

Бајесовски класификатори

Ненад Митић

Математички факултет
`nenad.mitic@matf.bg.ac.rs`

Увод

- Скуп Бајесовских класификатора је заснован на теорији вероватноће, односно Бајесовој теореме одређивања условних вероватноћа
- Условне вероватноће

$$P(C | A) = \frac{P(A, C)}{P(A)}$$

$$P(A | C) = \frac{P(A, C)}{P(C)}$$

- Бајесова теорема:

$$P(C | A) = \frac{P(A | C)P(C)}{P(A)}$$

Пример Бајесове теореме

- Дато је:
 - Доктор зна да менингитис у 50% случајева проузрокује кочење врата
 - Претходна (позната) вероватноћа да било који пацијент има менингитис је $1/50,000$
 - Претходна вероватноћа да било који пацијент има укочен врат је $1/20$
- Ако пацијент има укочен врат, која је вероватноћа да има и менингитис?

$$P(M | S) = \frac{P(S | M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Бајесовски класификатори

- Посматрајмо сваки атрибут и сваку ознаку класе као независне променљиве
- За дати слог са атрибутима $(A_1, A_2, \dots, A_n, C)$ где је C ознака класе
 - Циљ је предвидети класу C којој припада
 - Желимо да нађемо вредност C која максимизира $P(C | A_1, A_2, \dots, A_n)$
- Да ли се $P(C | A_1, A_2, \dots, A_n)$ може проценити директно на основу података?
- У овом начину разматрања проблема основна претпоставка је да су класе међусобно независне. Ако су независне, тада може да се примени Бајесова теорема.

Бајесовски класификатори

- Приступ

- израчунати последичну (енг. *posterior*) вероватноћу $P(C | A_1, A_2, \dots, A_n)$ за сваку класу C користећи Бајесову теорему

$$P(C | A_1, A_2, \dots, A_n) = \frac{P(C) \prod_{i=1}^n P(A_i | C)}{P(A_1, A_2, \dots, A_n)}$$

- Изабрати вредност C која максимизује $P(C | A_1, A_2, \dots, A_n)$
 - Еквивалентно, може да се узме и вредност C која максимизује $P(A_1, A_2, \dots, A_n | C)P(C)$
(услов - атрибути су међусобно независни за сваку од класа C)
- Како проценити $P(A_1, A_2, \dots, A_n | C)$ за различите типове атрибута?

Како проценити вероватноће на основу података?

| <i>Tid</i> | Refund | Marital Status | Taxable Income | Evade |
|------------|--------|----------------|----------------|-------|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

■ Класа: $P(C) = N_C/N$

- нпр. $P(\text{No}) = 7/10$,
 $P(\text{Yes}) = 3/10$

■ За дискретне атрибуте:

$$P(A_i | C_k) = |A_{ik}| / N_C^k$$

- где је $|A_{ik}|$ број инстанци које имају атрибут A_i и припадају класи C_k
- Пример:
 $P(\text{Status}=\text{Married}|\text{No}) = 4/7$
 $P(\text{Refund}=\text{Yes}|\text{Yes})=0$

Како проценити вероватноће на основу података?

За непрекидне атрибуте

- Дискретизација у групе
 - један редни атрибут (A_i) по групи
 - $P(A_i | C = c)$ се рачуна као проценат тренинг слогова класе c који припадају интервалу A_i
- Процена густине вероватноће
 - Претпоставимо да атрибути имају нормалну расподелу
 - Користити податке за процену параметара дистрибуције (нпр. средине или стандардне девијације)
 - Када је расподела вероватноћа позната она се може користити за процену условних вероватноћа

Како проценити вероватноће на основу података?

Може се претпоставити одређена расподела непрекидних променљивих и извршити процена параметара дистрибуције користећи тренинг податке. Обично се за ту процену бира нормална расподела

| Tid | Refund | Marital Status | Taxable Income | Evide |
|-----|--------|----------------|----------------|-------|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

- Нормална расподела:

$$P(A_i | c_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(A_i - \mu_j)^2}{2\sigma_j^2}}$$

- По једна за сваки пар (A_i, c_j)

- За $(Income, Class=No)$:

- Ако је $Class=No$
 - узорачка средина = 110
 - Узорачка варијанса = 2975

$$P(Income = 120 | No) = \frac{1}{\sqrt{2\pi(54.54)}} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Пример наивних Бајесовских класификатора

За дати тест слог $X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$

наивни Бајесов класификатор

$P(\text{Refund}=\text{Yes}|\text{No}) = 3/7$
 $P(\text{Refund}=\text{No}|\text{No}) = 4/7$
 $P(\text{Refund}=\text{Yes}|\text{Yes}) = 0$
 $P(\text{Refund}=\text{No}|\text{Yes}) = 1$
 $P(\text{Marital Status}=\text{Single}|\text{No}) = 2/7$
 $P(\text{Marital Status}=\text{Divorced}|\text{No})=1/7$
 $P(\text{Marital Status}=\text{Married}|\text{No}) = 4/7$
 $P(\text{Marital Status}=\text{Single}|\text{Yes}) = 2/7$
 $P(\text{Marital Status}=\text{Divorced}|\text{Yes})=1/7$
 $P(\text{Marital Status}=\text{Married}|\text{Yes}) = 0$

For taxable income:

If class=No: sample mean=110
 sample variance=2975
 If class=Yes: sample mean=90
 sample variance=25

- $P(X|\text{Class}=\text{No}) = P(\text{Refund}=\text{No}|\text{Class}=\text{No})$
 $\times P(\text{Married}|\text{Class}=\text{No})$
 $\times P(\text{Income}=120\text{K}|\text{Class}=\text{No})$
 $= 4/7 \times 4/7 \times 0.0072 = 0.0024$
- $P(X|\text{Class}=\text{Yes}) = P(\text{Refund}=\text{No}|\text{Class}=\text{Yes})$
 $\times P(\text{Married}|\text{Class}=\text{Yes})$
 $\times P(\text{Income}=120\text{K}|\text{Class}=\text{Yes})$
 $= 1 \times 0 \times 1.2 \times 10^{-9} = 0$

Пошто $P(X|\text{No})P(\text{No}) > P(X|\text{Yes})P(\text{Yes})$

Тада је $P(\text{No}|X) > P(\text{Yes}|X) \Rightarrow \text{Class} = \text{No}$

Процена квалитета наивних Бајесовски класификатора

Ако је једна од условних вероватноћа једнака нули, тада је и целокупна вероватноћа нула. Да би и у том случају могла да се класификује инстанца користи се једна од следећих мера

- Laplas $P(A_i = a | C = c) = \frac{n_a + 1}{n + v}$
- M-процена $P(A_i = a | C = c) = \frac{n_a + mp}{n + m}$

где је

- n укупан број инстанци у класи $C = c$
- n_a број тренинг инстанци из класе $C = c$ са $A_i = a$
- v укупан број вредности које A_i може да узима
- m еквивалент величини класа
- p иницијална процена $P(A_i = a | C = c)$ (позната унапред)

Наивни Бајесовски класификатори - резиме

- Робусни су у односу на изоловани шум
- Баратају недостајућим вредностима игноришући инстанцу при израчунавању процене вероватноће
- Робусни су у односу на ирелевантне атрибуте
- Претпоставка независност не мора да важи за све атрибуте
 - када постоје зависни атрибути или атрибути у корелацији добијене резултате треба узети са опрезом
 - друге технике као Бајесовске мреже поверења (енг. *Bayesian Belief Networks, BBN*)

Пример наивних Бајесовских класификатора

| Name | Give Birth | Can Fly | Live in Water | Have Legs | Class |
|---------------|------------|---------|---------------|-----------|-------------|
| human | yes | no | no | yes | mammals |
| python | no | no | no | no | non-mammals |
| salmon | no | no | yes | no | non-mammals |
| whale | yes | no | yes | no | mammals |
| frog | no | no | sometimes | yes | non-mammals |
| komodo | no | no | no | yes | non-mammals |
| bat | yes | yes | no | yes | mammals |
| pigeon | no | yes | no | yes | non-mammals |
| cat | yes | no | no | yes | mammals |
| leopard shark | yes | no | yes | no | non-mammals |
| turtle | no | no | sometimes | yes | non-mammals |
| penguin | no | no | sometimes | yes | non-mammals |
| porcupine | yes | no | no | yes | mammals |
| eel | no | no | yes | no | non-mammals |
| salamander | no | no | sometimes | yes | non-mammals |
| gila monster | no | no | no | yes | non-mammals |
| platypus | no | no | no | yes | mammals |
| owl | no | yes | no | yes | non-mammals |
| dolphin | yes | no | yes | no | mammals |
| eagle | no | yes | no | yes | non-mammals |

A: attributes

M: mammals

N: non-mammals

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

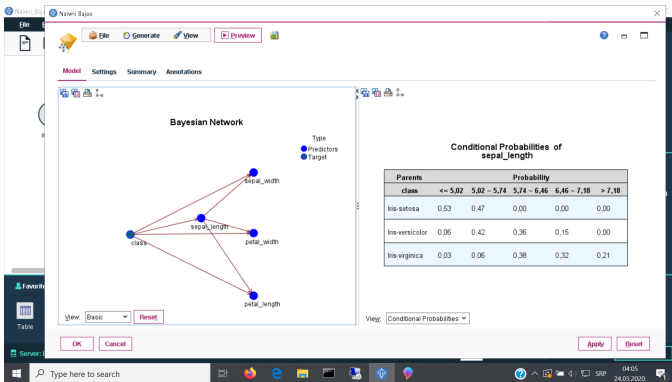
$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

| Give Birth | Can Fly | Live in Water | Have Legs | Class |
|------------|---------|---------------|-----------|-------|
| yes | no | yes | no | ? |

$P(A|M)P(M) > P(A|N)P(N)$

=> Mammals

Пример наивних Бајесовских класификатора



- Пример на скупу ИРИС
- SPSS Modeler, чвор Bayes Net
- У опцијама изабрати TAN (Tree Augmented Naive Bayes) модел
- Комплетан материјал је на сајту (слике, поток, резултати,...)