6.3 Naïve Bayes classifier

6.3.1 Inleiding

In vele toepassingen is de relatie tussen de attributenverzameling en de klasse variabele niet deterministisch. In andere woorden, een klasse label van een test record kan niet met zekerheid worden voorspeld, zelfs als de bijhorende attributenverzameling identiek is aan dat van een training voorbeeld. Neem bijvoorbeeld de taak waarin men probeert te voorspelen of een bepaalde persoon risico heeft op hartziekten gebaseerd op deze persoon zijn eetgewoonten en sportpatroon. De meeste mensen die gezond eten en regelmatig sporten zullen minder kans hebben op het ontwikkelen van hartziekten, maar er zijn natuurlijk nog tal van andere factoren die hun invloed hebben, denk maar aan roken, alcoholgebruik, …

In het hoofdstuk 'kansrekening' hebben we al kennis gemaakt met de regel van Bayes, een statistisch principe om voorafgaande kennis m.b.t. de klassen te combineren met nieuwe vergaarde data.

We hebben echter nood aan een "verkorte" vorm van de regel, hiervoor definiëren we eerst een nieuwe wet namelijk de **wet van de totale kans**:
Gegeven twee onafhankelijke s.v. X en Y, dan geldt

$$P(X) = \sum_{i=1}^{k} P(X, Y_i) = \sum_{i=1}^{k} P(X|Y_i).P(Y_i)$$

Herschrijven we de regel van Bayes, gebruik makend van de wet van de totale kans, dan:

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

6.3.2 Gebruik maken van de regel van Bayes bij het classificeren

Stel X de verzameling van attributen en Y de klasse variabele. Als de klasse variabele een niet-deterministische relatie heeft met de attributen, dan kunnen X en Y als stochastische variabelen behandeld worden en hun relatie probabilistisch bepalen via $P(Y|X)$. Deze voorwaardelijke kans wordt de **a-posteriori**[1] kans van Y genoemd en $P(Y)$ de **a-priori** kans. In de noemer staat $P(X)$, dit wordt ook wel '*evidence*' genoemd.
Tijdens de training fase moeten we de a-posteriori kansen $P(Y|X)$ leren kennen voor elke mogelijke combinatie van X en Y, gebaseerd op vergaarde informatie uit de training data. Op die manier kan een test record X' geclassificeerd worden door het vinden van de klasse Y' die de a-posteriori kans maximaliseert.

[1] A posteriori-kennis is kennis afgeleid uit de ervaring. Dit in tegenstelling tot a priori-kennis, die voorafgaat aan de ervaring of er niet afhankelijk van is. Een munt lijkt op het oog zuiver, daarom nemen we vooraf, a priori, aan dat de kans op kop ½ is. Bij 100 worpen met de munt blijkt 80 keer kop gegooid te zijn. Achteraf, a posteriori, stellen we onze aanname bij, en nemen aan dat de a posteriori-kans op kop 0,8 is.

6.3.3 Naive Bayes classifier

Een naive Bayes classifier schat de klasse-voorwaardelijke kans door te veronderstellen dat de attributen voorwaardelijk onafhankelijk zijn, gegeven het klasse label y. De voorwaardelijke onafhankelijkheidsaanname wordt formeel weergegeven (zie ook Hoofdstuk 2) als:

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

waarin elke attributenverzameling $\vec{X} = \{X_1, X_2, \ldots, X_d\}$ bestaat uit d attributen.

Om een test record te classificeren berekent de naive Bayes classifier de a-posteriori kans voor elke klasse Y:

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

Omdat $P(\vec{X})$ vastligt voor elke Y, is het voldoende om de klasse te kiezen die de teller maximaliseert.

7.3.2 Voorbeeld van de naive Bayes classifier

Beschouw volgende dataset in figuur 1-a met klasse voorwaardelijke kansen in figuur 1-b

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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

(a)

P(Home Owner=Yes|No) = 3/7
P(Home Owner=No|No) = 4/7
P(Home Owner=Yes|Yes) = 0
P(Home Owner=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) = 2/3
P(Marital Status=Divorced|Yes) = 1/3
P(Marital Status=Married|Yes) = 0

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

(b)

Figuur 1 - De naive Bayes classifier voor een leningprobleem. *Overgenomen uit Introduction to Data Mining (p. 235) door Tan P., Steinbach M., Kumar V., 2006, Pearson Education. Copyright 2006 Pearson Education Inc.*

Om het klasselabel van een test record $\vec{X}$ te voorspellen

$$\vec{X} = (\texttt{Home Owner=No, Marital Status=Married, Income=\$120K})$$

moeten we de a-posteriori kansen $P(\texttt{No}|\vec{X})$ en $P(\texttt{Yes}|\vec{X})$ berekenen. Deze kansen kunnen geschat worden door het product te berekenen tussen de a-priori kans $P(Y)$ en de klassevoorwaardelijke kansen $\prod_i P(X_i|Y)$.

De a-priori kansen van elke klasse kunnen geschat worden door de fractie van training records die behoren tot elke klasse te bepalen. Er zijn drie records die behoren tot de klasse `Yes` en zeven die behoren tot de klasse `No`, $P(\texttt{Yes}) = 0{,}3$ en $P(\texttt{No}) = 0{,}7$. De klasse).

De a-priori kansen voor discrete attributen zijn eenvoudig af te lezen uit figuur 58-b, voor continue attributen zoals "`annual income`" gebruiken we:

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

Alles samen geeft dit dan,

$$P(\vec{X}|\texttt{No}) = P(\texttt{Howe Owner=No}|\texttt{No}) * P(\texttt{Status=Married}|\texttt{No}) * P(\texttt{Annual Income=\$120K}|\texttt{No})$$
$$= 4/7 * 4/7 * 0{,}0072 = 0{,}0024$$

$$P(\vec{X}|\texttt{Yes}) = P(\texttt{Howe Owner=No}|\texttt{Yes}) * P(\texttt{Status=Married}|\texttt{Yes}) * P(\texttt{Annual Income=\$120K}|\texttt{Yes})$$
$$= 1 * 0 * 1{,}2 * 10^{-9} = 0$$

Alles samen wordt de a-posteriori kans voor de klasse `No` $P(\texttt{No}|\vec{X}) = \alpha * 7/10 * 0{,}0024 = 0{,}0016\alpha$ met $\alpha = \frac{1}{P(\vec{X})}$ een constante term. Analoog kunnen we aantonen dat de a-posteriori kans voor de klasse `Yes` gelijk is aan nul.
Omdat $P(\texttt{No}|X) > P(\texttt{Yes}|X)$ wordt het record geclassificeerd als `No`.

6.3.4 M-schatting van de voorwaardelijke kans

In 7.3.2 kwam een probleem naar boven bij het schatting van a-posteriori kansen uit de training data. Als de klasse voorwaardelijke kans voor één van de attributen gelijk is aan nul, dan zal de a-posteriori kans voor de klasse verdwijnen.
Een nog meer extreem geval, als de training records niet veel van de attribuutwaarden omvatten zal het onmogelijk zijn om een test record te classificeren. Neem als voorbeeld P(`Marital Status=Divorced`|`No`) gelijk aan 0 in plaats van 1/7, dan zal een record met attributenverzameling $\vec{X} = (\texttt{Home Owner=Yes, Marrital Status=Divorced, Income=\$120K})$ de volgende klasse voorwaardelijke kansen hebben:

$$P(\vec{X}|\texttt{No}) = 3/7 * 0 * 0{,}0072 = 0$$
$$P(\vec{X}|\texttt{Yes}) = 0 * 1/3 * 1{,}2 * 10^{-9} = 0$$

De naive Bayes classifier zal dit record niet kunnen classificeren!

Maar er bestaat hier een oplossing voor, we gebruiken **de m-schatting** voor het schatten van de voorwaardelijke kansen:

$$P(x_i|y_j) = \frac{n_c + mp}{n + m}$$

met n het totale aantal instanties van klasse y_j, n_c is het aantal trainingsvoorbeelden van klasse y_j die de waarde x_i aannemen, m is een parameter gekend als de equivalente steekproefgrootte en p is een gebruikersafhankelijke parameter. Als er geen training set beschikbaar is ($n = 0$), dan zal $P(x_i|y_j) = p$. p kan dus beschouwd worden als de a-priori kans dat we de attribuutwaarde x_i

waarnemen tussen records van klasse y_j. De equivalentie steekproefgrootte bepaalt de uitwisseling tussen de a-priori kans p en de waargenomen kans n_c/n.

Maken we gebruik van de m-schatting (met $m = 3$ en $p = 1/3$) dan wordt de voorwaardelijke kans

$P($`Marital Status=Married`|`Yes`$) = (0 + 3 * 1/3) + (3 + 3) = 1/6.$

Veronderstellen we $p = 1/3$ voor alle attributen van de klasse `Yes` en $p = 2/3$ voor alle attributen van de klasse `No`, dan

$$P(\vec{X}|\texttt{No}) = P(\texttt{Howe Owner=No}|\texttt{No}) * P(\texttt{Status=Married}|\texttt{No}) * P(\texttt{Annual Income=\$120K}|\texttt{No})$$
$$= 6/10 * 6/10 * 0{,}0072 = 0{,}0026$$

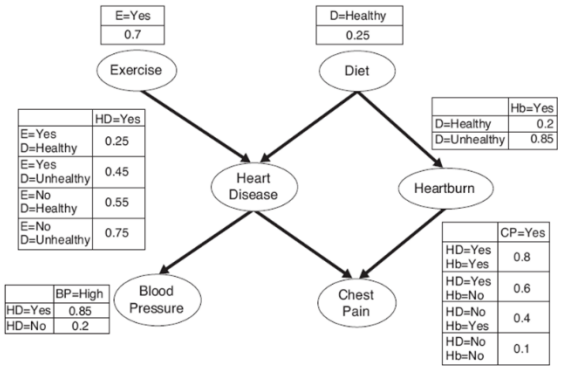
$$P(\vec{X}|\texttt{Yes}) = P(\texttt{Howe Owner=No}|\texttt{Yes}) * P(\texttt{Status=Married}|\texttt{Yes}) * P(\texttt{Annual Income=\$120K}|\texttt{Yes})$$
$$= 4/6 * 1/6 * 1{,}2 * 10^{-9} = 1{,}3 * 10^{-10}$$

De a-posteriori kans voor de klasse No is dan $P(\texttt{No}|\vec{X}) = \alpha * 7/10 * 0{,}0026 = 0{,}0018\alpha$, terwijl de a-posteriori kans voor de klasse Yes gelijk is aan $P(\texttt{Yes}|\vec{X}) = \alpha * 3/10 * 1{,}3 * 10^{-10} = 4{,}0 * 10^{-11}\alpha$.
De classificatiebeslissing wijzigt niet, maar de m-schattingsmethode levert in het algemeen een meer robuuste weg om kansen te schatten als het aantal training records klein is.

6.3.5 *Bayesian Belief Network (BBN)

Een **Bayesian belief network (BBN)**, of eenvoudigweg een Bayesiaans netwerk (probabilistisch netwerk), levert een grafische voorstelling van de kansrelaties tussen een verzameling stochastische variabelen. Er zijn twee kernelementen in een BNN:
1) Een gerichte acyclische graaf die de afhankelijkheids codeert tussen een verzameling variabelen.
2) Een kanstabel die elke node associeert met zijn direct parent node.

Onderstaande figuur toont een voorbeeld BNN gebruikt als model voor hartpatiënten. Elke variabele in het schema is binair gewaardeerd. De parent nodes voor `Heart Disease` (HD) corresponderen met risicofactoren zoals `Exercise` (E) en `Diet` (D). De child nodes van `Heart Disease` corresponderen met symptomen zoals `Chest Pain` (CP) en `High Blood Pressure` (BP).



Figuur 2 - BNN voor het detecteren van hartziekten en maagzuur bij hartpatiënten. *Overgenomen uit Introduction to Data Mining (p. 242) door Tan P., Steinbach M., Kumar V., 2006, Pearson Education. Copyright 2006 Pearson Education Inc.*

De nodes met de risicofactoren bevatten uitsluitend de a-priori kansen, terwijl de nodes voor Heart Disease, Heartburn en hun symptomen bevatten de voorwaardelijke kansen. Om plaats te sparen zijn sommige van de kansen weggelaten. De weggelaten kansen kunnen eenvoudig via de complementregel gevonden worden. Bijvoorbeeld:

$$P(\texttt{Heart Disease=No}|\texttt{Exercise=No, Diet=Healthy})$$
$$= 1 - P(\texttt{Heart Disease=Yes}|\texttt{Exercise=No, Diet=Healthy})$$

$$= 1 - 0{,}55 = 0{,}45$$

6.3.6 Oefeningen

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 naive 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?

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 naive 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=+).

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 1 is. Een student die het antwoord niet kent maar toch correct antwoordt is $1/m$ (met m het aantal antwoordmogelijkheden.
Wat is de kans dat een student het antwoord kent als hij correct antwoordde.

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:

```
age<=30, income=medium, student=yes, credit-rating=fair
```

5) 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.
a) Classificeer het record $(X_1 = 0, X_2 = 0)$.
b) Bereken $P(Y = 1|X_1 = 0, X_2 = 0)$

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