



EVROPSKÁ UNIE  
Evropské strukturální a investiční fondy  
Operační program Výzkum, vývoj a vzdělávání



MINISTERSTVO ŠKOLSTVÍ,  
MLÁDEŽE A TĚLOVÝCHOVY

Název projektu	Rozvoj vzdělávání na Slezské univerzitě v Opavě
Registrační číslo projektu	CZ.02.2.69/0.0./0.0/16_015/0002400

**Dolování dat**

**Bayesovská klasifikace**

**Jan Górecki**



**SLEZSKÁ  
UNIVERZITA**  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

# Obsah přednášky

---



**SLEZSKÁ  
UNIVERZITA**  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

- Podmíněná pravděpodobnost
- Bayesova věta
- Použití pro klasifikaci
- Příklad
- Naivní bayesovský klasifikátor
- Příklad
- Shrnutí



# Podmíněná pravděpodobnost

- **Zajímáme se o situace, kdy nastane jev  $H$  za **PODMÍNKY**, že nastal jev  $E$** 
  - např.  $H$  = splatil,  $E$  = příjem(vysoký)
  - podmíněná pravděpodobnost jevu  $H$  vzhledem k jevu  $E$  se značí  $P(H|E)$
  - *průnik jevů  $H$  a  $E$*  je současný výskyt jevů  $H$  a  $E$ , značí se  $H \cap E$ , pravděpodobnost  $P(H \cap E)$
  - podmíněná pravděpodobnost je definována:

$$P(H|E) = \frac{P(H \cap E)}{P(E)}$$

## Příklad:

$H$  – klient splatil úvěr

$E$  – klient má vysoký příjem

Víme, že osm ze sta klientů zároveň splatilo *úvěr* a má *vysoký příjem*. Dále víme, že každý *desátý* klient ze sta má *vysoký příjem*. Jaká je pravděpodobnost, že klient s vysokým příjmem splatí úvěr?

- Pokud  $P(H|E) = P(H)$ , potom je jev  $H$  **nezávislý** na jevu  $E$ .
-

# Bayesova věta

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Např. (bankovní úloha)  
 $H$  = splatil úvěr  
 $E$  = příjem(vysoký)

$P(H|E)$  - *Aposteriorní* pravděpodobnost hypotézy  $H$  při pozorování evidence  $E$  - vyjadřuje, jak se změní pravděpodobnost hypotézy, pokud víme, že nastalo  $E$ .

$P(H)$  - *Apriorní* pravděpodobnost hypotézy  $H$  - odpovídá znalostem o zastoupení jednotlivých hypotéz (tříd) bez ohledu na nějaké další informace.

$P(E|H)$  - podmíněná pravděpodobnost evidence  $E$  při pozorování hypotézy  $H$  – pravděpodobnost  $E$  ve třídě  $H$

$P(E)$  - pravděpodobnost evidence (pozorování).

Při více ( $T$ ) hypotézách  $H_1, \dots, H_T$  (např.  $H_1 = \text{splatil}$ ,  $H_2 = \text{nesplatil}$ )

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \longrightarrow P(H_t|E) = \frac{P(E|H_t)P(H_t)}{\sum_{i=1}^T P(E|H_i)P(H_i)}$$

Hledáme hypotézu s maximální aposteriori pravděpodobností ( $H_{\text{MAP}}$ ), tedy:

$$H_{\text{MAP}} = H_J \text{ právě když } P(H_J|E) = \max_t \frac{P(E|H_t)P(H_t)}{\sum_{i=1}^T P(E|H_i)P(H_i)}$$

Hledání lze zjednodušit:

$$H_{\text{MAP}} = H_J \text{ právě když } P(H_J|E) = \max_t P(E|H_t)P(H_t)$$

---

- Pokud bychom měli k dispozici skutečné rozdělení veličin  $P(E|H_t)$  a  $P(H_t)$ , potom bayesovská klasifikace je **optimální** vzhledem ke **skutečné chybě** klasifikace (nelze zkonstruovat lepší klasifikátor než právě ten bayesovský)
  - zjednodušená verze důkazu zde:  
<https://cs.stackexchange.com/questions/72295/showing-that-bayes-classifier-is-optimal>

# Příklad

---



- poskytování úvěru, tentokrát ale pouze na základě výše příjmu
- banka vyhoví u 2/3 žádosti o úvěr; tedy apriorní pravděpodobnosti budou  $P(\text{půjčit})=0.667$  a  $P(\text{nepůjčit})=0.333$
- vysoký příjem mělo 91% klientů, kterým banka půjčila a nízký příjem mělo 88% klientů, kterým banka nepůjčila

$$P(\text{příjem(vysoký)}|\text{půjčit}) = 0.91$$

$$P(\text{příjem(vysoký)}|\text{nepůjčit}) = 0.12$$

$$P(\text{příjem(nízký)}|\text{půjčit}) = 0.09$$

$$P(\text{příjem(nízký)}|\text{nepůjčit}) = 0.88$$

---

# Příklad

---



- Předpokládejme, že posuzujeme klienta s vysokým příjmem.
- Bude větší pravděpodobnost, že banka půjčí nebo že nepůjčí?

Podle Bayesovy věty:

$$P(\text{příjem(vysoký)}|\text{půjčit}) \times P(\text{půjčit}) = 0.607$$

$$P(\text{příjem(vysoký)}|\text{nepůjčit}) \times P(\text{nepůjčit}) = 0.040$$

Tedy  $H_{\text{MAP}} = \text{půjčit}$ .

---



# Naivní bayesovský klasifikátor

- vychází z předpokladu, že jednotlivé evidence  $E_1, \dots, E_d$  jsou **podmíněně nezávislé** při platnosti hypotézy  $H$
- zjednodušující předpoklad umožňuje spočítat aposteriorní pravděpodobnost hypotézy při platnosti všech evidencí

$$P(H|E_1, \dots, E_d) = \frac{P(E_1, \dots, E_d|H) \times P(H)}{P(E_1, \dots, E_d)}$$

jako

$$P(H|E_1, \dots, E_d) = \frac{P(H)}{P(E_1, \dots, E_d)} \times \prod_{i=1}^d P(E_i|H)$$

Např. (bankovní úloha)  
 $H$  = půjčit  
 $E_1$  = příjem(vysoký)  
 $E_2$  = konto(střední)

Opět hledáme hypotézu s největší aposteriorní pravděpodobností  $H_{\text{MAP}}$

$$H_{\text{MAP}} = H_J \text{ právě když } P(H_J|E_1, \dots, E_d) = \max_t P(H_t) \times \prod_{i=1}^d P(E_i|H_t)$$

Evidence  $E_i$  jsou pak hodnoty jednotlivých vstupních atributů (tedy  $E_i = A_j(v_i)$ ) a hypotézy  $H_t$  jsou cílové třídy (tedy  $H_t = C(v_t)$ ).

Např. (bankovní úloha)  
 $A_j(v_i) = \text{příjem}(\text{vysoký})$   
 $C(v_t) = \text{úvěr}(\text{půjčit})$

---

# Naivní bayesovský klasifikátor

---



- Neprovádí prohledávání prostoru hypotéz.
- Stačí jen spočítat příslušné pravděpodobnosti na základě četnosti výskytů hodnot jednotlivých atributů.

$$P(H_t) = P(C(v_t)) = \frac{n_t}{n}$$
$$P(E_i|H_t) = P(A_j(v_i) | C(v_t)) = \frac{n_t(A_j(v_i))}{n_t}$$

---

# Příklad



**SLEZSKÁ  
UNIVERZITA**  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVÍNĚ

klient	příjem	konto	pohlaví	nezaměstnaný	úvěr
k1	vysoký	vysoké	žena	ne	ano
k2	vysoký	vysoké	muž	ne	ano
k3	nízký	nízké	muž	ne	ne
k4	nízký	vysoké	žena	ano	ano
k5	nízký	vysoké	muž	ano	ano
k6	nízký	nízké	žena	ano	ne
k7	vysoký	nízké	muž	ne	ano
k8	vysoký	nízké	žena	ano	ano
k9	nízký	střední	muž	ano	ne
k10	vysoký	střední	žena	ne	ano
k11	nízký	střední	žena	ano	ne
k12	nízký	střední	muž	ne	ano

$$P(\text{úvěr(ano)}) = 8/12 = 0.667$$

$$P(\text{úvěr(ne)}) = 4/12 = 0.333$$

$$P(\text{konto(střední)}|\text{úvěr(ano)}) = 2/8 = 0.25$$

$$P(\text{konto(střední)}|\text{úvěr(ne)}) = 2/4 = 0.5$$

$$P(\text{nezaměstnaný(ne)}|\text{úvěr(ano)}) = 5/8 = 0.625$$

$$P(\text{nezaměstnaný(ne)}|\text{úvěr(ne)}) = 1/4 = 0.25$$

Uchazeč o úvěr, který má střední konto a není nezaměstnaný:

$$P(\text{úvěr}(ano)) * P(\text{konto}(\text{střední})|\text{úvěr}(ano)) * P(\text{nezaměstnaný}(ne) | \text{úvěr}(ano)) = 0.1042$$

$$P(\text{úvěr}(ne)) * P(\text{konto}(\text{střední})|\text{úvěr}(ne)) * P(\text{nezaměstnaný}(ne)|\text{úvěr}(ne)) = 0.0416$$

Naivní bayesovský klasifikátor zařadí tohoto uchazeče do třídy *úvěr(ano)*.

- Klasifikovali jsme *neúplně popsaný* případ, který by zůstal nezařazen dříve vytvořenými rozhodovacími stromy nebo neuronovými sítěmi.
-

# Příklad

---



Pro ženu, která má nízký příjem, střední konto a není nezaměstnaná:

$$P(\text{úvěr(ano)}) P(\text{příjem(nízký)}|\text{úvěr(ano)}) P(\text{konto(střední)}|\text{úvěr(ano)}) \\ P(\text{pohlaví(žena)}|\text{úvěr(ano)}) P(\text{nezaměstnaný(ne)}|\text{úvěr(ano)}) = 0.0195$$

$$P(\text{úvěr(ne)}) P(\text{příjem(nízký)}|\text{úvěr(ne)}) P(\text{konto(střední)}|\text{úvěr(ne)}) \\ P(\text{pohlaví(žena)}|\text{úvěr(ne)}) P(\text{nezaměstnaný(ne)}|\text{úvěr(ne)}) = 0.0208$$

Uchazečka bude zařazena do třídy *úvěr(ne)*.

- Tento příklad nebyl v trénovacích datech, klasifikátor má tedy *schopnost generalizovat*
-

- Schopnost klasifikovat příklady do tříd s určitou pravděpodobností.
  - Tuto pravděpodobnost můžeme interpretovat jako spolehlivost rozhodnutí.
  - I přes naivitu je Naivní Bayes překvapivě přesný
  - Lze jednoduše naimplementovat (naprogramovat)
  - Pro nezkušené uživatele však může být reprezentace znalostí ve formě (podmíněných) pravděpodobností méně srozumitelná
-

# Děkuji za pozornost

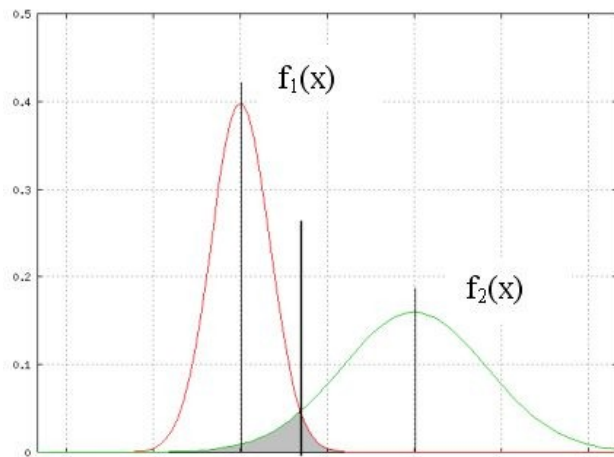
Některé snímky převzaty od:  
prof. Ing. Petr Berka, CSc. [berka@vse.cz](mailto:berka@vse.cz)



# Pokud $E$ je nekonečně mnoho

- Pokud  $E$  neodpovídá kategoriálnímu, ale numerickému atributu, např.  $E \in \mathbf{R} \rightarrow$  Diskriminační analýza ( $E = x$ )

$$P(H_j|E) = \max_t P(E|H_t)P(H_t) \rightarrow \max_t f(x|H_t)P(H_t)$$



příjem	uver
5000	ano
10000	ano
10000	ano
15000	ano
17000	ano
17000	ano
20000	ano
20000	ano
2000	ne
3000	ne
5000	ne
10000	ne