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 KARVINÉ

Obsah přednášky

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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

Podmíněná pravděpodobnost

SLEZSKÁ

- **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

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

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).

Použití pro klasifikaci

Při více (T) hypotézách $H_1, \ldots, H_T$ (např. H_1 = splatil, H_2 = 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 aposteriorní pravděpodobností (H_{MAP}), tedy:

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

Hledání lze zjednodušit:

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

Optimalita bayesovské klasifikace vzhledem ke správnosti klasifikace

- 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

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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

P(příjem(vysoký)|půjčit) = 0.91

P(příjem(vysoký)|nepůjčit) = 0.12

P(příjem(nízký)|půjčit) = 0.09

P(příjem(nízký)|nepůjčit) = 0.88

Příklad

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

Podle Bayesovy věty:

P(příjem(vysoký)|půjčit) × P(půjčit) = 0.607
P(příjem(vysoký)|nepůjčit) × P(nepůjčit) = 0.040

Tedy H_{MAP} = **půjčit**.

Naivní bayesovský klasifikátor

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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

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

jako

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

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

Naivní bayesovský klasifikátor

Opět hledáme hypotézu s největší aposteriorní pravděpodobností H_{MAP}

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

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)
$A_j(v_i)$ = příjem(vysoký)
$C(v_t)$ = úvěr(půjčit)

Naivní bayesovský klasifikátor

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

$$\mathrm{P}(H_t) = \mathrm{P}(C(v_t)) = \frac{n_t}{n}$$

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

Příklad

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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(úvěr(ano)) = 8/12 = 0.667

P(úvěr(ne)) = 4/12 = 0.333

P(konto(střední)|úvěr(ano)) = 2/8 = 0.25

P(konto(střední)|úvěr(ne)) = 2/4 = 0.5

P(nezaměstnaný(ne)|úvěr(ano)) = 5/8 = 0.625

P(nezaměstnaný(ne)|úvěr(ne)) = 1/4= 0.25

Příklad

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

P(úvěr(ano)) * P(konto(střední)|úvěr(ano)) * P(nezaměstnaný(ne) |úvěr(ano)) = 0.1042

P(úvěr(ne)) * P(konto(střední)|úvěr(ne)) * P(nezaměstnaný(ne)|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(úvěr(ano)) P(příjem(nízký)|úvěr(ano)) P(konto(střední)|úvěr(ano))
P(pohlaví(žena)|úvěr(ano)) P(nezaměstnaný(ne)|úvěr(ano)) = 0.0195

P(úvěr(ne)) P(příjem(nízký)|úvěr(ne)) P(konto(střední)|úvěr(ne))
P(pohlaví(žena)|úvěr(ne)) P(nezaměstnaný(ne)|úvěr(ne)) = 0.0208

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

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

Shrnutí

- 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

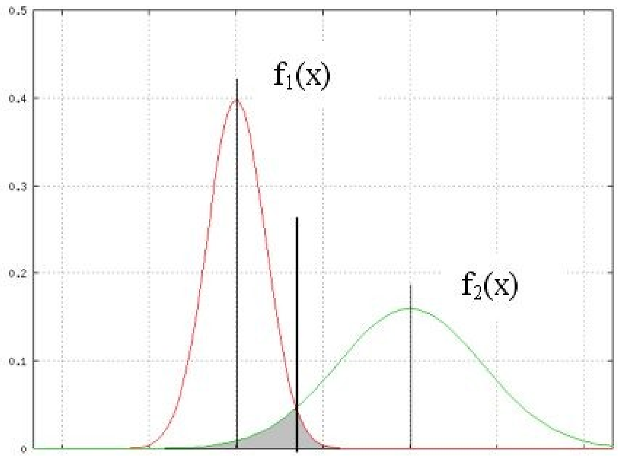
Některé snímky převzaty od:

prof. Ing. Petr Berka, CSc. berka@vse.cz

Pokud *E* je nekonečně mnoho

- Pokud *E* neodpovídá kategoriálnímu, ale numerickému atributu, např. $E \in \mathbb{R}$ -> 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