Werkblad week 2

d.r.m.langers@pl.hanze.nl

23 september 2021

1 De formule van Bayes

In de statistiek is de kans P op een uitkomst U gelijk aan de relatieve frequentie van het voorkomen van deze uitkomst in een grote hoeveelheid uitkomsten. Dit wordt genoteerd als $P(U)$. Zo is bijvoorbeeld de kans dat een willekeurige letter uit een Nederlandstalige tekst gelijk is aan een "u" ongeveer gelijk aan 2%, omdat uit tellingen blijkt dat bijna één op de vijftig letters in het Nederlands een "u" is. Een bekender voorbeeld is dat de kans op het gooien van vijf ogen met een eerlijke dobbelsteen is gelijk aan $\frac{1}{6}$, omdat gemiddeld één op de zes worpen een vijftal ogen zal opleveren.

Opgave 1. Wat is de kans dat een willekeurige nucleotide uit een DNA-sequentie thymine betreft als het GC-gehalte van die sequentie 40% is?

Naast deze normale *onvoorwaardelijke* kansen komen ook *voorwaardelijke* kansen voor. Deze meten de waarschijnlijkheid van een uitkomst U_1 gegeven een zekere voorwaarde U_2, genoteerd $P(U_1 \mid U_2)$. Je kunt dit kortweg uitspreken als "P van U_1 gegeven U_2". Zo neemt bijvoorbeeld de kans op de letter "u" toe tot vrijwel 100% onder de voorwaarde dat de voorgaande letter een "q" is, omdat in het Nederlands de "q" nagenoeg altijd gevolgd wordt door een "u". Voor een dobbelsteen hangt het aantal ogen van een worp niet af van het aantal ogen in de vorige worp, dus deze blijft $\frac{1}{6}$.

De kans $P(U_1)$ wordt ook wel de *a priori* kans genoemd en $P(U_1 \mid U_2)$ de *a posteriori* kans, omdat de eerste de waarschijnlijkheid van de uitkomst U_1 geeft vóórdat bekend is of aan de voorwaarde U_2 is voldaan en de tweede de kans geeft dat dit nog het geval is nádat de voorwaarde U_2 is toegepast.

Als de kans op de ene uitkomst niet wordt beïnvloed door een andere uitkomst geldt dat $P(U_1 \mid U_2) = P(U_1)$ voor alle mogelijke waarden van U_1 en U_2. Dan zeggen we dat de uitkomst U_1 *onafhankelijk* is van de voorwaarde U_2. Het al dan niet optreden van de uitkomst hangt dan niet af van de voorwaarde. Zo zijn opeenvolgende worpen van een dobbelsteen onafhankelijk van elkaar, want de kans op een worp hangt niet af van wat de vorige worp opleverde. Opeenvolgende letters in een Nederlandstalige tekst zijn duidelijk niet onafhankelijk, want de kans op een "u" wordt enorm verhoogd als de voorgaande letter een "q" is.

Opgave 2. Hoe groot is de a posteriori kans $P(\text{vijf} \mid \text{oneven})$ op het gooien van een vijf met een eerlijke dobbelsteen, gegeven dat je tijdens die worp een oneven aantal ogen gooit?

Opgave 3. Probeer grofweg de voorwaardelijke kans in te schatten dat in het Nederlands een willekeurige letter een "q" is, onder de voorwaarde dat de volgende letter een "u" is. Is $P(\text{'q'} \mid \text{'u'})$ hetzelfde als $P(\text{'u'} \mid \text{'q'})$?

Gebruikmakend van voorwaardelijke kansen kan ook de kans berekend worden dat meerdere uitkomsten U_1 en U_2 allebei samen plaatsvinden. We noteren dat als $P(U_1, U_2)$. Voor onafhankelijke kansen geldt dat $P(U_1, U_2) = P(U_1) \cdot P(U_2)$. Oftewel, kansen op onafhankelijke uitkomsten mogen vermenigvuldigd worden. Zo is bijvoorbeeld de kans dat je met twee dobbelstenen "snake eyes" gooit (allebei één oog) gelijk aan $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$. Echter, deze product-regel geldt niet langer wanneer de uitkomsten afhankelijk van elkaar zijn.

Opgave 4. De onderstaande twee dobbelstenen zijn aan elkaar vastgelijmd. Wat is de kans dat je 3 gooit met de rechter oranje dobbelsteen onder de voorwaarde dat je twee gooit met de linker blauwe? Zijn de kansen op de uitkomsten van beide dobbelstenen onafhankelijk van elkaar?

Om een samengestelde kans correct te berekenen kunnen we gebruik maken van de uitdrukking

$$P(U_1, U_2) = P(U_1) \cdot P(U_2 \mid U_1)$$

In woorden staat hier dat de kans op de uitkomsten U_1 en U_2 tezamen gelijk is aan het product van de kansen dat:

- ten eerste, uitkomst U_1 optreedt, en
- ten tweede, uitkomst U_2 optreedt, gegeven dat U_1 reeds plaatsvindt.

Deze formule werkt zowel voor afhankelijke als onafhankelijke uitkomsten.

Opgave 5. De kans dat een willekeurige Nederlander een man met kleurenblindheid is bedraagt 4.2%. Als je mag aannemen dat mannen en vrouwen in even grote getalen voorkomen, wat is dan de voorwaardelijke kans om kleurenblind te zijn gegeven dat je een man bent?

Neem bijvoorbeeld de onderstaande kruistabel voor de attributen Outlook en Temperature, afgeleid uit Weka's "weather.nominal" dataset.

	Cool	Mild	Hot	Totaal
Sunny	1	2	2	5
Overcast	1	1	2	4
Rainy	2	3	0	5
Totaal	4	6	4	14

Stel, we willen op basis van deze data de kans op een milde, bewolkte dag bepalen. Lezen we dit rechtstreeks uit de tabel af, dan zien we dat er één instance is met zowel de attributen Mild als Overcast, uit een totaal van 14 instances, wat resulteert in $P(\text{mild}, \text{overcast}) = \frac{1}{14} = 0.071$. Tegelijkertijd volgt uit de randtotalen in de tabel dat $P(\text{mild}) = \frac{6}{14}$ en $P(\text{overcast}) = \frac{4}{14}$. Als de attributen Outlook en Temperature onafhankelijk waren geweest, dan was de kans op een milde, bewolkte dag gelijk geweest aan hun product $P(\text{mild}) \cdot P(\text{overcast}) = \frac{24}{196} = 0.122$. Dit is duidelijk niet in overeenstemming. Outlook en Temperature zijn dus afhankelijk van elkaar.

Wel kunnen we gebruik maken van de correcte uitdrukking $P(\text{mild}, \text{overcast}) = P(\text{mild}) \cdot P(\text{overcast} \mid \text{mild})$. De voorwaardelijke kans $P(\text{overcast} \mid \text{mild})$ kan worden afgelezen door je in de tabel te beperken tot de instances die aan de voorwaarde voldoen, dat wil zeggen de milde dagen in de tweede datakolom. Hierin komen in totaal zes instances voor, waarvan er één het attribuut overcast heeft. Daarom is $P(\text{overcast} \mid \text{mild}) = \frac{1}{6}$. We vinden nu $P(\text{mild}, \text{overcast}) = P(\text{mild}) \cdot P(\text{overcast} \mid \text{mild}) = \frac{6}{14} \cdot \frac{1}{6} = \frac{1}{14} = 0.071$, in overeenstemming met wat we rechtstreeks uit de tabel afgeleid hadden.

Natuurlijk hadden we de condities ook om kunnen keren: de volgorde waarin we de uitkomsten opsommen doet er niet toe. Dit levert voor het gegeven voorbeeld op $P(\text{overcast}, \text{mild}) = P(\text{overcast}) \cdot P(\text{mild} \mid \text{overcast}) = \frac{4}{14} \cdot \frac{1}{4} = \frac{1}{14} = 0.071$. We vinden opnieuw dezelfde uitkomst.

Ga voor jezelf na dat je deze voorwaardelijke en onvoorwaardelijke kansen correct uit de tabel kunt aflezen.

Opgave 6. Bepaal op twee manieren de kans op een koele zonnige dag: enerzijds rechtstreeks op grond van de tabel hierboven; anderzijds gebruikmakend van de product formule voor samengestelde kansen.

Opgave 7. Bepaal uit de tabel de kansen $P(\text{hot})$, $P(\text{rainy})$, $P(\text{hot}, \text{rainy})$, $P(\text{hot} \mid \text{rainy})$, en $P(\text{rainy} \mid \text{hot})$.

In het algemeen geldt altijd dat de kans dat de ene en de andere uitkomst samen optreden dezelfde is als de kans dat de andere en de ene uitkomst samen optreden. In formulevorm betekent dit dat $P(U_1, U_2) = P(U_2, U_1)$. Vullen we hierin de formule voor samengestelde kansen in, dan krijgen we $P(U_1) \cdot P(U_2 \mid U_1) = P(U_2) \cdot P(U_1 \mid U_2)$. Hieruit volgt onmiddellijk

$$P(U_1 \mid U_2) = \frac{P(U_1) \cdot P(U_2 \mid U_1)}{P(U_2)}$$

Deze formule staat bekend als de *formule van Bayes*; hoewel vrij simpel van opbouw is deze formule buitengewoon veelzijdig qua toepassingen!

Zo bestaat bijvoorbeeld ongeveer 0.01% van een nederlandstalige tekst uit de letter "q". Gebruikmakend van de eerdere gegevens kunnen we nu eenvoudig de kans bepalen

dat een willekeurig gekozen letter "u" voorafgegaan wordt door een "q". Neem voor U_1 de uitkomst dat een voorafgaande letter gelijk is aan een "q" en voor U_2 de uitkomst dat een volgende letter gelijk is aan een "u", dan leert de formule van Bayes ons dat $P(\text{'q'} \mid \text{'u'}) = \frac{P(\text{'q'}) \cdot P(\text{'u'} \mid \text{'q'})}{P(\text{'u'})} = \frac{0.0001 \cdot 1.00}{0.02} = 0.005$. Oftewel, een half procent van alle u's wordt voorafgegaan door een q!

Opgave 8. In een patiënten support-groep heeft 10% van alle deelnemers een leverziekte, terwijl 5% van de deelnemers alcoholisten zijn. Uit wetenschappelijk onderzoek is bekend dat 7% van alle mensen met een leverziekte alcoholisten zijn. Bepaal de kans dat een deelnemer in deze groep een leverziekte heeft als je weet dat het een alcoholist betreft.

Opgave 9. Bij een opleiding verschijnen in 4% van alle lessen studenten te laat in de klas. Op dagen wanneer er een staking in het openbaar vervoer plaatsvindt neemt dit aantal toe tot 60%. Dergelijke stakingen vinden gemiddeld eens in de 100 dagen plaats. Wat is de kans dat er op een zekere dag een staking plaatsvindt als er laatkomers in de les blijken zijn?

2 Het "Naive Bayes" algoritme

In de context van het *Naive Bayes* classificatie-algoritme kunnen we de formule van Bayes gebruiken om de kans te berekenen voor de verschillende klasselabels C, gegeven de gemeten waarde van een zeker attribuut A.

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

Laten we opnieuw het Outlook attribuut bekijken van de weather.nominal dataset, nu in samenhang met de te voorspellen klasse (Yes/No) gegeven in het attribuut Play.

	Yes	No	Totaal
Sunny	2	3	5
Overcast	4	0	4
Rainy	3	2	5
Totaal	9	5	14

We willen nagaan welk klasselabel het meest waarschijnlijk is, afhankelijk van of het een zonnige, bewolkte, of regenachtige dag is. Daartoe berekenen we de kans op beide klasselabels afzonderlijk, gegeven de waarde van Outlook.

Stel bijvoorbeeld dat het een zonnige dag is. De formule van Bayes zegt dan

$$P(\text{yes} \mid \text{sunny}) = \frac{P(\text{yes}) \cdot P(\text{sunny} \mid \text{yes})}{P(\text{sunny})}$$

en

$$P(\text{no} \mid \text{sunny}) = \frac{P(\text{no}) \cdot P(\text{sunny} \mid \text{no})}{P(\text{sunny})}$$

Voor bewolkte en regenachtige dagen komen we tot een soortgelijk stel uitdrukkingen.

Nu kunnen we de kansen aan de rechterkant van deze vergelijkingen aflezen uit de data.

- Betreffende de a priori kansen $P(\text{yes})$ en $P(\text{no})$: er zijn negen instances gelabeld "yes" en vijf instances gelabeld "no". Derhalve, $P(\text{yes}) = \frac{9}{14}$ en $P(\text{no}) = \frac{5}{14}$. Soortgelijk is de a priori kans $P(\text{sunny}) = \frac{5}{14}$, aangezien vijf van de veertien instances in totaal een "sunny" Outlook tellen.
- Betreffende de a posteriori kansen $P(\text{sunny} \mid \text{yes})$ en $P(\text{sunny} \mid \text{no})$: van de negen "yes" instances zijn er twee gelabeld "sunny", dus $P(\text{sunny} \mid \text{yes}) = \frac{2}{9}$; op dezelfde wijze is $P(\text{sunny} \mid \text{no}) = \frac{3}{5}$, want drie van de vijf "no" instances hebben het label "sunny".

Vullen we deze waarden in en vereenvoudigen we de uitkomst, dan krijgen we

$$P(\text{yes} \mid \text{sunny}) = \frac{\frac{9}{14} \cdot \frac{2}{9}}{\frac{5}{14}} = \frac{2}{5} = 0.40$$

en

$$P(\text{no} \mid \text{sunny}) = \frac{\frac{5}{14} \cdot \frac{3}{5}}{\frac{5}{14}} = \frac{3}{5} = 0.60$$

Oftewel, voor een willekeurige zonnige dag is er een 40% kans dat er wel gespeeld wordt en een 60% kans dat er niet gespeeld wordt. Merk op dat de som van deze kansen 100% is, zoals verwacht, aangezien de klasse ofwel "yes" ofwel "no" moet zijn!

Kortom, als we niets weten behalve dat een zekere instance een "sunny" Outlook heeft, dan kunnen we dit het beste als "no" labelen. We behalen dan naar verwachting een accuracy van 60% en een error rate van 40%.

Opgave 10. Gebruikmakend van de tabel, wat zou je voorspelling zijn als de Outlook "overcast" is? En wat als deze "rainy" is?

Tot dusverre hebben we maar één attribuut tegelijkertijd beschouwd (hier: Outlook). In dit geval konden we de uitkomsten eenvoudiger rechtstreeks afleiden uit de tabel door naar de rijen te kijken. In de rij met het label "sunny" komen vijf instances voor, waarvan 2 met het label "yes" en 3 met het label "no", zodat $P(\text{yes} \mid \text{sunny}) = \frac{2}{5}$ en $P(\text{no} \mid \text{sunny}) = \frac{3}{5}$. De bovenstaande berekening met de formule van Bayes was dus alleen maar omslachtig. Dit verandert echter wanneer we met diverse attributen tegelijk rekening willen houden!

Als we informatie over meerdere attributen $A_1, A_2, \ldots, A_n$ beschikbaar hebben, dan kan de kans op een klasselabel C op basis van al deze attributen tezamen worden genoteerd als $P(C \mid A_1, A_2, \ldots, A_n)$. Dan kan de formule van Bayes worden herschreven tot

$$P(C \mid A_1, A_2, \ldots, A_n) = \frac{P(C) \cdot P(A_1 \mid C) \cdot P(A_2 \mid C) \cdot \ldots \cdot P(A_n \mid C)}{P(A_1, A_2, \ldots, A_n)}$$

De bovenstaande formule neemt aan dat alle attributen onderling onafhankelijk zijn. Dit is eigenlijk best een onrealistische aanname is. Echter, omdat we de kansen alleen willen gebruiken om te kiezen welk klasselabel het meest waarschijnlijk is, werkt deze formule in de praktijk meestal toch redelijk goed. Desalniettemin is dit de reden waarom het algoritme "naief" wordt genoemd.

Stel, een dag is zonnig en heet, met normale luchtvochtigheid, en er staat wel wind. De formule van Bayes gaat er dan als volgt uitzien

$$P(\text{yes} \mid \text{sunny}, \text{hot}, \text{normal}, \text{true}) = \frac{P(\text{yes}) \cdot P(\text{sunny} \mid \text{yes}) \cdot P(\text{hot} \mid \text{yes}) \cdot P(\text{normal} \mid \text{yes}) \cdot P(\text{true} \mid \text{yes})}{P(\text{sunny}, \text{hot}, \text{normal}, \text{true})}$$

en

$$P(\text{no} \mid \text{sunny}, \text{hot}, \text{normal}, \text{true}) = \frac{P(\text{no}) \cdot P(\text{sunny} \mid \text{no}) \cdot P(\text{hot} \mid \text{no}) \cdot P(\text{normal} \mid \text{no}) \cdot P(\text{true} \mid \text{no})}{P(\text{sunny}, \text{hot}, \text{normal}, \text{true})}$$

Een paar joekels van formules. Voor een computer nauwelijks een probleem, maar voor ons wel, dus vanaf hier gaan we vereenvoudigen. Door nauwkeurig alle instances in de data te tellen om de voorwaardelijke en onvoorwaardelijke kansen te bepalen, vergelijkbaar als hierboven, verkrijgen we

$$P(\text{yes} \mid \text{sunny}, \text{hot}, \text{normal}, \text{true}) = \frac{\frac{9}{14} \cdot \frac{2}{9} \cdot \frac{2}{9} \cdot \frac{6}{9} \cdot \frac{3}{9}}{P(\text{sunny}, \text{hot}, \text{normal}, \text{true})} = \frac{0.007}{P(\text{sunny}, \text{hot}, \text{normal}, \text{true})}$$

en

$$P(\text{no} \mid \text{sunny}, \text{hot}, \text{normal}, \text{true}) = \frac{\frac{5}{14} \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{3}{5}}{P(\text{sunny}, \text{hot}, \text{normal}, \text{true})} = \frac{0.010}{P(\text{sunny}, \text{hot}, \text{normal}, \text{true})}$$

Merk op dat we de waarde van de noemer van de breuk $P(\text{sunny}, \text{hot}, \text{normal}, \text{true})$ niet hebben ingevuld. Voor allebei de formules is deze noemer hetzelfde. In dit geval is er geen enkele instance in de dataset met precies deze combinatie van attributen, maar als we de kans gelijk zouden stellen aan nul krijgen we problemen wegens delen door nul. Dat werkt dus niet. Echter, we blijken deze waarde helemaal niet nodig te hebben!

Om te beginnen, als we slechts geïnteresseerd zijn in welk klasselabel we dienen toe te kennen, dan kan ook uit bovenstaande twee uitkomsten al worden afgeleid dat de kans op "no" groter is dan die op "yes". Namelijk, 0.010 is groter dan 0.007, en dat blijft zo, door wat voor kans je ook deelt.

Maar zelfs als we de kansen exact willen bepalen kunnen we met deze uitkomsten volstaan. Immers, omdat er slechts twee mogelijke klasselabels zijn moeten de kansen samen optellen tot 1.0 (of 100%). Dit kunnen we eenvoudig bereiken door de uitdrukkingen als volgt te schalen.

$$P(\text{yes} \mid \text{sunny}, \text{hot}, \text{normal}, \text{true}) = \frac{0.007}{0.007 + 0.010} = 0.407$$

en

$$P(\text{no} \mid \text{sunny}, \text{hot}, \text{normal}, \text{true}) = \frac{0.010}{0.007 + 0.010} = 0.593$$

We concluderen dat de kans dat er gespeeld wordt op een dergelijke dag 40.7% is, en de kans dat er niet gespeeld wordt 59.3% is. Wederom luidt de conclusie dat het klasselabel "no" het meest waarschijnlijk is, gegeven de waarden van de attributen van deze instance.

Opgave 11. Herhaal de bovenstaande berekening voor regenachtige en koele dagen met hoge luchtvochtigheid, waarbij onbekend is of het waait of niet. Dat wil zeggen, bereken $P(\text{yes} \mid \text{rainy}, \text{cool}, \text{high})$ en $P(\text{no} \mid \text{rainy}, \text{cool}, \text{high})$ en leid hieruit het meest waarschijnlijke klasselabel af.

Opgave 12. Bekijk de inhoud van Weka's "contact-lenses" arff-datafile. Hierin wordt aan de hand van vier kenmerken van patiënten aangegeven welk type contactlens het meest geschikt is. Komt de voorspelling van Naive Bayes overeen met de werkelijke data voor een patiënt met de kenmerken Presbyopic, Myope, No, en Normal? Bereken voor elk van de drie mogelijke klasselabels de voorspelde kans.

In de tabel hierboven zie je dat er geen enkele instance is die een Outlook "overcast" combineert met Play gelijk aan "no". Het gevolg zou zijn dat Naive Bayes aan elke nieuwe instance die "overcast" bevat automatisch het label "yes" wil toekennen, ongeacht wat de andere attributen zeggen. Het algoritme heeft uit de trainingsdata geleerd dat op bewolkte dagen altijd gespeeld wordt. In de praktijk kan dit tot problemen leiden: op basis van het ene attribuut zou het ene label een kans nul kunnen krijgen, en op basis van een ander attribuut zou het andere label een kans nul kunnen krijgen, met als gevolg dat er geen enkel label is waaraan wel een positieve kans wordt toegekend.

Om dit te voorkomen wordt soms gebruik gemaakt van een *Laplace-estimator*. Je telt dan een constante waarde op bij elke cel uit de kruistabel. Of, anders gezegd, je begint de aantallen instances niet te tellen vanaf nul, maar vanaf een andere gekozen waarde. Dit houdt er eigenlijk rekening mee dat de trainingsdata een beperkte omvang hebben. Het idee is dat als je maar voldoende data zou blijven verzamelen, vroeg of laat elke cel in de tabel niet meer nul zou zijn.

Opgave 13. Herhaal de berekening voor de "contact-lenses" arff-datafile in de voorgaande opgave, maar gebruik nu overal een Laplace-estimator van 1 om het "zero-frequency problem" op te lossen.

Opgave 14. Ga voor jezelf na dat Naive Bayes noodzakelijk exact dezelfde resultaten oplevert als OneR voor alle datasets waarin maar één attribuut beschikbaar is om het klasselabel te voorspellen. Evenzo, ga na dat naive Bayes hetzelfde oplevert als ZeroR voor alle datasets waarin nul attributen beschikbaar zijn.