Vytěžování dat

Filip Železný

Katedra kybernetiky
skupina Inteligentní Datové Analýzy (IDA)

24. února 2010

Jednoduchá úloha klasifikace

Datová tabulka

Vysoké příjmy	**Splácí úvěr**
ano	ano
ano	ne
ne	ano
ano	ano
ne	ne
ne	ne
ne	ano
ano	ano
ano	ano
ano	ano
ne	ne

Jednoduchá úloha klasifikace

Datová tabulka

Vysoké příjmy	**Splácí úvěr**
ano	ano
ano	ne
ne	ano
ano	ano
ne	ne
ne	ne
ne	ano
ano	ano
ano	ano
ano	ano
ne	ne

Kontingenční tabulka

		Splácí úvěr		
		ano	ne	$\sum$
vysoké	ano	5	1	6
příjmy	ne	2	3	5
	$\sum$	7	4	11

Nový žadatel o úvěr, má vysoké příjmy.

- Bude splácet úvěr?
- S jakou pravděpodobností?

Příznaková klasifikace

- **vysoké příjmy**: příznak
- **Splácí úvěr**: cílová veličina (třída)

Frekvence a pravděpodobnost

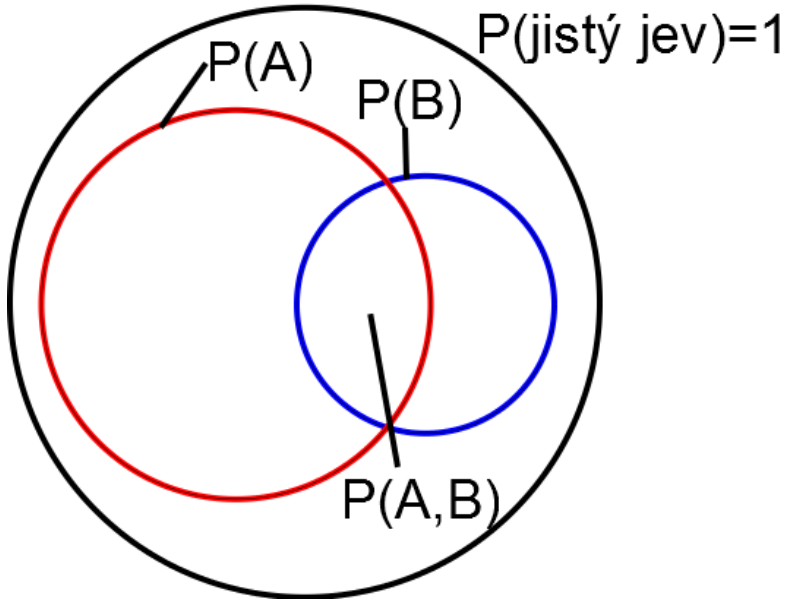
	S	$\neg S$	$\sum$
V	a	b	r
$\neg V$	c	d	s
$\sum$	k	l	n

- V (vys. příjmy), S (splácí úvěr): **náhodné jevy**
- $\Pr(V) \approx r/n$, $\Pr(S) \approx k/n$: **marginální** pravděpodobnosti
- $\Pr(V, S) \approx a/n$, $\Pr(V, \neg S) \approx b/n$, atd.: **sdružené** pravděpodobnosti
- $\Pr(V|S) \approx a/k$, $\Pr(V|\neg S) \approx b/l$, atd.: **podmíněné** pravděpod.
- Frekvence konvergují k pravděpodobnostem s roustoucím n. Např.

$$\lim_{n \to \infty} r/n = \Pr(V)$$

Základní vlastnosti pravděpodobnosti

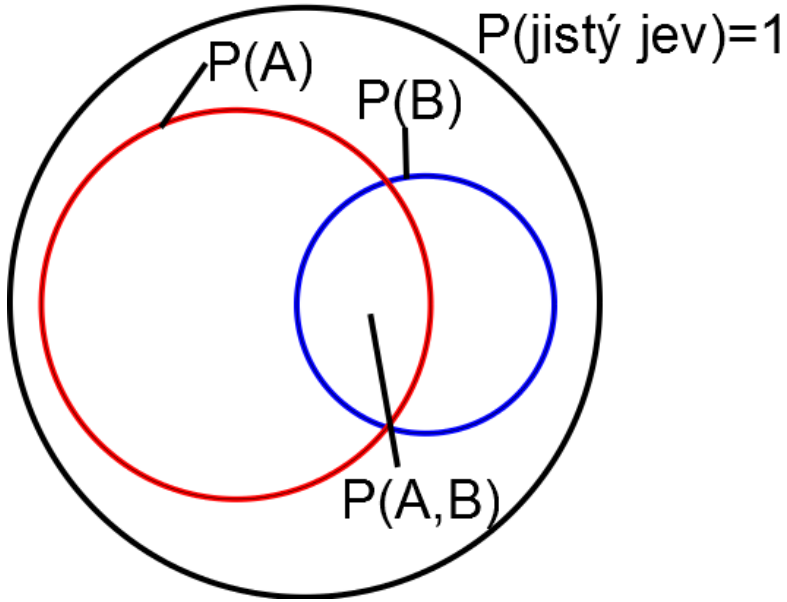
- Zřejmé z geometrické představy



- $0 \leq \Pr(\dots) \leq 1$
- $\Pr(\neg A) =$

Základní vlastnosti pravděpodobnosti

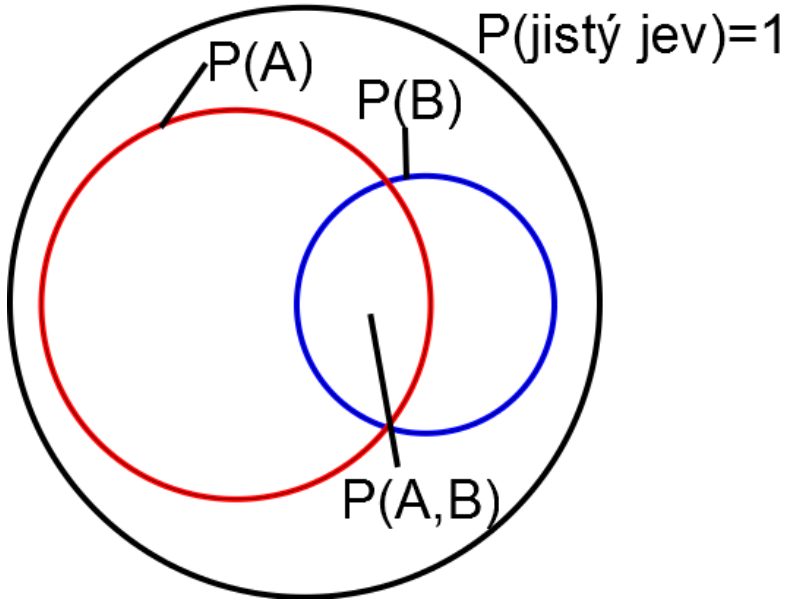
- Zřejmé z geometrické představy



- $0 \leq \Pr(\dots) \leq 1$
- $\Pr(\neg A) = 1 - \Pr(A)$

Základní vlastnosti pravděpodobnosti

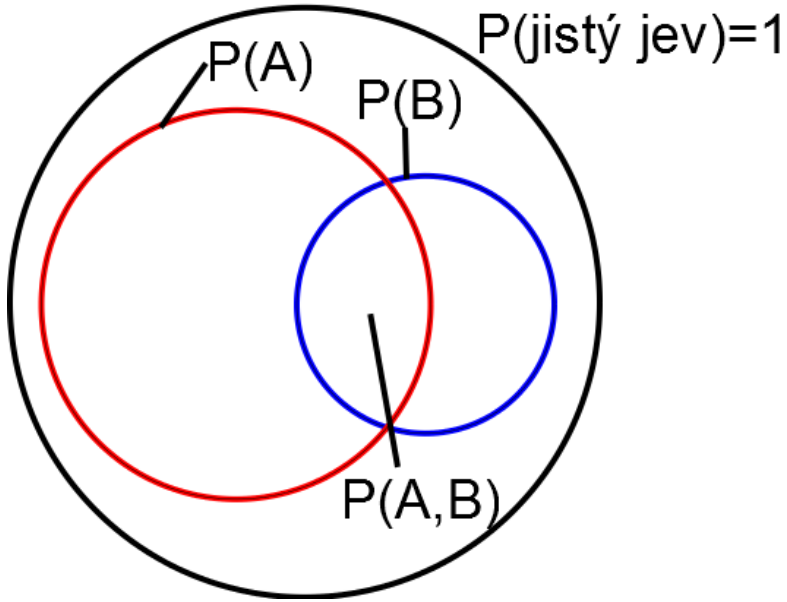
- Zřejmé z geometrické představy



- $\Pr(A|B) =$
- $0 \leq \Pr(\dots) \leq 1$
- $\Pr(\neg A) = 1 - \Pr(A)$

Základní vlastnosti pravděpodobnosti

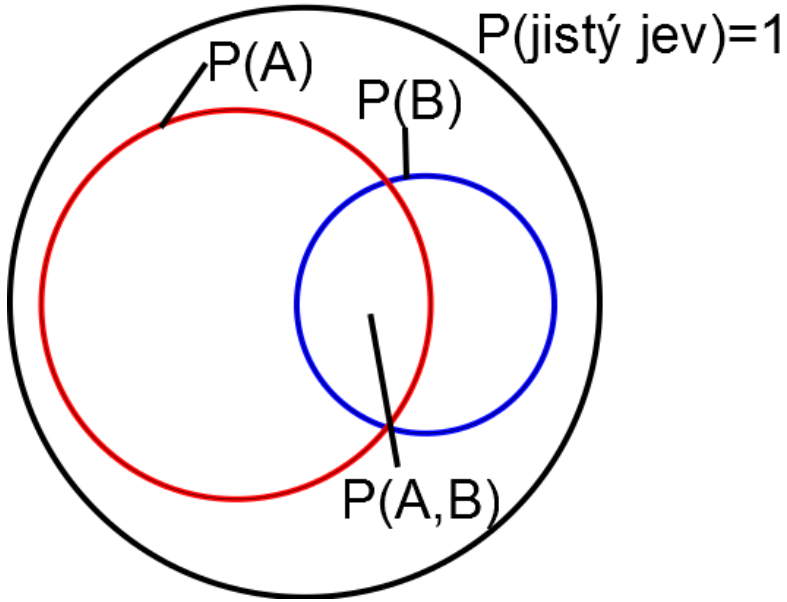
- Zřejmé z geometrické představy



- $\Pr(A|B) = \Pr(A, B)/\Pr(B)$
- $0 \leq \Pr(\dots) \leq 1$
- $\Pr(\neg A) = 1 - \Pr(A)$

Základní vlastnosti pravděpodobnosti

- Zřejmé z geometrické představy



- $0 \leq \Pr(\dots) \leq 1$
- $\Pr(\neg A) = 1 - \Pr(A)$
- $\Pr(A|B) = \Pr(A, B)/\Pr(B)$
- $\Pr(\neg A|\dots) = 1 - \Pr(A|\dots)$
- $\Pr(A \text{ nebo } B) =$
  $\Pr(A) + \Pr(B) - \Pr(A, B)$

Nezávislost jevů

- Pokud platí

$$\Pr(A, B) = \Pr(A) \cdot \Pr(B)$$

neboli $\Pr(A|B) = \Pr(A)$, tak jsou jevy A a B **nezávislé**.

- Jsou splácení úvěru (S) a vysoké příjmy (V) nezávislé?

	S	$\neg S$	$\sum$
V	5	1	6
$\neg V$	2	3	5
$\sum$	7	4	11

Nezávislost jevů

- Pokud platí

$$\Pr(A, B) = \Pr(A) \cdot \Pr(B)$$

  neboli $\Pr(A|B) = \Pr(A)$, tak jsou jevy A a B **nezávislé**.
- Jsou splácení úvěru (S) a vysoké příjmy (V) nezávislé?

	S	$\neg S$	$\sum$
V	5	1	6
$\neg V$	2	3	5
$\sum$	7	4	11

- $\Pr(V, S) \approx 5/11 = 0.45\ldots$
- $\Pr(V) \cdot \Pr(S) \approx 6/11 \cdot 7/11 = 0.34\ldots$
- Z dat se zdá, že jsou závislé. Proč to nemůžeme říci s jistotou?

Pokračování klasifikační úlohy

	S	$\neg S$	$\sum$
V	5	1	6
$\neg V$	2	3	5
$\sum$	7	4	11

- "Bude vysokopříjmový klient splácet úvěr?"
  - Jaký typ pravděpodobnosti odpovídá na tuto otázku?

Pokračování klasifikační úlohy

	S	$\neg S$	$\sum$
V	5	1	6
$\neg V$	2	3	5
$\sum$	7	4	11

- "Bude vysokopříjmový klient splácet úvěr?"
  - Jaký typ pravděpodobnosti odpovídá na tuto otázku?
  - S pravděpodobností $\Pr(S|V) \approx 5/6$ bude splácet.
- "Bude klient, o kterém nic nevíme, splácet úvěr?"

Pokračování klasifikační úlohy

	S	$\neg S$	$\sum$
V	5	1	6
$\neg V$	2	3	5
$\sum$	7	4	11

- "Bude vysokopříjmový klient splácet úvěr?"
  - Jaký typ pravděpodobnosti odpovídá na tuto otázku?
  - S pravděpodobností $\Pr(S|V) \approx 5/6$ bude splácet.
- "Bude klient, o kterém nic nevíme, splácet úvěr?"
  - Jaký typ pravděpodobnosti odpovídá na tuto otázku?

Pokračování klasifikační úlohy

	S	$\neg S$	$\sum$
V	5	1	6
$\neg V$	2	3	5
$\sum$	7	4	11

- "Bude vysokopříjmový klient splácet úvěr?"
  - Jaký typ pravděpodobnosti odpovídá na tuto otázku?
  - S pravděpodobností $\Pr(S|V) \approx 5/6$ bude splácet.
- "Bude klient, o kterém nic nevíme, splácet úvěr?"
  - Jaký typ pravděpodobnosti odpovídá na tuto otázku?
  - S pravděpodobností $\Pr(S) \approx 7/11$ bude splácet.
- $\Pr(S|V) > \Pr(S)$ (nejsou nezávislé!)
  - $\Pr(S|V)$ též **apriorní** pravděpodobnost
  - $\Pr(S|V)$ též **aposteriorní** pravděpodobnost

Náhodná veličina

- Náhodný jev je binární pojem (nastane / nenastane)
- Pro modelování dat potřebujeme širší škály hodnot. Např.
  - příjmy: $p \in \{\text{vysoké}, \text{střední}, \text{nízké}\}$
  - splácení úvěru: $u \in \{\text{splácí}, \text{problémy}, \text{nesplácí}\}$

Příjmy (p)	**Úvěr (u)**
vysoké	splácí
nízké	nesplácí
střední	problémy
nízké	problémy
...	...

- p a u jsou (diskrétní) **náhodné veličiny** (n.v.)
- N.v. charakterizuje tzv. **rozdělení pravděpodobnosti**

Rozdělení pravděpodobnosti n.v.

- Rozdělení n.v. v je funkce $P_v(x) = \Pr(v = x)$.
- K hodnotám rozdělení opět konvergují frekvence v kontingenční tabulce:

p↓ u→	splácí	problémy	nesplácí	$\sum$
vysoké	2	1	0	3
střední	2	2	2	6
nízké	0	1	1	2
$\sum$	4	4	3	11

- P_p, P_u: **marginální** rozdělení p (příjmy) resp. u (splácení úvěru)
  - např. $P_p(\text{střední}) \approx 6/11$, $P_u(\text{problémy}) \approx 4/11$
- $P_{p,u}$: **sdružené** rozdělení p a u
  - např. $P_{p,u}(\text{střední}, \text{splácí}) \approx$

Rozdělení pravděpodobnosti n.v.

- Rozdělení n.v. v je funkce $P_v(x) = \Pr(v = x)$.
- K hodnotám rozdělení opět konvergují frekvence v kontingenční tabulce:

p↓ u→	splácí	problémy	nesplácí	$\sum$
vysoké	2	1	0	3
střední	2	2	2	6
nízké	0	1	1	2
$\sum$	4	4	3	11

- P_p, P_u: **marginální** rozdělení p (příjmy) resp. u (splácení úvěru)
  - např. $P_p(\text{střední}) \approx 6/11$, $P_u(\text{problémy}) \approx 4/11$
- $P_{p,u}$: **sdružené** rozdělení p a u
  - např. $P_{p,u}(\text{střední}, \text{splácí}) \approx 2/11$

Rozdělení pravděpodobnosti n.v.

- Rozdělení n.v. v je funkce $P_v(x) = \Pr(v = x)$.
- K hodnotám rozdělení opět konvergují frekvence v kontingenční tabulce:

p↓ u→	splácí	problémy	nesplácí	$\sum$
vysoké	2	1	0	3
střední	2	2	2	6
nízké	0	1	1	2
$\sum$	4	4	3	11

- P_p, P_u: **marginální** rozdělení p (příjmy) resp. u (splácení úvěru)
  - např. $P_p(\text{střední}) \approx 6/11$, $P_u(\text{problémy}) \approx 4/11$
- $P_{p,u}$: **sdružené** rozdělení p a u
  - např. $P_{p,u}(\text{střední}, \text{splácí}) \approx 2/11$
- $P_{p,u}$: **podmíněné** rozdělení p a u
  - např. $P_{p|u}(\text{střední}|\text{splácí}) \approx$

Rozdělení pravděpodobnosti n.v.

- Rozdělení n.v. v je funkce $P_v(x) = \Pr(v = x)$.
- K hodnotám rozdělení opět konvergují frekvence v kontingenční tabulce:

p↓ u→	splácí	problémy	nesplácí	$\sum$
vysoké	2	1	0	3
střední	2	2	2	6
nízké	0	1	1	2
$\sum$	4	4	3	11

- P_p, P_u: **marginální** rozdělení p (příjmy) resp. u (splácení úvěru)
  - např. $P_p(\text{střední}) \approx 6/11$, $P_u(\text{problémy}) \approx 4/11$
- $P_{p,u}$: **sdružené** rozdělení p a u
  - např. $P_{p,u}(\text{střední}, \text{splácí}) \approx 2/11$
- $P_{p,u}$: **podmíněné** rozdělení p a u
  - např. $P_{p|u}(\text{střední}|\text{splácí}) \approx 2/4$

Rozdělení pravděpodobnosti n.v.

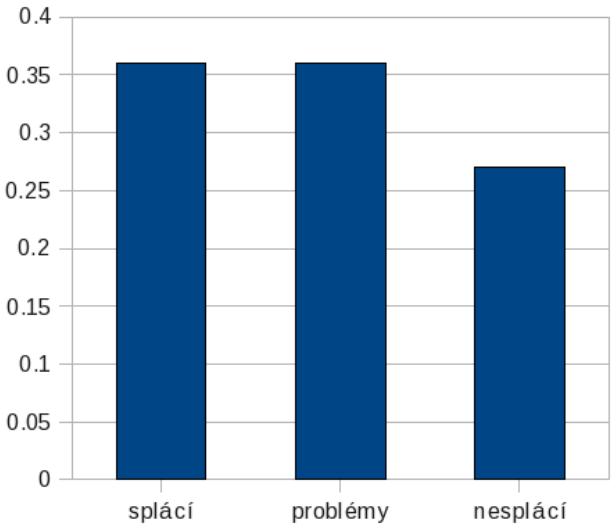
- Rozdělení n.v. v je funkce $P_v(x) = \Pr(v = x)$.
- K hodnotám rozdělení opět konvergují frekvence v kontingenční tabulce:

p↓ u→	splácí	problémy	nesplácí	$\sum$
vysoké	2	1	0	3
střední	2	2	2	6
nízké	0	1	1	2
$\sum$	4	4	3	11

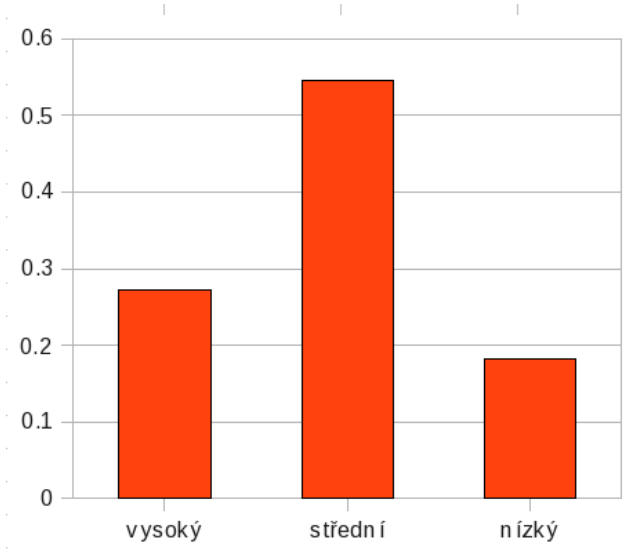
- P_p, P_u: **marginální** rozdělení p (příjmy) resp. u (splácení úvěru)
  - např. $P_p(\text{střední}) \approx 6/11$, $P_u(\text{problémy}) \approx 4/11$
- $P_{p,u}$: **sdružené** rozdělení p a u
  - např. $P_{p,u}(\text{střední}, \text{splácí}) \approx 2/11$
- $P_{p,u}$: **podmíněné** rozdělení p a u
  - např. $P_{p|u}(\text{střední}|\text{splácí}) \approx 2/4$

Histogramy

p↓ u→	splácí	problémy	nesplácí	$\sum$
vysoké	2	1	0	**3**
střední	2	2	2	**6**
nízké	0	1	1	**2**
$\sum$	**4**	**4**	**3**	11



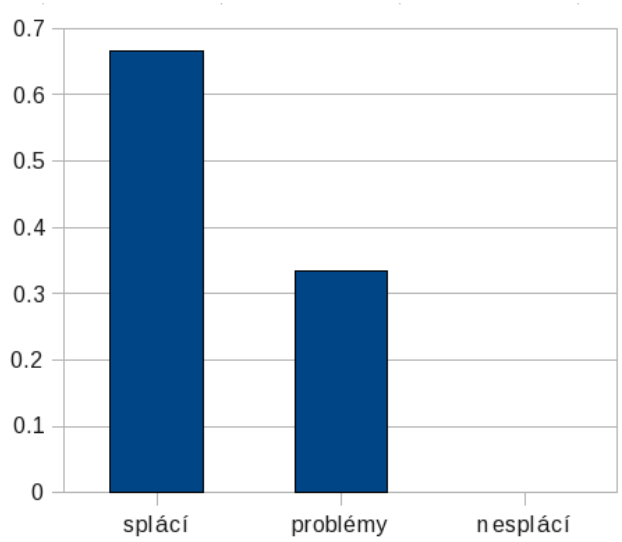
$P_u(x)$



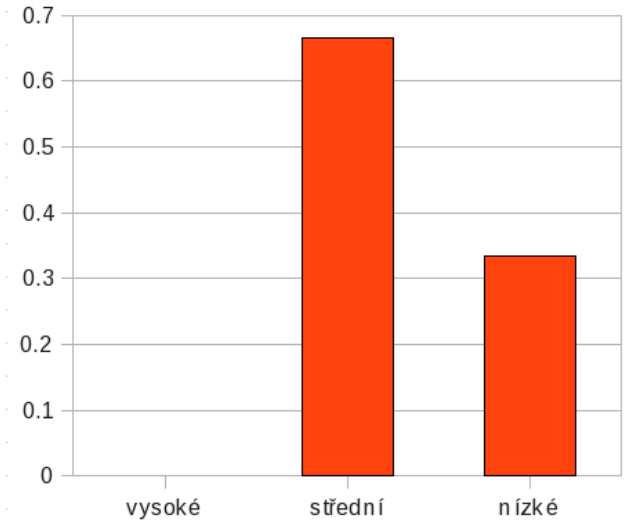
$P_p(x)$

Histogramy

p↓ u→	splácí	problémy	nesplácí	$\sum$
vysoké	**2**	**1**	**0** **0**	3
střední	2	2	**2**	6
nízké	0	1	**1**	2
$\sum$	4	4	3	11



$P_{u|p}(x|\text{vysoké})$



$P_{p|u}(x|\text{nesplácí})$

Součty rozdělení

Vždy platí

$$\sum_x P_v(x) = 1$$

Sčítáme přes všechny hodnoty x, kterých může n.v. v nabývat.

Např. $P_u(\text{splácí}) + P_u(\text{problémy}) + P_u(\text{nesplácí}) = 4/11 + 4/11 + 3/11 = 1$

Součty rozdělení

Vždy platí

$$\sum_x P_v(x) = 1$$

Sčítáme přes všechny hodnoty x, kterých může n.v. v nabývat.

Např. $P_u(\text{splácí}) + P_u(\text{problémy}) + P_u(\text{nesplácí}) = 4/11 + 4/11 + 3/11 = 1$

Analogicky pro podmíněné rozdělení

$$\sum_x P_{v|w}(x|y) = 1$$

Pro jakoukoliv hodnotu y n.v. w.

Např. $P_{u|v}(\text{splácí}|\text{nízké}) + P_{u|v}(\text{problémy}|\text{nízké}) + P_{u|v}(\text{nesplácí}|\text{nízké})$
$= 0/2 + 1/2 + 1/2 = 1$

Klasifikace dle aposteriorní pravděpodobnosti

- Klient má vysoké příjmy. Jak bude splácet úvěr?
- Tedy z $p =$ vysoké urči nejpravděpodobnější hodnotu u (třídu).
- Hledáme y^* vyhovující

$$y^* = \arg\max_y P_{u|p}(y|\text{vysoké})$$

- Řešení je $y^* =$ splácí

$$P_{u|p}(y^*|\text{vysoké}) \approx 2/3$$

- S pravděpodobností $1 - P_{u|p}(y^*|\text{vysoké})$ klasifikujeme chybně.
- Klasifikací $y^* =$ splácí tedy minimalizujeme chybu.
- 'Kritérium minimální chyby' (UI6)

Ztrátová funkce

- Každá chyba klasifikace je jinak drahá.
- Např. příliš optimististické hodnocení klienta stojí víc než příliš skeptické.
- Klasifikace dle aposteriorní pravděpodobnosti toto nerespektuje.
- **Ztrátová funkce** $L(u, y)$ zachycuje ztrátu pro každou kombinaci
  - u - skutečná třída
  - y - třída, do které klasifikujeme
- Pro náš příklad $L(u, y)$ např.:

u↓ y⟶	splácí	problémy	nesplácí
splácí	0	1	2
problémy	5	0	1
nesplácí	10	5	0

Střední hodnota

Definujeme **střední hodnotu** číselné n.v. v

$$\sum_x x \cdot P_v(x)$$

Sčítáme přes všechny hodnoty x, kterých může n.v. v nabývat.

Střední hodnota

Definujeme **střední hodnotu** číselné n.v. v

$$\sum_x x \cdot P_v(x)$$

Sčítáme přes všechny hodnoty x, kterých může n.v. v nabývat.

Analogicky **střední hodnotu podmíněnou** $w = y$

$$\sum_x x \cdot P_{v|w}(x|y)$$

Pro jakoukoliv hodnotu y n.v. w

Riziko

- Ztráta $L(u, y)$ je n.v., protože její argumenty jsou n.v.

Riziko klasifikace do y za podmínky $p = x$

$$r_{u|p}(y, x) = \sum_t L(t, y) P_{u|p}(t|x)$$

Riziko je tedy střední hodnota ztráty podmíněná $p = x$.

Klasifikace jako minimalizace rizika

- Jak klasifikovat vysokopříjmového klienta?

$P_{u|p}$

p↓ u→	splácí	problémy	nesplácí
vysoké	**2/3**	**1/3**	**0/3**
střední	2/6	2/6	2/6
nízké	0/2	1/2	1/2

L

u↓ y→	splácí	problémy	nesplácí
splácí	**0**	1	2
problémy	**5**	0	1
nesplácí	**10**	5	0

- Klasifikace $y =$ splácí

skut. třída	**ztráta**	**s pravděp.**
splácí	0	2/3
problémy	5	1/3
nesplácí	10	0

- Riziko při této klasifikaci:

$$0 \cdot 2/3 + 5 \cdot 1/3 + 10 \cdot 0 = 5/3$$

Klasifikace jako minimalizace rizika

- Jak klasifikovat vysokopříjmového klienta?

$P_{u|p}$

p↓ u→	splácí	problémy	nesplácí
vysoké	**2/3**	**1/3**	**0/3**
střední	2/6	2/6	2/6
nízké	0/2	1/2	1/2

L

u↓ y→	splácí	**problémy**	nesplácí
splácí	0	**1**	2
problémy	5	**0**	1
nesplácí	10	**5**	0

- Klasifikace do $y =$ problémy

skut. třída	**ztráta**	**s pravděp.**
splácí	1	2/3
problémy	0	1/3
nesplácí	5	0

- Riziko při této klasifikaci:

$$1 \cdot 2/3 + 0 \cdot 1/3 + 5 \cdot 0 = 2/3$$

Klasifikace jako minimalizace rizika

- Jak klasifikovat vysokopříjmového klienta?

$P_{u|p}$

p↓ u→	splácí	problémy	nesplácí
vysoké	**2/3**	**1/3**	**0/3**
střední	2/6	2/6	2/6
nízké	0/2	1/2	1/2

L

u↓ y→	splácí	problémy	nesplácí
splácí	0	2	**2**
problémy	5	1	**1**
nesplácí	10	0	**0**

- Klasifikace $y =$ nesplácí

skut. třída	**ztráta**	**s pravděp.**
splácí	2	2/3
problémy	1	1/3
nesplácí	0	0

- Riziko při této klasifikaci:

$$2 \cdot 2/3 + 1 \cdot 1/3 + 0 \cdot 0 = 4/3$$

Klasifikace jako minimalizace rizika

- Klasifikujeme do

$$y^* = \arg\min_y r_{u|p}(y, \text{vysoké}) = \text{problémy}$$

Klasifikace jako minimalizace rizika

- Klasifikujeme do

$$y^* = \arg\min_y r_{u|p}(y, \text{vysoké}) = \text{problémy}$$

- Pozor, jiný výsledek než dle aposteriorní pravděpodobnosti

$$y^* = \arg\max_y P_{u|p}(y|\text{vysoké}) = \text{splácí}$$

- Při jaké ztrátové funkci $L(u, y)$ by výsledky vyšly stejně?

Klasifikace jako minimalizace rizika

- Klasifikujeme do

$$y^* = \arg\min_y r_{u|p}(y, \text{vysoké}) = \text{problémy}$$

- Pozor, jiný výsledek než dle aposteriorní pravděpodobnosti

$$y^* = \arg\max_y P_{u|p}(y|\text{vysoké}) = \text{splácí}$$

- Při jaké ztrátové funkci $L(u, y)$ by výsledky vyšly stejně?

u↓ y→	splácí	problémy	nesplácí
splácí	0	1	1
problémy	1	0	1
nesplácí	1	1	0

- Tzv. L_{01} ztrátová funkce. Je-li použita, je $r_{u|p}(y, x)$ pravděpodobnost chybné klasifikace instance s příznakem x.

Klasifikace s několika příznaky

- Zatím jsme klasifikovali pouze dle jediného příznaku (p - výše příjmů)
- O klientech toho víme obvykle více.

Příjmy (p)	**Rok narození (n)**	**Úvěr (u)**
vysoké	1969	splácí
nízké	1974	nesplácí
střední	1940	problémy
nízké	1985	problémy
...	...	...

- Třírozměrná kontingenční tabulka
  - p vs. n vs. u

Klasifikace s několika příznaky

- Na principech klasifikace se nic nemění. Např. jak klasifikovat nízkopříjmového klienta narozeného v r. 1985?
  - Maximalizací aposteriorní pravděpodobnosti

$$y^* = \arg\max_y P_{u|p,n}(y|\text{nízké}, 1974)$$

  - Minimalizací rizika

$$y^* = \arg\min_y r_{u|p,n}(y, \text{nízké}, 1974)$$

- Z kontingenční tabulky

$$P_{u|p,n}(y|\text{nízké}, 1974) \approx \frac{\text{počet klientů s } u = y, p = \text{nízké}, n = 1974}{\text{počet klientů s } p = \text{nízké}, n = 1974}$$

Prokletí rozměrnosti

$$P_{u|p,n}(y|\text{nízké}, 1974) \approx \frac{\text{počet klientů s } u = y, p = \text{nízké}, n = 1974}{\text{počet klientů s } p = \text{nízké}, n = 1974}$$

- Čím více příznaků, tím větší nebezpečí výsledku "0/0"!
- Kolik dat potřebujeme, aby odhady dobře konvergovaly k pravděpodobnostem?
- Kontingenční tabulka musí být 'dostatečně zaplněna'.
- V předchozím příkladě

p↓ u→	splácí	problémy	nesplácí	$\sum$
vysoké	2	1	0	3
střední	2	2	2	6
nízké	0	1	1	2
$\sum$	4	4	3	11

v průměru 11/9 případů na kolonku tabulky.

Prokletí rozměrnosti

- Předpokládejme, že 11/9 je dostatečný poměr. Kolik případů (m) potřebujeme pro jeho zachování se dvěma příznaky p a n?
- Přepodkládejme 100 možných roků narození. Kontingenční tabulka má $100 \cdot 3 \cdot 3 = 900$ kolonek.

$$\frac{m}{900} = \frac{11}{9}$$

  tedy nyní již potřebujeme $11 \cdot 900/9 = 1100$ dat.
- Po přidání dalšího příznaku, např. roku ukončení studia už potřebujeme $11 \cdot 90000/9 = 110000$ dat!
- Obecně pro odhady z kontingenční tabulky (tzv. neparametrické odhady) roste potřebný počet dat **exponenciálně** s počtem příznaků.
  - "Prokletí rozměrnosti"

Podmíněně nezávislé příznaky

- Situace se zjednoduší, jsou-li výše příjmů a rok narození **podmíněně nezávislé**, tj. platí

$$P_{p,n|u}(x, x'|y) = P_{p|u}(x|y) \cdot P_{n|u}(x'|y)$$

pro každou z hodnot $y \in \{\text{splácí}, \text{problémy}, \text{nesplácí})$

- Využijeme tzv. Bayesova pravidla

$$P_{u|p,n}(y|x, x') = \frac{P_{p,n|u}(x, x'|y)P_u(y)}{P_{p,n}(x, x')}$$

Podmíněně nezávislé příznaky

- Z Bayesova pravidla platí pro klasifikaci maximalizací aposteriorní pravděpodobnosti

$$\arg\max_y P_{u|p,n}(y|x,x') = \arg\max_y P_{p,n|u}(x,x'|y)P_u(y)$$

- Podobně pro klasifikaci minimalizací rizika

$$\arg\min_y \sum_t L(t,y)P_{u|p,n}(t|x,x')$$

$$= \arg\min_y \sum_t L(t,y)P_{p,n|u}(x,x'|y)P_u(y)$$

- Proč 'zmizelo' $P_{p,n}(x,x')$?

Podmíněně nezávislé příznaky

- Z Bayesova pravidla platí pro klasifikaci maximalizací aposteriorní pravděpodobnosti

$$\arg\max_y P_{u|p,n}(y|x,x') = \arg\max_y P_{p,n|u}(x,x'|y)P_u(y)$$

- Podobně pro klasifikaci minimalizací rizika

$$\arg\min_y \sum_t L(t,y)P_{u|p,n}(t|x,x')$$

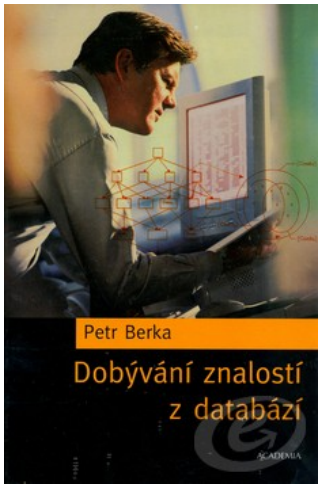
$$= \arg\min_y \sum_t L(t,y)P_{p,n|u}(x,x'|y)P_u(y)$$

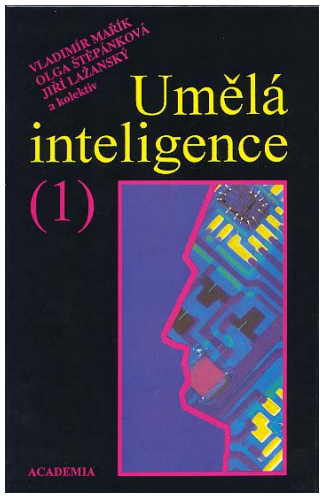
- Proč 'zmizelo' $P_{p,n}(x,x')$? Nezávisí na y!
- K oběma typům klasifikace tedy potřebujeme odhady dvou rozdělení: $P_{p,n|u}$ a P_u.

Podmíněně nezávislé příznaky

- K oběma typům klasifikace tedy potřebujeme odhady dvou rozdělení: $P_{p,n|u}$ a P_u.
- P_u odhadneme z jednorozměrné kontingenční tabulky
- Z podmíněné nezávislosti plyne $P_{p,n|u}(x, x'|y) = P_{p|u}(x|y) \cdot P_{n|u}(x'|y)$
- $P_{p|u}$ odhadneme z dvourozměrné tabulky 3×3
- $P_{n|u}$ odhadneme z dvourozměrné tabulky 100×3.
  - Nejnáročnější na počet dat, pro zachování poměru 11/9 vyžaduje cca 367 ($<<$ 1100).
- Obecně: při podmíněně nezávislých příznacích neroste potřebný počet dat exponenciálně s počtem příznaků.
  - Je určen příznakem s největším oborem hodnot.
- Příznaky obvykle nejsou podmíněně nezávislé!
  - Je-li přesto použita tato metoda, mluvíme o **naivní Bayesovské klasifikaci**.

Literatura k této přednášce





- Pokrývá: kontingenční tabulky
- Navíc: χ^2 test nezávislosti
- Pokrývá: klasifikaci dle aposteriorní pravděpodobnosti a dle etalonů
- Navíc: pojem diskriminační funkce
- Nepokryto: klasifikace minimalizací ztráty, podmíněná nezávislost příznaků
- Je třeba pochopit tyto slidy, tedy chodit na přednášky.