Machine learning a data mining 1

Jan Outrata

KATEDRA INFORMATIKY
UNIVERZITA PALACKÉHO V OLOMOUCI

přednášky

Klasifikace

Regrese

= **prediktivní modelování spojitého** cílového atributu (**závislé proměnné**) y jako **(cílové) funkce**, **regresního modelu**, f jiných atributů (**nezávislých proměnných**) $x_1, \ldots, x_m$

$\rightarrow$ predikce (neznámé) hodnoty atributu y objektu z jeho (známých) hodnot atributů $x_1, \ldots, x_m$

- **univariate** = jeden atribut x
- **multivariate** = více atributů $x_1, \ldots, x_m$
- mnoho aplikací: modelování a předpovědi počasí, cen, prodeje, vývoje atd.

Lineární regrese

= **lineární** cílová funkce (model): $f(x_1, \ldots, x_m) = a_0 + a_1 x_1 + \ldots + a_m x_m \approx y$, $a_0, \ldots, a_m$ **regresní koeficienty**

- n objektů $\Rightarrow$ maticová reprezentace: objekty $\sim$ řádky, atributy $\mathbf{x_j} = (x_{1j}, \ldots, x_{nj})^T$, $j = 1, \ldots, m \sim$ sloupce (maticová data), cílový atribut $\mathbf{y} = (y_1, \ldots, y_n)^T \sim$ sloupec

$\rightarrow$ soustava lineárních rovnic $f(\mathbf{x_1}, \ldots, \mathbf{x_m}) = \mathbf{Xa} \approx \mathbf{y}$, $\mathbf{X} = (\mathbf{1}\ \mathbf{x_1}\ \ldots\ \mathbf{x_m})$ **design matrix**, $\mathbf{1} = (1, \ldots, 1)^T, \mathbf{a} = (a_0, \ldots, a_m)^T$:

$$\begin{aligned} f(x_{11}, \ldots, x_{1m}) = a_0 + a_1 x_{11} + \ldots + a_m x_{1m} &\approx y_1 \\ &\vdots \\ f(x_{n1}, \ldots, x_{nm}) = a_0 + a_1 x_{n1} + \ldots + a_m x_{nm} &\approx y_n \end{aligned}$$

- právě jedno (přesné) řešení $\mathbf{a} = \mathbf{X}^{-1}\mathbf{y}$, pokud existuje $\mathbf{X}^{-1}$ (při $n = m + 1$), jinak žádné nebo nekonečně mnoho – nelze použít, obvykle $n > m + 1$

Lineární regrese

→ nejbližší řešení $\mathbf{a}$ . . . minimalizace chyby mezi predikovanou $f(\mathbf{x_1}, \ldots, \mathbf{x_m})$ a skutečnou hodnotou cílového atributu $\mathbf{y}$ = **regresní problém**: **chybová funkce** – typicky squared error $E = \sum_i^n (y_i - f(x_{i1}, \ldots, x_{im}))^2 = \sum_i^n (y_i - (a_0 + a_1 x_{i1} + \ldots + a_m x_{im}))^2 \rightarrow$ **least squares problem + method**

- univariate: $E = \sum_i^n (y_i - (a_0 + a_1 x_i))^2 \rightarrow \frac{\partial E}{\partial a_0} = -2 \sum_i^n (y_i - (a_0 + a_1 x_i)) = 0$, $\frac{\partial E}{\partial a_1} = -2 \sum_i^n (y_i - (a_0 + a_1 x_i)) x_i = 0 \rightarrow$

$$\begin{pmatrix} n & \sum_i^n x_i \\ \sum_i^n x_i & \sum_i^n x_i^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_i^n y_i \\ \sum_i^n x_i y_i \end{pmatrix}$$

$a_0 = \overline{y} - a_1 \overline{x}, \quad a_1 = \sigma_{xy}/\sigma_{xx}, \qquad \sigma_{xy} = \sum_i^n (x_i - \overline{x})(y_i - \overline{y}), \sigma_{xx} = \sum_i^n (x_i - \overline{x})^2$
funkce (model): $f(x) = a_0 + a_1 x = \overline{y} + \sigma_{xy}/\sigma_{xx}(x - \overline{x}) \approx y$

- multivariate:

$$\begin{pmatrix} n & \sum_i^n x_i \\ \sum_i^n x_i & \sum_i^n x_i^2 \end{pmatrix} = \begin{pmatrix} \mathbf{1}^T\mathbf{1} & \mathbf{1}^T\mathbf{x} \\ \mathbf{x}^T\mathbf{1} & \mathbf{x}^T\mathbf{x} \end{pmatrix} = \mathbf{X}^T\mathbf{X}, \qquad \begin{pmatrix} \sum_i^n y_i \\ \sum_i^n x_i y_i \end{pmatrix} = \begin{pmatrix} \mathbf{1}^T\mathbf{y} \\ \mathbf{x}^T\mathbf{y} \end{pmatrix} = \mathbf{X}^T\mathbf{y}$$

$$\mathbf{X}^T\mathbf{X}\mathbf{a} = \mathbf{X}^T\mathbf{y} \rightarrow \mathbf{a} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Nelineární regrese

= nelineární cílová funkce (model): $f(x_1, \ldots, x_m) = a_0 + \sum_k a_k g_k(x_1, \ldots, x_m) \approx y$, g_k jakékoliv (diferencovatelné) funkce, typicky polynomické

- např. kvadratická: $f(x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1 x_2 + a_4 x_1^2 + a_5 x_2^2$
  $\mathbf{X} = (\mathbf{1}\ \mathbf{x_1}\ \mathbf{x_2}\ \mathbf{x_1 x_2}\ \mathbf{x_1^2}\ \mathbf{x_2^2})$
- lze použít stejnou metodu (least squares) pro určení regresních koeficientů a_0, a_k:
  $\mathbf{X}^T\mathbf{X}\mathbf{a} = \mathbf{X}^T\mathbf{y} \rightarrow \mathbf{a} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

Vyhodnocení přesnosti (goodness of fit) regresního modelu

1. **chybová funkce** – squared error $E = \sum_i^n (y_i - f(x_{i1}, \ldots, x_{im}))^2 \in [0, \infty] \rightarrow$ minimální (0)
2. **koeficient determinace**: $\rightarrow$ maximální (1)

$$R^2 = \frac{S_r}{S_t} = \frac{\sum_i^n (f(x_{i1}, \ldots, x_{im}) - \overline{y})^2}{\sum_i^n (y_i - \overline{y})^2} \in [0, 1]$$

S_r regression (explained) sum of squares, S_t total sum of squares
$E = S_t - S_r$
univariate: $R^2 = \sigma_{xy}^2 / \sigma_{xx}\sigma_{yy}$, $corr(x, y) = \sigma_{xy} / \sqrt{\sigma_{xx}\sigma_{yy}}$

Klasifikace

= modelování **diskrétního** cílového atributu, **tříd klasifikace (class labels)**, jako **(cílové) funkce**, **klasifikačního modelu**, jiných atributů

$\sim$ rozřazení objektů (**instancí**) záznamových dat do jedné z definovaných **kategorií** (class labels), na základě (ostatních) atributů, např.

$\rightarrow$ **prediktivní modelování**: predikce class label pro objekt (s neznámou class label) na základě jeho (známých) hodnot atributů = **prediktorů**, např.

$\rightarrow$ **deskriptivní modelování**: rozlišení/kategorizace objektů (s neznámými class labels) do různých tříd na základě (známých) hodnot atributů objektů $\sim$ zjištění, jaké atributy (jejich hodnoty) určují třídy, např.

- typicky **nominální (binární)** class labels, pro ordinální **rank classification** – zohledňování pořadí class labels, také jiné vztahy, např. inkluzivní, mezi třídami/kategoriemi objektů
- mnoho aplikací: diagnóza nemocí, rozhodování o klientech (banky apod.), klasifikační atlasy, call centrum aj.

Klasifikace: příklad dat

ID	jméno	věk	partner	dětí	zaměstnání	příjem Kč	nemovitost	dávka
1	Adriana Dobrovičová	32	ano	5	ne	6 335	ne	ANO
2	František Smutný	58	ne	0	ano	9 055	ne	NE
3	Zuzana Nesmělá	25	ne	2	ne	4 625	ano	ANO
4	Diego Horváth	46	ano	6	ano	15 500	ne	NE
5	Eva Vopršálková	28	ne	1	ano	9 290	ne	ANO
6	Denisa Šťastná	34	ne	1	ne	14 836	ano	NE
7	Zdeněk Pokorný	57	ne	2	ne	6 060	ano	ANO
8	Esmeralda Balogová	19	ano	3	ne	1 420	ne	ANO
9	Radovan Bobek	43	ano	2	ne	14 600	ano	NE
10	Vladimíra Komínková	62	ne	0	ne	6 214	ne	ANO
11	Mojmír Zatuchlý	49	ano	3	ano	16 150	ne	NE
12	Nikola Gáborová	31	ano	7	ne	10 375	ne	?

atributy: věk, partner ve společné domácnosti, počet nezletilých dětí, zaměstnání, průměrný měsíční příjem za poslední kvartál v Kč, vlastní nemovitost

cílový atribut (třídy klasifikace, class labels, kategorie): přidělení sociální dávky ANO/NE

Klasifikační model

- vytvářený ze vstupních dat klasifikační metodou/technikou (**klasifikátor, classifier**) = **učení** (indukce)
- např. rozhodovací stromy, pravidlový (rule-based), naivní bayesovský (naıve Bayes), umělé neuronové sítě, support vector machines aj.

! **klasifikační problém**: co nejlepší modelování vstupních dat, ale zároveň i správná predikce class label pro nové objekty = **aplikování** (dedukce) – **schopnost generalizace**

- data (objekty se známou class label) pro učení = **training set**, data (objekty s neznámou class label) pro aplikování = **test set**
- vyhodnocení výkonnosti (performance) na základě počtů správných a nesprávných predikcí → **confusion matrix** (matice záměn, chybová), např. pro binární class labels:

		predikovaná třída	
		0	1
skutečná třída	0	f_{00}	f_{01}
	1	f_{10}	f_{11}

f_{ij} počet objektů s class label i predikovanou jako j

Klasifikační model

Základní ukazatele výkonnosti (performance)

- **accuracy (přesnost)**: $\rightarrow$ nejvyšší

$$\frac{\text{počet správných predikcí}}{\text{počet všech predikcí}} = \frac{\sum_i f_{ii}}{\sum_{i,j} f_{ij}}$$

- **error rate (chybovost)**: $\rightarrow$ nejnižší

$$\frac{\text{počet nesprávných predikcí}}{\text{počet všech predikcí}} = \frac{\sum_{i \neq j} f_{ij}}{\sum_{i,j} f_{ij}}$$

Rozhodovací stromy

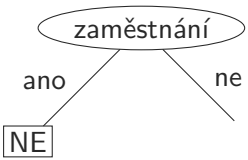
? predikce class label/zařazení do třídy pro (nový) objekt = klasifikace: série vhodných otázek, v závislosti na odpovědích na předchozí otázky, na hodnoty atributů objektu, např. zaměstnání?: ano , ne

→ organizace otázek a odpovědí ve formě (rozhodovacího) stromu, např.

Rozhodovací stromy

? predikce class label/zařazení do třídy pro (nový) objekt = klasifikace: série vhodných otázek, v závislosti na odpovědích na předchozí otázky, na hodnoty atributů objektu, např. zaměstnání?: ano ⇒ dávka NE, ne

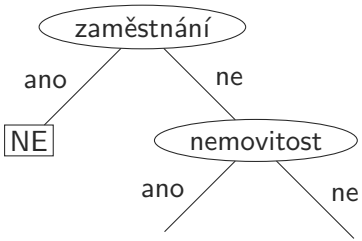
→ organizace otázek a odpovědí ve formě (rozhodovacího) stromu, např.



Rozhodovací stromy

? predikce class label/zařazení do třídy pro (nový) objekt = klasifikace: série vhodných otázek, v závislosti na odpovědích na předchozí otázky, na hodnoty atributů objektu, např. zaměstnání?: ano $\Rightarrow$ dávka NE, ne $\rightarrow$ nemovitost?: ne , ano

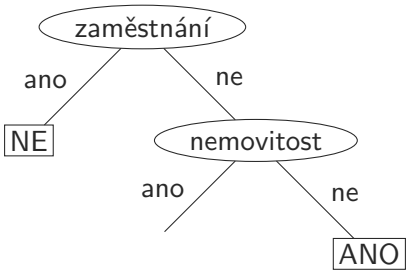
$\rightarrow$ organizace otázek a odpovědí ve formě (rozhodovacího) stromu, např.



Rozhodovací stromy

? predikce class label/zařazení do třídy pro (nový) objekt = klasifikace: série vhodných otázek, v závislosti na odpovědích na předchozí otázky, na hodnoty atributů objektu, např. zaměstnání?: ano $\Rightarrow$ dávka NE, ne $\rightarrow$ nemovitost?: ne $\Rightarrow$ ANO, ano

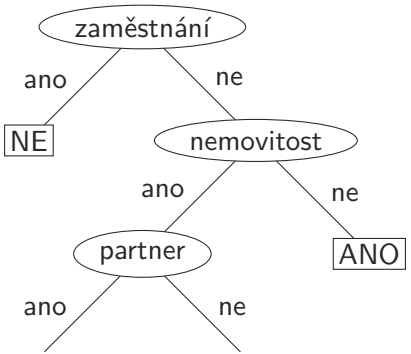
$\rightarrow$ organizace otázek a odpovědí ve formě (rozhodovacího) stromu, např.



Rozhodovací stromy

? predikce class label/zařazení do třídy pro (nový) objekt = klasifikace: série vhodných otázek, v závislosti na odpovědích na předchozí otázky, na hodnoty atributů objektu, např. zaměstnání?: ano ⇒ dávka NE, ne → nemovitost?: ne ⇒ ANO, ano → partner?: ano , ne

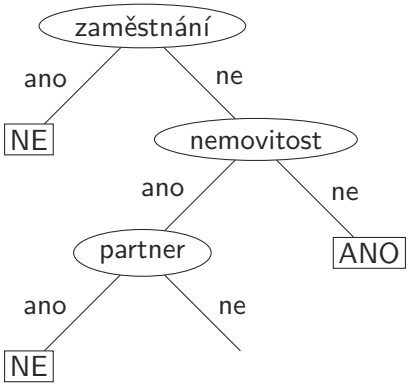
→ organizace otázek a odpovědí ve formě (rozhodovacího) stromu, např.



Rozhodovací stromy

? predikce class label/zařazení do třídy pro (nový) objekt = klasifikace: série vhodných otázek, v závislosti na odpovědích na předchozí otázky, na hodnoty atributů objektu, např. zaměstnání?: ano ⇒ dávka NE, ne → nemovitost?: ne ⇒ ANO, ano → partner?: ano ⇒ NE, ne

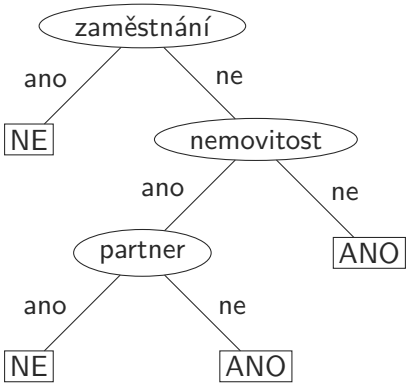
→ organizace otázek a odpovědí ve formě (rozhodovacího) stromu, např.



Rozhodovací stromy

? predikce class label/zařazení do třídy pro (nový) objekt = klasifikace: série vhodných otázek, v závislosti na odpovědích na předchozí otázky, na hodnoty atributů objektu, např. zaměstnání?: ano $\Rightarrow$ dávka NE, ne $\rightarrow$ nemovitost?: ne $\Rightarrow$ ANO, ano $\rightarrow$ partner?: ano $\Rightarrow$ NE, ne $\Rightarrow$ ANO

$\rightarrow$ organizace otázek a odpovědí ve formě (rozhodovacího) stromu, např.



Rozhodovací stromy

= hierarchická (stromová) struktura:

- **vnitřní uzly** – jedna příchozí hrana, dvě a více odchozích hran, přiřazená **podmínka/test nad atributy**, typicky jedním
- **listové uzly** – jedna příchozí hrana, žádná odchozí hrana, přiřazená class label
- **kořenový uzel** = vnitřní nebo listový, žádná příchozí hrana
- **orientované hrany** mezi uzly – od rodičovského k potomkům, přiřazený **výsledek testu** u rodičovského uzlu, typicky hodnota(y) atributu

→ klasifikace objektu: počínaje kořenovým uzlem opakovaně aplikace na objekt testu u aktuálního uzlu a přesun po hraně k uzlu potomka vybrané/ho dle výsledku testu, až přesun do listového uzlu s class label, např.

Vytvoření (indukce) rozhodovacího stromu

- exponenciálně mnoho různě přesných (accurate) stromů pro danou množinu atributů

→ obvyklá **greedy strategie** pro suboptimálně přesný strom – série lokálně optimálních stanovení testu u uzlu, typicky výběru atributu

Huntův algoritmus

- základ mnoha používaných, např. ID3, C4.5, CART

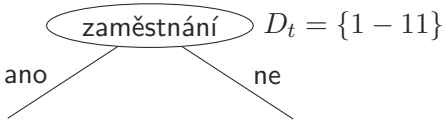
= rekurzivní rozklad (**split**) množiny vstupních (trénovacích) objektů na (disjunktní) podmnožiny s cílem stejné class label u objektů

- D_t množina vstupních (trénovacích) objektů asociovaných s uzlem t stromu, na počátku všechny ($D_t \neq \emptyset$)

1. jestliže všechny objekty z D_t mají stejnou class label y, vrať listový uzel t s y,
2. jinak stanov **test nad atributy** (typicky jedním), rozkládající (splitting) D_t na podmnožiny objektů se stejným výsledkem testu (typicky hodnota(y) atributu) a
   - vytvoř vnitřní uzel t s tímto testem a, pro každý různý výsledek testu, s odchozí hranou s výsledkem testu,
   - na každou hranu napoj výsledek tohoto algoritmu pro $D_{t'}$ = podmnožina objektů s výsledkem testu u hrany a
   - vrať t

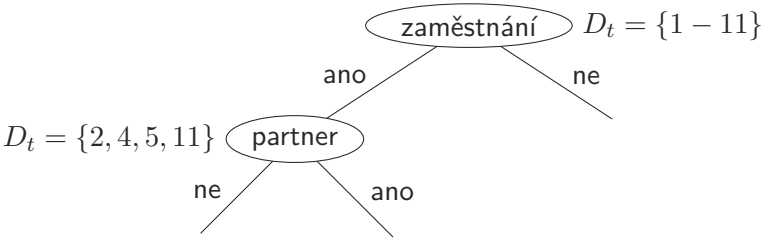
Vytvoření (indukce) rozhodovacího stromu

Huntův algoritmus: příklad



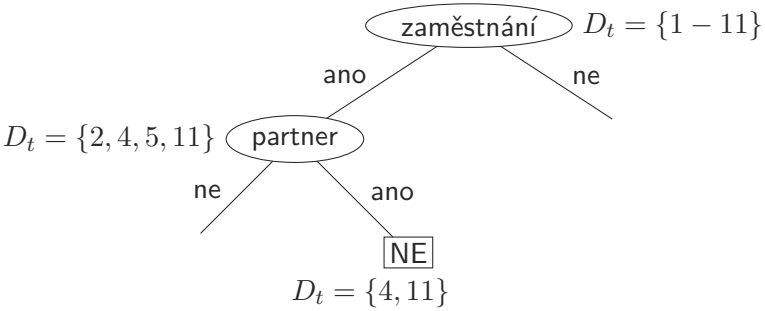
Vytvoření (indukce) rozhodovacího stromu

Huntův algoritmus: příklad



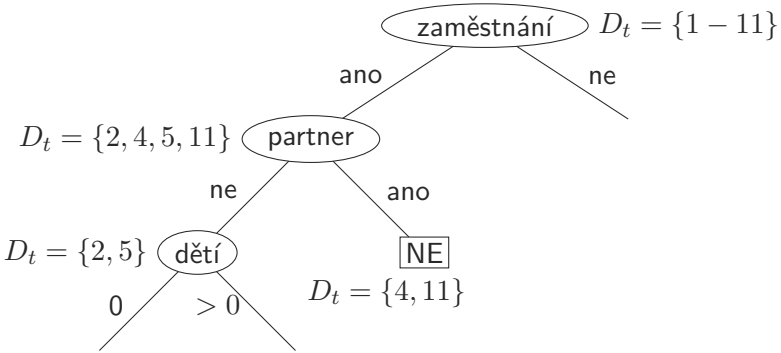
Vytvoření (indukce) rozhodovacího stromu

Huntův algoritmus: příklad



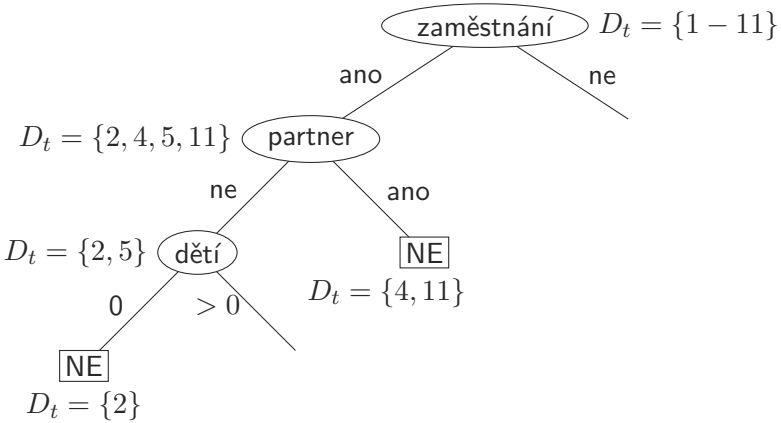
Vytvoření (indukce) rozhodovacího stromu

Huntův algoritmus: příklad



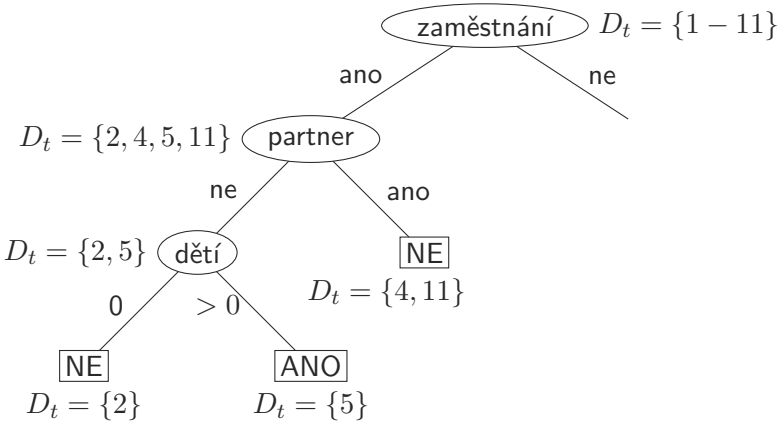
Vytvoření (indukce) rozhodovacího stromu

Huntův algoritmus: příklad



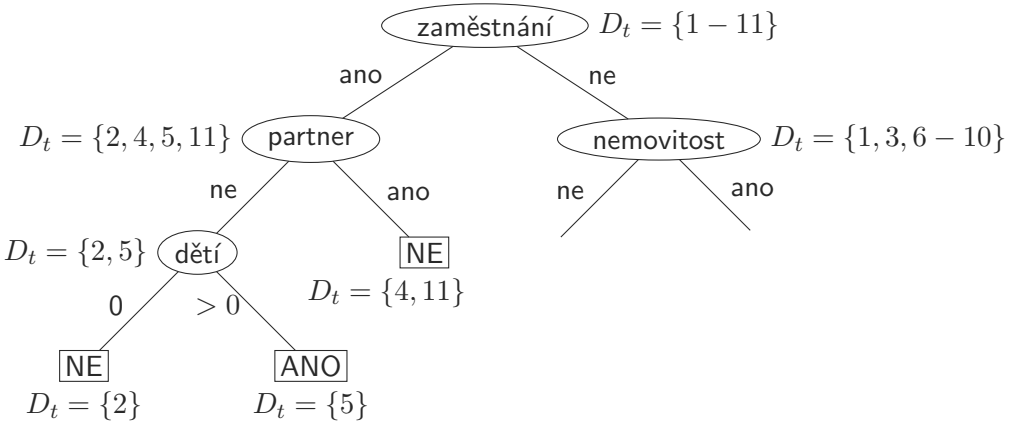
Vytvoření (indukce) rozhodovacího stromu

Huntův algoritmus: příklad



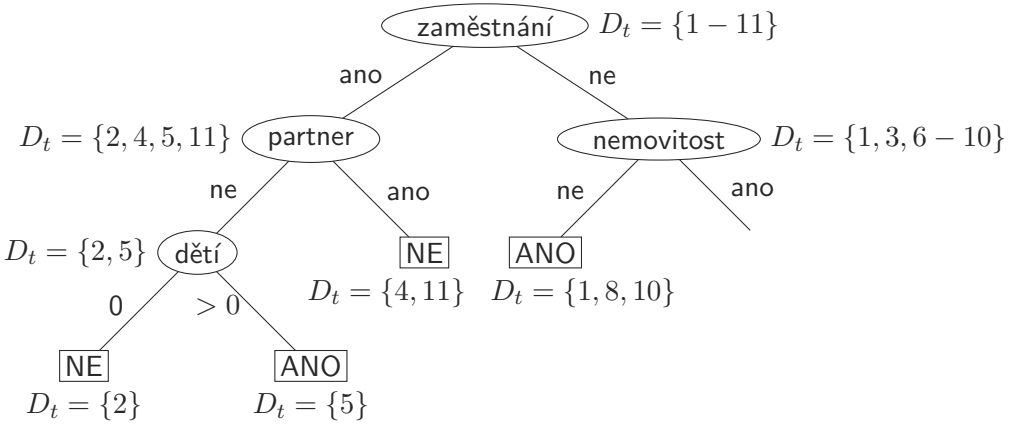
Vytvoření (indukce) rozhodovacího stromu

Huntův algoritmus: příklad



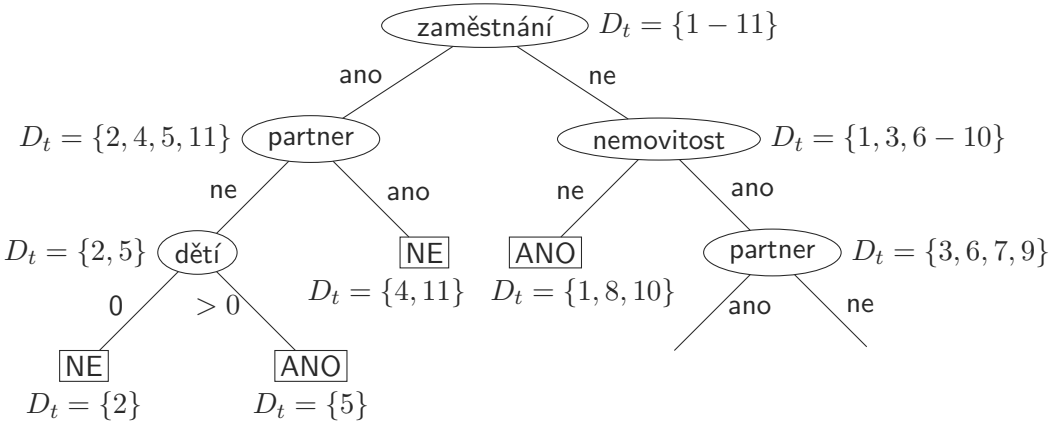
Vytvoření (indukce) rozhodovacího stromu

Huntův algoritmus: příklad



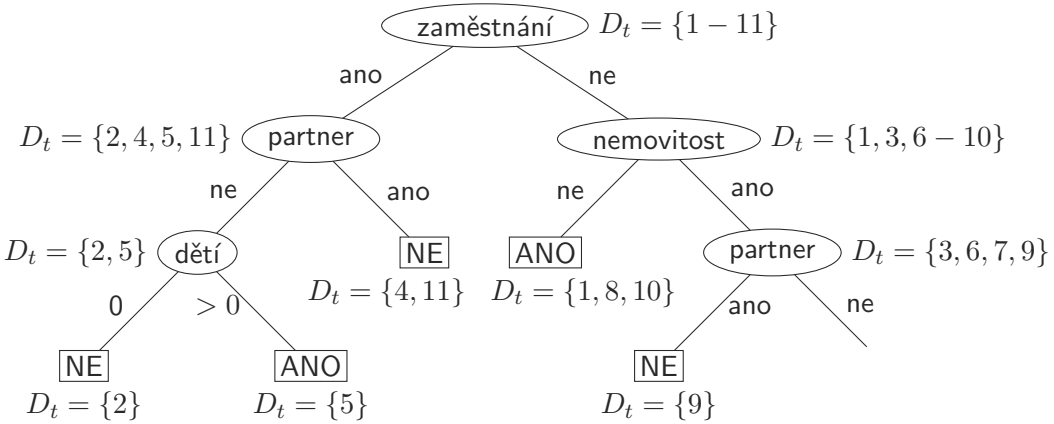
Vytvoření (indukce) rozhodovacího stromu

Huntův algoritmus: příklad



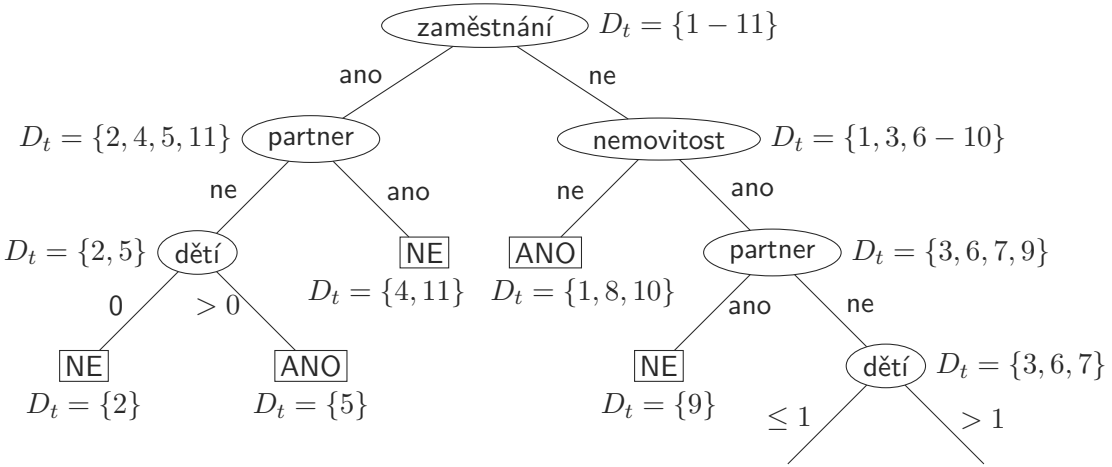
Vytvoření (indukce) rozhodovacího stromu

Huntův algoritmus: příklad



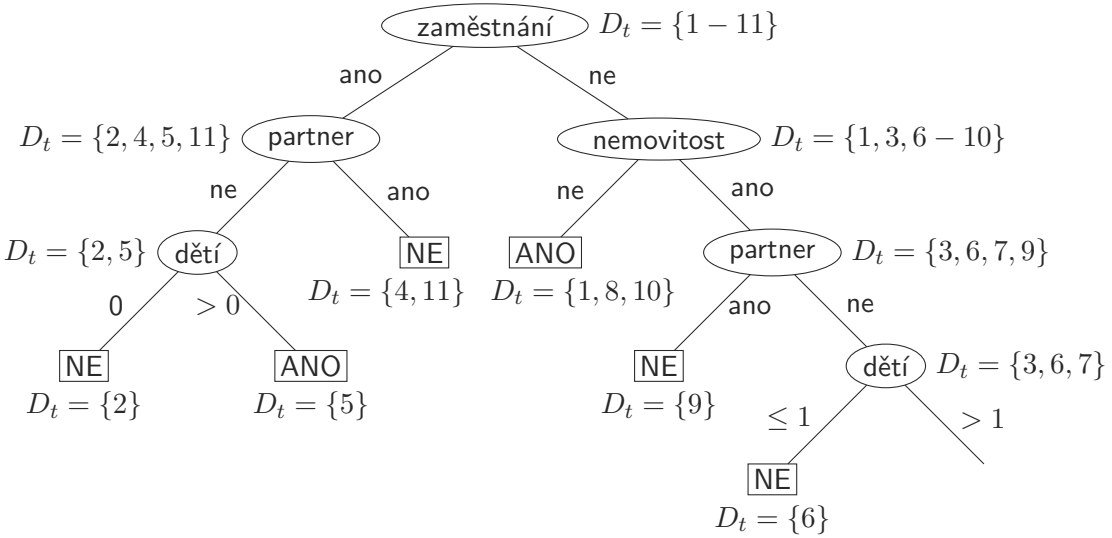
Vytvoření (indukce) rozhodovacího stromu

Huntův algoritmus: příklad



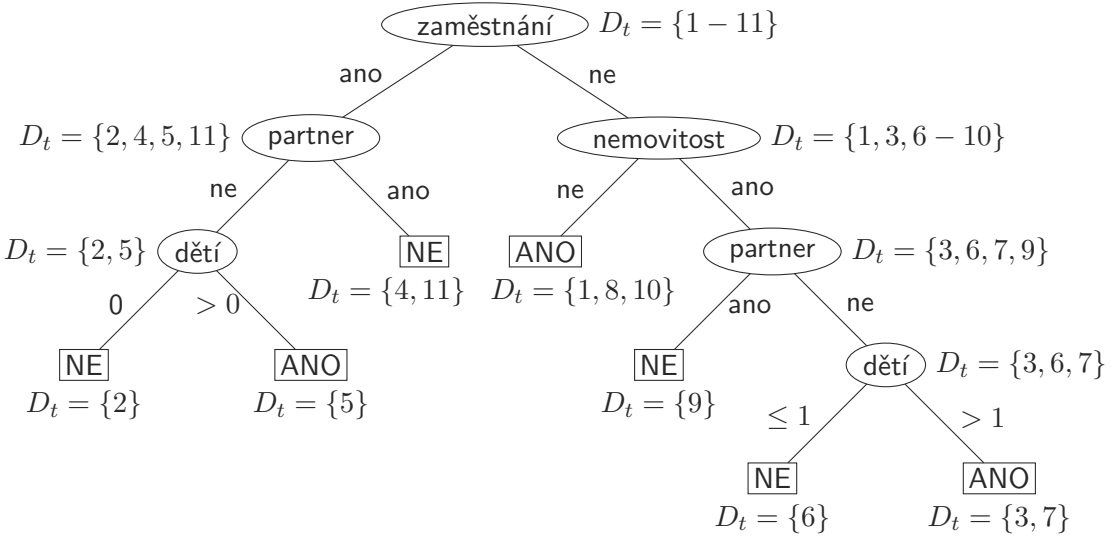
Vytvoření (indukce) rozhodovacího stromu

Huntův algoritmus: příklad



Vytvoření (indukce) rozhodovacího stromu

Huntův algoritmus: příklad



Vytvoření (indukce) rozhodovacího stromu

Huntův algoritmus

!! test nad atributy pokrývající všechny hodnoty atributů v testu, nejen ty v (aktuální) D_t – generalizace

? D_t z kroku 1 prázdná $\rightarrow$ class label y přiřazená $t =$ (obvykle) majoritní class label objektů asociovaných s rodičovským uzlem

? jediný výsledek testu nad atributy (stejný pro všechny objekty z D_t), tj. ne rozklad D_t $\rightarrow$ uzel t listový s (obvykle) majoritní class label objektů z D_t

! stanovení testu nad atributy – pro různé typy atributů, vyhodnocení „výhodnosti“ testu

- jiné podmínky pro listový uzel, např. minimální velikost D_t – listové uzly (dříve) místo některých vnitřních $\rightarrow$ zlepšení schopnosti generalizace
- po vytvoření „prořezání“ (**pruning**) pro zamezení problému **přeučení (overfitting)** $\rightarrow$ zlepšení schopnosti generalizace, viz dále

Test nad atributy

- binární: která hodnota? $\rightarrow$ dva výsledky = dvě hodnoty
- nominální:
  1. která hodnota? $\rightarrow$ k výsledků = k hodnot (multiway split)
  2. které hodnoty (výběr)? $\rightarrow$ méně než k výsledků = disjunktních podmnožin k hodnot, typicky dvou (binary split), např. CART – $2^{k-1} - 1$ možností
- ordinální: jako nominální, podmnožiny „zachovávající pořadí hodnot“ (není $x < y$ a současně $x' > y'$ pro $x, x' \in X$ a $y, y' \in Y$)
- spojité: jaké hodnoty (porovnání s konkrétními)? $\rightarrow$ výsledky = disjunktní intervaly setřízených hodnot vymezené dělicími body, typicky dva (binary split) – kolik?, jaké? $\rightsquigarrow$ supervised diskretizace atributu a ordinální (s příp. jinou měrou impurity intervalu, viz dále)

? které atributy, binary nebo multiway split, kolik intervalů (spojitý atribut), kombinace (testů)

$\rightarrow$ typicky jeden atribut (dělicí, **splitting attribute**) – který?

Test nad atributy $\sim$ rozřazení objektů

* rozklad (split) množiny D_t objektů asociovaných s uzlem t stromu

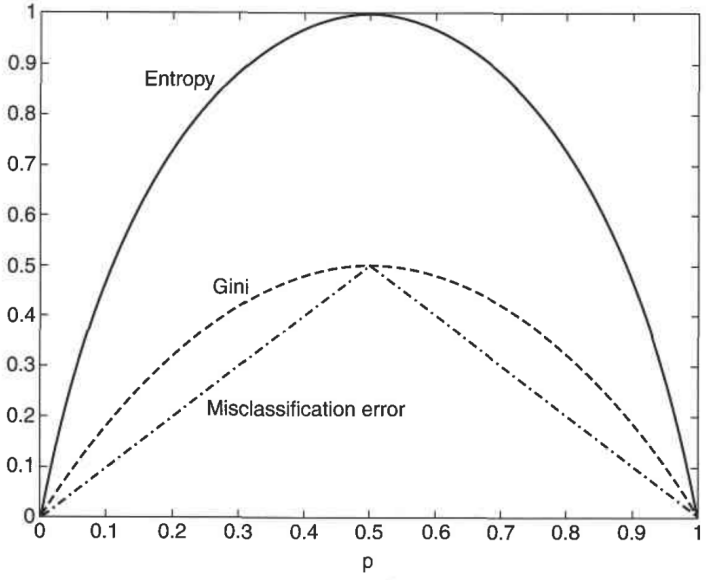
$\rightarrow$ „výhodnost“ testu $\sim$ „výhodnost“ rozkladu (dělicí kritérium, splitting criterion)

$+$ cíl: stejná class label u všech objektů v podmnožinách $D_{t'}$ rozkladu (asociovaných s novými uzly t' stromu)

$\rightarrow$ měření jedinečnosti class label u objektů asociovaných s uzlem stromu $=$ „čistota“ (purity) uzlu $\rightarrow$ maximalizace

- $p(i|t) = m(i|t)/m(t)$ podíl ($\approx$ pravděpodobnost) objektů asociovaných s uzlem t s class label i, $m(i|t)$ počet objektů asociovaných s uzlem t s class label i, $m(t)$ počet objektů asociovaných s uzlem t – purity $t \approx$ rozdílnost (neuniformnost) $p(i|t)$
- **míry impurity** $I(t)$ uzlu $t \rightarrow$ minimalizace, nejčastější: např. pro $D_t = \{1 - 11\}$
  - **entropie**$(t) = -\sum_i p(i|t) \log_2 p(i|t)$ $\quad -(\frac{6}{11} \log_2 \frac{6}{11} + \frac{5}{11} \log_2 \frac{5}{11}) \doteq 0.994$
  - **Gini index**$(t) = 1 - \sum_i [p(i|t)]^2$ $\quad 1 - ((\frac{6}{11})^2 + (\frac{5}{11})^2) \doteq 0.496$
  - **klasifikační chyba**$(t) = 1 - \max_i p(i|t)$ $\quad 1 - \max(\frac{6}{11}, \frac{5}{11}) \doteq 0.364$

Test nad atributy $\sim$ rozřazení objektů



Test nad atributy $\sim$ rozřazení objektů

- „výhodnost“ testu $\sim$ „výhodnost“ rozkladu D_t – dělicí kritérium **gain** $\rightarrow$ maximalizace:

$$\Delta(t) = I(t) - \sum_{t'} \frac{N(t')}{N(t)} I(t')$$

  $N(t)$ počet objektů asociovaných s uzlem t stromu
  ... rozdíl impurity (rodičovského) uzlu t, před rozkladem, s váženým součtem impurity nových uzlů t' (potomků), po rozkladu
  information gain Δ_{info} = gain pro entropii jako míru $I(t)$ impurity uzlu

$\Rightarrow$ splitting attribute = atribut s maximální gain

Např. pro $D_t = \{1 - 11\}$, dělicí atribut = zaměstnání: hodnoty ano, ne $\rightarrow$
$D_{t'_{ano}} = \{2, 4, 5, 11\}, D_{t'_{ne}} = \{1, 3, 6 - 10\}$:
$I(t'_{ano}) = -(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}) \doteq 0.811$, $I(t'_{ne}) = -(\frac{5}{7}\log_2\frac{5}{7} + \frac{2}{7}\log_2\frac{2}{7}) \doteq 0.863$
$\Delta_{info}(t) = 0.994 - (\frac{4}{11} \cdot 0.811 + \frac{7}{11} \cdot 0.863) = 0.15$

Vytvoření (indukce) rozhodovacího stromu: příklad

$D_t = \{1 - 11\}$

věk:	19	25	28	32	34	43	46	49	57	58	62	$\rightarrow \leq 33, > 33$
	ANO	ANO	ANO	ANO	NE	NE	NE	NE	ANO	NE	ANO	

$D_{t'_{\leq 33}} = \{1, 3, 5, 8\}, D_{t'_{>33}} = \{2, 4, 6, 7, 9 - 11\}$

$I(t'_{\leq 33}) = -(\frac{4}{4}\log_2\frac{4}{4} + \frac{0}{4}\log_2\frac{0}{4}) = 0,\ I(t'_{>33}) = -(\frac{2}{7}\log_2\frac{2}{7} + \frac{5}{7}\log_2\frac{5}{7}) \doteq 0.863$

$\Delta_{info}(t) = 0.994 - (\frac{4}{11}\cdot 0 + \frac{7}{11}\cdot 0.863) = 0.445$

Vytvoření (indukce) rozhodovacího stromu: příklad

$D_t = \{1 - 11\}$
věk: $\Delta_{info}(t) = 0.445$
partner: ano, ne
$D_{t'_{ano}} = \{1, 4, 8, 9, 11\}, D_{t'_{ne}} = \{2, 3, 5 - 7, 10\}$
$I(t'_{ano}) = -(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}) \doteq 0.971,\ I(t'_{ne}) = -(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}) \doteq 0.918$
$\Delta_{info}(t) = 0.994 - (\frac{5}{11} \cdot 0.971 + \frac{6}{11} \cdot 0.918) = 0.052$

Vytvoření (indukce) rozhodovacího stromu: příklad

$D_t = \{1 - 11\}$
věk: $\Delta_{info}(t) = 0.445$
partner: $\Delta_{info}(t) = 0.052$
děti: $\{0\}$, $\{1, 2\}$, > 2
$D_{t'_{\{0\}}} = \{2, 10\}, D_{t'_{\{1,2\}}} = \{3, 5 - 7, 9\}, D_{t'_{>2}} = \{1, 4, 8, 11\}$
$I(t'_{\{0\}}) = -(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) = 1$, $I(t'_{\{1,2\}}) = -(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}) \doteq 0.971$,
$I(t'_{>2}) = -(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}) = 1$
$\Delta_{info}(t) = 0.994 - (\frac{2}{11}\cdot 1 + \frac{5}{11}\cdot 0.971 + \frac{4}{11}\cdot 1) = 0.007$

Vytvoření (indukce) rozhodovacího stromu: příklad

$D_t = \{1 - 11\}$
věk: $\Delta_{info}(t) = 0.445$
partner: $\Delta_{info}(t) = 0.052$
děti: $\Delta_{info}(t) = 0.007$
zaměstnání: ano, ne
$D_{t'_{ano}} = \{2, 4, 5, 11\}, D_{t'_{ne}} = \{1, 3, 6 - 10\}$:
$I(t'_{ano}) = -(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}) \doteq 0.811$, $I(t'_{ne}) = -(\frac{5}{7}\log_2\frac{5}{7} + \frac{2}{7}\log_2\frac{2}{7}) \doteq 0.863$
$\Delta_{info}(t) = 0.994 - (\frac{4}{11} \cdot 0.811 + \frac{7}{11} \cdot 0.863) = 0.15$

Vytvoření (indukce) rozhodovacího stromu: příklad

$D_t = \{1 - 11\}$
věk: $\Delta_{info}(t) = 0.445$
partner: $\Delta_{info}(t) = 0.052$
děti: $\Delta_{info}(t) = 0.007$
zaměstnání: $\Delta_{info}(t) = 0.15$
nemovitost: ano, ne
$D_{t'_{ano}} = \{3, 6, 7, 9\}, D_{t'_{ne}} = \{1, 2, 4, 5, 8, 10, 11\}$:
$I(t'_{ano}) = -(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}) = 1,\ I(t'_{ne}) = -(\frac{4}{7}\log_2\frac{4}{7} + \frac{3}{7}\log_2\frac{3}{7}) \doteq 0.985$
$\Delta_{info}(t) = 0.994 - (\frac{4}{11}\cdot 1 + \frac{7}{11}\cdot 0.985) = 0.004$

Vytvoření (indukce) rozhodovacího stromu: příklad

$D_t = \{1 - 11\}$

věk: $\Delta_{info}(t) = 0.445$

partner: $\Delta_{info}(t) = 0.052$

děti: $\Delta_{info}(t) = 0.007$

zaměstnání: $\Delta_{info}(t) = 0.15$

nemovitost: $\Delta_{info}(t) = 0.004$

příjem:

1 420	4 625	6 060	6 214	6 335	9 055	9 290	14 600	14 836	15 500	16 150
ANO	ANO	ANO	ANO	ANO	NE	ANO	NE	NE	NE	NE

$\rightarrow \leq 9,000, > 9,000$

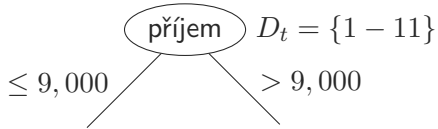
$D_{t'_{\leq 9,000}} = \{1, 3, 7, 8, 10\}, D_{t'_{>9,000}} = \{2, 4 - 6, 9, 11\}$

$I(t'_{\leq 9,000}) = -(\frac{5}{5}\log_2\frac{5}{5} + \frac{0}{5}\log_2\frac{0}{5}) = 0,\ I(t'_{>9,000}) = -(\frac{1}{6}\log_2\frac{1}{6} + \frac{5}{6}\log_2\frac{5}{6}) \doteq 0.65$

$\Delta_{info}(t) = 0.994 - (\frac{5}{11} \cdot 0 + \frac{6}{11} \cdot 0.65) = 0.639$

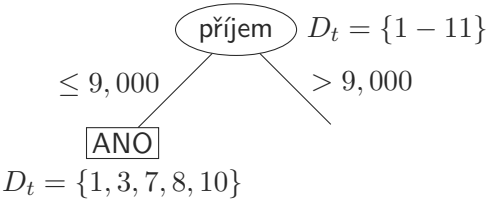
Vytvoření (indukce) rozhodovacího stromu: příklad

$D_t = \{1 - 11\}$
věk: $\Delta_{info}(t) = 0.445$
partner: $\Delta_{info}(t) = 0.052$
děti: $\Delta_{info}(t) = 0.007$
zaměstnání: $\Delta_{info}(t) = 0.15$
nemovitost: $\Delta_{info}(t) = 0.004$
příjem: $\Delta_{info}(t) = 0.639$



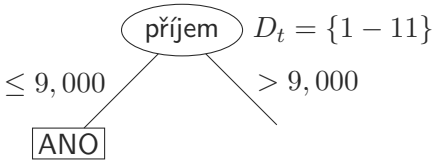
Vytvoření (indukce) rozhodovacího stromu: příklad

příjem: $\leq 9,000$
$D_t = \{1, 3, 7, 8, 10\}$
dávka: ANO



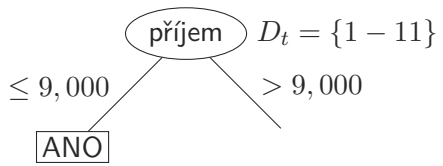
Vytvoření (indukce) rozhodovacího stromu: příklad

příjem: $> 9,000$
$D_t = \{2, 4-6, 9, 11\}$
věk: $\leq 33, > 33$
$D_{t'_{\leq 33}} = \{5\}, D_{t'_{>33}} = \{2, 4, 6, 9, 11\}$
$I(t'_{\leq 33}) = -(\frac{1}{1}\log_2\frac{1}{1} + \frac{0}{1}\log_2\frac{0}{1}) = 0$, $I(t'_{>33}) = -(\frac{0}{5}\log_2\frac{0}{5} + \frac{5}{5}\log_2\frac{5}{5}) = 0$
$\Delta_{info}(t) = 0.639 - (\frac{1}{6}\cdot 0 + \frac{5}{6}\cdot 0) = 0.639$



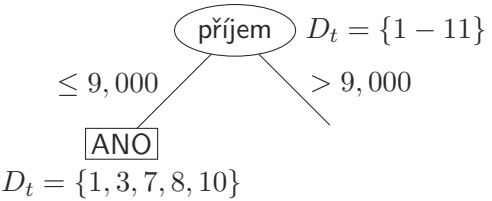
Vytvoření (indukce) rozhodovacího stromu: příklad

příjem: $> 9,000$
$D_t = \{2, 4-6, 9, 11\}$
věk: $\Delta_{info}(t) = 0.639$
partner: ano, ne
$D_{t'_{ano}} = \{4, 9, 11\}, D_{t'_{ne}} = \{2, 5, 6\}$
$I(t'_{ano}) = -(\frac{0}{3}\log_2\frac{0}{3} + \frac{3}{3}\log_2\frac{3}{3}) =$ 0, $I(t'_{ne}) = -(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}) \doteq 0.918$
$\Delta_{info}(t) = 0.639 - (\frac{3}{6}\cdot 0 + \frac{3}{6}\cdot 0.918) = 0.18$



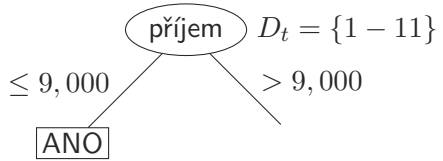
Vytvoření (indukce) rozhodovacího stromu: příklad

příjem: $> 9,000$
$D_t = \{2, 4-6, 9, 11\}$
věk: $\Delta_{info}(t) = 0.639$
partner: $\Delta_{info}(t) = 0.18$
děti: $\{0\}$, $\{1,2\}$, > 2
$D_{t'_{\{0\}}} = \{2\}, D_{t'_{\{1,2\}}} = \{5,6,9\}, D_{t'_{>2}} = \{4,11\}$
$I(t'_{\{0\}}) = -(\frac{0}{1}\log_2\frac{0}{1} + \frac{1}{1}\log_2\frac{1}{1}) = 0$, $I(t'_{\{1,2\}}) = -(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}) \doteq 0.918$, $I(t'_{>2}) = -(\frac{0}{2}\log_2\frac{0}{2} + \frac{2}{2}\log_2\frac{2}{2}) = 0$
$\Delta_{info}(t) = 0.639 - (\frac{1}{6}\cdot 0 + \frac{3}{6}\cdot 0.918 + \frac{2}{6}\cdot 0) = 0.18$



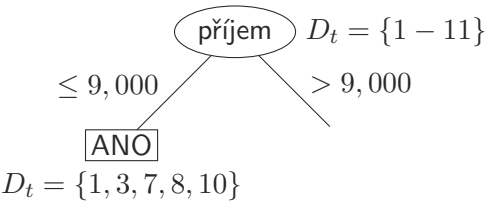
Vytvoření (indukce) rozhodovacího stromu: příklad

příjem: $> 9,000$
$D_t = \{2, 4-6, 9, 11\}$
věk: $\Delta_{info}(t) = 0.639$
partner: $\Delta_{info}(t) = 0.18$
děti: $\Delta_{info}(t) = 0.18$
zaměstnání: ano, ne
$D_{t'_{ano}} = \{2, 4, 5, 11\}, D_{t'_{ne}} = \{6, 9\}$: $D_t = \{1, 3, 7, 8, 10\}$
$I(t'_{ano}) = -(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}) \doteq 0.811$, $I(t'_{ne}) = -(\frac{0}{2}\log_2\frac{0}{2} + \frac{2}{2}\log_2\frac{2}{2}) = 0$
$\Delta_{info}(t) = 0.639 - (\frac{4}{6}\cdot 0.811 + \frac{2}{6}\cdot 0) = 0.098$



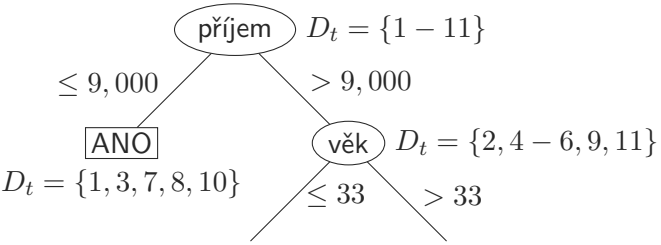
Vytvoření (indukce) rozhodovacího stromu: příklad

příjem: $> 9,000$
$D_t = \{2, 4-6, 9, 11\}$
věk: $\Delta_{info}(t) = 0.639$
partner: $\Delta_{info}(t) = 0.18$
děti: $\Delta_{info}(t) = 0.18$
zaměstnání: $\Delta_{info}(t) = 0.098$
nemovitost: ano, ne
$D_{t'_{ano}} = \{6, 9\}, D_{t'_{ne}} = \{2, 4, 5, 11\}$:
$I(t'_{ano}) = -(\frac{0}{2}\log_2\frac{0}{2} + \frac{2}{2}\log_2\frac{2}{2}) = 0$, $I(t'_{ne}) = -(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}) \doteq 0.811$
$\Delta_{info}(t) = 0.639 - (\frac{2}{6}\cdot 0 + \frac{4}{6}\cdot 0.811) = 0.098$



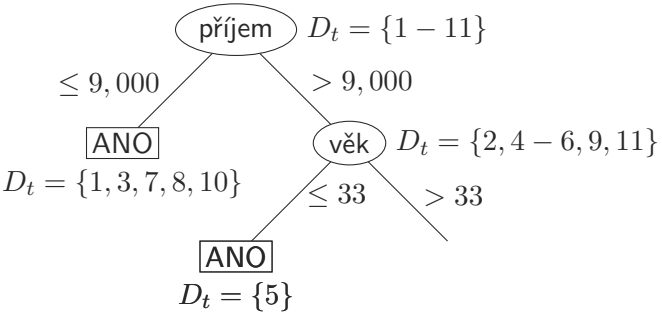
Vytvoření (indukce) rozhodovacího stromu: příklad

příjem: $> 9,000$
$D_t = \{2, 4-6, 9, 11\}$
věk: $\Delta_{info}(t) = 0.639$
partner: $\Delta_{info}(t) = 0.18$
děti: $\Delta_{info}(t) = 0.18$
zaměstnání: $\Delta_{info}(t) = 0.098$
nemovitost: $\Delta_{info}(t) = 0.098$



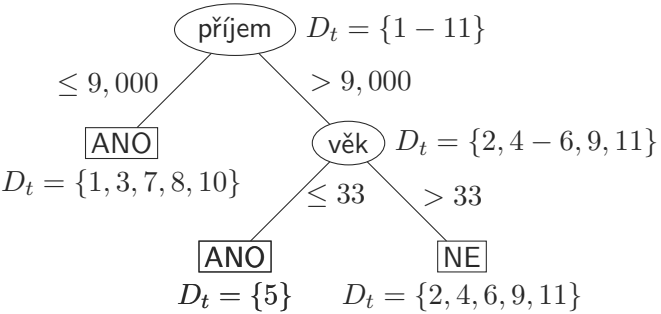
Vytvoření (indukce) rozhodovacího stromu: příklad

příjem: $> 9,000$, věk: ≤ 33
$D_t = \{5\}$
dávka: ANO



Vytvoření (indukce) rozhodovacího stromu: příklad

příjem: $> 9,000$, věk: > 33
$D_t = \{2, 4, 6, 9, 11\}$
dávka: NE



Test nad atributy $\sim$ rozřazení objektů

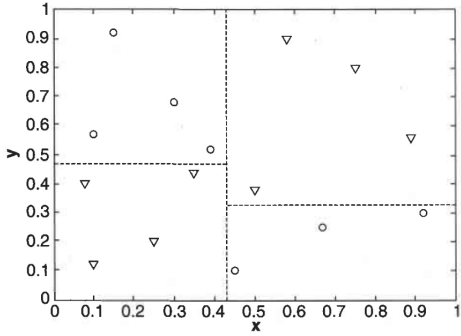
- testy nad jedním atributem = rozklady množiny / separace objektů na (disjunktní) podmnožiny s hranicemi (**decision boundaries**) nezávislými na jiných atributech (tzv. rektilineárními), např. $x < x_1$ a $y < y_1, y_2$

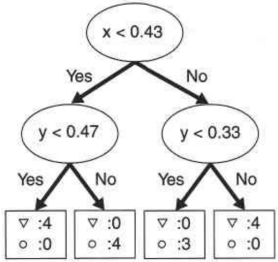
$\Rightarrow$ problém s klasifikací (separací) objektů s různými class labels s hranicemi mezi nimi závislými současně na více atributech, např. $x + y < 1$ nebo některé booleovské funkce, např. parita $\rightarrow$ špatná generalizace (potřeba „plný“ strom)

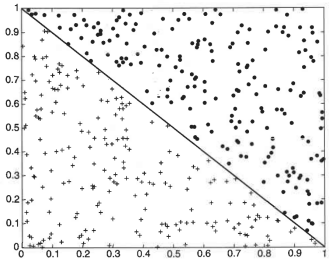
$\rightarrow$ **oblique rozhodovací stromy** = test nad více atributy (aritmetické, logické funkce atributů), ale výpočetně náročnější

$\rightarrow$ **constructive induction** = před indukcí tvorba nových, pro klasifikace lepších, atributů jako (aritmetických, logických) kombinací původních atributů (= feature creation/construction), ale potenciálně redundantních

Test nad atributy $\sim$ rozřazení objektů







Test nad atributy $\sim$ rozřazení objektů

? binary nebo multiway split, kolik intervalů (spojitý atribut) – jaká gain?

Test nad atributy $\sim$ rozřazení objektů

? binary nebo multiway split, kolik intervalů (spojitý atribut) – jaká gain?

- více výsledků testu $\Rightarrow$ menší podmnožiny rozkladu množiny objektů asociovaných s uzlem stromu $\rightsquigarrow$ častěji nižší (nulové) impurity uzlů (potomků) s asociovanými objekty v podmnožinách (objekty mají častěji stejnou class label) $\Rightarrow$ vyšší gain, např.
  pro $D_t = \{1 - 11\}$, dělicí atribut = dětí: hodnoty $\{0\}$, $> 0 \rightarrow D_{t'_{\{0\}}} = \{2, 10\}$,
  $D_{t'_{>0}} = \{1, 3 - 9, 11\}$
  $I(t'_{\{0\}}) = -(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) = 1$, $I(t'_{>0}) = -(\frac{5}{9}\log_2\frac{5}{9} + \frac{4}{9}\log_2\frac{4}{9}) \doteq 0.991$
  $\Delta_{info}(t) = 0.994 - (\frac{2}{11} \cdot 1 + \frac{9}{11} \cdot 0.991) = 0.001$

Test nad atributy $\sim$ rozřazení objektů

? binary nebo multiway split, kolik intervalů (spojitý atribut) – jaká gain?

- více výsledků testu $\Rightarrow$ menší podmnožiny rozkladu množiny objektů asociovaných s uzlem stromu $\rightsquigarrow$ častěji nižší (nulové) impurity uzlů (potomků) s asociovanými objekty v podmnožinách (objekty mají častěji stejnou class label) $\Rightarrow$ vyšší gain, např.
  pro $D_t = \{1 - 11\}$, dělicí atribut = dětí: hodnoty $\{0\}, > 0$ $\Delta_{info}(t) = 0.001$
  hodnoty $\{0\}$, $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, $\geq 6 \rightarrow D_{t'_{\{0\}}} = \{2, 10\}, D_{t'_{\{1\}}} = \{5, 6\}, D_{t'_{\{2\}}} = \{3, 7, 9\}, D_{t'_{\{3\}}} = \{8, 11\}, D_{t'_{\{4\}}} = \emptyset, D_{t'_{\{5\}}} = \{1\}, D_{t'_{\geq 6}} = \{4\}$:
  $I(t'_{\{0\}}) = -(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) = 1$, $I(t'_{\{1\}}) = -(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) = 1$,
  $I(t'_{\{2\}}) = -(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}) \doteq 0.918$, $I(t'_{\{3\}}) = -(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) = 1$,
  $I(t'_{\{4\}}) = 0$, $I(t'_{\{5\}}) = -(\frac{1}{1}\log_2\frac{1}{1} + \frac{0}{1}\log_2\frac{0}{1}) = 0$, $I(t'_{\geq 6}) = -(\frac{0}{1}\log_2\frac{0}{1} + \frac{1}{1}\log_2\frac{1}{1}) = 0$
  $\Delta_{info}(t) = 0.994 - (\frac{2}{11}\cdot 1 + \frac{2}{11}\cdot 1 + \frac{3}{11}\cdot 0.918 + \frac{2}{11}\cdot 1 + \frac{0}{11}\cdot 0 + \frac{1}{11}\cdot 0 + \frac{1}{11}\cdot 0) = 0.198$

Test nad atributy $\sim$ rozřazení objektů

? binary nebo multiway split, kolik intervalů (spojitý atribut) – jaká gain?

- více výsledků testu $\Rightarrow$ menší podmnožiny rozkladu množiny objektů asociovaných s uzlem stromu $\rightsquigarrow$ častěji nižší (nulové) impurity uzlů (potomků) s asociovanými objekty v podmnožinách (objekty mají častěji stejnou class label) $\Rightarrow$ vyšší gain, např.
  pro $D_t = \{1 - 11\}$, dělicí atribut = dětí: hodnoty $\{0\}$, > 0 $\Delta_{info}(t) = 0.001$
  hodnoty $\{0\}$, $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, ≥ 6 $\Delta_{info}(t) = 0.198$
- ovšem také $\rightsquigarrow$ klasifikace na základě menšího (statisticky nevýznamného) počtu objektů asociovaných s listovými uzly (data fragmentation problem) = nižší schopnost generalizace – extrém jeden objekt (atribut „typu ID“)

$\rightarrow$ pouze binary split, např. CART

$\rightarrow$ zahrnutí počtu výsledků testu do dělicího kritéria, např. v C4.5 **gain ratio**:

$$\frac{\Delta_{info}}{split\ info}, \quad split\ info(t) = -\sum_{t'} \frac{N(t')}{N(t)} \log_2 \frac{N(t')}{N(t)}$$

Test nad atributy $\sim$ rozřazení objektů

např. pro $D_t = \{1 - 11\}$, dělicí atribut $=$ dětí:
hodnoty $\{0\}$, $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, ≥ 6

Test nad atributy $\sim$ rozřazení objektů

např. pro $D_t = \{1 - 11\}$, dělicí atribut = dětí:
hodnoty $\{0\}$, $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, $\geq 6 \rightarrow D_{t'_{\{0\}}} = \{2, 10\}, D_{t'_{\{1\}}} = \{5, 6\}, D_{t'_{\{2\}}} = \{3, 7, 9\}, D_{t'_{\{3\}}} = \{8, 11\}, D_{t'_{\{4\}}} = \emptyset, D_{t'_{\{5\}}} = \{1\}, D_{t'_{\geq 6}} = \{4\}$:
$split\ info(t) = -(\frac{2}{11}\log_2\frac{2}{11} + \frac{2}{11}\log_2\frac{2}{11} + \frac{3}{11}\log_2\frac{3}{11} + \frac{2}{11}\log_2\frac{2}{11} + \frac{0}{11}\log_2\frac{0}{11} + \frac{1}{11}\log_2\frac{1}{11} + \frac{1}{11}\log_2\frac{1}{11}) \doteq 2.482$
$\Delta_{info}(t)/split\ info = 0.198/2.482 \doteq 0.08$

Test nad atributy $\sim$ rozřazení objektů

např. pro $D_t = \{1 - 11\}$, dělicí atribut = dětí:
hodnoty $\{0\}$, $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, ≥ 6 $\Delta_{info}(t)/split\ info \doteq 0.08$

hodnoty $\{0\}$, $\{1,2\}$, $> 2 \rightarrow D_{t'_{\{0\}}} = \{2,10\}$, $D_{t'_{\{1,2\}}} = \{3,5-7,9\}$, $D_{t'_{>2}} = \{1,4,8,11\}$
$split\ info(t) = -(\frac{2}{11}\log_2\frac{2}{11} + \frac{5}{11}\log_2\frac{5}{11} + \frac{4}{11}\log_2\frac{4}{11}) \doteq 1.495$
$\Delta_{info}(t)/split\ info = 0.007/1.495 \doteq 0.005$

Test nad atributy $\sim$ rozřazení objektů

např. pro $D_t = \{1 - 11\}$, dělicí atribut = dětí:
hodnoty $\{0\}$, $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, ≥ 6 $\Delta_{info}(t)/split\ info \doteq 0.08$

hodnoty $\{0\}$, $\{1, 2\}$, > 2 $\Delta_{info}(t)/split\ info \doteq 0.005$

hodnoty $\{0\}$, $> 0 \rightarrow D_{t'_{\{0\}}} = \{2, 10\}$, $D_{t'_{>0}} = \{1, 3 - 9, 11\}$
$split\ info(t) = -(\frac{2}{11} \log_2 \frac{2}{11} + \frac{9}{11} \log_2 \frac{9}{11}) \doteq 0.684$
$\Delta_{info}(t)/split\ info = 0.001/0.684 \doteq 0.001$

Vlastnosti rozhodovacích stromů

- vytvoření optimálního stromu = **NP-těžký problém** $\rightarrow$ heuristický přístup, např. greedy, top-down rekurzivní rozklad – výpočetně nenáročné algoritmy, klasifikace objektu rychlá (daná maximální hloubkou stromu)
- relativně **snadná interpretace**, **přesnost (accuracy) srovnatelná** s jinými klasifikátory (pro běžná data, pro hůře klasifikovatelná nižší)
- **robustní vůči šumu** v datech (zejména při řešení přeučení), redundantní (korelované) atributy nemají vliv na přesnost, příliš mnoho irelevantních atributů strom zbytečně zvětšuje $\rightarrow$ feature selection
- výběr míry impurity má malý vliv na přesnost (chovají se podobně), větší vliv má strategie prořezání stromu, viz dále
- neparametrická klasifikační metoda – nepředpokládá určitá pravděpodobností rozložení class labels a atributů

Přeučení (overfitting) modelu

- **chyba učení (training, resubstitution error)** = počet chyb klasifikace trénovacích dat
- **chyba generalizace (generalization error)** = množství chyb klasifikace (neznámých) testovacích dat

* klasifikační problém $\sim$ dobrý klasifikační model: co nejlepší modelování trénovacích dat, ale zároveň i správná klasifikace neznámých dat (generalizace) $\Rightarrow$ nízké obě chyby

= vyšší chyba generalizace v důsledku příliš přesného modelování trénovacích dat a špatné klasifikace neznámých dat

- zpočátku učení s malým modelem chyby na trénovacích i testovacích datech vysoké = **„podučení“ (underfitting)** modelu, s narůstajícím modelem (učení) klesají, avšak od určité velikosti modelu chyba na testovacích datech začne růst (na trénovacích datech dál klesá) = přeučení

Přeučení (overfitting) modelu

- příčina: šum a chyby v trénovacích datech – např. chybná class label, výjimečné testovací objekty s jinou class label než stejné trénovací objekty, např.
- příčina: nedostatek reprezentativních objektů v trénovacích datech (pro určité kombinace hodnot atributů), např.
- důvod (?): opakovaně výběr nejlepší (maximalizující nějaké kritérium) možnosti z několika a její přidání do modelu, pokud je dostatečná (tzv. multiple comparison procedure) – čím více možností, tím větší šance a zvětšování modelu, zvláště při malém počtu trénovacích objektů (velký rozptyl hodnot kritéria výběru), např. splitting atribute a gain „hlouběji“ v rozhodovacích stromech $\rightarrow$ zahrnout počet možností do kritéria výběru, výběr možnosti jen při dostatečném počtu trénovacích objektů

Odhad chyby generalizace

- cíl = **model selection**: velikost/složitost modelu odpovídající nejnižší chybě generalizace – jaká?
- při učení dostupná jen trénovací data, ne neznámá (testovací) $\Rightarrow$ odhad chyby

Přeučení (overfitting) modelu

Odhad chyby generalizace

= chyba učení – předpoklad: trénovací data dobře reprezentují všechna, ovšem minimální $\Rightarrow$ složitý model $\Rightarrow$ přeučení

$\rightarrow$ preference jednoduššího modelu – dle principu **Ockhamovy břitvy** (úspornosti) nebo **KISS**: „z více modelů se stejnou chybou má být preferován ten nejjednodušší“ $\rightarrow$ zahrnutí složitosti modelu do odhadu chyby:

- \+ penalizace za nárůst modelu, např. za každý listový uzel stromu, ve vztahu k chybě klasifikace, neboli jak moc se chyba nárůstem modelu musí snížit
- \+ velikost modelu, např. stromu – minimum description length (MDL) princip

$\rightarrow$ statistická korekce – horní mez na základě předpokládané distribuce chyb klasifikace a počtu trénovacích objektů klasifikovaných např. v listových uzlech stromu

= chyba na **validation set** = část trénovacích dat nepoužitých k učení, obvykle třetina, viz dále vyhodnocení výkonnosti, použití u metod s parametrem pro složitost modelu, např. úroveň prořezání stromu – výběr toho s nejnižší touto chybou

Přeučení (overfitting) modelu

Řešení v rozhodovacích stromech

- **prepruning** = zastavení větvení stromu před dosažením plného (přeučeného) bezchybně modelujícího celá vstupní data → přísnější podmínka pro listový uzel (místo vnitřního) – jaká?, např.
  - minimální počet objektů asocionavých s listovým uzlem – jaký?
  - minimální hodnota dělícího kritéria (gain) – jaká?, další větvení může vést k lepšímu modelu (menší chyba generalizace)
  - minimální zlepšení odhadu chyby generalizace – jaké?, aj.
- **post-pruning** = „prořezání“ stromu po vytvoření plného, dokud zlepšení od listů ke kořenu nahrazování podstromů:
  - listovým uzlem s majoritní class label objektů asociovaných s uzly podstromu (subtree replacement)
  - nejčastěji používanou větví (podstromem) podstromu = s nejvíce objekty asociovanými s uzly (subtree raising)
- post-pruning obvykle lepší (z plného stromu), za cenu vytvoření plného stromu

Vyhodnocení výkonnosti (performace)

- vytvořeného modelu na testovacích/validation datech – nezkreslený odhad chyby generalizace (oproti odhadu z učení), také pro porovnání výkonnosti různých metod

* základní ukazatele: accuracy (přesnost) a error rate (chybovost)

! potřeba znát class labels testovacích objektů

Holdout

- původní data (s class labels) náhodně disjunktně rozdělena na trénovací a testovací množiny – typicky 50-50 nebo 2/3 trénovací, a
- učení na trénovací, vyhodnocení modelu na testovací
- nevýhody: méně dat pro učení $\Rightarrow$ horší model, závislost modelu na velikosti množin – kvalita modelu vs. důvěryhodnost odhadu výkonnosti (malý rozptyl)

Random subsampling

= opakované holdout pro zlepšení odhadu výkonnosti

- přesnost modelu = průměr přesností z opakování
- nevýhody: stále méně dat pro učení než maximum, různé objekty různě často pro učení a testování

Vyhodnocení výkonnosti (performace)

Cross-validation

- **k-fold**: původní data (s class labels) náhodně disjunktně rozdělena na k stejně velkých množin, typicky 10
- vyhodnocení modelu na jedné (testovací), učení na sjednocení zbývajících $k-1$ (trénovací)
- k opakování pro každou množinu jako testovací
- chybovost modelu $=$ součet chybovostí z opakování
- **leave-one-out**: $k =$ celkový počet objektů – maximum dat pro učení
- výhody: každý objekt stejně často $(k-1)$ pro učení a jednou pro testování, testovací množiny disjunktní a obsahující všechny objekty
- nevýhody: výpočetní náročnost, maximum dat pro učení vs. důvěryhodnost odhadu výkonnosti (malý rozptyl)

Vyhodnocení výkonnosti (performace)

Bootstrap

= část původních dat (s class labels) pro učení (trénovací) vybraná náhodně (sampling) s ponecháním, tj. objekt může být vybrán vícekrát

- v průměru konverguje k 63.2 % objektů původních dat (N, pravděpodobnost výběru objektu $1 - (1 - 1/N)^N$, $\lim_{N \to \infty} 1 - (1 - 1/N)^N = 1 - e^{-1} \doteq 0.632$)
- zbývající část pro vyhodnocení modelu (testovací)
- opakování, více variant pro přesnost modelu z přesností z opakování, např. pro .632 bootstrap $= \frac{1}{b} \sum_{i=1}^{b} 0.632 \cdot a_i + 0.368 \cdot a$, b počet opakování, a_i přesnost z opakování, a přesnost modelu naučeného z celých původních dat

Class imbalance problem

= významně různý počet objektů vstupních dat s různými class labels (třídami klasifikace)

- v aplikacích běžné, zejména u binární klasifikace (dvě třídy), např. významně méně vadných výrobků, podvodných finančních transakcí, pozitivních případů nemoci, ... různé poměry
- správná klasifikace méně zastoupené třídy důležitější – problém, pro metody všechny třídy stejně významné → zaměření se, specializace, problém šumu

→ **cost-sensitive learning** = zahrnutí ceny za (mis)klasifikaci do kritérií tvorby modelu s cílem nejnižší celkové ceny, např. gain u rozhodovacích stromů

→ **vzorkování (sampling)** = podvzorkování více zastoupené nebo nadvzorkování méně zastoupené třídy (objektů s ní), nebo obojí, u dat pro učení pro stejné zastoupení

- přesnost (accuracy) pro měření výkonnosti (performance) nevhodná – všechny třídy stejně významné: např. při 1 % pozitivních má model klasifikující vše jako negativní přesnost 99 %

Class imbalance problem

Alternativní ukazatele výkonnosti (performace) – pro binární klasifikaci

- méně zastoupená třída (objekty s ní) $\sim$ pozitivní $(+)$, více (majoritní) $\sim$ negativní $(-)$
- confusion matrix:

		predikovaná třída	
		$+$	$-$
skutečná třída	$+$	f_{++} (TP)	f_{+-} (FN)
	$-$	f_{-+} (FP)	f_{--} (TN)

TP = true positive, FN = false negative, FP = false positive, TN = true negative

- **true positive rate / sensitivity** = podíl správně predikovaných pozitivních objektů: $TPR = TP/(TP + FN)$
- **true negative rate / specificity** = podíl správně predikovaných negativních objektů: $TNR = TN/(TN + FP)$
- **false positive rate** = podíl negativních objektů nesprávně predikovaných jako pozitivní: $FPR = FP/(TN + FP)$
- **false negative rate** = podíl pozitivních objektů nesprávně predikovaných jako negativní: $FNR = FN/(TP + FN)$

Class imbalance problem

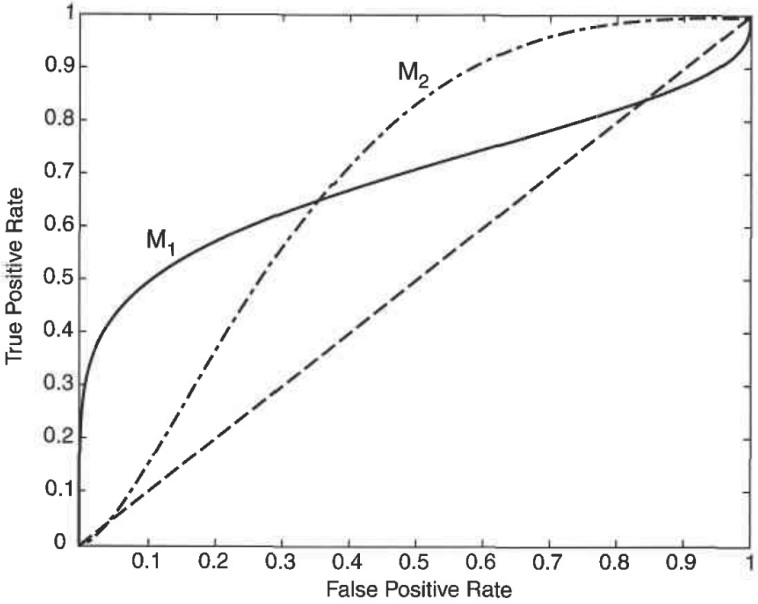
Alternativní ukazatele výkonnosti (performace) – pro binární klasifikaci

- **precision** = podíl správně predikovaných pozitivních objektů z predikovaných jako pozitivní: $p = TP/(TP + FP)$
- **recall** = podíl správně predikovaných pozitivních objektů (ze všech pozitivních): $r = TP/(TP + FN) = TPR$

$\rightarrow$ (oba) nejvyšší – pouze jeden snadné, např. vše predikované jako pozitivní

- $\mathbf{F_1} = 2pr/(p + r) = 2TP/(2TP + FP + FN) \sim$ harmonický průměr p a $r = 2/(1/p + 1/r) \rightsquigarrow$ menší z nich $\Rightarrow$ vysoký $F_1 \sim$ vysoké oba
- $\mathbf{F}_\beta = (\beta^2 + 1)pr/(\beta^2 p + r) = (\beta^2 + 1)TP/((\beta^2 + 1)TP + \beta^2 FP + FN)$, $=$ precision pro $\beta = 0$, recall pro $\beta \rightarrow \infty$
- **weighted accuracy (vážená přesnost)** $= (w_1 TP + w_4 TN)/(w_1 TP + w_2 FP + w_3 FN + w_4 TN)$
- **ROC** (receiver operating characteristic) **křivka**: vztah mezi TPR ($[0, 1]$ osa y) a FPR ($[0, 1]$ osa x) pro různé modely (s různými parametry), ..., např., plocha pod ní, ...

Class imbalance problem



Pravidlová (rule-based) klasifikace

= pomocí množiny **klasifikačních pravidel „jestliže-pak“** = model:
$R = r_1 \vee r_2 \vee \ldots r_n$ **rule set**
$\mathbf{r_i} : ((\mathbf{x_{i1}\ op\ v_{i1}}) \wedge (\mathbf{x_{i2}\ op\ v_{i2}}) \wedge \ldots (\mathbf{x_{ik}\ op\ v_{ik}})) \longrightarrow \mathbf{y_i}$ (antecedent/prekondice $\longrightarrow$ konsekvent)
x_{ij} atribut, v_{ij} hodnota atributu x_{ij}, $op \in \{=, \neq, <, >, \leq, \geq\}$ relační operátor, y_i predikovaná class label, např.
$r_1 : ((zaměstnání = ne) \wedge (nemovitost = ne)) \longrightarrow ANO$
$r_2 : ((zaměstnání = ne) \wedge (nemovitost = ano) \wedge (partner = ne)) \longrightarrow ANO$
$r_3 : ((zaměstnání = ano) \wedge (partner = ano)) \longrightarrow NE$
$r_4 : ((dětí > 1)) \longrightarrow ANO$

- pravidlo r **pokrývá** objekt x: prekondice r je pravdivá pro hodnoty atributů objektu x, např. r_1 pokrývá objekt 1 z příkladu ke klasifikaci, nepokrývá objekt 3
- základní ukazatele kvality pravidla r, pro množinu D objektů (dataset):
  - **coverage (pokrytí)** = podíl objektů z D pokrytých r
  - **accuracy (přesnost)** = podíl objektů majících class label rovnu konsekventu r z objektů pokrývaných r

  např. r_2 má pokrytí 3/11 a přesnost 2/3 na příkladu ke klasifikaci

Klasifikace objektu

= predikovaná class label pravidla pokrývajícího objekt, např. objekt 12 z příkladu ke klasifikaci pokrývají pravidla r_1 a r_4 $\Rightarrow$ class label ANO – pravidel může být víc nebo žádné

- vyčerpávající (exhaustive) pravidla v R: v R existuje pravidlo pro každou kombinaci hodnot atributů $\Rightarrow$ každý objekt je pokrývaný alespoň jedním pravidlem z R, např.
  $r'_1 : ((zaměstnání = ne) \wedge (nemovitost = ne)) \longrightarrow ANO$
  $r'_2 : ((zaměstnání = ne) \wedge (nemovitost = ano)) \longrightarrow NE$
  $r'_3 : ((zaměstnání = ano)) \longrightarrow NE$
  - jinak navíc v R **výchozí pravidlo** $() \longrightarrow \mathbf{y}$ – pokrývající objekt když ne žádné jiné, y typicky majoritní class label trénovacích objektů nepokrývaných jinými pravidly
- vzájemně výlučná pravidla v R: žádná dvě nepokrývají stejný objekt $\Rightarrow$ každý objekt je pokrývaný nejvýše jedním pravidlem z R, např. výše, jinak při více s různými class labels:
- neuspořádaná pravidla: klasifikace objektu (obvykle) většinovou class label pokrývajících pravidel, příp. vážené pomocí accuracy

Klasifikace objektu

- uspořádaná pravidla (R **decision list**): sestupně dle priority, např. accuracy, coverage, počet atributů, pořadí vytvoření, klasifikace objektu prvním pokrývajícím
  - dle pravidel: dle ukazatele kvality pravidla – méně prioritní pravidla hůře interpretovatelná (předpokládaná negace antecedentu všech předchozích pravidel), např. interpretace r_4 (v $r_1 - r_4$): jestliže (má zaměstnání a nemá partnera) nebo (nemá zaměstnání, má partnera a má nemovitost) a má více než jedno dítě, pak přidělit dávku
  - dle class labels: se stejnou class label neuspořádaně za sebou – pořadí class labels?, většina používaných, např. C4.5rules, RIPPER

Vytvoření (indukce) pravidel

- přímé metody: přímo z (trénovacích) dat, rozkládání množiny objektů (prostoru atributů) podle hodnot atributů na podmnožiny klasifikovatelné (pokrytelné) jedním pravidlem, např. sekvenční pokrývání, RIPPER
- nepřímé metody: z jiných (složitějších) klasifikačních modelů, např. rozhodovacího stromu, např. C4.5rules, neuronové sítě

Sekvenční pokrývání

- greedy strategie, pravidla postupně pro uspořádané class labels vyjma poslední – pořadí např. dle podílu objektů s class label (vzestupně), ceny za misklasifikaci (sestupně)

Input : množina D trénovacích objektů, uspořádané class labels $Y = \{y_1, y_2, \ldots, y_k\}$
Output: rule set R
$R = \emptyset$;
foreach $y \in Y \setminus \{y_k\}$ **do**
  while neplatí ukončovací podmínka **do**
    $r \leftarrow$ Learn-One-Rule(D, y);
    $D \leftarrow D \setminus$ objekty pokrývané r;
    $R \leftarrow R \vee r$;
$R \leftarrow R \vee () \longrightarrow y_k$;

Sekvenční pokrývání

LEARN-ONE-RULE(D, y)

$\rightarrow$ nejlepší pravidlo r : antecedent $\longrightarrow y$ pokrývající nejvíce objektů z D s class label y (= **pozitivní objekty**) a nejméně ostatních (= **negativní**)

- greedy postup tvorby (antecedentu) pravidla, dokud se např. zlepšuje jeho kvalita nebo pokrývá jen pozitivní objekty, strategie:
  - od obecného ke specifickému: počáteční pravidlo $() \longrightarrow y$, přidávání porovnání atributů do antecedentu, např. $() \longrightarrow NE$, $((zaměstnání = ano)) \longrightarrow NE$, $((zaměstnání = ano) \wedge (partner = ano)) \longrightarrow NE$
  - od specifického k obecnému: náhodně vybraný pozitivní objekt (atributy s hodnotami pro objekt) jako počáteční antecedent, odebírání porovnání atributů z něj pro větší pokrytí, např. pro objekt 3 z příkladu na klasifikaci $((zaměstnání = ne) \wedge (nemovitost = ano) \wedge (partner = ne)) \longrightarrow ANO$, $((zaměstnání = ne) \wedge (partner = ne)) \longrightarrow ANO$, $((zaměstnání = ne)) \longrightarrow ANO$
  - beam search: více nezávisle vytvářených nejlepších pravidel současně
- kvalita pravidla: nejen přesnost (accuracy), ale i pokrytí (coverage), např. pro 60/100 pozitivních/negativních objektů lepší r_x pokrývající 50/5 než r_y pokrývající 2/0
- poté možné odstranění (pruning) pravidla pro snížení chyby generalizace – metody odhadu viz dříve (část přeučení (overfitting) modelu)

Kvalita pravidla

- statistický test likelihood ratio – vyšší při více správných predikcích než náhodných:

$$2 \sum_c n_c \log_2(n_c/e_c)$$

  c class label, n_c počet pokrývaných objektů s class label c, e_c očekávaný n_c při náhodných predikcích pravidlem (tj. počet pokrývaných · počet s class label c / počet všech objektů), např. pro r_x je $2 \cdot (50 \cdot \log_2(50/(55 \cdot 60/160)) \ldots) \approx 99.9$, pro r_y je 5.7
- míra Laplace $\rightsquigarrow$ accuracy při vyšší coverage:

$$\frac{n_+ + 1}{n + k}$$

  n_+ počet pozitivních pokrývaných objektů (= support pravidla), n počet pokrývaných objektů, k počet class labels, např. pro r_x je $51/57 \approx 89.5\,\%$, pro r_y je $3/4 = 75\,\%$
- **FOIL's information gain** – vyšší pro vyšší n_+^r a accuracy, pro nové pravidlo r a předchozí r':

$$n_+^r \cdot \left(\log_2 \frac{n_+^r}{n_+^r + n_-^r} - \log_2 \frac{n_+^{r'}}{n_+^{r'} + n_-^{r'}} \right)$$

  např. při r' pokrývajícím 60/100 pro r_x je $50 \cdot (\log_2(50/55) - \ldots) \approx 63.9$, pro r_y je 2.8

Algoritmus RIPPER

- široce používaný
- pravidla vytvářena sekvenčním pokrývání – pořadí class labels vzestupně dle podílu objektů s class label, ukončovací podmínka (s nepřidáním pravidla) maximální navýšení velikosti rule set (description length) nebo chybovost pravidla na validation set přes 50 %
- tvorba pravidla od obecného ke specifickému – dokud pokrývá jen pozitivní objekty, ukazatel kvality pravidla FOIL's information gain
- pruning pravidla postupným odebíráním posledně přidaného porovnání atributu – při vyšším $(n_+ - n_-)/(n_+ + n_-)$ ($\sim$ accuracy) na validation set, poté může pokrývat i negativní objekty
- dále nahrazování pravidel jinými lepšími

Pravidla z rozhodovacího stromu

- pravidlo $\sim$ cesta z kořenového do listového uzlu: konjunkce testů nad atributy u vnitřních uzlů na cestě = antecedent, class label u listového uzlu = konsekvent, např. pro strom z příkladu dříve:
  $((příjem \leq 9,000)) \longrightarrow ANO$
  $((příjem > 9,000) \wedge (věk \leq 33)) \longrightarrow ANO$
  $((příjem > 9,000) \wedge (věk > 33)) \longrightarrow NE$
- pravidla pro všechny cesty vyčerpávající a vzájemně výlučná $\Rightarrow$ každý objekt je pokrývaný právě jedním pravidlem
- některá pravidla lze zjednodušit (nemusí pak být vzájemně výlučná)

Algoritmus C4.5rules

- pravidla pro všechny cesty z kořenového do listového uzlu
- pruning pravidel odebíráním porovnání atributu – pro co nejnižší odhad chyby generalizace (dle preference jednoduššího modelu, viz dříve)
- pravidla uspořádaná dle class labels – pořadí class labels vzestupně dle velikosti pravidel se stejnou class label plus počet misklasifikovaných objektů (description length)

Vlastnosti pravidlové klasifikace

- vytvoření, interpretace, kvalita (přesnost), robustnost pravidel **podobné jako u rozhodovacích stromů**
- rozklad množiny / separace objektů (prostoru atributů) na podmnožiny s hranicemi (decision boundaries) nezávislými na jiných atributech (rektilineárními) – podobně jako rozhodovací stromy, při pokrývání objektu více pravidly i jiné hranice ($\sim$ oblique stromy)
- typicky používaná pro **deskriptivní modelování**
- s uspořádanými pravidly dle class labels vhodná pro data s významně různými počty objektů s různými class labels (**class imbalanced data**)

Klasifikace nejbližšími sousedy (nearest-neighbor)

* klasifikace = (1) vytvoření (učení, indukce) modelu ze vstupních dat, (2) aplikování modelu na nová data (dedukce)
- **eager klasifikace/learner** = nejdříve (1) pro celý model, pak (2), např. rozhodovací stromy, pravidlová
- **lazy klasifikace/learner** = odložení (1) pro část modelu až do její potřeby v (2) nebo i vynechání (1), např. **Rote klasifikace**: model = vstupní data a klasifikace dle vstupního objektu = nový objekt – nemusí existovat

→ klasifikace dle vstupních objektů podobných novému = **nejbližší sousedi** – „když to kváká jako kachna, . . . “

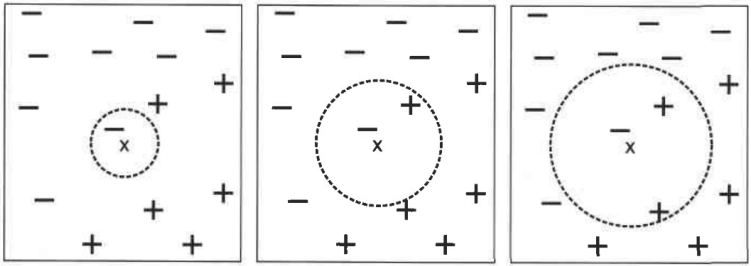
- **k-nejbližší sousedi** = k vstupních objektů nejbližších novému, dle míry (ne)podobnosti (vzdálenosti) = **model dal** – objekt $\sim$ bod v n-rozměrném prostoru, kde n je počet atributů, např.

= majoritní class label nejbližších sousedů, nebo kterákoliv

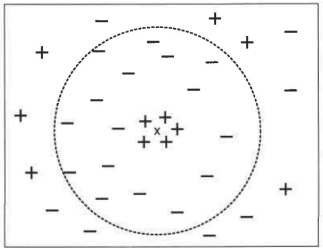
! volba k: malé $\rightsquigarrow$ přeučení kvůli šumu ve vstupních datech, velké $\rightsquigarrow$ chybná klasifikace dle většiny vzdálenějších objektů, např.

→ váhy class label sousedů $\mathbf{x_i}$ dle jejich vzdálenosti: $w_i = 1/d(\mathbf{x}, \mathbf{x_i})^2$

Klasifikace nejbližšími sousedy (nearest-neighbor)



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor



Vlastnosti klasifikace nejbližšími sousedy

- jedna z **instance-based klasifikací/learning** = klasifikace **dle** určitých **vstupních objektů bez** udržování **modelu** (abstrakce) dat – potřeba míry (ne)podobnosti mezi objekty a určení (klasifikační funkce) class label nového objektu dle (ne)podobností
- lazy klasifikátory nevytváří model dat, ale klasifikace objektu může být výpočetně náročná – výpočet (ne)podobností s objekty vstupních dat, eager klasifikátory opačně
- klasifikace na základě „lokální informace“ – vliv šumu (s malým k)
- libovolné proměnlivé hranice (decision boundaries) separace objektů (prostoru atributů) – závisí na konkrétním umístění objektů v prostoru atributů
- zavisí na volbě míry (ne)podobnosti a předzpracování dat, např. normalizaci

Bayesovská klasifikace

- vztah mezi atributy a class label může být neurčitý (nedeterministický), class label nového objektu dle atributů nemusí být možné určit jednoznačně – kvůli šumu v datech, nejednoznačné interpretaci atributů nebo dalším neznámým faktorům mimo atributy, např. predikce zdravotních problémů

$\rightarrow$ pravděpodobnostní modelování vztahu

Bayesova věta

$=$ statistický princip kombinace (v klasifikaci) znalostí class labels a hodnot atributů

- X, Y náhodné proměnné, sdružená (joint) pravděpodonost $P(X = x, Y = y)$, $P(X) = \sum_i P(X, Y = y_i)$ (zákon o úplné pravděpodobnosti), podmíněná pravděpodobnost $P(Y = y|X = x)$: $P(X, Y) = P(Y|X)P(X) = P(X|Y)P(Y)$ (Bayesovo pravidlo)

$$\mathbf{P}(\mathbf{Y}|\mathbf{X}) = \frac{\mathbf{P}(\mathbf{X}|\mathbf{Y})\mathbf{P}(\mathbf{Y})}{\mathbf{P}(\mathbf{X})}$$

Bayesovská klasifikace

Bayesova věta: příklad

Nakažených 10 %, z nich pozitivně testovaných 90 %, ale z nenakažených pozitivně testovaných 20 %. Nakažený, když pozitivně testovaný?

Bayesova věta: příklad

Nakažených 10 %, z nich pozitivně testovaných 90 %, ale z nenakažených pozitivně testovaných 20 %. Nakažený, když pozitivně testovaný?

$X \in 0, 1 \ldots$ test negativní (0) / pozitivní (1),
$Y \in 0, 1 \ldots$ nákaza ne (0) / ano (1),
$P(Y = 1) = 0.1$, $P(Y = 0) = 1 - P(Y = 1) = 0.9$, $P(X = 1|Y = 1) = 0.9$,
$P(X = 1|Y = 0) = 0.2$

$$\begin{aligned} P(Y = 1|X = 1) &= P(X = 1|Y = 1)P(Y = 1)/P(X = 1) \\ &= P(X = 1|Y = 1)P(Y = 1)/(P(X = 1, Y = 1) + P(X = 1, Y = 0)) \\ &= P(X = 1|Y = 1)P(Y = 1)/(P(X = 1|Y = 1)P(Y = 1) + P(X = 1|Y = 0)P(Y = 0)) \\ &= 0.9 \cdot 0.1/(0.9 \cdot 0.1 + 0.2 \cdot 0.9) \\ &\doteq 0.333 \end{aligned}$$

Bayesovská klasifikace

- $\mathbf{X} = (X_1, \ldots, X_m)$ vektor m atributů, Y class labels, náhodné proměnné
- $P(Y)$ **prior pravděpodobnost** class labels, $P(Y|\mathbf{X})$ **posterior pravděpodobnost** class labels v závislosti na atributech = **model dat** (pro class labels a hodnoty atributů pro objekty)

$=$ class label nového objektu $\mathbf{x} = (x_1, \ldots, x_m)$: y s maximální $P(Y = y|\mathbf{X} = \mathbf{x})$ – její (přesnější) odhad vyžaduje mnoho (ideálně všechny) kombinací class labels y a hodnot x_i atributů $\mathbf{X}$

$\rightarrow$ vyjádření $P(Y|\mathbf{X})$ pomocí $P(Y)$, pravděpodobnosti $P(\mathbf{X}|Y)$ v závislosti na class labels a $P(\mathbf{X})$ – s využitím Bayesovy věty

- $P(\mathbf{X})$ pro různé y konstantní $\Rightarrow$ netřeba, odhad $P(Y)$ z četností y pro objekty vstupních dat

? odhad $P(\mathbf{X}|Y)$ – stále vyžaduje mnoho (ideálně všechny) kombinací y a x_i, ale možné vyjádřit snadněji

- např. X_i výskyt určitého slova ano/ne (nebo i počet), Y spam ano/ne, objekty $\sim$ maily, $P(Y)$, $P(Y|\mathbf{X})$, $P(\mathbf{X}|Y)$

Naïvní bayesovská klasifikace

= atributy X_i jsou, při dané class label y, jako náhodné proměnné **nezávislé**:

$$P(\mathbf{X}|Y = y) = \prod_{i=1}^{m} P(X_i|Y = y)$$

- nezávislost X_1 a X_2 při Y: $P(X_1|X_2, Y) = P(X_1|Y)$, např. X_1 velikost hlavy, X_2 inteligence, Y ?, $P(X_1, X_2|Y) = \ldots = P(X_1|Y)P(X_2|Y)$

→ místo odhadu $P(\mathbf{X}|Y)$ odhady $P(X_i|Y)$ – stačí méně kombinací y a hodnot x_i (jednoho) atributu X_i

= class label y s maximální

$$P(Y = y|\mathbf{X} = \mathbf{x}) = \frac{P(Y = y)\prod_{i=1}^{m} P(X_i = x_i|Y = y)}{\cancel{P(\mathbf{X} = \mathbf{x})}}$$

? odhad $P(X_i|Y)$

Naïvní bayesovská klasifikace

- diskrétní atribut X: odhad $P(X|Y)$ z relativních četností hodnot X vzhledem k y pro objekty vstupních dat
- spojitý atribut X:
  - diskretizace na ordinální (nahrazení hodnot intervaly) – chyba odhadu závisí na metodě a počtu intervalů (malé unikátní vs. velké více class labels)
  - předpokládáná určitá distribuce $P(X|Y)$ a odhad parametrů funkce hustoty pravděpodobnosti f z hodnot X vzhledem k y pro objekty vstupních dat, např. normální (gaussovská) s parametry střední hodnota μ (průměr $\overline{x}$) a rozptyl σ^2 (s^2):

$$P(x \leq X \leq x + \epsilon | Y = y) \approx \epsilon f(X, \mu_{xy}, \sigma_{xy}) = \epsilon \frac{1}{\sigma_{xy}\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu_{xy})^2}{\sigma_{xy}^2}}$$

Naïvní bayesovská klasifikace: příklad

$P(partner = ano|dávka = ANO) = 2/6$, $P(partner = ne|dávka = ANO) = 4/6$,
$P(partner = ano|dávka = NE) = 3/5$, $P(partner = ne|dávka = NE) = 2/5$
$P(zaměstnání = ano|dávka = ANO) = 1/6$, $P(zaměstnání = ne|dávka = ANO) = 5/6$,
$P(zaměstnání = ano|dávka = NE) = 3/5$, $P(zaměstnání = ne|dávka = NE) = 2/5$
$P(nemovitost = ano|dávka = ANO) = 2/6$, $P(nemovitost = ne|dávka = ANO) = 4/6$,
$P(nemovitost = ano|dávka = NE) = 2/5$, $P(nemovitost = ne|dávka = NE) = 3/5$

$X = věk$, $dávka = ANO$: $\overline{x} = \frac{32+...+62}{6} \doteq 31.9$, $s^2 = \frac{(32-31.9)^2+...+(62-31.9)^2}{5} \doteq 353.1$
$X = věk$, $dávka = NE$: $\overline{x} = \frac{58+...+49}{5} = 46$, $s^2 = \frac{(58-46)^2+...+(49-46)^2}{4} \doteq 76.5$

$X = dětí$, $dávka = ANO$: $\overline{x} = \frac{5+...+0}{6} \doteq 2.2$, $s^2 = \frac{(5-2.2)^2+...+(0-2.2)^2}{5} \doteq 3$
$X = dětí$, $dávka = NE$: $\overline{x} = \frac{0+...+3}{5} = 2.4$, $s^2 = \frac{(0-2.4)^2+...+(3-2.4)^2}{4} = 5.3$

$X = příjem$, $dávka = ANO$: $\overline{x} = \frac{6335+...+6214}{6} \doteq 5597.3$,
$s^2 = \frac{(6335-5597.3)^2+...+(6214-5597.3)^2}{5} \doteq 6799900.7$
$X = příjem$, $dávka = NE$: $\overline{x} = \frac{9055+...+16150}{5} = 14028.2$,
$s^2 = \frac{(9055-14028.2)^2+...+(16150-14028.2)^2}{4} = 8095111.2$

Naïvní bayesovská klasifikace: příklad

class label ($dávka = ANO/NE$) objektu $\mathbf{X} = (věk = 31, partner = ano, dětí = 7, zaměstnání = ne, příjem = 10375, nemovitost = ne)$?
$\rightarrow P(dávka = ANO|\mathbf{X})$? $P(dávka = NE|\mathbf{X})$?

$P(dávka = ANO) = 6/11$, $P(dávka = NE) = 5/11$
$P(\mathbf{X}|dávka = ANO) = P(věk = 31|ANO)P(partner = ano|ANO)P(dětí = 7|ANO)P(zaměstnání = ne|ANO)P(příjem = 10375|ANO)P(nemovitost = ne|ANO) \doteq$

$$P(věk = 31|ANO) = \epsilon \frac{1}{\sqrt{353.1}\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(31-31.9)^2}{353.1}} \doteq 0.0212\epsilon$$

$$P(dětí = 7|ANO) = \epsilon \frac{1}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(7-2.2)^2}{3}} \doteq 0.00495\epsilon$$

$$P(příjem = 10375|ANO) = \epsilon \frac{1}{\sqrt{6799900.7}\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(10375-5597.3)^2}{6799900.7}} \doteq 0.0000286\epsilon$$

Naïvní bayesovská klasifikace: příklad

class label ($dávka = ANO/NE$) objektu $\mathbf{X} = (věk = 31, partner = ano, dětí = 7, zaměstnání = ne, příjem = 10375, nemovitost = ne)$?
$\rightarrow P(dávka = ANO|\mathbf{X})$? $P(dávka = NE|\mathbf{X})$?

$P(dávka = ANO) = 6/11$, $P(dávka = NE) = 5/11$
$P(\mathbf{X}|dávka = ANO) = P(věk = 31|ANO)P(partner = ano|ANO)P(dětí = 7|ANO)P(zaměstnání = ne|ANO)P(příjem = 10375|ANO)P(nemovitost = ne|ANO) \doteq 0.0212 \cdot 2/6 \cdot 0.00495 \cdot 5/6 \cdot 0.0000286 \cdot 4/6 \doteq 10^{-9}\epsilon^3$

$P(věk = 31|ANO) = \epsilon \frac{1}{\sqrt{353.1}\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(31-31.9)^2}{353.1}} \doteq 0.0212\epsilon$

$P(dětí = 7|ANO) = \epsilon \frac{1}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(7-2.2)^2}{3}} \doteq 0.00495\epsilon$

$P(příjem = 10375|ANO) = \epsilon \frac{1}{\sqrt{6799900.7}\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(10375-5597.3)^2}{6799900.7}} \doteq 0.0000286\epsilon$

Naïvní bayesovská klasifikace: příklad

class label ($dávka = ANO/NE$) objektu $\mathbf{X} = (věk = 31, partner = ano, dětí = 7, zaměstnání = ne, příjem = 10375, nemovitost = ne)$?
$\rightarrow P(dávka = ANO|\mathbf{X})$? $P(dávka = NE|\mathbf{X})$?

$P(dávka = ANO) = 6/11,\ P(dávka = NE) = 5/11$
$P(\mathbf{X}|dávka = ANO) = P(věk = 31|ANO)P(partner = ano|ANO)P(dětí = 7|ANO)P(zaměstnání = ne|ANO)P(příjem = 10375|ANO)P(nemovitost = ne|ANO) \doteq 0.0212 \cdot 2/6 \cdot 0.00495 \cdot 5/6 \cdot 0.0000286 \cdot 4/6 \doteq 10^{-9}\epsilon^3$
$P(\mathbf{X}|dávka = NE) = P(věk = 31|NE)P(partner = ano|NE)P(dětí = 7|NE)P(zaměstnání = ne|NE)P(příjem = 10375|NE)P(nemovitost = ne|NE) \doteq$
$P(věk = 31|NE) = \epsilon\frac{1}{\sqrt{76.5}\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(31-46)^2}{76.5}} \doteq 0.0105\epsilon$
$P(dětí = 7|NE) = \epsilon\frac{1}{\sqrt{5.3}\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(7-2.4)^2}{5.3}} \doteq 0.0235\epsilon$
$P(příjem = 10375|NE) = \epsilon\frac{1}{\sqrt{8095111.2}\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(10375-14028.2)^2}{8095111.2}} \doteq 0.0000615\epsilon$

Naïvní bayesovská klasifikace: příklad

class label ($dávka = ANO/NE$) objektu $\mathbf{X} = (věk = 31, partner = ano, dětí = 7, zaměstnání = ne, příjem = 10375, nemovitost = ne)$?
$\rightarrow P(dávka = ANO|\mathbf{X})$? $P(dávka = NE|\mathbf{X})$?

$P(dávka = ANO) = 6/11,\ P(dávka = NE) = 5/11$
$P(\mathbf{X}|dávka = ANO) = P(věk = 31|ANO)P(partner = ano|ANO)P(dětí = 7|ANO)P(zaměstnání = ne|ANO)P(příjem = 10375|ANO)P(nemovitost = ne|ANO) \doteq 0.0212 \cdot 2/6 \cdot 0.00495 \cdot 5/6 \cdot 0.0000286 \cdot 4/6 \doteq 10^{-9}\epsilon^3$
$P(\mathbf{X}|dávka = NE) = P(věk = 31|NE)P(partner = ano|NE)P(dětí = 7|NE)P(zaměstnání = ne|NE)P(příjem = 10375|NE)P(nemovitost = ne|NE) \doteq 0.0105 \cdot 3/5 \cdot 0.0235 \cdot 2/5 \cdot 0.0000615 \cdot 3/5 \doteq 2 \cdot 10^{-9}\epsilon^3$

$$P(věk = 31|NE) = \epsilon \frac{1}{\sqrt{76.5}\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(31-46)^2}{76.5}} \doteq 0.0105\epsilon$$

$$P(dětí = 7|NE) = \epsilon \frac{1}{\sqrt{5.3}\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(7-2.4)^2}{5.3}} \doteq 0.0235\epsilon$$

$$P(příjem = 10375|NE) = \epsilon \frac{1}{\sqrt{8095111.2}\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(10375-14028.2)^2}{8095111.2}} \doteq 0.0000615\epsilon$$

Naïvní bayesovská klasifikace: příklad

class label ($dávka = ANO/NE$) objektu $\mathbf{X} = (věk = 31, partner = ano, dětí = 7, zaměstnání = ne, příjem = 10375, nemovitost = ne)$?
$\rightarrow P(dávka = ANO|\mathbf{X})$? $P(dávka = NE|\mathbf{X})$?

$P(dávka = ANO) = 6/11$, $P(dávka = NE) = 5/11$
$P(\mathbf{X}|dávka = ANO) = P(věk = 31|ANO)P(partner = ano|ANO)P(dětí = 7|ANO)P(zaměstnání = ne|ANO)P(příjem = 10375|ANO)P(nemovitost = ne|ANO) \doteq 0.0212 \cdot 2/6 \cdot 0.00495 \cdot 5/6 \cdot 0.0000286 \cdot 4/6 \doteq 10^{-9}\epsilon^3$
$P(\mathbf{X}|dávka = NE) = P(věk = 31|NE)P(partner = ano|NE)P(dětí = 7|NE)P(zaměstnání = ne|NE)P(příjem = 10375|NE)P(nemovitost = ne|NE) \doteq 0.0105 \cdot 3/5 \cdot 0.0235 \cdot 2/5 \cdot 0.0000615 \cdot 3/5 \doteq 2 \cdot 10^{-9}\epsilon^3$

$P(dávka = ANO|\mathbf{X}) \doteq 6/11 \cdot 10^{-9}\epsilon^3/P(\mathbf{X}) < P(dávka = NE|\mathbf{X}) \doteq 5/11 \cdot 2 \cdot 10^{-9}\epsilon^3/P(\mathbf{X}) \Rightarrow$ class label NE

Naïvní bayesovská klasifikace

! PROBLÉM: $P(X = x|Y = y) = 0$ pro některý atribut X a jeho hodnotu x a class label $y \Rightarrow P(Y = y|\mathbf{X} = \mathbf{x}) = 0$

- při $P(Y = y|\mathbf{X}) = 0$ pro všechny y nemožnost klasifikace některých nových objektů!

→ **Laplace-odhad** $P(X = x|Y = y)$ (místo relativní četnosti x vzhledem k y pro objekty vstupních dat):

$$P(X = x|Y = y) = \frac{n_{x,y} + 1}{n_y + v}$$

$n_{x,y}$ počet vstupních objektů s hodnotou x atributu X a class label y, n_y počet vstupních objektů s class label y, v počet všech možných hodnot atributu X

→ m-**odhad** $P(X = x|Y = y)$:

$$P(X = x|Y = y) = \frac{n_{x,y} + mp}{n_y + m}$$

p prior pravděpodobnost hodnoty x atributu X pro objekty s class label y (při $n_y = 0$), m tzv. ekvivalentní velikost vzorku pro p („důvěra“ v p při malém n_y) $\approx$ kompromis mezi p a $n_{x,y}/n_y$

- robustnější odhad $P(X|Y)$ při malém počtu vstupních objektů

Vlastnosti naïvní bayesovské klasifikace

- **robustní vůči šumu** (izolované body jsou „zprůměrovány“) i **přeučení**, při chybějících hodnotách objekt vynechán nebo hodnoty odhadnuty z pravděpodobnostního rozložení atributu
- robustní vůči irelevantním atributům ($P(X|Y)$ rozložena uniformně)
- **korelované atributy snižují výkonnost** – přestává platit předpoklad nezávislosti atributů: $P(\mathbf{X}|Y=y) = \prod_{i=1}^{m} P(X_i|Y=y)$ se mění na $P(\mathbf{X}|Y=y) = P(X_i|Y=y), i = 1, \ldots, m$

→ **Bayesovské sítě (Bayesian (belief) networks)** – uvažování závislostí mezi atributy v modelu a při klasifikaci

- z pravděpodobnostních rozložení $P(X|Y)$ a $P(Y)$ je možné určit ideální decision boundaries (= hodnota $\mathbf{x}$ pro kterou $P(Y=y_i)P(\mathbf{X}=\mathbf{x}|Y=y_i) = P(Y=y_j)P(\mathbf{X}=\mathbf{x}|Y=y_j)$), např., a minimální chybovost (**Bayes error rate**, = suma ploch pod částmi grafů $P(Y=y|\mathbf{X})$ odpovídajícími misklasifikaci) jakéhokoliv klasifikačního modelu

Vlastnosti naïvní bayesovské klasifikace