Ochutnávka strojového učení

Úvod do problematiky

Barbora Hladká
http://ufal.mff.cuni.cz/bvh

Univerzita Karlova
Matematiko-fyzikální fakulta
Ústav formální a aplikované lingvistiky

Ochutnávka strojového učení

Cílem přednášky je ilustrovat základní myšlenku strojového učení pomocí několika reálných příkladů. Řešení jednoho z nich bude tématem praktické části, kdy projdeme hlavní kroky učení jednoho vybraného algoritmu v systému R. Obě dvě části jsou úvodní, proto nevyžadují žádné speciální znalosti a dovednosti posluchačů. Jak název napovídá, v první části přičichneme k nalité sklence a ve druhé části sklenkou zatočíme a její obsah poválíme lehce na jazyku.

Páteční večer – ilustrační příklad

- vybrat filmy na víkend
- ručně (téměř) nemožné
- $\rightarrow$ **automaticky**

  Pro uživatele ohodnoťte, jak moc se mu bude film líbit.
  Použijte známky 1-5 (1-znechucení, 5-nadšení).

  Např: Určete pro Jarmilu známku filmu *Někdo to rád horké*

  Co nás zajímá

  - které filmy už Jarmila viděla a jak se jí líbily
  - kteří uživatelé mají podobný vkus
  - které filmy jsou podobné

Co je strojové učení

Připravme příklady a naučme počítač učit se z příkladů.

To je strojové učení!

Strojové učení potřebuje příklady

	Příběh hraček (1995)	Hvězdné války (1977)	Někdo to rád horké (1959)
Vilém	–	**5**	**4**
Hynek	**2**	**5**	–
Jarmila	**2**	**4**	–

Jak dostat Jarmilu do počítače?

Popíšeme ji **příznaky** (atributy). Tím vznikne **vektor příznaků**:

jméno	věk	pohlaví	profese	PSČ
Jarmila	50	F	zdravotní sestra	60657

Popíšeme i filmy.

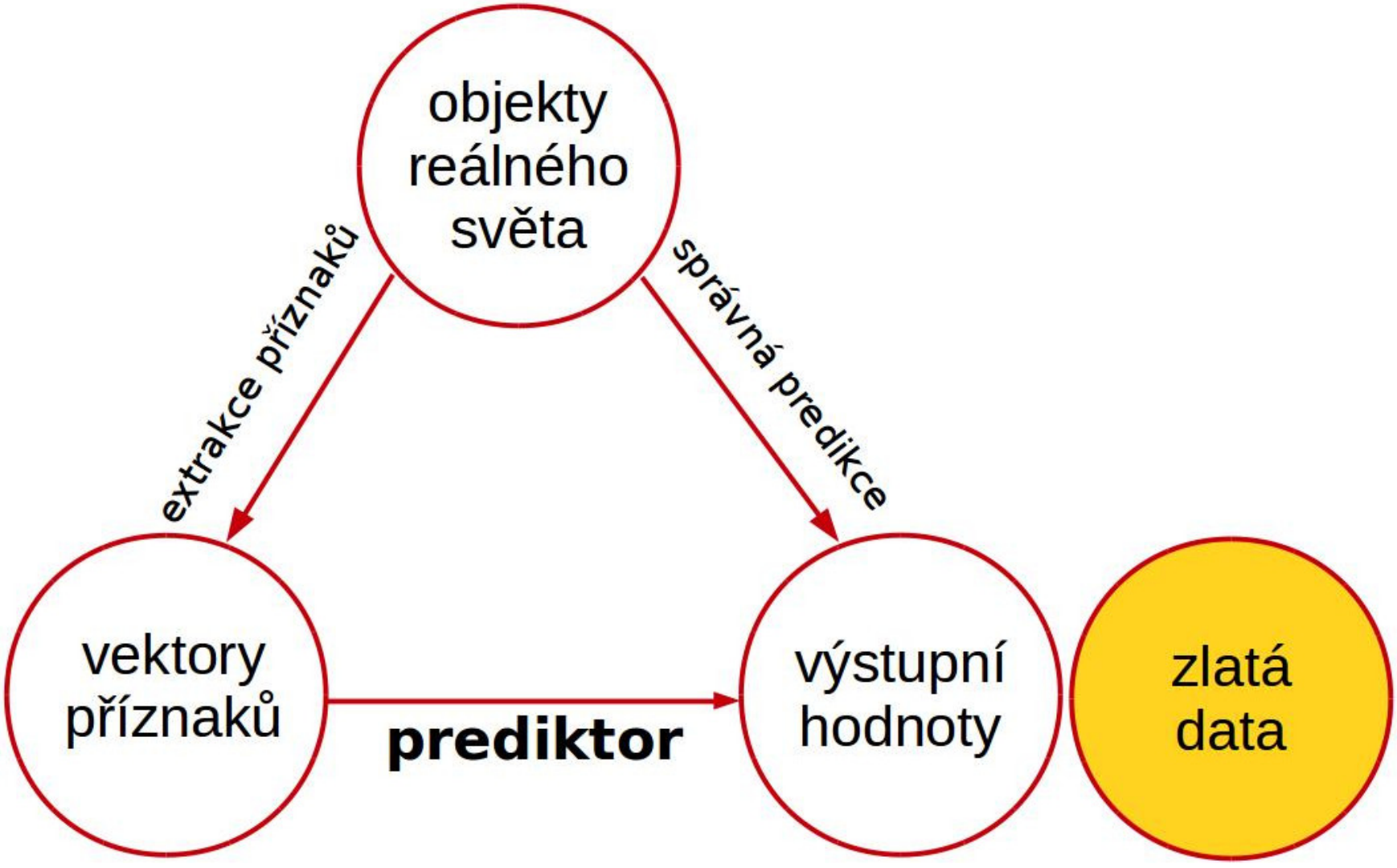
název	animovaný	hodnocení IMDb	režisér
Někdo to rád horké	ne	8.3	Billy Wilder
Příběh hraček	ano	8.3	John Lasseter

Příklady pro strojové učení

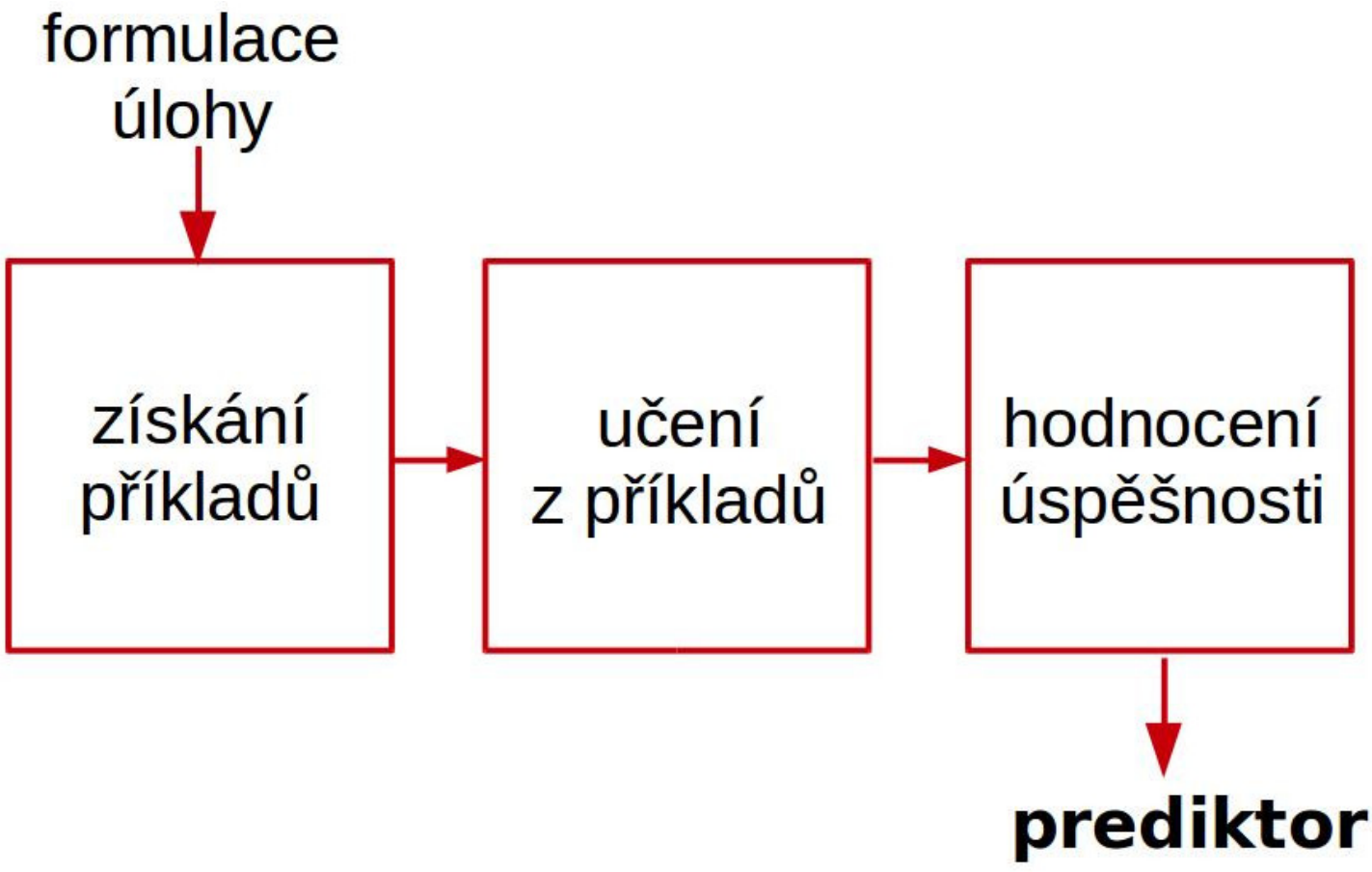
příklad = vektor příznaků + výstupní hodnota (pokud je známa)

<Jarmila,50,F,zdravotní sestra,60657,Příběh hraček,ano,8.3,John Lasseter> + 4

Proces strojového učení



Proces strojového učení



Proces strojového učení
Formulace úlohy

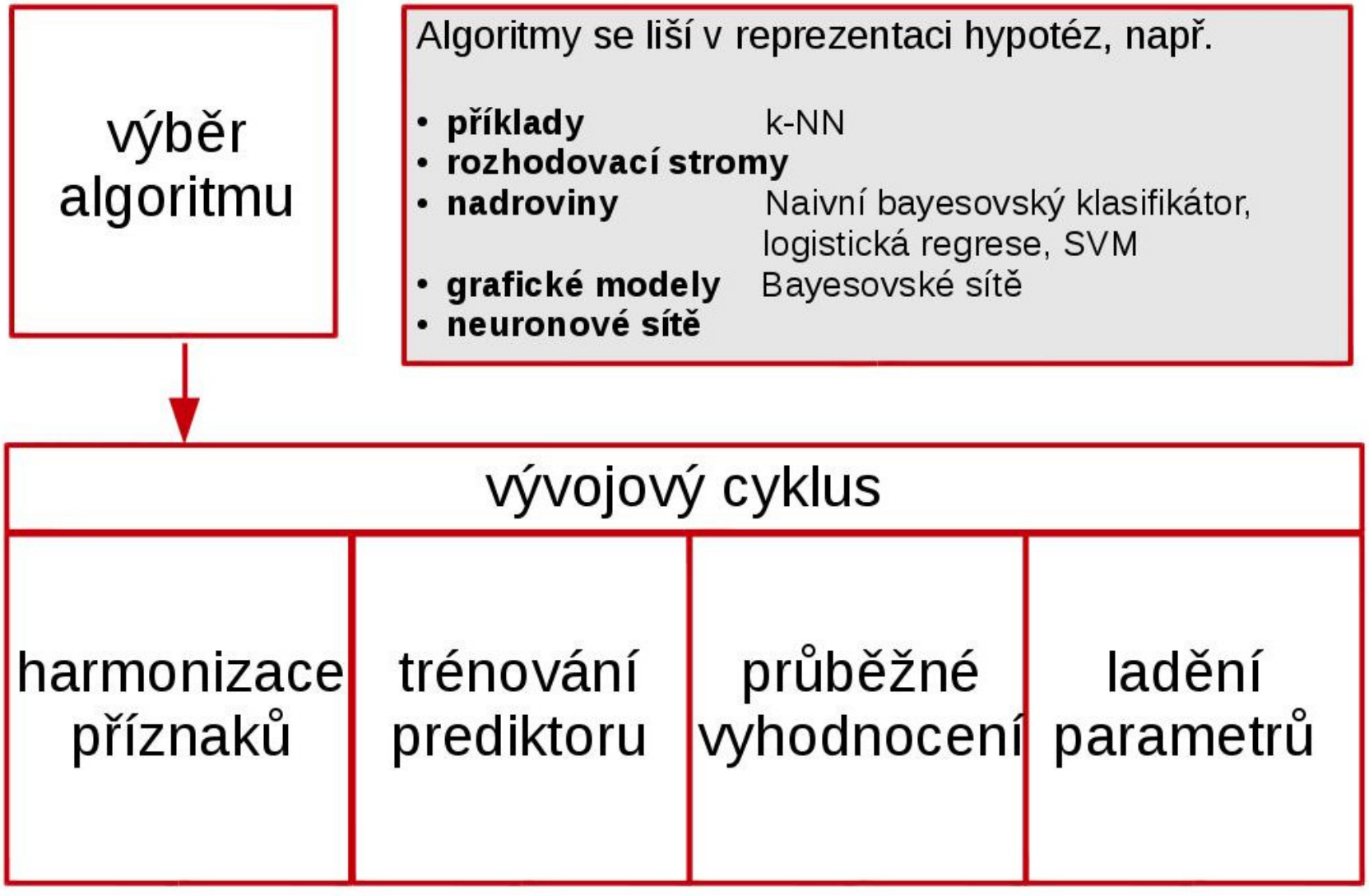
1. Ohodnoťte film pro uživatele.
   Výstupní hodnoty: 1,2,3,4,5
2. Odhadněte prodeje produktů v následujícím měsíci.
   Výstupní hodnoty: celá čísla
3. Rozpoznejte na obrázku psa, nebo kočku.
   Výstupní hodnoty: pes, kočka
4. Určete význam slova *line* v anglické větě.
   Př. I've got Inspector Jackson on the *line* for you.
   Výstupní hodnoty: cord, phone, division, formation, product, text

Proces strojového učení
Získání příkladů, extrakce příznaků

1. Zorganizujte sběr hodnocení.
2. Zorganizujte sběr informací.
   Příznaky: informace o prodejci, o zboží, prodeje v minulosti, ...
3. Shromážděte obrázky psů a koček a ručně je oštítkujte.
   Příznaky: informace opixelech
4. Vyberte věty s *line* a ručně přiřaďte významy.
   Příznaky: předcházející/následující slovo, typické slovo, ...

Rozdělte příklady na trénovací a testovací.

Proces strojového učení
Učení z trénovacích příkladů



Proces strojového učení
Hodnocení úspěšnosti

- testovací data jsou pro vývojáře během učení skryta
- test prediktoru na testovacích datech

Jde o co nejlepší **generalizaci**!

Zabránit **přetrénování**!

Strojové učení a umělá inteligence

Strojové učení a umělá inteligence

Marvin Minsky, 1967

Umělá inteligence je věda o vytváření strojů nebo systémů, které budou při řešení určitého úkolu užívat takového postupu, který – kdyby ho dělal člověk – bychom považovali za projev jeho inteligence.

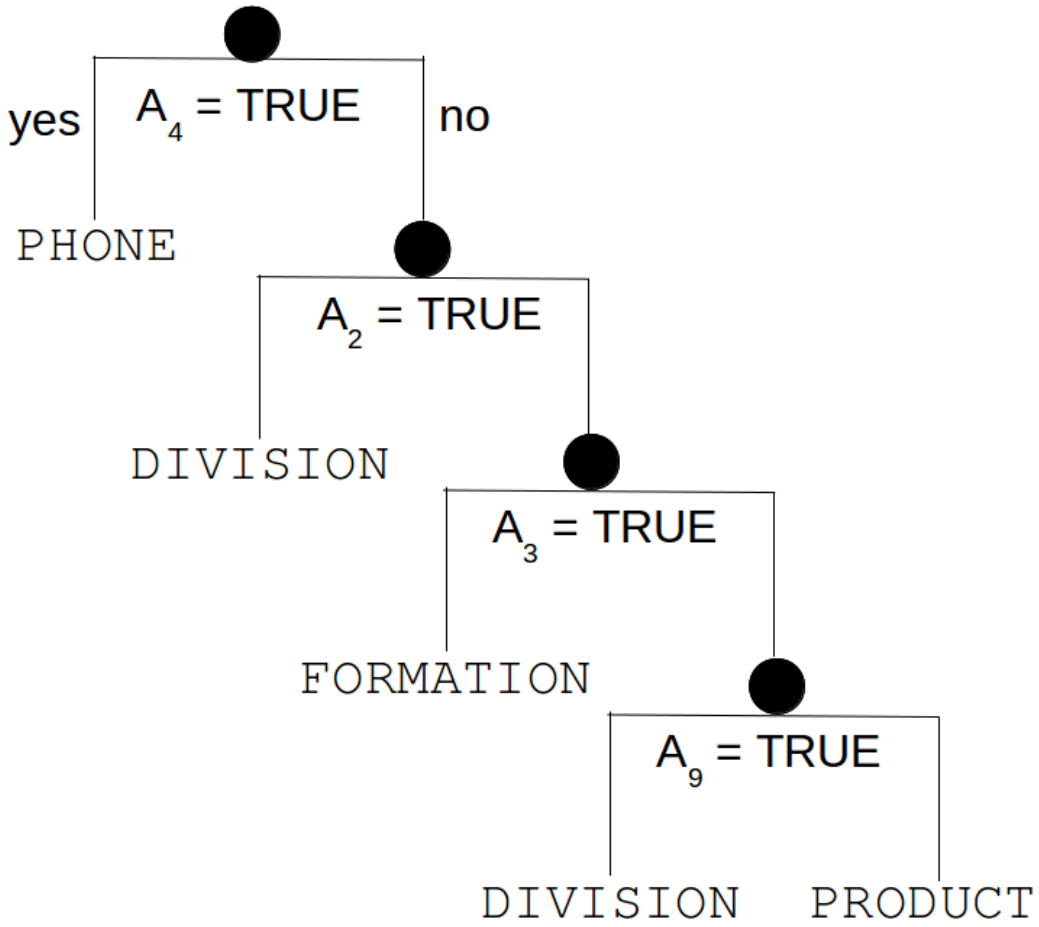
Zkáza Titaniku v rozhodovacích stromech

- Úloha: Předpovězte, zda-li cestující na Titaniku jeho zkázu přežije
- Výstupní hodnoty: 0 (ne), 1 (ano)
- Zdroj: `https://www.kaggle.com/c/titanic`
- Algoritmus rozhodovacích stromů v kostce
- Cvičení v R, viz data a kód na `http://ufal.mff.cuni.cz/bvh`

Co si myslí zákazníci

- Facebook Data for Sentiment Analysis
- `http://hdl.handle.net/11858/00-097C-0000-0022-FE82-7`
- Vyzkoušejte přípravu dat `https://tinyurl.com/ycdfrpbg`

Rozhodovací strom pro určování významu slova *line*



Klasifikace pomocí stromu

Příklad

Určete význam slova *line* ve větě *Draw a line between the points P and Q.* (Narýsujte čáru mezi body P a Q.)

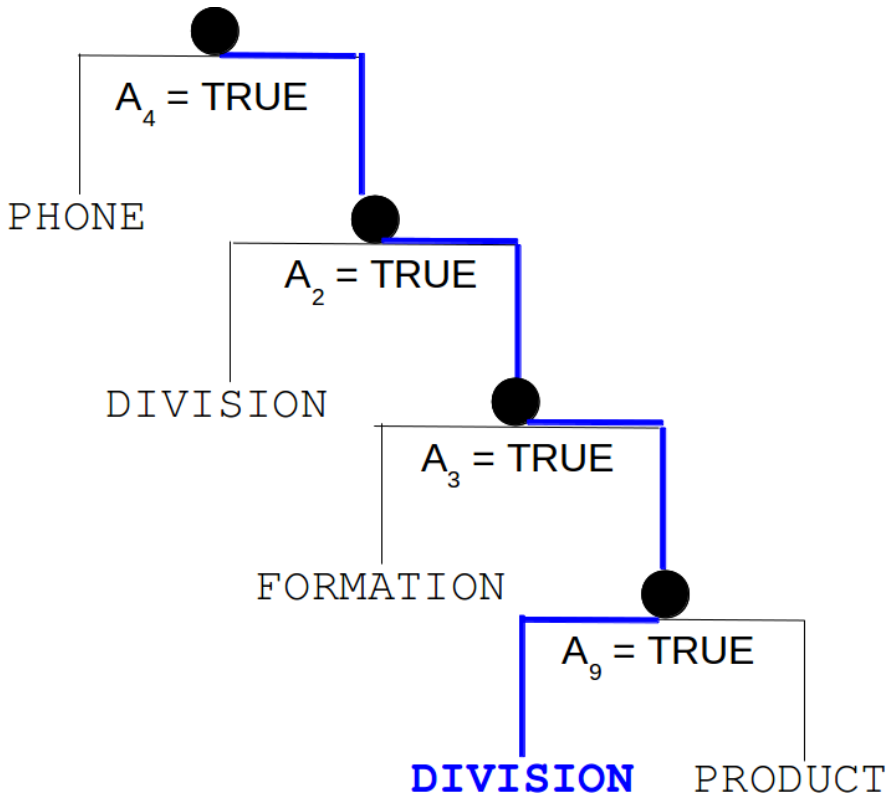
Zaprvé, extrahujte hodnoty dvaceti příznaků

A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}
0	0	0	0	0	0	0	0	1	0	0

A_{12}	A_{13}	A_{14}	A_{15}	A_{16}	A_{17}	A_{18}	A_{19}	A_{20}
a	draw	X	between	DT	IN	DT	line	dobj

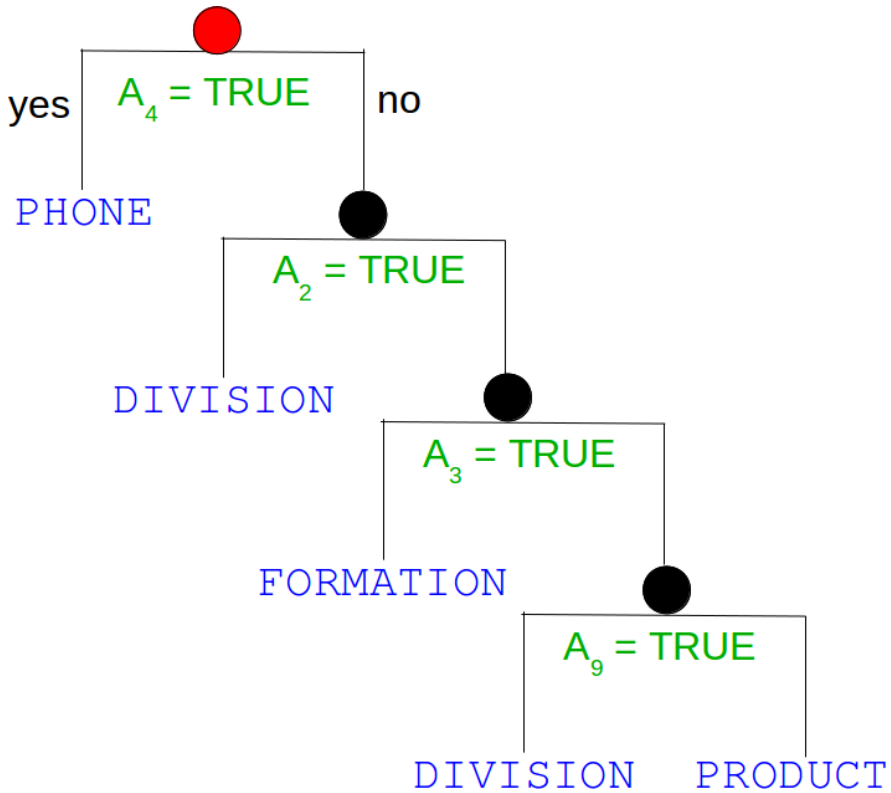
Klasifikace pomocí stromu

Zadruhé, klasifikujte příklad.



Popis stromové struktury

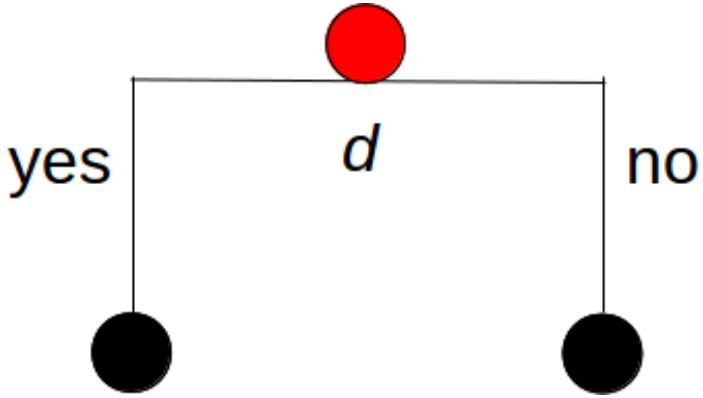
- Uzly
  - Kořen
  - Vnitřní uzly
  - Listy s `VÝSTUPNÍMI HODNOTAMI`
- Rozhodnutí
  - Binární otázky na hodnoty jednoho příznaku, tj. každý vnitřní uzel má dvě děti



Konstrukce rozhodovacího stromu z trénovacích příkladů

- **Krok 1** Vytvoř kořen

- **Krok 2** Vyber rozhodnutí d a přidej dvě děti k existujícímu uzlu



Konstrukce rozhodovacího stromu z trénovacích příkladů

Jak vybrat rozhodnutí d?

Spoj kořen s trénovací množinou t.

Příklad

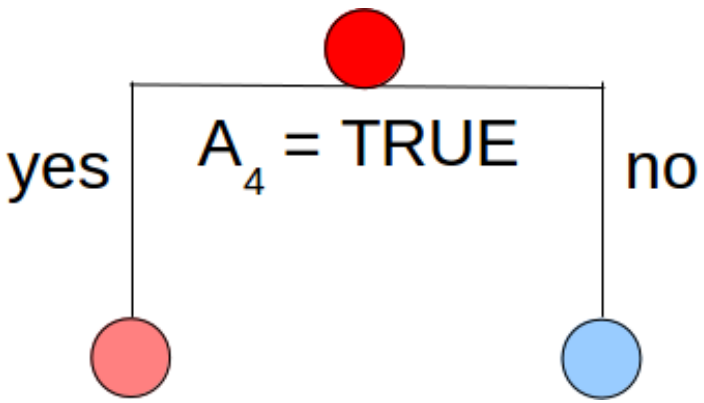
1. Uvažuj rozhodnutí `if` $\mathrm{A}_4 = \mathit{TRUE}$.
2. Rozděl trénovací množinu t podle tohoto rozhodnutí na dvě podmnožiny – "růžová" a "modrá"

t

VÝZNAM	...	A4	...
FORMATION		TRUE	
FORMATION		FALSE	
PHONE		TRUE	
CORD		TRUE	
DIVISION		FALSE	
...	...	...	

Konstrukce rozhodovacího stromu z trénovacích příkladů

3. Ke kořeni přidej dvě děti, "růžové" a "modré". Spoj každé z nich s odpovídající podmnožinou t_L, t_R, resp.



t_L

VÝZNAM	...	A4	...
FORMATION		TRUE	
CORD		TRUE	
PHONE		TRUE	
...	...	...	

t_R

VÝZNAM	...	A4	...
FORMATION		FALSE	
DIVISION		FALSE	
...	...	...	

Konstrukce rozhodovacího stromu z trénovacích příkladů

Jak vybrat rozhodnutí d?

Pracovat s více příznaky, formulovat více rozhodnutí.

Které rozhodnutí je nejlepší?

Zaměřit se na distribuci výstupních hodnot v příslušných množinách trénovacích příkladů.

Konstrukce rozhodovacího stromu z trénovacích příkladů

Příklad pro úlohu s významem slov

- Uvažujme 120 trénovacích příkladů.
- Rozhodnutím je rozdělíme na dvě podmnožiny (1) a (2) s následující distribucí výstupních hodnot:

	CORD	DIVISION	FORMATION	PHONE	PRODUCT	TEXT	
(1)	0	0	0	120	0	0	"čistá"
(2)	20	20	20	20	20	20	"nečistá"

"čistá" podmnožina obsahuje většinu příkladů z jedné výstupní třídy

Konstrukce rozhodovacího stromu z trénovacích příkladů

Které rozhodnutí je nejlepší?

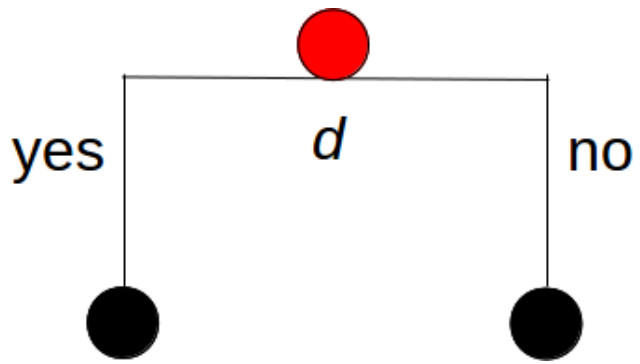
Rozhodnutí, které dělí trénovací příklady do "čistých"podmnožin.

Konstrukce rozhodovacího stromu z trénovacích příkladů

Algoritmus rozhodovacích stromů – základní formulace

- **Krok 1** Vytvoř kořen

- **Krok 2** Vyber rozhodnutí d a přidej dvě děti k existujícímu uzlu



- **Krok 3** Rozděl trénovací příklady spojené s rodičovským uzlem t podle rozhodnutí d do podmnožin t_L a t_R.
- **Krok 4** Rekurzivně opakuj kroky (2) a (3) pro oba dětské uzly a s nimi spojené trénovací příklady.
- **Krok 5** Zastav, jakmile je uzel spojený s trénovacími příklady z jedné výstupní třídy. Vyvoř list s touto výstupní hodnotou.

Křížová validace

test	train	train	train
train	test	train	train
train	train	test	train
train	train	train	test

Přetrénování