

PLIN021 Sémantická analýza v praxi

OP VK Mezi bohemistikou a informatikou
www.projekt-inova.cz

Zuzana Nevěřilová
xpopelk@fi.muni.cz

Centrum zpracování přirozeného jazyka, B203
Fakulta informatiky, Masarykova univerzita

11. března 2013

Algoritmy strojového učení

Pravidlové algoritmy ML

Matematické algoritmy ML

„Promluvné“ systémy ML

Word Sense Desambiguation

úkolem WSD je zjistit, jaký význam (z inventáře významů) má slovo ve vstupním textu

minule jsme mluvili o metodách založených na znalostech (Leskův algoritmus pracující se slovníkovými definicemi a příklady užití)

Algoritmy strojového učení

Strojové učení (machine learning, ML) = algoritmy a techniky, které způsobí změnu stavu počítačového systému tak, že zefektivní schopnost přizpůsobení se . . .

Učím se, že pokud je blízko „kočka“ i „pes“, má „kočka“ význam 1. . .

- s učitelem – pro zadaný vstup máme i správný výstup (trénovací data)
- bez učitele – pro zadaný vstup neznáme správný výstup
- kombinace – pro část vstupu máme i správný výstup

Typické úlohy strojového učení jsou **klasifikační úlohy**.

Algoritmy strojového učení

- „pravidlové“
 - rozhodovací seznamy
 - rozhodovací stromy
- „matematické“
 - pravděpodobnostní: naivní Bayesovský (Duda et Hart, 1973)
 - maximální entropie: (Berger, 1996)
 - podobnostní: k-NN ve vektorovém prostoru (Ng et Lee, 1996)
- „promluvané“
 - předpoklad *one sense per discourse* (Gale 1992)
 - předpoklad *one sense per collocation* (Yarowsky, 1995)

Rozhodovací seznam

```
if (zvíře má chobot) then output(slon)
if (zvíře má pruhy) then output(zebra)
if (zvíře má ploutve & zvíře není ryba) then
output(žralok)
```

Rozhodovací strom

savec?

žije ve vodě?

žije na souši?

žije v moři?

žije v řece?

býložravec?

masožravec?



Seznam je jednodušší na implementaci, ale vidíme, že strom je přehlednější při stejné i vyšší složitosti.

Častěji pracujeme se stromy.



V této hře jsou Myslím si zvíře aspekty:

- Jak poznám z množiny otázek $\mathcal{O} = \{o_1, o_2, \dots, o_n\}$, kde o_i je např. „Má zvíře srst?“, o jaké zvíře jde? Redukcí. Pokud odpověď na o_i je „ne“, vyloučím ze správných odpovědí všechna zvířata z_j , která mají srst. Podobně pro další otázky, dokud nezůstane (ideálně) 1 zvíře.
- Jaká je strategie kladení otázek? Cílem je minimalizovat n . Prostředkem k dosažení tohoto cíle je neklást otázky, které dělí množinu možných zvířat stejným způsobem. Např. otázky „Má zvíře srst?“ a „Má zvíře 4 nohy?“ dělí \mathcal{Z} na dvě téměř stejné části.

Na celou hru můžeme pohlížet jako na množinu zvířat (která známe) a rozhodovací strom, který nás „dovede“ k myšlenému zvířeti.

Algoritmy strojového učení

- „pravidlové“
 - rozhodovací seznamy
 - rozhodovací stromy
- „matematické“
 - pravděpodobnostní: naivní Bayesovský (Duda et Hart, 1973)
 - maximální entropie: (Berger, 1996)
 - podobnostní: k-NN ve vektorovém prostoru (Ng et Lee, 1996)
- „promluvané“
 - předpoklad *one sense per discourse* (Gale 1992)
 - předpoklad *one sense per collocation* (Yarowsky, 1995)

- „pravidlové“
 - rozhodovací stromy
 - rozhodovací stromy
- „matematické“
 - pravděpodobnosti: naivní Bayesovský (Duda et Hart, 1973)
 - maximální entropie (Baecker, 1996)
 - podobnostní: b-NM ve vektorovém prostoru (Ng et Lee, 1996)
- „proměnlivé“
 - předpoklad *sense per discourse* (Gale 1992)
 - předpoklad *sense per collocation* (Yarowsky, 1995)

„Matematické“ algoritmy zde uvedené jsou každý úplně jiný, spíš jde o reprezentanty různých skupin algoritmů.

Naivní Bayesovský klasifikátor

Naivní Bayesovský alg. předpokládá nezávislost znaků (což nemusí být správně), ale je rychlý.

$$P(C|F_1, \dots, F_n) = \frac{P(C) \cdot P(F_1, \dots, F_n|C)}{P(F_1, \dots, F_n)}$$

zvíře	velikost	barva	potrava
slon	velký	šedý	býložravec
slon	střední	šedý	býložravec
kráva	velká	černá	býložravec
kráva	velká	strakatá	býložravec
kráva	malá	strakatá	býložravec
kráva	velká	bílá	býložravec
vlk	velký	černý	masožravec
vlk	malý	šedý	masožravec

Naivní Bayes pro **velkého černého býložravce**

zvíře	velikost	barva	potrava
slon	velký	šedý	býložravec
slon	střední	šedý	býložravec
kráva	velká	černá	býložravec
kráva	velká	strakatá	býložravec
kráva	malá	strakatá	býložravec
kráva	velká	bílá	býložravec
vlk	velký	černý	masožravec
vlk	malý	šedý	masožravec

Na základě těchto dat můžeme vypočítat, že zvíře, které vidíme, bude: na 25 % slon, na 50 % kráva a na 25 % vlk, tj. $P(\text{slon}) = \frac{2}{8}$, $P(\text{kráva}) = \frac{4}{8}$ a $P(\text{vlk}) = \frac{2}{8}$.

Podmíněné pravděpodobnosti jsou $P(\text{černá barva}|\text{slon}) = 0$, $P(\text{černá barva}|\text{kráva}) = \frac{1}{4}$, $P(\text{černá barva}|\text{vlk}) = \frac{1}{2}$.

Naivní Bayes pro **velkého černého býložravce**

$$P(C|F_1, \dots, F_n) = \frac{P(C) \cdot P(F_1, \dots, F_n|C)}{P(F_1, \dots, F_n)}$$

$$P(\text{slon}) = \frac{2}{8}, P(\text{kráva}) = \frac{4}{8}, P(\text{vlk}) = \frac{2}{8}, P(\text{černý}|\text{slon}) = 0,$$

$$P(\text{černý}|\text{kráva}) = \frac{1}{4}, P(\text{černý}|\text{vlk}) = \frac{1}{2},$$

$$P(\text{velký}|\text{slon}) = \frac{1}{2}, P(\text{velký}|\text{kráva}) = \frac{3}{4},$$

$$P(\text{velký}|\text{vlk}) = \frac{1}{2}, P(\text{býložravec}|\text{slon}) = \frac{2}{2},$$

$$P(\text{býložravec}|\text{kráva}) = \frac{4}{4}, P(\text{býložravec}|\text{vlk}) = 0$$

$$P(\text{slon}|\text{velký, černý, býložravec}) = P(\text{slon})P(\text{velký}|\text{slon}) \cdot$$

$$P(\text{černý}|\text{slon}) \cdot P(\text{býložravý}|\text{slon}) = 0.25 \cdot 0.25 \cdot 0 \cdot 1 = 0$$

$$P(\text{kráva}|\text{velký, černý, býložravec}) = P(\text{kráva})P(\text{velký}|\text{kráva}) \cdot$$

$$P(\text{černý}|\text{kráva}) \cdot P(\text{býložravý}|\text{kráva}) = 0.5 \cdot 0.75 \cdot 0.25 \cdot 1 = \mathbf{0.09375}$$

$$P(\text{kráva}|\text{velký, černý,$$

$$\text{býložravec}) = P(\text{vlk})P(\text{velký}|\text{vlk}) \cdot P(\text{černý}|\text{vlk}) \cdot P(\text{býložravý}|\text{vlk}) =$$

Algoritmy strojového učení

- „pravidlové“
 - rozhodovací seznamy
 - rozhodovací stromy
- „matematické“
 - pravděpodobnostní: naivní Bayesovský (Duda et Hart, 1973)
 - maximální entropie: (Berger, 1996)
 - podobnostní: k-NN ve vektorovém prostoru (Ng et Lee, 1996)
- „promluvané“
 - předpoklad *one sense per discourse* (Gale 1992)
 - předpoklad *one sense per collocation* (Yarowsky, 1995)

Algoritmus strojového učení [Yarowsky, 1995]

hledáme význam slova w

- 1. vezmi všechny výskyty slova w z korpusu včetně jejich **kontextů**
- 2. pro každý možný **význam** slova, vytvoř malou sadu příkladů (buď ručně, nebo pomocí kolokací)
- 3. vytvoř **rozhodovací seznam** s pravděpodobnostmi pro další slova, která se vyskytují v kontextech
- 4. aplikuj tento seznam na celý korpus (s prahem pro pravděpodobnost)
- 5. nově zařazená slova obsahují **další slova** v kontextech
- 6. algoritmus můžeme upravit pomocí zařazení předpokladu one-sense-per-discourse
- 7. opakuj kroky 3–6
- 8. jakmile množiny přestanou narůstat, zastav
- 9. systém je nyní natrénovaný i na jiný korpus!

Algoritmus strojového učení

závisí na:

- první volbě kolokací
- způsobu určení pravděpodobnosti: typicky log likelihood
 $\log \frac{P(\text{senseA}, \text{collocateA})}{P(\text{senseB}, \text{collocateA})}$
- prahu pro pravděpodobnost
- správnosti předpokladu one-sense-per-discourse



Yarowsky, D. (1995).

Unsupervised word sense disambiguation rivaling supervised methods.

In Proceedings of the 33rd annual meeting on Association for Computational Linguistics, ACL '95, pages 189–196, Stroudsburg, PA, USA. Association for Computational Linguistics.