Oplossingen oefeningen hoofdstuk 6 - 3.7

Oefening 1

Beschouw de onderstaande tabel.

Table 5.1.

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

a) Schat de voorwaardelijke kansen voor P(A|+), P(B|+), P(C|+), P(A|-), P(B|-) en P(C|-).

b) Maak gebruik van de geschatte voorwaardelijke kansen in de vorige vraag om het class label toe te kennen voor het testrecord (A=0, B=1, C=0) gebruik makend van naïve Bayes classificeren.

c) Schat de voorwaardelijke kansen met de *m*-schattingsmethode (p=1/2 en m=4)

d) Herhaal (b) gebruik makend van de voorwaardelijke kansen uit (c).

e) Vergelijk de twee schattingsmethodes. Welke is beter? Waarom?

Oplossing oefening 1

a) De voorwaardelijke kansen
P(A=1|-) = 2/5 = 0,4
P(B=1|-) = 2/5 = 0,4
P(C=1|-) = 5/5 = 1
P(A=0|-) = 3/5 = 0,6
P(B=0|-) = 3/5 = 0,6
P(C=0|-) = 0/5 = 0
P(A=1|+) = 3/5 = 0,6
P(B=1|+) = 1/5 = 0,2
P(C=1|+) = 4/5 = 0,8
P(A=0|+) = 2/5 = 0,4
P(B=0|+) = 4/5 = 0,8
P(C=0|+) = 1/5 = 0,2

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

Hier stelt Y de waarde van class voor (en is dus ofwel +, ofwel -)

We berekenen eerste de voorwaardelijke kans dat de record tot de class – behoort:

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

In bovenstaande formule is α een onbekende constante die we niet kunnen vervangen. De rest kunnen we vervangen.

$P(-|\vec{X}) = \alpha.\frac{5}{10}. P(A = 0|-). P(B = 1|-). P(C = 0|-) = \alpha.\frac{5}{10}.\left(\frac{3}{5}\right).\left(\frac{2}{5}\right).\left(\frac{0}{5}\right) = 0$

We berekenen vervolgens de voorwaardelijke kans dat de record tot de class + behoort:

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

In bovenstaande formule is α een onbekende constante die we niet kunnen vervangen. De rest kunnen we vervangen.

$P(+|\vec{X}) = \alpha.\frac{5}{10}. P(A = 0|+). P(B = 1|+). P(C = 0|+) = \alpha.\frac{5}{10}.\left(\frac{2}{5}\right).\left(\frac{1}{5}\right).\left(\frac{1}{5}\right) = 0{,}008.\alpha$

Omdat $P(+|\vec{X}) > P(-|\vec{X})$ zullen we aan testrecord $\vec{X}$ het klasselabel + toekennen.

c) M-schatting

$$P(x_i|y_i) = \frac{n_i + mp}{n + m}$$

P(A=0|+) = (2+2)/(5+4) = 4/9
P(B=0|+) = (4+2)/(5+4) = 6/9
P(C=0|+) = (1+2)/(5+4) = 3/9
P(A=0|-) = (3+2)/(5+4) = 5/9
P(B=0|-) = (3+2)/(5+4) = 5/9
P(C=0|-) = (0+2)/(5+4) = 2/9
P(A=1|+) = (3+2)/(5+4) = 5/9
P(B=1|+) = (1+2)/(5+4) = 3/9
P(C=1|+) = (4+2)/(5+4) = 6/9
P(A=1|-) = (2+2)/(5+4) = 4/9
P(B=1|-) = (2+2)/(5+4) = 4/9
P(C=1|-) = (5+2)/(5+4) = 7/9

d) We herhalen (b) gebruik makend van de m-schattingen.

$P(-|\vec{X}) = \alpha.\frac{5}{10}. P(A = 0|-). P(B = 1|-). P(C = 0|-) = \alpha.\left(\frac{5}{10}\right).\left(\frac{5}{9}\right).\left(\frac{4}{9}\right).\left(\frac{2}{9}\right) = 0{,}0274\alpha$

$P(+|\vec{X}) = \alpha.\frac{5}{10}. P(A = 0|+). P(B = 1|+). P(C = 0|+) = \alpha.\left(\frac{5}{10}\right).\left(\frac{4}{9}\right).\left(\frac{3}{9}\right).\left(\frac{3}{9}\right) = 0{,}0247\alpha$

Omdat $P(-|\vec{X}) > P(+|\vec{X})$ zullen we nu aan testrecord $\vec{X}$ het klasselabel - toekennen.

e) Bij het vergelijken van de 2 methodes is de m-schatting een betere methode, omwille van het klein aantal trainingsrecords.

Oefening 2

Beschouw onderstaande tabel.

Table 5.2.

Instance	A	B	C	Class
1	0	0	1	$-$
2	1	0	1	$+$
3	0	1	0	$-$
4	1	0	0	$-$
5	1	0	1	$+$
6	0	0	1	$+$
7	1	1	0	$-$
8	0	0	0	$-$
9	0	1	0	$+$
10	1	1	1	$+$

a) Schat de voorwaardelijke kansen voor P(A=1|+), P(B=1|+), P(C=1|+), P(A=1|-), P(B=1|-) en P(C=1|-) met dezelfde methode als in oefening 1.

b) Gebruik de voorwaardelijke kansen uit (a) om het class label te voorspellen voor een testrecord (A=1, B=1, C=1) met naïve Bayes classificeren.

c) Vergelijk P(A=1), P(B=1) en P(A=1,B=1). Is er een relatie tussen A en B?

d) Herhaal de analyse uit (c) met P(A=1), P(B=0) en P(A=1,B=0).

e) Vergelijk P(A=1,B=1|Class=+) met P(A=1|Class=+) en P(B=1|Class=+).

Oplossing oefening 2

a)
P(A=1|+) = 3/5 = 0,6
P(B=1|+) = 2/5 = 0,4
P(C=1|+) = 4/5 = 0,8
P(A=1|-) = 2/5 = 0,4
P(B=1|-) = 2/5 = 0,4
P(C=1|-) = 1/5 = 0,2

b) Formule van Bayes : $\mathrm{P}(\mathrm{Y}|\vec{\mathrm{X}}) = \frac{\mathrm{P}(\mathrm{Y}).\prod_{i=1}^{d}\mathrm{P}(\mathrm{X_i}|\mathrm{Y})}{\mathrm{P}(\vec{\mathrm{X}})}$

Hier stelt Y de waarde van class voor (en is dus ofwel +, ofwel -)

We berekenen eerste de voorwaardelijke kans dat de record tot de class – behoort:

$$\mathrm{P}(-|\vec{\mathrm{X}}) = \frac{\mathrm{P}(-).\prod_{i=1}^{d}\mathrm{P}(\mathrm{X_i}|-)}{\mathrm{P}(\vec{\mathrm{X}})} = \alpha.\mathrm{P}(-).\prod_{i=1}^{d}\mathrm{P}(\mathrm{X_i}|-) \quad \text{(met } \alpha = \frac{1}{\mathrm{P}(\vec{\mathrm{X}})}\text{)}$$

In bovenstaande formule is α een onbekende constante die we niet kunnen vervangen. De rest kunnen we vervangen.

$$\mathrm{P}(-|\vec{\mathrm{X}}) = \alpha.\frac{5}{10}.\mathrm{P}(A=1|-).\mathrm{P}(B=1|-).\mathrm{P}(C=1|-) = \alpha.\left(\frac{5}{10}\right).\left(\frac{2}{5}\right).\left(\frac{2}{5}\right).\left(\frac{1}{5}\right) = \frac{2}{125}\alpha$$

Analoog is:

$$\mathrm{P}(+|\vec{\mathrm{X}}) = \alpha.\frac{5}{10}.\mathrm{P}(A=1|+).\mathrm{P}(B=1|+).\mathrm{P}(C=1|+) = \alpha.\left(\frac{5}{10}\right).\left(\frac{3}{5}\right).\left(\frac{2}{5}\right).\left(\frac{4}{5}\right) = \frac{12}{125}\alpha$$

Omdat $\mathrm{P}(+|\vec{\mathrm{X}}) > \mathrm{P}(-|\vec{\mathrm{X}})$ zullen we aan testrecord $\vec{\mathrm{X}}$ het klasselabel + toekennen.

c)
P(A=1) = 0,5

P(B=1) = 0,4

P(A=1,B=1) = 0,2

→ P(A=1,B=1)= P(A=1).P(B=1)

→ A=1 en B=1 zijn onafhankelijk.

d)
P(A=1) = 0,5

P(B=0) = 0,6

P(A=1,B=0) = 0,3

→ P(A=1,B=0)= P(A=1).P(B=0)

→ A=1 en B=0 zijn onafhankelijk.

e)
P(A=1,B=1|Class=+) = 0,1

P(A=1|Class=+) = 0,3

P(B=1|Class=+) = 0,2

Omdat P(A=1|+).P(B=1|+) niet hetzelfde is als P(A=1,B=1|+) zijn A=1 en B=1 niet voorwaardelijk onafhankelijk (als gegeven is dat ze tot class + behoren)

Omdat ze niet voorwaardelijk onafhankelijk zijn, mogen we eigenlijk naïve Bayes niet toepassen op deze dataset.

Oefening 3

Op een meerkeuze-examen kent de student het antwoord met kans p, of gokt op het juiste antwoord met kans 1-p.

Veronderstel dat de kans op het correct antwoorden, als je het antwoord kent, gelijk is aan 1 is en dat de kans dat een student die het antwoord niet kent, toch correct antwoordt, gelijk is aan $1/m$ (met m het aantal antwoordmogelijkheden).

Wat is de kans dat een student het antwoord kent als hij correct antwoordde?

Oplossing oefening 3

Stel K = "Kent antwoord" en C = "Correct antwoorden"

$$
\begin{aligned}
P(K|C) &= \frac{P(C|K).P(K)}{P(C)} \\
&= \frac{P(C|K).P(K)}{P(C|K).P(K) + P(C|\overline{K}).P(\overline{K})} \\
&= \frac{1.p}{1.p + \frac{1}{m}.(1-p)} \\
&= \frac{p}{\frac{mp}{m} + \frac{1}{m} - \frac{p}{m}} \\
&= \frac{mp}{mp + 1 - p}
\end{aligned}
$$

Oefening 4

Gegeven de onderstaande dataset "*Buy computer data*",

RID	age	income	student	credit_rating	Class: buys_computer
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	no
3	31 . . . 40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31 . . . 40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31 . . . 40	medium	no	excellent	yes
13	31 . . . 40	high	yes	fair	yes
14	>40	medium	no	excellent	no

Voorspel de class die bij het volgende testrecord hoort aan de hand van Naive Bayes:

$\vec{X}$={age <= 30; income = medium; student = yes; credit-rating = fair}

Oplossing oefening 4

X_1: age <= 30; X_2: income = medium; X_3: student = yes; X_4: credit-rating = fair

$$P(X_1|yes) = \frac{2}{9} = 0{,}222$$
$$P(X_2|yes) = \frac{4}{9} = 0{,}444$$
$$P(X_3|yes) = \frac{6}{9} = 0{,}667$$
$$P(X_4|yes) = \frac{6}{9} = 0{,}667$$

$$P(yes) = \frac{9}{14} = 0{,}643$$

$$P(X_1|no) = \frac{3}{5} = 0{,}6$$
$$P(X_2|no) = \frac{2}{5} = 0{,}4$$
$$P(X_3|no) = \frac{1}{5} = 0{,}2$$
$$P(X_4|no) = \frac{2}{5} = 0{,}4$$

$$P(no) = \frac{5}{14} = 0{,}357$$

$$P(yes\,|\vec{X}) = \frac{(0{,}643).(0{,}222).(0{,}444).(0{,}667).(0{,}667)}{P(\vec{X})} = 0{,}028\alpha$$

$$P(no\,|\vec{X}) = \frac{(0{,}357).(0{,}6).(0{,}4).(0{,}2).(0{,}4)}{P(\vec{X})} = 0{,}007\alpha$$

$P(yes|\vec{X}) > P(no|\vec{X})$ → het record krijgt als label "yes".

Uitbreiding: Berekening van $P(\vec{X})$ en complementregel

$P(yes|\vec{X}) + P(no|\vec{X})$ moet gelijk zijn aan 1. Opdat dit gelijk zou zijn moeten wij natuurlijk ook de priorkans $P(\vec{X})$ kennen.

De priorkans berekenen wij als volgt:

$$P(yes).P(\vec{X}|yes) + P(no).P(\vec{X}|no) = P(\vec{X})$$
$$0{,}028 + 0{,}007 = 0{,}035 = P(\vec{X}) \quad (*)$$

Vullen wij dit in dan geldt:

$$P(yes|\vec{X}) = \frac{0{,}028}{0{,}035} = 0{,}8$$
$$P(no|\vec{X}) = \frac{0{,}007}{0{,}035} = 0{,}2$$

En dus

$$P(yes|\vec{X}) + P(no|\vec{X}) = 0{,}8 + 0{,}2 = 1$$

Oefening 5 (thuis)

Beschouw het volgende classificatieprobleem waarbij de tabel links wordt gegeven. X_1 en X_2 zijn twee binaire geobserveerde variabelen. Y is het class-label. Maak gebruik van de Naive Bayes Classifier om volgende vragen te beantwoorden.

X_1	X_2	Y	Counts
0	0	0	2
0	0	1	18
1	0	0	4
1	0	1	1
0	1	0	4
0	1	1	1
1	1	0	2
1	1	1	18

a) Classificeer het record $(X_1 = 0, X_2 = 0)$.

b) Bereken $P(Y = 1|X_1 = 0, X_2 = 0)$

Oplossing oefening 5

$\frac{1}{P(\vec{X})} = \alpha$

a)

P(Y=0|$\vec{X}$) = P(Y=0).P(X_1=0|Y=0). P(X_2=0|Y=0).α = $\frac{12}{50}.\frac{6}{12}.\frac{6}{12}.\alpha = \frac{3}{50}\alpha = \frac{6}{100}\alpha$

P(Y=1|$\vec{X}$) = P(Y=1).P(X_1=0|Y=1). P(X_2=0|Y=1).α = $\frac{38}{50}.\frac{19}{38}.\frac{19}{38}.\alpha = \frac{19}{100}\alpha$

Omdat P(Y=1|$\vec{X}$) > P(Y=0|$\vec{X}$) geven we de testrecord het label “1” mee.

b)

Wij maken gebruik van dezelfde redenering als in de uitbreiding van oefening 4.

Dan geldt:

P(Y=0|$\vec{X}$) + P(Y=1|$\vec{X}$) = 1.

De priorkans berekenen wij als volgt:

$$P(Y = 0).P(\vec{X}|Y = 0) + P(Y = 1); P(\vec{X}|Y = 1) = P(\vec{X})$$
$$\frac{6}{100} + \frac{19}{100} = \frac{25}{100} = \frac{1}{4} = P(\vec{X})$$

Vullen wij dit in dan geldt:

$$P(Y = 0\,|\vec{X}) = \frac{\frac{6}{100}}{\frac{1}{4}} = \frac{6}{100} * \frac{4}{1} = \frac{24}{100}$$
$$P(Y = 1|\vec{X}) = \frac{\frac{19}{100}}{\frac{1}{4}} = \frac{19}{100} * \frac{4}{1} = \frac{76}{100}$$

De gevraagde kans is dus 76/100.