Intuitivní úvod do pravděpodobnosti

Filip Železný

Katedra kybernetiky
skupina Inteligentní Datové Analýzy (IDA)

Evropský sociální fond Praha & EU:
Investujeme do vaší budoucnosti

Intuitivní úvod do pravděpodobnosti

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

Intuitivní úvod do pravděpodobnosti

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

Nový žadatel o úvěr, má vysoké příjmy.

- Bude splácet úvěr?
- S jakou pravděpodobností?

Frekvence a pravděpodobnost

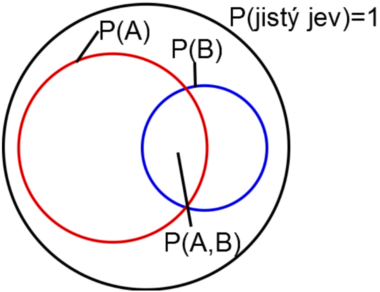
	S	$\neg S$	$\sum$
V	a	b	r
$\neg V$	c	d	s
$\sum$	k	l	n

- 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

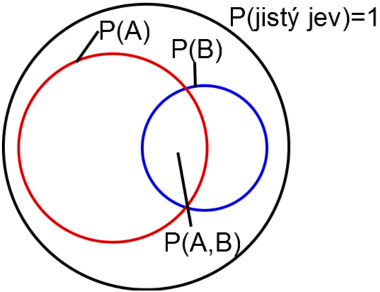
- Zřejmé z geometrické představy



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

Základní vlastnosti pravděpodobnosti

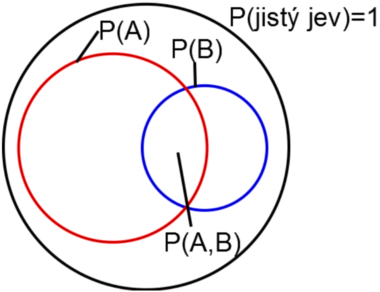
- Zřejmé z geometrické představy



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

Základní vlastnosti pravděpodobnosti

- Zřejmé z geometrické představy

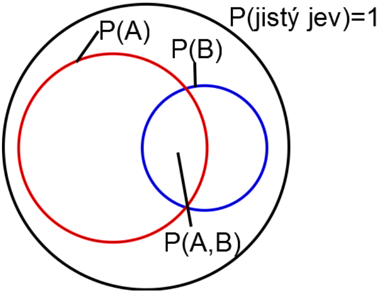


- $\Pr(A|B) =$

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

Základní vlastnosti pravděpodobnosti

- Zřejmé z geometrické představy

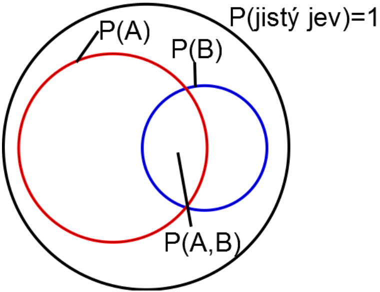


- $\Pr(A|B) = \Pr(A, B) / \Pr(B)$

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

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

- $\Pr(V,S) \approx 5/11 = 0.45\ldots$
- $\Pr(V)\cdot\Pr(S) \approx 6/11 \cdot 7/11 = 0.34...$
- 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.

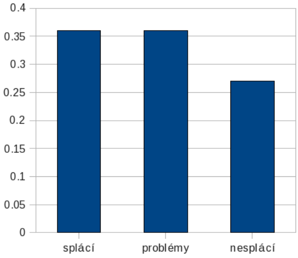
- Rozdělení n.v. v je funkce $P_v(x) = \Pr(v = x)$.
- K hodnotám rozdělení opět konvergují frekvence v kontingenční tabulce:

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

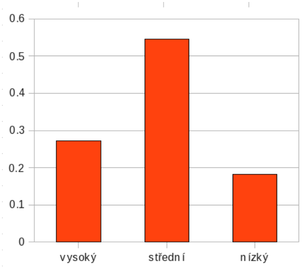
- P_p, P_u: **marginální** rozdělení p (příjmy) resp. u (splácení úvěru)
  - např. $P_p(\text{střední}) \approx 6/11$, $P_u(\text{problémy}) \approx 4/11$
- $P_{p,u}$: **sdružené** rozdělení p a u
  - např. $P_{p,u}(\text{střední}, \text{splácí}) \approx 2/11$
- $P_{p,u}$: **podmíněné** rozdělení p a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



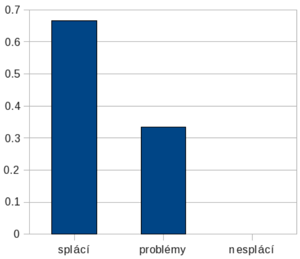
$P_u(x)$



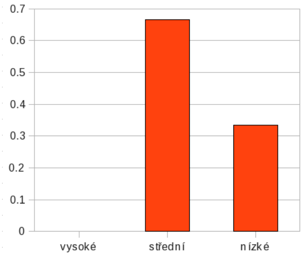
$P_p(x)$

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

Spojitá náhodná veličina

- Nabývá hodnot z R. Její rozdělení pravděpodobnosti je dáno *hustotou*.
- Hustota spojité n.v. X: $f(x)$ taková, že platí

$$\Pr(a \leq X < b) = \int_a^b f(x)dx$$

- Tedy $\Pr(a \leq X \leq b) =$ plocha pod grafem $f(x)$ mezi a a b.
- Proč ne jednoduše $f(x) \equiv \Pr(X = x)$ jako u diskrétní?

Spojitá náhodná veličina

- Nabývá hodnot z R. Její rozdělení pravděpodobnosti je dáno *hustotou*.
- Hustota spojité n.v. X: $f(x)$ taková, že platí

$$\Pr(a \leq X < b) = \int_a^b f(x)dx$$

- Tedy $\Pr(a \leq X \leq b) =$ plocha pod grafem $f(x)$ mezi a a b.
- Proč ne jednoduše $f(x) \equiv \Pr(X = x)$ jako u diskrétní?
- $\Pr(X = x) = 0$ pro jakékoliv x! (výběr z ∞ množství hodnot!)

Binomiální a normální rozdělení

- *Binomiální* rozdělení diskrétní n.v.:

$$P(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}$$

$P(x) =$ pravděpodobnost x orlů při n hodech mincí, kde $\Pr(\mathrm{orel}) = p$

- *Normální* hustota spojité n.v.:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

parametry: μ - střed, σ^2 - rozptyl (rozpětí "zvonu") příklad: obvyklé rozložení chyb měření kolem skutečné hodnoty μ.

Binomiální a normální rozdělení (pokr.)