# Бајесовски класификатори

Ненад Митић

Математички факултет
nenad@matf.bg.ac.rs

# Увод

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

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

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

- Бајесова теорема:

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

# 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?

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

# Бајесовски класификатори

- Посматрајмо сваки атрибут и сваку ознаку класе као независне променљиве
- За дати слог са атрибутима $(A_1, A_2, \ldots, A_n, C)$ где је $C$ ознака класе
  - Циљ је предвидети класу $C$ којој припада
  - Желимо да нађемо вредност $C$ која максимизира $P(C \mid A_1, A_2, ..., A_n)$
- Да ли се $P(C \mid A_1, A_2, ..., A_n)$ 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 $P(C \mid A_1, A_2, ..., A_n)$ за сваку класу $C$ користећи Бајесову теорему

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

  - Изабрати вредност $C$ која максимизује $P(C \mid A_1, A_2, ..., A_n)$
  - Еквивалентно, може да се узме и вредност $C$ која максимизује $P(A_1, A_2, ..., A_n \mid C)P(C)$

    (услов - атрибути су међусобно независни за сваку од класа $C$)

- Како проценити $P(A_1, A_2, ..., A_n \mid C)$ за различите типове атрибута?

# Како проценити вероватноће на основу података?

| *Tid* | 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(No) = 7/10,
    P(Yes) = 3/10
- За дискретне атрибуте:

$$P(A_i \mid C_k) = |A_{ik}| / N_c^k$$

  - где је $|A_{ik}|$ број инстанци које имају атрибут $A_i$ и припадају класи $C_k$
  - Пример:

    P(Status=Married|No) = 4/7
    P(Refund=Yes|Yes)=0

# Како проценити вероватноће на основу података?

За непрекидне атрибуте

- Дискретизација у групе
  - један редни атрибут ($A_i$) по групи
  - $P(A_i \mid C = c)$ 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 | 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 |

- Нормална расподела:

$$P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

  - По једна за сваки пар $(A_i, c_j)$

- За (Income, Class=No):
  - Ако је Class=No
    - узорачка средина = 110
    - Узорачка варијанса = 2975

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

# Пример наивних Бајесовских класификатора

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

наивни Бајесов класификатор

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|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|Class=No) = P(Refund=No|Class=No)
  × P(Married| Class=No)
  × P(Income=120K| Class=No)
  = 4/7 × 4/7 × 0.0072 = 0.0024

- P(X|Class=Yes) = P(Refund=No| Class=Yes)
  × P(Married| Class=Yes)
  × P(Income=120K| Class=Yes)
  = 1 × 0 × 1.2 × $10^{-9}$ = 0

Пошто P(X|No)P(No) > P(X|Yes)P(Yes)

Тада је P(No|X) > P(Yes|X) => Class = 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 $P(A_i = a \mid C = c) = \frac{n_a+1}{n+v}$
- M-procena $P(A_i = a \mid C = c) = \frac{n_a+mp}{n+m}$

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 $P(A_i = a \mid C = c)$ (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**

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

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

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

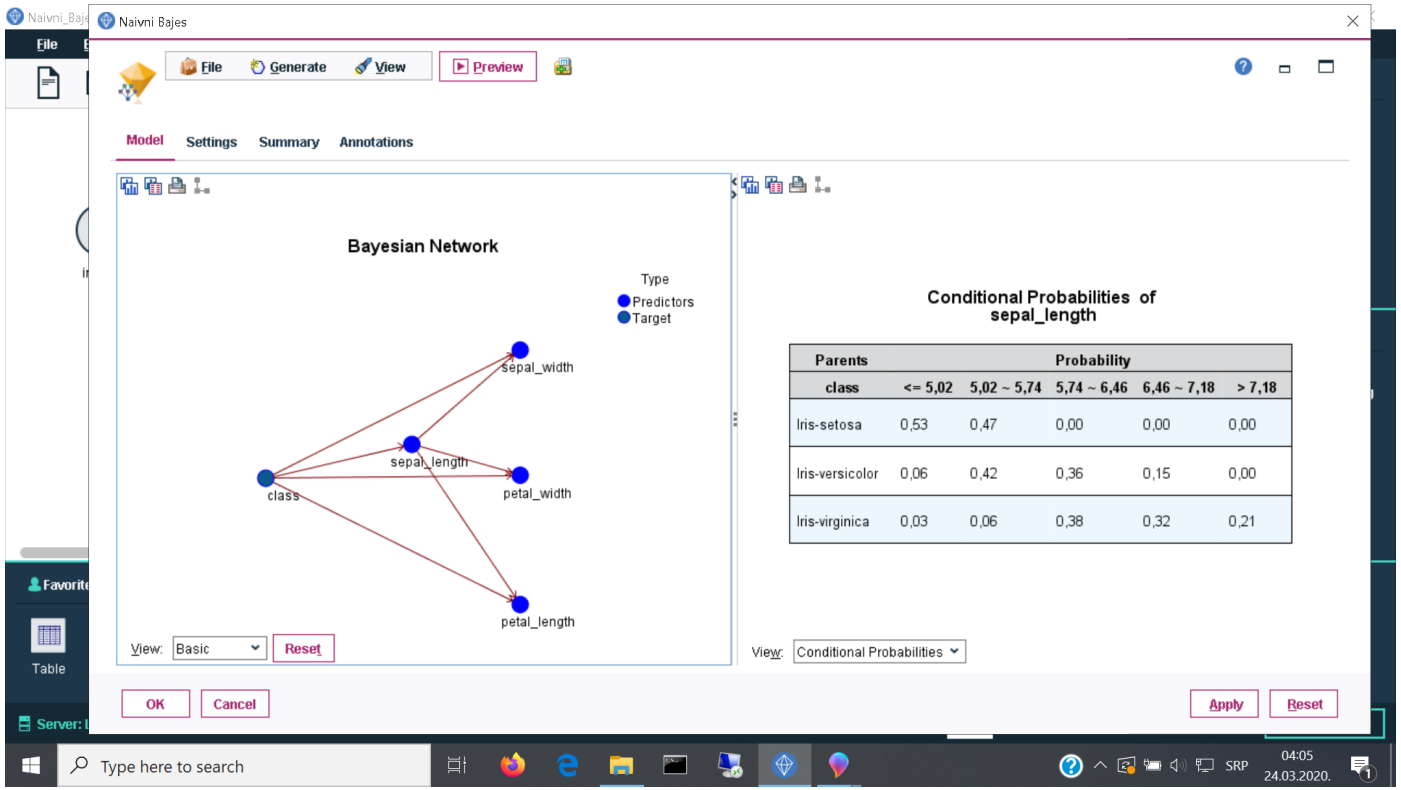
$$P(A \mid N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

| 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,...)