Universiteit van Amsterdam

UvA science park

Statistisch Redeneren: Week 2

Tessa Klunder & Vincent Hagen

Datum:
15 april 2014

1 Opgave 1

a) $F(x) = \int_3^x f(u)du$. Waarbij $f(u)$ de kansdichtheidsfunctie is. De verdeling functie is tot 3 gelijk aan 0. Vanaf 3 is hij $\frac{1}{6} = 9 - 3$, waarna hij tot en met 9, per integer stijgt met $\frac{1}{7}$. $F(x) = \left[\frac{1}{6}u\right]_3^x = \frac{1}{6}x - \frac{1}{2}$ waar F(x) tussen 0 en 1 ligt. Dus voor $x < 3$ is $F(x) = 0$, tussen 3 en 9 is $F(x) = \frac{1}{6}x - \frac{1}{2}$, voor $x > 9$ $F(x) = 1$.

b) De kans op $P([-10, 3]) = 0$, omdat de verdeling maar loopt van 3 tot en met 9 en deze continue is. Dus $P([-10, 3]) = 0$.

c) Sinds het uniform verdeeld is nemen we het aantal getallen wat binnen ons interval ligt delen door het totaal aantal mogelijke waardes in ons universum. Dus als $c(a, b)$ gedefinieerd is als de functie die het aantal elementen telt in onze universum tussen (a, b). Dan is de kans op $P([a, b]) = \frac{c(a,b)}{c(-\infty,\infty)} = \frac{c(a,b)}{6}$.

2 Opgave 2

a) $U = \{0...n\}$, waarbij n het aantal worpen is. In ons universum worden dus het aantal keren dat de worp munt was opgenomen.

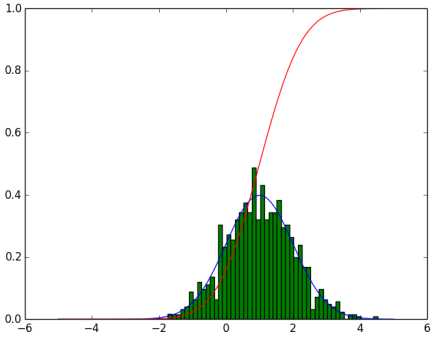
b) $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$, met n het aantal worpen.

c) Binomiale kansverdeling.

d) Zie som.py bijgeleverd bij de opgave.

3 Opgave 3

Code bijgeleverd bij opgave. File: verd.py. In de main kan code worden uitgecomment om bv. de dichtheidsfunctie te verbergen.



Figuur 1: Normaalverdeling verdelings functie en dichtheids functie. 1000 "willekeurige" trekkingen weergegeven als histogram

4 Practicum

4.1 Naive Bayes Classifier

De Naive Bayes Classifier is een methode om de kansverdeling over verschillende klassen te voorspellen. Hij doet dit door middel van "test samples". Dus hij gebruikt voorbeelden om van te leren. Waarom heet het de "Naive" Bayes Classifier? Dit is, omdat aangenomen wordt dat de de variabelen onafhankelijk zijn. Dus in ons geval worden lengte, gewicht en schoenmaat gezien als gescheiden variabelen, de één sluit de anderen niet uit. Een groot voordeel hiervan is dat de *Joint Distribution table* hier aanzienlijk eenvoudiger van wordt. Statistische onafhankelijkheid wordt geschreven als: $P(A \cap B) = P(A)P(B) \Leftrightarrow P(A) = \frac{P(A \cap B)}{P(B)} \Leftrightarrow P(A) = P(A|B)$. Hetzelfde is te doen voor $P(B) = P(B|A)$. De twee gebeurtenissen zijn dus onafhankelijk van elkaar. Dit kan ook gedaan worden voor conditionele onafhankelijkheden. Je krijgt dan: $P(A \cap B|C) = P(A|C)P(B|C)$.

Wat het theorema van Bayes beschrijft is de relatie tussen de kans op A en B en de conditionele kans op A gegeven B en visa versa. Er zijn twee axioma's die samen dit theorema vormen. De eerste is: $P(A \cap B) = P(B|A)P(A)$. Daarnaast kan de kans op A gegeven B geschreven worden als:

$P(A|B) = \frac{P(A \cap B)}{P(B)}$

Als we deze twee combineren krijgen we het theorema van Bayes:

$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

Nou hebben we voor deze opdracht verschillende stochasten: lengte(L), gewicht(G) en schoenmaat(S). In de onderstaande formules staan alleen de hoofdletters, dit staat echter voor: $L = l, G = g, S = s$. We hebben dit even weggelaten om het overzichtelijk te houden. Wat we uiteindelijk willen berekenen, is de kans (X) of iemand een man of een vrouw is, afhankelijk van die stochasten. Als we deze variabelen invullen in het theorema van Bayes krijgen we:

$P(X_i|L, G, S) = \frac{P(L,G,S|X_i)P(X_i)}{P(L,G,S)}$

Hierbij is $P(X_i)$ de a priori kans. Dus de kans op een man of een vrouw voordat er getest is. Vervolgens moeten we de kans dat iemand een vrouw is ($X = 1$) vergelijken met de kans dat die persoon een man is ($X = 0$). We vullen dus X in:

$\frac{P(L,G,S|X=0)P(X=0)}{\cancel{P(L,G,S)}} > \frac{P(L,G,S|X=1)P(X=1)}{\cancel{P(L,G,S)}}$

We komen dan op de bovenstaande vergelijking uit. De onderkant van de breuk is doorgestreept, omdat deze hetzelfde zijn. In de vergelijking vallen ze dus weg. Als we dit netjes omschrijven en de overgebleven kansen door elkaar delen, krijgen we de Naive Bayes Classifier:

$\frac{P(L=l,G=g,S=s|X=0)}{P(L=l,G=g,S=s|X=1)} > \frac{P(X=1)}{P(X=0)} \rightarrow 1$

De tweede breuk zijn de a priori kansen door elkaar gedeeld. De waarde hiervan is 1, omdat we uitgaan van de verdeling van mannen en vrouwen in de hele wereld. In onze test zijn er maar zes vrouwen, maar we kiezen voor een algemene 50/50 verdeling. Dit resulteert dus in de waarde 1.

4.2 Kansdichtheidsverdelingen

Om het programma te laten berekenen of iemand een man of een vrouw kan zijn hebben we twee functies gebruikt. De verdelingsfunctie F_X en de kansdichtheidsfunctie f_X. Voor deze functies geldt het volgende:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(u)du$$

De kansdichtheidsfunctie is de afgeleide van de kansverdelingsfunctie F_X:

$f_X(x) = \frac{d}{dx}F_X(x)$

Hierdoor krijgt de kansdichtheidsfunctie $f_X(x)$ een belangrijke eigenschap. Namelijk het totale oppervlak onder de grafiek is 1:

$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

Dit is handig, omdat je hierdoor eigenlijk de "verspreiding van de totale kans" ziet in een grafiek. $P(U) = 1$ is immers één van de axioma's van de kansrekening. Een voorbeeld hiervan is te zien in opdracht 3 op bladzijde 1. Met bovenstaande gegeven, kunnen we kijken hoe deze kunnen worden gebruikt. We kiezen voor als interval $[a, b]$ en vullen deze in:

$P(X > a) = 1 - P(X \leq a)$

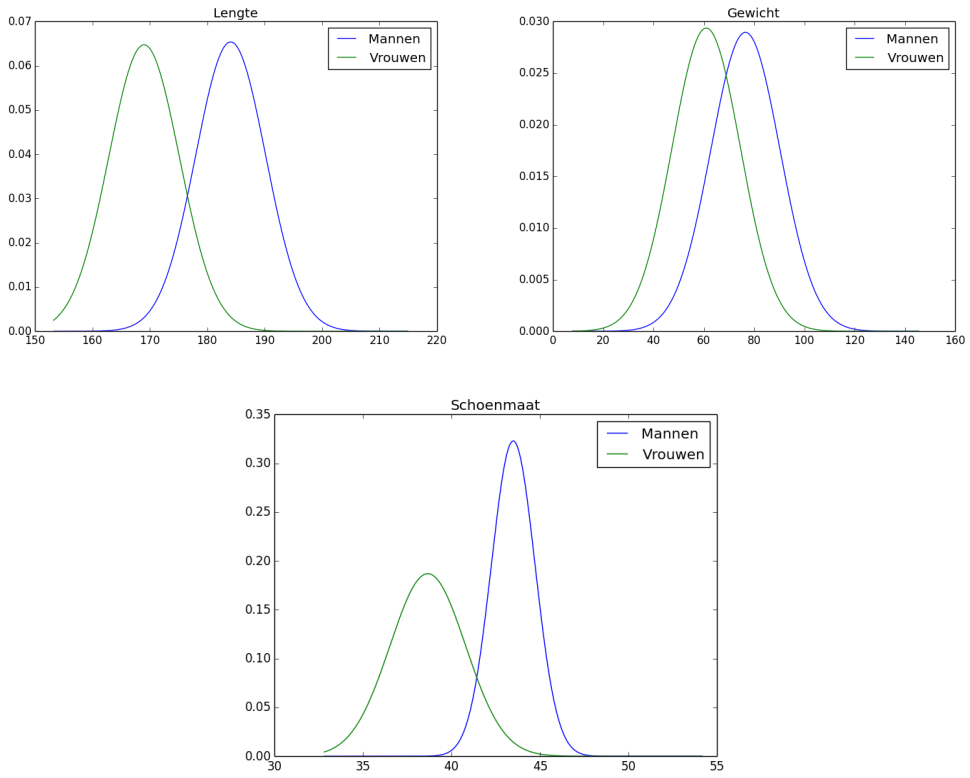
$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a)$

We hebben nu de twee verdelingsfuncties voor het interval $[a, b]$. Om het te berekenen, gebruiken we de kansdichtheidsfunctie. Dit is immers, zoals we eerder zeiden, de afgeleide functie van F_X. Uiteindelijk komen we op het volgende uit:

$$P(a < X \leq b) = F_X(b) - F_X(a) = \int_{a}^{b} f_X(x)dx$$

Maar om het programma nou te laten "leren" hebben we een normaal functie gemaakt. Deze is op basis van het gemiddelde en de standaard deviatie van de schoenmaat, lengte en het gewicht. Als test hebben we hiervoor een hele grote leerset genomen. In de eerdere pogingen, met een kleine leerset kwamen te veel fouten naar voren. Als experiment hebben wij nu dus gekozen voor een leerset van $n - 1$, waarbij n de totale set. We doen dit voor alle n om te bepalen of het een man of een vrouw is. Door dit voor alle waarden te doen, hebben wij onze *confusion matrix* kunnen opstellen. Deze staat onderaan de volgende bladzijde.

4.3 Resultaten



Figuur 2: Kansverdeling functies

In de kansverdeling functie van de schoenmaat is te zien dat de opgenomen sample vrouwen een uiteenlopende schoenmaat hebben. Dit is te verklaren door het kleine aantal vrouwen opgenomen. Daarnaast is te zien, dat ondanks het kleine aantal vrouwen in onze leer set, de lengte en het gewicht zich hetzelfde gedragen als die van een grotere leer set van de mannen.

4.4 Confusion Matrix

De matrix geeft weer hoeveel seksen er goed zijn voorspeld en hoeveel fout. Van beiden maar één fout. Als we gaan kijken naar de fouten en met de bovenstaande tabellen in ons achterhoofd, zien we dat vooral de lengte en het gewicht voor de fout zorgen. De fouten waren: $[F, 82, 180, 42]$ en $[M, 55, 170, 42]$. Hierbij werd de eerste als man gezien en de tweede als vrouw. De schoenmaat is hetzelfde, dus de andere twee variabelen hebben voor de doorslag gezorgd. Als er naar de grafieken gekeken wordt, is te zien dat dit inderdaad overeenkomt met de oppervlaktes onder de grafiek op de des betreffende punten. Bij een gewicht van $82kg$ en een lengte van $182cm$ is de grafiek inderdaad hoger voor de kans op een man.

		Voorspeld	
		man	vrouw
Werkelijk	man	26	1
	vrouw	1	5

Figuur 3: Confusion matrix van ons programma. n keer leerset $n-1$ genomen en de ene getest.