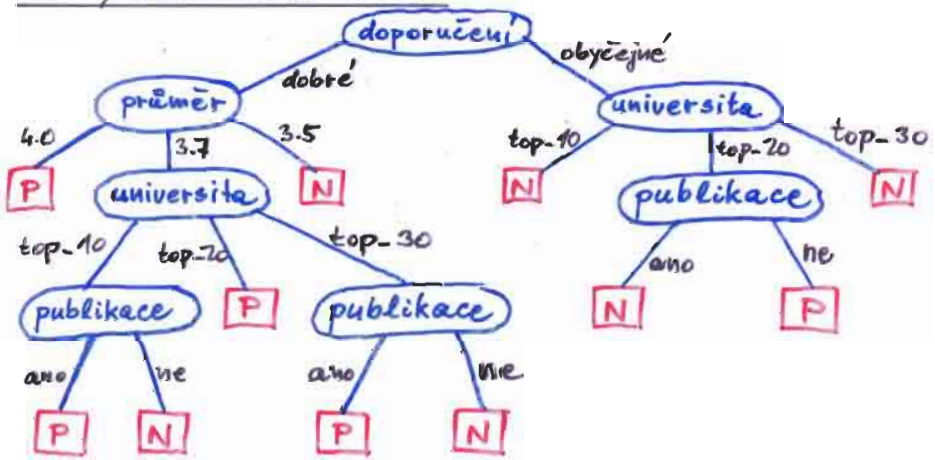
Jeden z možných rozhodovacích stromů ukazuje obrázek:



Příklad dat a klasifikací:

č.	ATRIBUTY				klasifikace
	průměr	universita	publikace	doporučení	
	4.0	top-10	ano	dobré	P
	4.0	top-10	ne	dobré	P
	4.0	top-20	ne	obyčejné	P
	3.7	top-10	ano	dobré	P
	3.7	top-20	ne	dobré	N
	3.7	top-30	ano	dobré	P
	3.7	top-30	ne	dobré	N
	3.7	top-10	ne	dobré	N
	3.5	top-20	ano	obyčejné	N
	3.5	top-10	ne	obyčejné	N
	3.5	top-30	ano	obyčejné	N
	3.5	top-30	ne	dobré	N

Složitý rozhodovací strom:



Kategorizace dat prostřednictvím rozhodovacích stromů je přímočará: na začátku je údaj testován v kořeni stromu. V závislosti na výsledku testu je údaj předán přes příslušnou hranu na uzel ležící na konci této hrany, atd. až je dosažen list, který přímo určuje třídu do níž zkoumaný údaj spadá.

ID3

V kontextu metody indukce rozhodovacích stromů ID3 mají data a testy partikulární formu. Na počátku se určí konečný soubor atributů a každému atributu se přiřadí množina možných hodnot. Dále se stanoví soubor kategorií, do nichž musí údaje spadat. U binárních rozhodovacích stromů se jedná vždy o 2 kategorie.

Každý údaj je tvořen dvojicí (kategorie, vektor vlastností), kde vektorem vlastností rozumíme soubor hodnot atributů (zde 1 hodnotu na každý atribut).

Test v každém vnitřním uzlu zkoumá právě jeden z atributů. Hrany vycházející z těchto uzlů jsou označeny možnými hodnotami atributu.

Stanovení problému: Je dán nějaký soubor dat a jejich požadovaná kategorizace do několika tříd. Cílem je nalezení rozhodovacího stromu, který bude data klasifikovat správně.

Jednou z rozšířených a úspěšně aplikovaných metod generování rozhodovacích stromů je tzv. **ID3** (Ross Quinlan z Austrálie). ID3 konstruuje strom rekurzivním způsobem:

1. Na počátku jsou data považována za členy jediné ekvivalentní třídy. Pokud by tomu tak skutečně bylo, procedura končí, strom se pak skládá z jediného uzlu.

2. Nevyhovuje-li identická klasifikace, ID3 vybere jeden atribut a rozdělí data podle různých hodnot tohoto atributu (tzn. že data mající tutéž hodnotu atributu vytvoří novou ekvivalentní třídu).

3. Každá ekvivalentní třída je opakovaně rozdělena uvedeným způsobem (za použití dalších atributů), a proces končí tehdy když je každá ekvivalentní třída klasifikována identicky.

Jediným nezřejmým krokem je výběr děliciho atributu. Tento výběr je kritický, protože různé volby tohoto atributu mohou vést k radikálně různým stromům.

Intuitivně nás vede snaha vybrat takový atribut, který vede k co nejjednodušším stromu. ID3 podchycuje tuto intuici odkazem na formální pojem **entropie** z informační teorie.

Pro jednoduchost uvažme pouze binární kategorizace (princip je ovšem platný pro libovolné n-ární kategorizace).

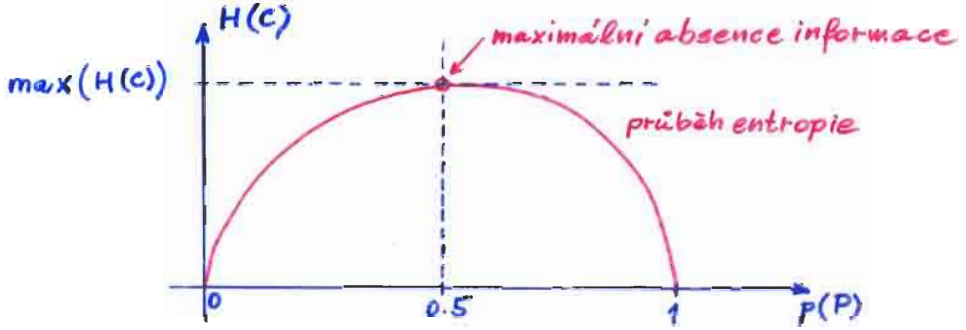
Existuje možnost, jak stanovit množství informace přítomné v daném souboru (kategorizovaných) dat. Intuitivně: čím uniformnější jsou data, tím vyšší je informační obsah v souboru. Jsou-li veškerá data kategorizována identicky, je informační obsah nejvyšší. Je-li polovina dat kategorizována jako „P" a polovina jako „N", je informační obsah nejnižší.

Obvykle se zjišťuje opak, tj. místo (míry) obsahu informace se používá míra absence informace, zvaná entropie.

Nechť $p(P)$ označuje část dat (v souboru) kategorizovaných jako P, a $p(N)$ obdobně data kategorizovaná jako N. Platí $p(P)+p(N)=1$ (celek = 1).
Entropie souboru dat, $H(C)$, je definována vztahem:

$$H(C) = -\left(p(P)\log p(P) + p(N)\log p(N)\right)$$

$p(P)=0$ a $p(N)=1$ vede k nejnižší možné entropii ($=0$), zatímco $p(P)=0.5$ vede k nejvyšší možné entropii, jak lze snadno ověřit:



Rozšíření definice entropie na soubor disjunktních souborů je snadné: Celková entropie „souboru souborů" je jednodušše součet entropií jednotlivých souborů:

$$H(C) = -\sum_n p_n \log_2 p_n$$

Použití entropie v ID3: Předpokládejme, že některá vstupní data skončí v jistém uzlu, přičemž některá jsou označena „P" a některá „N" (jinak by nebylo zapotřebí uzel dělit). Dále předpokládejme, že zvolíme některý atribut pro další dělení. Rozdělením uzlu stromu vznikne několik nových uzlů-potomků, a data přidružená předtím k tomuto uzlu jsou nyní rovnoměrně rozdělena mezi potomky. Každý soubor, přidružený nyní k novému potomku, má svou partikulární entropii. Celková entropie potomků je tedy součtem jejich individuálních entropií.

Vraťme se nyní k původní otázce: který atribut použít k dělení daného uzlu?

Odpověď je jednoduchá: Zvolí se atribut, který vede k nejnižší celkové entropii $H(C)$, tj. k maximální celkové informaci, ve výsledných uzlech potomků.

viz R. Quinlan: Program C4.5 pro ML (implementace v C pro Unix/Sun)

Úplný algoritmus ID3 ovšem počítá i s jinými faktory, např.:
- data mohou být dělena do více než dvou tříd
- některá data mohou postrádat hodnoty některých atributů
- trénovací data mohou obsahovat šum
- některé klasifikace mohou být chybné

Příklad: Předpokládejme, že nevíme, které faktory způsobují, že někteří lidé jsou po krátkém pobytu na pláži spálení, zatímco ostatní jsou vpořádku a opálení. Můžeme zajít na pláž a příčiny uvedených jevů studovat. Pozorováním zjistíme, že lidé se liší barvou vlasů, výškou, váhou. Někteří se natírají krémem, jiní ne. Někteří zčervenají, někteří ne. Chceme zjistit, které vlastnosti nám umožní předpovědět, zda nový příchozí na pláž (tj. osoba, jež nenáleží do trénovací množiny) se spálí či nikoliv.

Lze kupř. hledat shodu mezi vlastnostmi nově přicházejícího a již studovaných příkladů, ale šance na přesnou shodu jsou malé. Dejme tomu, že naše pozorování dalo následující údaje:

	jméno	vlasy	výška	váha	krém	výsledek
1.	Zuzana	blond	přům.	nízká	ne	spálená
2.	Dana	blond	vysoká	přům.	ano	nic
3.	Pepa	hnědé	malá	přům.	ano	nic
4.	Anna	blond	malá	přům.	ne	spálená
5.	Jana	zrzavé	přům.	vysoká	ne	spálená
6.	Petr	hnědé	vysoká	vysoká	ne	nic
7.	Pavel	hnědé	přům.	vysoká	ne	nic
8.	Kateřina	blond	malá	nízká	ano	nic

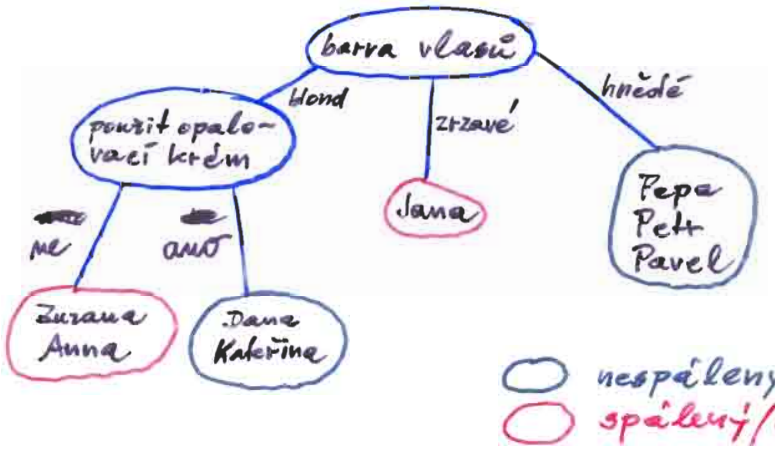
Celkem je 54 možných kombinací hodnot atributů ($3 \times 3 \times 3 \times 2 = 54$). Je-li nový příchozí náhodným vzorkem, pak pravděpodobnost přesné shody s někým z tabulky je $8/54 = 0.15$ (pouze 15%). V praxi může být pravděpodobnost ještě nižší, např. 12 atributů po 5 hodnotách s rovnoměrným rozložením dává $5^{12} = 2.44 \times 10^8$. Pro tabulku s 10^6 řádky lze tedy očekávat pouze 0.4% přesné shody.

Z uvedených důvodů může být nepraktické klasifikovat neznámý objekt hledáním přesné shody mezi měřenými hodnotami tohoto objektu a vzorků se známou klasifikací.

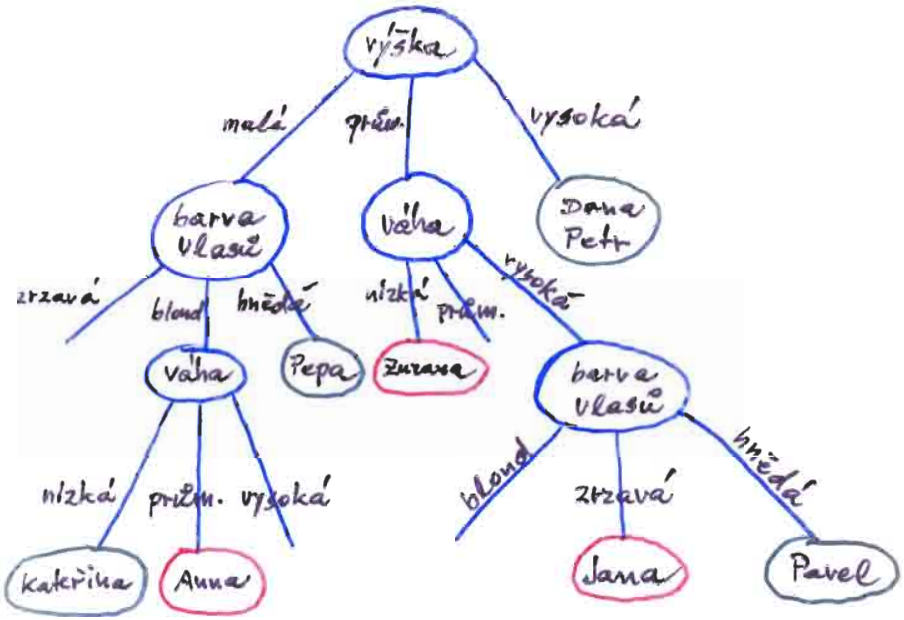
~~(Je možné použít např. prostá věta?)~~

Použití identifikačních stromů: Identifikační strom je rozhodovací strom, v němž každý soubor možných závěrů je implicitně vytvořen seznamem příkladů, jejichž klasifikace je známa.

Problém: často předem nevíme, které z atributů jsou rozhodující pro klasifikaci a které atributy jsou irelevantní:



Příklad identifikačního stromu, jenž je konsistentní s databází (tabulkou). Strom je konsistentní s přirozenou intuicí o „spálení sluncem".



Jiný příklad identifikačního stromu konsistentního s databází. Tento strom je ovšem rozsáhlejší a není konsistentní s přirozenou intuicí.

Identifikační strom z předchozího obrázku se zdá být (a také je) lepší, ale jak má počítačový program dojít ke stejným závěrům bez a priorní znalosti o účincích opalovacího krému či o tom, jak se vztahuje barva vlasů k možnosti být spálen sluncem?

Jedna možná odpověď je založena na aplikaci tzv. Occamova ostří:

> Svět je v principu jednoduchý. Tudíž ten nejmenší identifikační strom konsistentní s příklady je nejpravděpodobnější ten nejvhodnější k identifikaci neznámých objektů.

Otázka se nyní mění na Jak zkonstruovat nejmenší identifikační strom ?

Je-li zapotřebí mnoho testů ke klasifikaci, je výpočetně nepraktické hledat zaručeně nejmenší strom. Proto je lepší sestavit postup, který má tendenci vytvářet malé stromy, byť není teoreticky zaručeno, že najde ten nejmenší.

Jedna z možností, jak začít, je v kořeni zvolit takový test, který co nejlépe rozdělí databázi příkladů na podmnožiny, v nichž co nejvíce prvků má tuléž klasifikaci. Pak pro každý soubor, obsahující více než jeden druh příkladů, vybereme další test, jenž by měl rozdělit množinu nehomogenní na homogenní podmnožiny.

Z předchozího obrázku (složitý strom) je patrno, že test na váhu je zřejmě nejhorší vzhledem k tomu, jak mnoho lidí skončí v homogenních množinách. Po aplikaci váhového testu nikdo není v homogenním souboru. Test na výšku je o něco lepší (2 lidé skončí v homogenní množině). Ještě lepší je test na opalovací krém (3 vzorky v homogenním souboru). Nejlepší je však test na barvu vlasů, neboť 4 vzorky jsou zařazeny do homogenní množiny. Proto by měl být jako první použit test na barvu vlasů.

Příklad: Mějme množinu obsahující členy dvou tříd, A a B. Jsou-li počty obou tříd vyváženy, pak neuspořádanost (tj. její míra) = 1 (maximum):

$$H = \sum_c -\frac{n_{bc}}{n_b}\log_2\frac{n_{bc}}{n_b} = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} =$$

$$= \frac{1}{2} + \frac{1}{2} = \underline{1}$$

Existují-li však pouze členové A (nebo jen B), pak

$$H = -1\cdot\log_2 1 - 0\cdot\log_2 0 = -0-0 = \underline{0}$$

tj. žádná neuspořádanost (= dokonalá uspořádanost).

Můžeme-li měřit neuspořádanost v jedné množině, lze měřit průměrnou neuspořádanost množin na konci větví vedoucích z testu: Neuspořádanost množiny každé větvě se váhuje rozměrem množiny n_b relativně vzhledem k celkovému rozměru množin ve všech větvích n_t.

$$\overline{H} = \sum_b \frac{n_b}{n_t} \times (\text{Neuspořádanost větve } \underline{b})$$

Závěr: Dobrý test minimalizuje neuspořádanost.

U reálných databází je nepravděpodobné, že libovolný test dá zcela homogenní množiny. Proto je nutné určit míru nehomogenity, tj. míru neuspořádanosti v podmnožinách produkovaných jednotlivými testy.

Teorie informace poskytuje vztah pro průměrnou neuspořádanost:

$$\overline{H} = \sum_b \left(\frac{n_b}{n_t}\right) \times \left(\sum_c - \frac{n_{bc}}{n_b} \log_2 \frac{n_{bc}}{n_b}\right)$$

n_b ... počet příkladů ve větvi b
n_t ... celkový počet příkladů ve všech větvích
n_{bc} ... celkový počet příkladů ve větvi b třídy c

Vztah, zahrnující n_{bc} a n_b pro nějakou větev b je

vztah pro neuspořádanost:

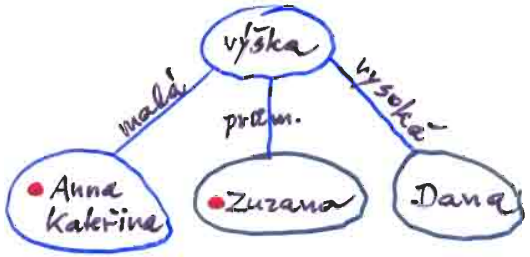
$$H = \sum_c - \frac{n_{bc}}{n_b} \log_2 \frac{n_{bc}}{n_b}$$

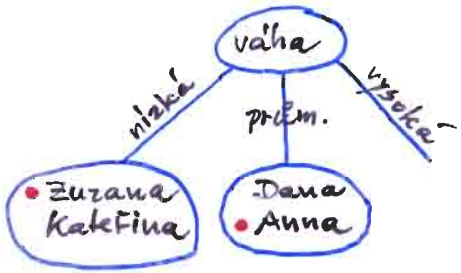
poměrná část náležející do třídy c

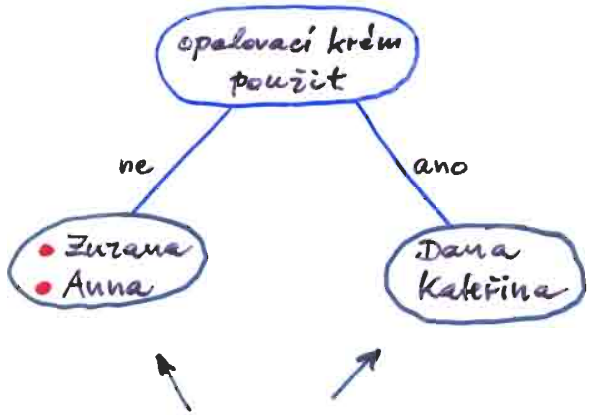
(Tento vztah není posvátný, ale ukázal se být jako velmi užitečný, proto se používá - lze ovšem zvolit i jiný vztah, pokud bude dávat dobré výsledky.)

(Pozn.: $0 \log_2 0$ je definován pro entropii jako 0.)

Pro výběr dalšího testu na rozdělení nehomogenní množiny se opět vybere ten, který je nejúčinnější:





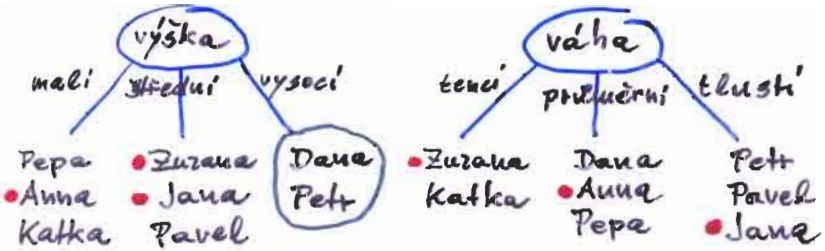


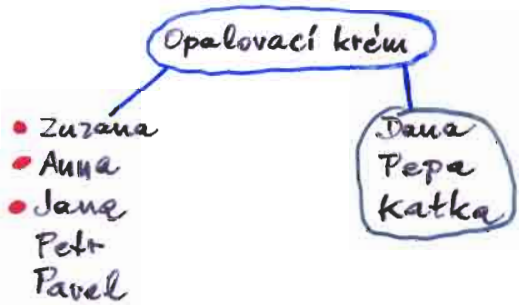
2 homogenní a žádná nehomogenní množina, proto je test na opalovací krém nejlepší.

Zpátky k příkladu slunečního spálení:

Test na barvu vlasů dělí vzorky do 3 množin (blond, zrzaví, hnědovlasí). Průměrná neuspořádanost:

$$\bar{H} = \underbrace{\frac{4}{8}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right)}_{\text{blond}} + \underbrace{\frac{1}{8}\times 0}_{\text{zrzaví}} + \underbrace{\frac{3}{8}\times 0}_{\text{hnědovlasí}} = 0.5$$





$\bar{H}$(vlasy) = 0.50 ← nejmenší prům. neuspořádanost
$\bar{H}$(výška) = 0.69 (3.)
$\bar{H}$(váha) = 0.94 (4.)
$\bar{H}$(krém) = 0.61 (2.)

Měl by tedy jako první být použit test na barvu vlasů.

Pro výběr druhého testu provedeme obdobné výpočty:

$\bar{H}$(výška) = 0.5
$\bar{H}$(váha) = 1.0
$\bar{H}$(krém) = 0.0 ← zřejmý vítěz

Pro náš konkrétní příklad tedy sestavíme rozhodovací strom jako posloupnost testů:

(1) barva vlasů
(2) použití opalovacího krému

Pro generování identifikačního stromu lze použít následující proceduru zvanou SPROUTER*:

- Opakuj dokud každý list není co nejhomogennější:
  - Vyber list s nehomogenní množinou příkladů.
  - Nahraď list testovacím uzlem tak, aby test rozdělil nehomogenní soubor do souborů s minimální nehomogenitou v souladu se zvolenou metodou výpočtu míry neuspořádanosti.

* to sprout [spraut] = vyhánět výhonky, klíčit, rašit...

Od rozhodovacích stromů k pravidlům

Je-li jednou sestaven rozhodovací strom, je velmi jednoduché ho zkonvertovat na soustavu ekvivalentních pravidel:

Sleduje se každá cesta od kořene stromu k listu, výsledky testů se zaznamenávají jako antecedenty a listové klasifikace jako konsekventy.

Náš rozhodovací strom o slunečním spálení dá tedy 4 ekvivalentní pravidla:

1. IF barva_vlasů = blond AND
   použití_opalovacího_krému = ano
   THEN nic_se_nestane

2. IF barva_vlasů = blond AND
   použití_opalovacího_krému = ne
   THEN osoba_se_spálí

3. IF barva_vlasů = zrzavá
   THEN osoba_se_spálí

4. IF barva_vlasů = hnědá
   THEN ~~osoba_se_nespálí~~ nic_se_nestane

Nepotřebné antecedenty by měly být z pravidel odstraněny:

Po vytvoření souboru pravidel by pravidla měla být co nejvíce zjednodušena a posléze by měla být eliminována nepotřebná pravidla.

V našem příkladu mají dvě pravidla dva antecedenty. (viz pravidlo č. 1 a 2.) Je dobré zjistit, zda jsou oba antecedenty nezbytné. Např. pravidlo

antecedent 1 — antecedent 2

IF barva_vlasů = blond AND použití_opal._krému = ano
THEN nic_se_nestane
konsekvent

Odstraněním 1. antecedentu (o blond vlasech) je pravidlo aktivováno pro každou osobu používající opalovací krém. Tři vzorky (Dana, Pepa, Katka) používají krém a žádný z nich se nespálí. Z toho plyne, že spálení zde nezávisí na barvě vlasů a dané pravidlo lze redukovat na:

IF použití_opal._krému = ano THEN nic_se_nestane

Pozn.: Metodu SPROUTER např. úspěšně použila firma Westinghouse při zvyšování výtěžnosti přeměny plynu hexafluoridu uranu na dioxid uranu (30 parametrů procesu). Někdy to šlo lépe, jindy hůře. Experimentovat s jadernou elektrárnou nebylo možné. Pomocí SPROUTERu byl vytvořen nejjednodušší rozhodovací strom, který z parametrů je pro řízení procesu podstatnější a který ne. Investice do změny řízení se vrátila za 1/2 dne.