HOWEST TOEGEPASTE INFORMATICA, 2022-2023, © BRIAN BAERT

DATA ANALYTICS

HOOFDSTUK 6 (DEEL 2)
NAIVE BAYES CLASSIFIER

howest.be

1

HERHALING CLASSIFICEREN VIA BESLISSINGSBOMEN

- Classificatieprobleem = "Hoe kunnen we data indelen in een klasse op basis van een aantal **training records**?"
- Voorbeeld (zie figuur hiernaast): tot welke `Evade` klasse (`Yes` of `No`) behoort een record, als je weet dat voor dit nieuwe record `No`, `Married`, `120K` ?
- Methode van vorige week = classificatie o.b.v. een **beslissingsboom** (**decision tree**) → opsplitsen van de training data mbv een algoritme (algoritme van Hunt)
- Splitsingsvoorkeur laten afhangen van de onzuiverheid van de informatie: beste splitsing = deze met de laagste onzuiverheidsgraad (= hoogste informatiewinst)
- Maten voor onzuiverheid: Gini-coëfficiënt of Entropie

Tabel met training records

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
	No	Married	120K	???

2

NAIVE BAYES CLASSIFIER

- Classificatiemethode van deze week = Naive Bayes
- Deze methode is gebaseerd op **voorwaardelijke kansen**
- Kansen worden berekend m.b.v. de formule van Bayes

Tabel met training records

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
	No	Married	120K	????

3

FORMULE VAN BAYES (HERHALING)

- Volgens de definitie van voorwaardelijke kans, geldt:

$$P(Y|X) = \frac{P(X \cap Y)}{P(X)} \quad (1)$$

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} \Rightarrow P(X \cap Y) = P(X|Y).P(Y) \quad (2)$$

- Als je nu in (1) $P(X \cap Y)$ vervangt door (2), dan bekom je onderstaande formule van Bayes:

$$P(Y|X) = \frac{P(Y).P(X|Y)}{P(X)}$$

4

NAIVE BAYES CLASSIFIER

- Stel dat we het te classificeren record voorstellen als een vector $\vec{X} = (X_1, X_2, \ldots, X_d)$ waarbij $X_1, X_2, \ldots, X_d$ gekende attribuutwaarden zijn. Dan kunnen we onderstaande formule van Bayes gebruiken om de voorwaardelijke kans te berekenen dat die record tot een bepaalde klasse Y behoort:

$$P(Y|\vec{X}) = \frac{P(Y)P(\vec{X}|Y)}{P(\vec{X})}$$

- Het gebruik van de Naive Bayes classifier, gaat uit van de (naïeve) veronderstelling van "**class conditional independence**", hetgeen betekent dat een *attribuut* van een gegeven klasse, *onafhankelijk* is van de *andere attributen* van die gegeven klasse. Hierdoor mag je de voorwaardelijke kans $P(\vec{X}|Y)$ in de teller van bovenstaande formule vervangen door het product van voorwaardelijke kansen, hetgeen leidt tot onderstaande formule:

$$P(Y|\vec{X}) = \frac{P(Y).\prod_{i=1}^{d} P(X_i|Y)}{P(\vec{X})}$$

5

NAIVE BAYES CLASSIFIER SAMENGEVAT

Bij gebruik van de naive Bayes classifier, moet je m.b.v. onderstaande formule de voorwaardelijke (**a-posteriori)** kans voor elke mogelijke klasse (= onbekende attribuutwaarde) Y van een record $\vec{X}$ berekenen.

De classifier zal dan de record $\vec{X}$ aan de klasse Y met de hoogste voorwaardelijke kans toekennen.

$$P(Y|\vec{X}) = \frac{P(Y).\prod_{i=1}^{d} P(X_i|Y)}{P(\vec{X})}$$

Omdat $P(\vec{X})$ constant is (stel = α) volstaat het de kans te bepalen van de teller.

$$P(Y|\vec{X}) = \frac{P(Y).\prod_{i=1}^{d} P(X_i|Y)}{\alpha}$$

6

VOORBEELD VAN BEREKENINGEN MET DE NAIVE BAYES CLASSIFIER

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
	No	Married	120K	????

We herformuleren de vraag uit dia 2:
Tot welke *Evade-klasse* (Yes of No) behoort onderstaande record $\vec{X}$, als we gebruikmaken van naïve Bayes?
$\vec{X}$ (Refund=No, Marital Status=Married, Income=120K)

Aangezien het Evade-attribuut 2 mogelijke waarden kan aannemen (Yes of No), zullen we **P(Evade=No $|\vec{X}$)** en **P(Evade=Yes $|\vec{X}$)** moeten berekenen en de grootste van die 2 kansen moeten bepalen.

Als de eerste kans de grootste is, dan zullen we aan het record de *Evade-klasse* "No" toekennen; in het andere geval kennen we de *Evade-klasse* "Yes" toe.

7

VOORBEELD VAN BEREKENINGEN MET DE NAÏVE BAYES CLASSIFIER (2)

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
	No	Married	120K	????

We maken gebruik van de formule op dia 6:

$$P(\text{Evade} = \text{No}|\vec{X}) = \frac{P(\text{Evade} = \text{No}).\prod_{i=1}^{d} P(X_i|\text{Evade} = \text{No})}{P(\vec{X})}$$

met:
X1 = "Refund=No"
X2 = "Marital Status = Married"
X3 = "Income=120K"

We berekenen de kansen uit de teller van de breuk:

P(Evade=No) = $\frac{7}{10}$ (af te leiden uit de Evade-kolom in de tabel)

8

VOORBEELD VAN BEREKENINGEN MET DE NAÏVE BAYES CLASSIFIER (3)

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
	No	Married	120K	????

$P(X_1 | \text{Evade=No}) = P(\text{Refund=No} | \text{Evade=No}) = \frac{4}{7}$

$P(X_2 | \text{Evade=No}) = P(\text{Marital Status =Married} | \text{Evade=No}) = \frac{4}{7}$

$P(X_3 | \text{Evade=No}) = P(\text{Income=120K} | \text{Evade=No}) = ???$

De eerste 2 kansen zijn gemakkelijk af te leiden uit de tabel omdat het om Refund en Marital Status discrete variabelen zijn en hun waarden dus gemakkelijk kunnen geteld worden.

De derde kans is moeilijker te berekenen omdat Income (theoretisch) een continue variabele is.

9

VOORBEELD VAN BEREKENINGEN MET DE NAIVE BAYES CLASSIFIER (4)

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
	No	Married	120K	????

Voor continue variabelen, gebruiken we onderstaande formule (benadering mbv de **normaalverdeling**):

$$P(X_i = x_i | Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

De parameter μ_{ij} kan geschat worden op basis van X_i $(\bar{x})$, het steekproefgemiddelde voor alle training records die behoren tot klasse y_j.
Analoog kan ook σ_{ij}^2 geschat worden uit (s^2), de steekproef-variantie van de training records.

10

VOORBEELD VAN BEREKENINGEN MET DE NAIVE BAYES CLASSIFIER (5-A)

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
	No	Married	120K	????

Om de kans P(Income=120K|Evade=No) te berekenen, moeten we dus eerst het gemiddelde en de standaardafwijking van de kolom Income berekenen (voor de klasse "No").

$$\bar{x} = \frac{(125 + 100 + 70 + \cdots + 75)}{7} = 110$$

$$s^2 = \frac{(125 - 110)^2 + (100 - 110)^2 + \cdots + (75 - 110)^2}{6} = 2975$$

$$s = \sqrt{2975} \approx 54{,}54$$

We passen vervolgens de formule van de normaalverdeling toe (of gebruik je GRM of Excel):

$$P(\text{Income=120}|\text{No}) = \frac{1}{\sqrt{2\pi}.(54{,}54)} e^{-\frac{(120-110)^2}{2*2975}} = 0{,}0072$$

11

VOORBEELD VAN BEREKENINGEN MET DE NAIVE BAYES CLASSIFIER (5-B)

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
	No	Married	120K	????

Om de kans P(Income=120K|Evade=No) te berekenen, moeten we dus eerst het gemiddelde en de standaardafwijking van de kolom Income berekenen (voor de klasse "Yes").

$$\bar{x} = \frac{(95 + 85 + 90)}{3} = 90$$

$$s^2 = \frac{(95 - 90)^2 + (85 - 90)^2 + (90 - 90)^2}{2} = 25$$

$$s = \sqrt{25} = 5$$

We passen vervolgens de formule van de normaalverdeling toe (of gebruik je GRM of Excel):

$$P(\text{Income=120}|\text{Yes}) = \frac{1}{\sqrt{2\pi}.(5)} e^{-\frac{(120-90)^2}{2*25}} = 1{,}2\,.\,10^{-9}$$

12

VOORBEELD VAN BEREKENINGEN MET DE NAIVE BAYES CLASSIFIER (5)

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
	No	Married	120K	????

We vullen nu in onderstaande formule de berekende kansen in:

$$P(\text{Evade} = \text{No}|\vec{X}) = \frac{P(\text{Evade} = \text{No}).\prod_{i=1}^{d} P(X_i|\text{Evade} = \text{No})}{P(\vec{X})}$$

$$= \frac{\frac{7}{10} \cdot \frac{4}{7} \cdot \frac{4}{7} \cdot 0{,}0072}{P(\vec{X})} \approx \frac{0{,}00165}{P(\vec{X})}$$

$$P(\text{Evade} = \text{Yes}|\vec{X}) = \frac{P(\text{Evade} = \text{Yes}).\prod_{i=1}^{d} P(X_i|\text{Evade} = \text{Yes})}{P(\vec{X})}$$

$$= \frac{\frac{3}{10} \cdot 1.0.1{,}2x10^{-9}}{P(\vec{X})} = 0$$

Omdat $P(\text{Evade} = \text{No}|\vec{X}) > P(\text{Evade} = \text{Yes}|\vec{X})$ zullen we dus de waarde No toekennen aan het Evade-attribuut.

WAT ALS EEN VOORWAARDELIJKE KANS 0 IS?

- **Al**s de klasse-voorwaardelijke **kans** voor één van de attributen **nul** is, zal de volledige voorwaardelijke kans voor de klasse verdwijnen! Dit was in het vorig voorbeeld het geval bij de berekening van $P(\texttt{Evade=Yes}|\vec{X})$, want P(`Marital Status = Married|Evade = Yes`) =0.
- Mogelijke oplossing is de **m-schatting**

$$P(X_i|Y_j) = \frac{n_i + mp}{n + m}$$

Hierin is n het totale aantal records van klasse y_j, n_i is het aantal trainingsrecords van klasse y_i die de waarde x_i aannemen, m en p zijn (gegeven) parameters.

- Maken we gebruik van deze m-schatting (met m=3 en p=1/3), dan krijgen we:

$$P(\text{Marital Status=Married}|\text{Evade=Yes}) = \left(0 + 3 * \frac{1}{3}\right)/(3 + 3) = \frac{1}{6}$$

VOORBEELD 2

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

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

Omdat de noemer van de voorwaardelijke kansen toch steeds dezelfde is en niet berekend hoeft te worden, stellen we deze voor door de Griekse letter α. Verder duiden we Mammals aan met de letter M en Non-mammals met de letter N.

$P(Class=M|\vec{X}) = \alpha \times P(Class=M) \times P(GiveBirth=Yes|Class=M) \times P(Can\ Fly=No|Class=M) \times P(Live\ in\ Water=Yes|Class=M) \times P(Have\ Legs=No|Class=M)$

$= \alpha \times \frac{7}{20} \times \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} \approx 0{,}021\alpha$

$P(Class=N|\vec{X}) = \alpha \times P(Class=N) \times P(GiveBirth=Yes|Class=N) \times P(Can\ Fly=No|Class=N) \times P(Live\ in\ Water=Yes|Class=N) \times P(Have\ Legs=No|Class=N)$

$= \alpha \times \frac{13}{20} \times \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} \approx 0{,}0027\alpha$

Omdat $P(Class=M|\vec{X}) > P(Class=N|\vec{X})$
delen we het dier in bij de Mammals.

15

OEFENINGENREEKS

- 6.3.7 Oefeningen
  1-5

16