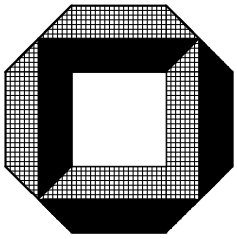
Proseminar

Künstliche Intelligenz



Universität Karlsruhe (TH)
Fakultät für Informatik
Institut für Algorithmen und Kognitive Systeme

Prof. Dr. J. Calmet
Dipl.-Inform. A. Daemi

Wintersemester 2003/2004

Institut für Algorithmen und Kognitive Systeme
Fakultät für Informatik
Universität Karlsruhe
Am Fasanengarten 5
76 128 Karlsruhe

Unsicherheit und Fuzzy Logik

Jan Niehues
Kay Rottmann

Inhaltsverzeichnis

Kapitel 3

Unsicherheit und Fuzzy Logik

3.1 Einleitung

In unserer Ausarbeitung gehen wir zunächst auf die Grundlagen der Wahrscheinlichkeitsrechnung ein, die im folgenden benutzt werden. Dabei soll insbesondere ein Einblick in die Themen bedingte und unbedingte Zufallsvariablen gegeben werden und dargestellt werden wie man aus gegebenen Abhängigkeiten die Wahrscheinlichkeiten von beliebigen Ereignissen berechnen kann.
Anschließend soll das Bayes Netzwerk als Darstellungsmöglichkeit eingeführt werden. Dabei wird verstärkt sowohl auf exakte als auch approximative Inferenz eingegangen.
Der dann folgende Teil beschäftigt sich mit der Dempster Shafer Theorie, um eine weitere Möglichkeit der Entscheidungsfindung zu geben.
Abschließend wird auf Fuzzy Sets und Fuzzy Logic eingegangen, die eine Verallgemeinerung der klassischen Mengenlehre und Booleschen Logik darstellen.

3.2 Grundlagen der Wahrscheinlichkeitstheorie

3.2.1 Begriffserklärung

In der Wahrscheinlichkeitstheorie arbeitet man mit *Wahrscheinlichkeiten* für bestimmte *Ereignisse*. Diese Wahrscheinlichkeiten repräsentieren den *Grad der Sicherheit*, dass das zugehörige Ereignis in dem betrachteten System eintrifft. Die Wahrscheinlichkeiten sind Werte aus dem Intervall $[0, 1]$, wobei ein Ereignis mit der Wahrscheinlichkeit 0 einem unmöglichen Ereignis und 1 einem immer erfüllten Ereignis entspricht.
Zufallsvariablen bilden das Hauptinstrument, um mit Wahrscheinlichkeiten zu arbeiten. Sie stehen dabei für Ereignisse bzw. Zustände, deren Wahrscheinlichkeiten man nicht von vornherein weiß.
Zufallsvariablen können Werte aus einer bestimmten Menge annehmen. So unterscheidet man meistens drei verschiedene Typen von Zufallsvariablen.

- *boolean* Variablen: Diese können nur Werte aus der Menge {Wahr, Falsch}[1] annehmen. Beispielsweise die Variable *istStudent* ist entweder *true* oder *false*.

- *diskrete* Variablen: Diese Variablen nehmen Werte aus einer endlichen (abzählbaren) Menge an. Die Werte in der Menge dürfen sich dabei nicht in ihrer Bedeutung überschneiden. Ein Beispiel für eine diskrete Variable ist *inSemester*, welche Werte aus der Menge der Natürlichen Zahlen annehmen kann.

- *stetige* Variablen: Diese Variablen nehmen Werte aus einer Menge, die stetig ist, an. Zum Beispiel die Temperatur ist eine stetige Variable.

Im folgenden soll die Funktion P für die Wahrscheinlichkeit eines Ereignisses benutzt werden.
Beispiel: $P(StudentVD = true) = 0.3$
Interpretation des Beispiels: Die Wahrscheinlichkeit dafür, dass sich ein Student im Vordiplom befindet, beträgt genau 0.3. Im folgenden werden wir auch häufig auf *atomare Ereignisse* zurückgreifen. Atomare Ereignisse beschreiben exakt einen Zustand, in dem sich ein betrachtetes System befindet. Ein solches Ereignis ordnet jeder Zufallsvariable aus dem System einen Wert aus ihrer Umgebung zu. Ein atomares Ereignis hat dabei ebenfalls einige Eigenschaften.

- Es kann nur eines von zwei verschiedenen atomaren Ereignissen der Fall sein. Als Beispiel wieder die Zufallsvariable *StudentVD*. Ein Student kann nicht im Vordiplom und gleichzeitig doch im Vordiplom sein (wir beschränken uns auf den Fall, dass nur ein Fach studiert wird).

- von allen atomaren Ereignissen ist eines immer der Fall. Das heißt, es gibt in dem betrachteten System nie einen Zustand, dem kein atomares Ereignis zugeordnet ist.

- jedes atomare Ereignis sagt über eine Zufallsvariable für jeden Wert aus ihrer Menge aus, ob der Wert richtig oder falsch ist.

- die Wahrscheinlichkeit eines Ereignisses ist äquivalent zu der Disjunktion aller atomaren Ereignisse, die Elemente des Ereignisses sind.

3.2.2 Die Kolmogorov Axiome

Die Kolmogorov Axiome[2] bilden eine Grundlage der Wahrscheinlichkeitstheorie, aus der alle weiteren Formeln und Sätze geschlossen werden können.
Unter anderem sind dies:

- Für alle Ereignisse a gilt: $0 \leq P(a) \leq 1$ Die Wahrscheinlichkeit eines Ereignisses liegt also im Intervall $[0, 1]$

[1] im weiteren Dokument werden die englischen Bezeichner true und false anstelle von wahr und falsch verwendet

[2] benannt nach dem Russischen Mathematiker Andrei Kolmogorov

- $P(\Omega) = 1$, wobei Ω die Menge aller atomaren Ereignisse ist. Also die Wahrscheinlichkeit, dass eins von allen Ereignissen eintritt ist 1.
- Die Wahrscheinlichkeit für eine Disjunktion ist:
  $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

Ein wichtiges Beispiel für eine Folgerung aus den Kolmogorov Axiomen ist die Negation von Variablen. So gilt:

$$P(\neg a) = 1 - P(a) \tag{3.1}$$

3.2.3 Unbedingte Wahrscheinlichkeiten

Unbedingte Wahrscheinlichkeiten beschreiben die Wahrscheinlichkeit für ein Ereignis ohne weitere Kenntnisse über andere, dieses Ereignis beeinflussende, Ereignisse zu haben. Beispiel: $P(StudentVD = true) = 0.3$.
Häufig stellt man die Wahrscheinlichkeiten von allen Belegungen einer Zufallsvariable in einem Vektor dar, anstatt alle Wahrscheinlichkeiten in Gleichungen hinzuschreiben.
Also anstatt:
$P(StudentVD = true) = 0.3$
$P(StudentVD = false) = 0.7$
schreibt man: $P(StudentVD) =< 0.3, 0.7 >$
Dieser Vektor entspricht der Wahrscheinlichkeitsverteilung der Variable. Ebenso kann man die Wahrscheinlichkeitsverteilung von mehreren Zufallsvariablen zu einer Tabelle zusammenfassen. Diese Tabelle wird *joint probability distribution* genannt. Wenn diese Tabelle über alle Zufallsvariablen des Systems erstellt wird, nennt man sie auch *full joint probability distribution*. In einer full joint probability distribution sind die Wahrscheinlichkeiten von allen atomaren Ereignissen eingetragen. Beispiel:

		StudentVD	
		true	false
Alter	20-22	0.15	0.05
	23-25	0.1	0.2
	> 25	0.05	0.45

Tabelle 1: Beispiel für *full joint distribution*

Für stetige Zufallsvariablen ist eine Aufstellung der Verteilung als Tabelle oder Vektor nicht möglich. Anstatt der Tabelle benutzt man dann häufig Funktionen, mit deren Hilfe man die Wahrscheinlichkeiten berechnen kann. Dabei entspricht die Wahrscheinlichkeit für ein Intervall dem Integral der Funktion über diesem Intervall. Diese Funktionen werden dann auch Dichtefunktionen genannt.

3.2.4 Bedingte Wahrscheinlichkeiten

Was bis jetzt noch nicht geklärt ist, ist die Frage wie man Wahrscheinlichkeiten von Ereignissen berechnet, die in Zusammenhang mit anderen bereits bekannten Ereignissen stehen. Zum Beispiel die Wahrscheinlichkeit, dass ein Student

zwischen 20 und 22 ist unter der Bedingung, dass er im Vordiplom ist. Offensichtlich ist diese Wahrscheinlichkeit größer als die des unbedingten Ereignisses der Student ist zwischen 20 und 22 Jahre alt. Die Schreibweise für diesen Sachverhalt ist dabei:
$P(20 \leq Alter \leq 22|StudentVD = true)$
Diese sogenannten *bedingten Wahrscheinlichkeiten* sind dabei folgendermaßen über unbedingte Wahrscheinlichkeiten definiert:

$$P(a|b) = \frac{P(a \wedge b)}{P(b)} \tag{3.2}$$

Nach Umformen der Gleichung erhält man die *Produkt Regel*:

$$P(a \wedge b) = P(a|b)P(b) \tag{3.3}$$
$$P(a \wedge b) = P(b|a)P(a) \tag{3.4}$$

Aus 3.3 und 3.4 kann man nach Gleichsetzen und Division durch $P(a)$ die im folgenden noch sehr wichtige Regel von Bayes herleiten:

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)} \tag{3.5}$$

Diese Regel gibt eine Möglichkeit die Ursache (a) und Wirkung (b) in der Formel 3.2 zu vertauschen.
Wenn $P(a|b) = P(a)$ so heißen die Ereignisse a und b voneinander unabhängig.

3.2.5 Inferenz aus der Full Joint Distribution

Aus der *full joint distribution* kann man alle Wahrscheinlichkeiten, die das System betreffen, berechnen, was im folgenden gezeigt werden soll. Um die Wahrscheinlichkeit für den Fall, dass eine oder mehrere Zufallsvariablen einen bestimmten Wert annehmen, summiert man aus der *full joint distribution* alle Wahrscheinlichkeiten auf, in denen die Zufallsvariablen diesen bestimmten Wert haben. Also für den Fall, dass Y die betrachtete Zufallsvariable ist:

$$P(Y) = \sum_z P(Y, z) \tag{3.6}$$

Also in unserem Beispiel:
$P(20 \geq Alter \geq 22) = P(20-22 \wedge true) + P(20-22 \wedge false) = 0.15 + 0.05 = 0.2$

Häufig ist es jedoch wichtiger die bedingten Wahrscheinlichkeiten anstelle der unmittelbaren Wahrscheinlichkeiten zu benutzen, womit man dann die folgende Formel erhält:

$$P(Y) = \sum_z P(Y \mid z)P(z) \tag{3.7}$$

Die in der Formel 3.2 benutzten Wahrscheinlichkeiten kann man unmittelbar nach der Formel 3.6 aus der *full joint distribution* berechnen. Die Division durch $P(b)$ in der Formel für bedingte Wahrscheinlichkeiten nennt man auch Normalisieren und soll im folgenden durch einen der Division entsprechenden Faktor

α ersetzt werden. Wenn man nun also die Wahrscheinlichkeit von $P(X \mid e)$ berechnen möchte erhält man:

$$P(X \mid e) = \alpha P(X, e) = \alpha \sum_{y} P(X, e, y) \tag{3.8}$$

Dabei sind die y alle möglichen Kombinationen der nicht betrachteten Zufallsvariablen. Da die *full joint distribution* eine Tabelle mit mindestens 2^n Einträgen ist[3], kann man leicht sehen, dass die Laufzeit in $O(2^n)$ liegt - schließlich müssen alle Wahrscheinlichkeiten aus der Tabelle betrachtet und eventuell aufsummiert werden. Da in der Realität Systeme mit sehr vielen Zufallsvariablen betrachtet werden, ist dieses Verfahren der Inferenz im Normalfall unbrauchbar. Daher soll im folgenden ein anderer Weg aufgezeigt werden, die Informationen zu berechnen.

3.3 Bayes Netzwerke

3.3.1 Spezifikation von Bayes Netzwerken

Das Bayes Netzwerk ist eine Datenstruktur, in der sich Zusammenhänge der Zufallsvariablen aus der Anordnung ergeben. Das Netzwerk ist ein gerichteter Graph mit folgenden Eigenschaften:

- Die Zufallsvariablen des Systems, das man mit dem Netzwerk darstellen möchte, sind die Knoten im Graph. Dabei ist es egal, ob die Zufallsvariablen stetig oder diskret sind.
- Die Menge der gerichteten Kanten verbindet Knoten miteinander, die in direktem Zusammenhang stehen. Der Knoten von dem aus die Kante abgeht heißt Vaterknoten zu dem Knoten, zu dem die Kante führt.
- Zu jedem Knoten X gibt es eine Wahrscheinlichkeitsverteilung $P(X|Vater(X))$, die den Einfluss der Vaterknoten widerspiegelt.
- Der Graph ist azyklisch, das heißt es gibt keinen gerichteten Zyklus in dem Graph

Die Wahrscheinlichkeitsverteilung der Knoten lässt sich recht einfach als Tabelle, die sogenannte CPT[4] darstellen (siehe Beispiel). In der Tabelle ist für jede mögliche Kombination von Eigenschaften der Vaterknoten die Wahrscheinlichkeit für die Zufallsvariable in einer Zeile angegeben. Falls es sich um eine Boolean-Variable handelt, so liegt es nah, nur die Wahrscheinlichkeit für den Fall, dass das Ereignis erfüllt ist, anzugeben, da sich der andere Fall über das *Gegenereignis* $P(X) = 1 - P(\neg X)$ leicht berechnen lässt. Für Variablen ohne Vaterknoten hat man nur eine einzige Zeile, die aus den Wahrscheinlichkeiten für die verschiedenen Belegungen der Variable besteht. Abbildung 3.1 zeigt ein Bayes Netzwerk, in dem der Status eines Studenten (Vordiplom oder Hauptdiplom) von seinem Alter und seinem Fleiß abhängt. Sein Status wiederum

[3] für den Fall, dass alle n Variablen boolesch sind, im anderen Fall wären es noch mehr
[4] engl.: conditional probability Table

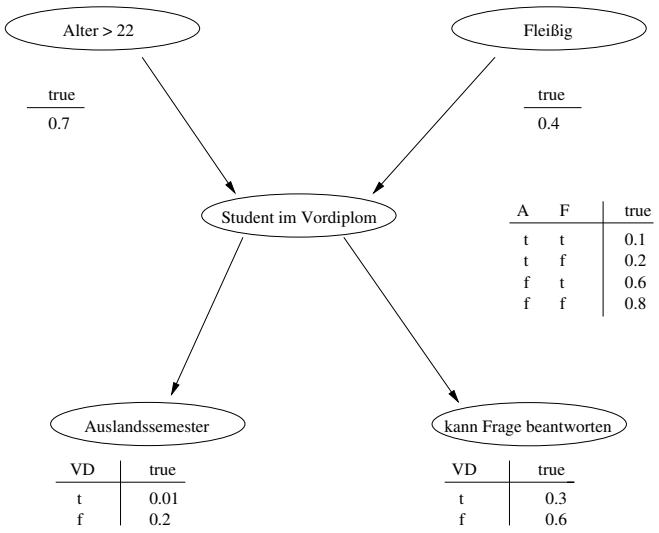


Abbildung 3.1:

beeinflusst, ob er ein Auslandssemester macht oder nicht, bzw. ob er eine Fachspezifische Frage beantworten kann oder nicht[5].

3.3.2 Konstruktion von Bayes Netzwerken

Hier soll jetzt eine Methode angegeben werden, wie man ein Bayes Netzwerk für ein System von Zufallsvariablen konstruieren kann. Man geht von den Zufallsvariablen aus, die Ursachen für andere Zufallsvariablen sind. Diese fügt man als Knoten in das Netzwerk ein. Nach und nach werden dann die Knoten zu dem Netzwerk hinzugefügt, die von den bereits aufgenommenen Zufallsvariablen *direkt* beeinflusst werden und fügt gerichtete Kanten von den Zufallsvariablen, die die Ursache repräsentieren zu der neu eingefügten Zufallsvariable ein. Zum Beispiel seien in dem obigen Netzwerk alle Knoten bis auf *Frage* eingezeichnet. Das Auslandssemester hat offensichtlich keinen (oder nur sehr geringen) Einfluss darauf, ob der Student die Fachfrage beantworten kann. *Alter* und *fleißig* haben sicher einen Einfluss, jedoch ist dieser nur indirekt, da das Wissen ja schließlich über das Studium an sich kommt. Daher ist *Student* der einzige Vaterknoten von *Frage*. An diesem Beispiel sieht man außerdem, dass Faktoren, die nur geringen Einfluss haben unter Umständen unbeachtet bleiben, um die Komplexität des Netzwerkes zu begrenzen (hier mit *Auslandssemester* geschehen). Allgemein sollte versucht werden, die Anzahl der Vorgänger zu begrenzen um die Komplexität der CPT's möglichst gering zu halten. Zum Beispiel hat man bei 30 Zufallsvariablen, die jeweils 5 Eltern haben in einem Bayes Netzwerk 960 Zahlenwert für die komplette Beschreibung. Im Gegensatz dazu hat man in der *full joint distribution* über eine Milliarde Einträge. Anhand der Abbildung

[5]Im folgenden werden dafür die Abkürzungen A für $Alter > 22$, FL für $Fleißig$, VD für $Vordiplom$, AS für $Auslandssemester$ und FR für $kannFragebeantworten$ benutzt. Wenn nur die Abkürzungen benutzt werden, heißt das, dass die Variable den Wert *true* annimmt.

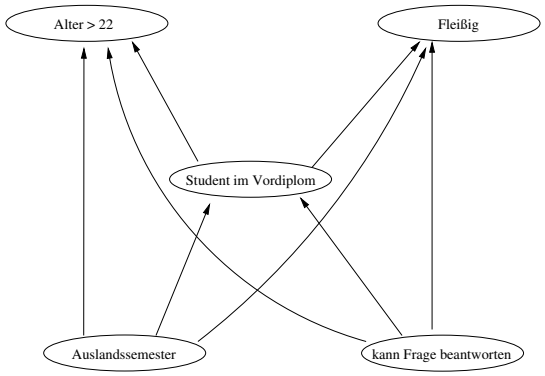


Abbildung 3.2:

3.2 soll nun noch gezeigt werden, was passiert, wenn man bei der Konstruktion von der Wirkung aus zur Ursache geht, was sich hier in 4 weiteren Kanten bemerkbar macht.

3.3.3 Eigenschaften von Bayes Netzwerken

Mit einem Bayes Netzwerk kann man ein System von Zufallsvariablen komplett beschreiben. Man kann über die CPT's jeden beliebigen Wert aus der *full joint distribution* berechnen. Als Beispiel sei hier gezeigt, wie man einen beliebigen Wert der *full joint distribution* aus dem Bayes Netzwerk berechnet.
Gesucht sei die Wahrscheinlichkeit für das atomare Ereignis
$P(X_1 = x_1 \wedge \ldots \wedge X_n = x_n) = P(x_1, \ldots, x_n)$,
wobei $x_1, \ldots, x_n$ alle vorkommenden Variablen sind. Der Wert dieses atomaren Ereignisses lässt sich über die folgende Formel berechnen:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid Vater(X_i)) \tag{3.9}$$

$Vater(X_i)$ steht dabei für die Werte der Vaterknoten von x_i. In dem oben gegebenen Bayes Netzwerk wäre das zum Beispiel:

$$P(A > 22, FL, VD, \neg AS, FR)$$

also die Wahrscheinlichkeit, dass der Student älter als 22, fleißig und im Vordiplom ist, sowie kein Auslandssemester gemacht hat und die Frage beantworten kann. Das entspricht dann nach obiger Formel:

$$\begin{aligned}
&= P(A > 22)P(FL)P(VD \mid A > 22, FL)P(\neg AS|VD)P(FR \mid VD) \\
&= 0.7 \cdot 0.4 \cdot 0.1 \cdot 0.99 \cdot 0.3 \\
&= 0.0091
\end{aligned}$$

Damit hat man mit dem Bayes Netzwerk eine Datenstruktur, aus der sich die *full joint distribution* berechnen lässt.

Um nun die Wahrscheinlichkeiten von Ereignissen unter bestimmten Vorbedingungen zu berechnen, gibt es zwei verschiedene Möglichkeiten, die sich in ihrer Effizienz und Genauigkeit unterscheiden und im folgenden besprochen werden.

3.3.4 Exakte Inferenz im Bayes Netzwerk

Wie in Abschnitt 3.2.5 beschrieben, kann man jede Wahrscheinlichkeit in einem System aus der *full joint distribution* berechnen (vgl. Formel 3.8). Außerdem haben wir im vorangehenden Abschnitt gezeigt, wie man beliebige Einträge in der *full joint* Tabelle aus dem Bayes Netzwerk über 3.9 berechnen kann.
Also kann man auch aus einem Bayes Netzwerk jede Wahrscheinlichkeit berechnen. Allerdings soll auch dafür die Laufzeit betrachtet werden. Oben hatten wir gesehen, dass die Laufzeit mittels der *full joint distribution* in $O(2^n)$ liegt, da unter Umständen jeder Wert aus der Tabelle betrachtet werden muss. Nun müssen die Werte der Tabelle beim gegebenen Bayes Netzwerk allerdings noch berechnet werden. In der Formel 3.9 sieht man, dass für jedes atomare Ereignis im schlimmsten Fall eine Laufzeit von n Schritten nötig ist. Also liegt man bei dem naiven Ansatz, einfach alle atomaren Ereignisse zu berechnen und dann wie bei der *full joint distribution* vorzugehen, in der Klasse $O(n2^n)$. Dies ist offensichtlich kein Vorteil gegenüber der *full joint distribution* in Hinsicht auf die Laufzeit (der Speicherbedarf ist offensichtlich geringer bei Bayes Netzwerken, da die *full joint distribution* hier nie ganz im Speicher ist). Eine Verbesserung der Laufzeit kann man erreichen, wenn man betrachtet, wie die Werte genau berechnet werden. Beispiel: Gesucht sei die Wahrscheinlichkeit, dass ein Student fleißig ist, wobei man weiß, dass er im Ausland war und die Frage beantworten kann. Also gesucht ist:

$$\begin{aligned} P(FL \mid AS, FR) &\overset{3.8}{=} \alpha P(FL, AS, FR) \\ &= \alpha \textstyle\sum_a \sum_s P(FL, s, a, AS, FR) \\ &\overset{3.9}{=} \alpha \textstyle\sum_a \sum_s P(a)P(FL)P(s \mid a, FL)P(AS \mid s)P(FR \mid s) \end{aligned}$$

Es fällt auf, dass es Zufallsvariablen in der Formel gibt, die keine Vaterknoten haben und damit Konstanten sind. Sind dies Zufallsvariablen, die von der gesuchten Wahrscheinlichkeit vorgegeben sind, so kann man sie ganz nach vorne vor die Formel ziehen. Falls es Zufallsvariablen sind, über deren Werte aufsummiert wird, so kann man sie bis zu dem zu ihnen gehörigen Summenzeichen ziehen. So erhält man die folgende Formel:

$$P(FL \mid AS, FR) = \alpha P(FL) \sum_a P(a) \sum_s P(s \mid A, FL)P(AS \mid s)P(FR \mid s)$$

Daraus erhält man eine Verbesserung der Laufzeit von $O(n2^n)$ zu $O(2^n)$. Dies bedeutet noch keine Verbesserung der Laufzeit im Vergleich zur *full joint distribution*, jedoch eine Verbesserung des Speicherbedarfs von $O(2^n)$ auf $O(n)$. Eine weitere Verbesserung ist noch möglich, wenn man mehrfache Berechnungen von Wahrscheinlichkeiten verhindert. In der Aufsummierung für die nicht festgelegten Zufallsvariablen werden nämlich die Wahrscheinlichkeiten mehrfach berechnet, die nicht von der Zufallsvariable abhängig sind, über die gerade aufsummiert wird. Lässt man diese unnötigen Berechnungen weg, so erreicht man auf jeden Fall eine Verbesserung der Laufzeit im Vergleich zu der Berechnung

aus der *full joint distribution*. Dennoch bleibt zu sagen, dass die Exakte Inferenz NP-schwer ist.

3.3.5 Approximative Inferenz im Bayes Netzwerk

Offensichtlich ist die exakte Inferenz im Bayes Netzwerk nur wenig von Hilfe, so dass ein anderer Weg aufgezeigt werden soll, der besser angewendet werden kann. Im folgenden werden nun mittels *Monte Carlo* Algorithmen Wahrscheinlichkeiten berechnet. Dazu werden Beispiele generiert und von diesen die Wahrscheinlichkeiten abgeleitet. Je mehr Beispiele generiert werden, umso genauer wird die gesuchte Wahrscheinlichkeit angenähert. Dabei soll nur auf die Beispielgenerierung eingegangen werden, da sich die Wahrscheinlichkeiten dann aus den Häufigkeiten der generierten Beispiele berechnen lassen. Eine sehr einfache Möglichkeit ein Beispiel für ein atomares Ereignis zu generieren ist die folgende.
Dazu geht man von den Knoten, die keine Vaterknoten haben, aus und generiert eine zufällige Belegung anhand der *CPT*. Nun geht man weiter zu den Knoten, von denen alle Werte der Vorfahren bekannt sind und generiert auch für diese Knoten einen Wert nach der in der *CPT* angegebenen Verteilung.

Damit erhält man dann nach und nach ein Beispiel. Die Wahrscheinlichkeit, dass ein bestimmtes Beispiel $S(x_1, \ldots, x_n)$ generiert wurde entspricht offensichtlich

$$S(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid Vater(X_i)) \tag{3.10}$$

Diese Formel ist identisch mit 3.9. Also kann man gesuchte Wahrscheinlichkeiten für atomare Ereignisse sehr einfach über Beispiele berechnen. Wenn insgesamt N Beispiele generiert wurden und davon $N_S(x_1, \ldots, x_n)$ das gesuchte Ereignis $x_1, \ldots, x_n$ beinhalteten, so kann man annehmen:

$$\lim_{N \to \infty} \frac{N_S(x_1, \ldots, x_n)}{N} = S(x_1, \ldots, x_n) = P(x_1, \ldots, x_n) \tag{3.11}$$

Also für genügend großes N nähert sich das durch Simulation berechnete Ergebnis immer näher an die tatsächliche Wahrscheinlichkeit an.
Bedingte Wahrscheinlichkeiten lassen sich mit diesem Algorithmus nicht direkt berechnen, sondern es werden zunächst Beispiele konstruiert, die losgelöst von den Bedingungen der gesuchten Wahrscheinlichkeit sind. Erst wenn das Beispiel konstruiert wurde, testet man, ob es auch ein Beispiel für das gesuchte Ereignis ist, und über diese wird dann aufsummiert. Das heißt, wenn man 100 Beispiele konstruiert hat, von denen jedoch nur 20 die Bedingungen der gesuchten Wahrscheinlichkeit erfüllen, dann lässt man die anderen 80 Beispiele wegfallen, was auch ein großer Schwachpunkt bei dieser Vorgehensweise ist, da unter Umständen sehr viele Beispiele nicht benutzt werden können, und so die Anzahl der verwertbaren Beispiele zu gering für eine genaue Aussage ist (die Standardabweichung beträgt dabei $1/\sqrt{n}$, wobei n die Anzahl der benutzten Beispiele ist).
Die nächste Möglichkeit, die hier besprochen werden soll, beruht auf einer Gewichtung der Ereignisse. Der Ansatz dabei ist, dass die Ausgangsbedingungen

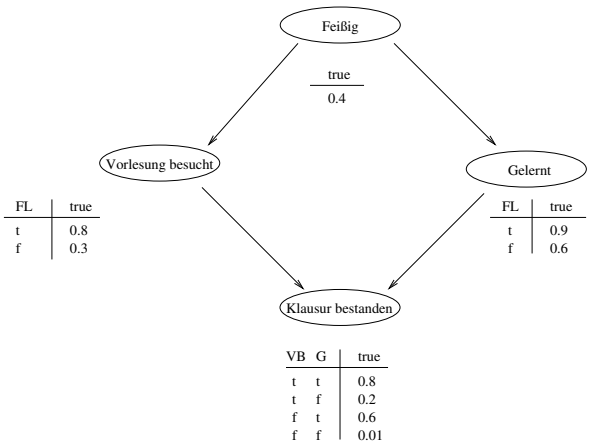


Abbildung 3.3:

festgehalten werden und nur für die verbleibenden Zufallsvariablen Beispiele generiert werden. Bevor die Beispiele ausgewertet werden, wird dann noch jedes Ereignis nach seiner Wahrscheinlichkeit gewichtet. Dabei wird das Gewicht über das Produkt der bedingten Wahrscheinlichkeiten der Bedingungen (die gegebenen Zufallsvariablen) unter der Kenntnis ihrer Vaterknoten berechnet.
Beispiel: Gegeben sei das Bayes Netzwerk aus Abbildung 3.3.
Nun soll ein Beispiel mit Gewichtung für die gesuchte Wahrscheinlichkeit $P(G \mid VB, KB)$ berechnet werden. Also die bedingte Wahrscheinlichkeit dafür, dass der Student gelernt hat unter der Bedingung er hat die Vorlesung besucht und bestanden.

1. Vor Beginn entspricht die Gewichtung $w_0 = 1.0$.
2. Es wird ein Wert für *fleißig* anhand der Wahrscheinlichkeitsverteilung ($< 0.4, 0.6 >$) generiert. Angenommen, das Ergebnis ist *false*.
3. Es geht weiter mit der nächsten Zufallsvariable *Vorlesungbesucht*. Der Wert ist Vorgegeben mit *true*.
   Nun muss das Gewicht für das Beispiel angepasst werden:
   $w_1 = w_0 \cdot P(VB \mid \neg FL) = 1.0 \cdot 0.3 = 0.3$
4. Ein Wert für *gelernt* aus der zu $\neg$*fleißig* passenden Verteilung ($< 0.6, 0.4 >$) wird generiert. Hierfür sei das Ergebnis auch *false*.
5. Abschließend muss noch die Zufallsvariable *bestanden* betrachtet werden. Da hier ein Wert vorgegeben ist (*true*) muss die Gewichtung wieder angepasst werden.
   $w_2 = w_1 \cdot P(B \mid VB, \neg G) = 0.3 \cdot 0.2 = 0.06$

Damit ergibt sich für das generierte Beispiel $(\neg FL, VB, \neg G, B)$ ein Gewicht von 0.06.

Da es nicht sofort einleuchtend ist, dass dieses Vorgehen auf die korrekten Wahrscheinlichkeiten führt, soll hier noch die Korrektheit gezeigt werden.
Alle Ereignisse, die erzeugt werden, unterscheiden sich nur in den nicht gegebenen Zufallsvariablen. Im Folgenden seien die Variablen deren Werte angegeben sind mit e, die Zufallsvariablen, deren Werte interessieren mit X und die, die nicht angegeben sind und auch nicht zu dem gesuchtem Ereignis gehören mit Y bezeichnet. Die Variablen, die nicht angegeben sind, sollen hier zu $Z = X \cup Y$ zusammengefasst werden. Bei dieser Bezeichnung erhält man also für die Wahrscheinlichkeit, dass ein bestimmtes Ereignis als Beispiel konstruiert wurde:

$$S_g(z, e) = \prod_{i=1}^{l} P(z_i \mid Vater(Z_i)) \tag{3.12}$$

Würde man diese Wahrscheinlichkeiten direkt übernehmen, so hätte man zwar die Bedingungen e in Z_i's Vaterknoten enthalten, jedoch werden die Bedingungen für jedes Z_i in den Knoten ignoriert, die nicht Vorfahren von Z_i sind. Das Gewicht w ist dafür da, um diesen Unterschied wieder auszugleichen. Wie oben bereits beschrieben, wird das Gewicht über das Produkt der Wahrscheinlichkeiten der Bedingungen e unter Kenntnis ihrer Vaterknoten berechnet. Also:

$$w(z, e) = \prod_{i=1}^{m} P(e_i \mid Vater(E_i)) \tag{3.13}$$

Wenn man nun 3.12 und 3.13 miteinander multipliziert, dann erhält man daraus:

$$S_g(z, e)w(z, e) = \prod_{i=1}^{l} P(y_i \mid Vater(Yi)) \prod_{i=1}^{m} P(e_i \mid Vater(E_i)) = P(y, e) \tag{3.14}$$

wenn man nun einen bestimmten Wert x betrachten möchte, so erhält man:

$$P(x \mid e) = \alpha \sum_{y} N_g(x, y, e)w(x, y, e)$$

wobei N_g die Aufsummierung von S_g ist. Für große N gilt:

$$\approx \alpha' \sum_{y} S_g(x, y, e)w(x, y, e) = \alpha' \sum_{y} P(y, x, e) = \alpha' P(x, e) = P(x \mid e)$$

Also liefert das Gewichtungsverfahren einen angenäherten Wert für die Wahrscheinlichkeit. Jedoch hat auch dieses Verfahren Nachteile. Wenn es mehr Ausgangsbedingungen gibt, bekommen die meisten Beispiele ein sehr niedriges Gewicht und die Laufzeit nimmt zu.

3.4 Dempster Shafer Theory

3.4.1 Einführung

Die Dempster Shafer Theory ist eine weitere Möglichkeit die Wahrscheinlichkeit für Ereignisse darzustellen. Sie beruht dabei auf zwei Ideen. Zum einen wird jeder Menge von Ereignissen ein Grad an Glaubwürdigkeit zugeordnet. Zum anderen lassen sich diese Glaubwürdigkeiten mit Hilfe der 'Dempster rule' kombinieren.

3.4.2 Ereignisse

In der Dempster Shafer Theorie wird jeder Menge von Ereignissen ein Tupel [Belief(Bel), Plausibility(Pl)] zugeordent. Belief und Plausibility sind dabei Werte zwischen 0 und 1. Belief gibt an, wie stark die Hinweise sind, die für das Eintreten dieser Ereignisse sprechen. Die Plausibility gibt an, wie Wahrscheinlich die Ereignisse sind, falls alle unbekannten Faktoren für diese Ereignisse sprechen. Es ist definiert:

$$Pl(s) = 1 - Bel(\neg s)$$

Falls man noch keine Information über die verschiedenen Ereignisse hat, ist somit jeder Menge von Ereignissen das Tupel [0,1] zugeordnet.
Die Größe des Intervalls sagt somit etwas über die Menge der Information aus, die man schon zu den möglichen Ereignissen gesammelt hat. Somit kann anhand dessen entschieden werden, ob zuerst noch weiter Informationen gesammelt werden sollen oder auf Grund der bisherigen Erkenntnisse gehandelt werden kann. Im Gegensatz dazu ist bei der Bayes Theorie jedem Ereignis ein fester Wert zugeordnet und es gibt keine Informationen darüber wie fundiert diese Aussage ist.
Um nun exakt mit der Dempster Shafer Theorie arbeiten zu können, braucht man eine Menge von Ereignissen, wobei die Menge disjunkt ist und ein Ereignis sicher eintritt. Diese Menge bezeichnen wir mit Θ und wird auch "Frame of Discernment" genannt. Außerdem wird eine Funktion benötigt, mit deren Hilfe jeder Teilmenge von Θ der zugehörige Beliefwert zugeordnet wird. Dazu wird eine Funktion m benutzt, die die folgenden Eigenschaften erfüllt:

$$m : \mathcal{P}(\Theta) \rightarrow [0, 1]$$

$$Bel(A) = \sum_{B \subseteq A} m(B)$$

Der Beliefwert für die Teilmenge $A \subseteq \Theta$ entspricht also der Summe über die Funktionswerte von m aller Teilmengen.Da sicher ein Ereignis aus Θ eintrifft, muss $Bel(\Theta) = 1$ gelten, woraus mit obiger Formel folgt:

$$\sum_{A \subseteq \Theta} m(A) = 1$$

Somit ist die Funktion m , falls keine Informationen vorhanden sind, wie folgt definiert (im Folgenden wird die Schreibweise "Menge" "zugeordneter Wert" benutzt):
Θ (1.0)
Dies bedeutet: $m(\Theta) = 1.0$ und für alle anderen Mengen ist der Funktionswert von m gleich Null.
In unserem Beispiel sei Θ nun {Student im Vordiplom (VD), Student im Hauptdiplom (HD), Doktorand (D), Professor (P)}.
Wissen wir jetzt zum Beispiel, dass eine Person mit einer Wahrscheinlichkeit von 0.6 aus der Menge {VD, HD} kommt (z.B. anhand seines Alters), ergibt sich folgende Funktion:

{VD, HD}	(0.6)
Θ	(0.4)

Dabei ist zu beachten, dass wir für den Rest keine Informationen haben und somit nicht sagen können, dass die Person mit einer Wahrscheinlichkeit von 0.4 aus der Menge {D,P} kommt, da es bis jetzt keine Hinweise gibt, dass die Person aus $\{D, P\}$ kommt. Es gilt aber in jedem Fall $Bel(\Theta) = 1.0$, da es sicher ist, dass die Person aus dieser Menge kommt.
Daraus ergibt sich nach obiger Formel für $Bel(\{VD, HD\}) = 0.6$ und
$Pl(\{VD, HD\}) = 1 - Bel(\{D, P\}) = 1 - 0 = 1$
Somit wird der Menge {VD,HD} das Tupel [0.6;1.0] zugeordnet.

3.4.3 Kombinieren von Aussagen - Dempsters Rule

Existieren zu einer Menge von Ereignissen zwei unabhängige Aussagen, so können diese mit der Dempster Rule zu einer Gesamtaussage kombiniert werden. Werden die beiden Aussagen durch die Funktionen m_1 und m_2 beschrieben, wobei m_1 und m_2 über dem gleichen Θ definiert sind, so ergibt sich die Funktion für die Gesamtaussage:

$$m_3(Z) = K \sum_{X \cap Y = Z} m_1(X) \cdot m_2(Y) \tag{3.15}$$

$$K^{-1} = 1 - \sum_{X \cap Y = \emptyset} m_1(X) \cdot m_2(Y) \tag{3.16}$$

Dabei gilt $X, Y, Z \subseteq \Theta$

In diesen Formeln entspricht das K einem Normalisator, der das Auftreten von leeren Mengen ausgleicht (siehe unten).
Man betrachte nun die folgenden beiden Funktionen:

m_1 :	$\{VD, HD\}$	(0.6)	m_2 :	$\{HD, D, P\}$	(0.7)
	Θ	(0.4)		Θ	(0.3)

Bei diesen beiden Funktionen ist der Schnitt zweier Mengen, denen nicht der Wert Null zugeordnet wird, niemals die leere Menge. Führt man nun die obige Regel zur Kombination von Funktionswerten durch (hier an einem Beispiel vorgeführt), so ergibt sich folgendes:

$$\begin{aligned} & \{VD, HD\} \cap \{HD, D, P\} && = \{HD\} \\ \Rightarrow \quad & m_3(\{HD\}) && = m_1(\{VD, HD\}) \cdot m_2(\{HD, D, P\}) \\ & && = 0.6 \cdot 0.7 = 0.42 \end{aligned}$$

Daraus ergibt sich dann folgende Tabelle:

		HD,D,P	0,7	Θ	0,2
VD,HD	0,6	HD	0,42	VD,HD	0,12
Θ	0,4	HD,D,P	0,28	Θ	0,08

aus der sich die Funktion m_3 ablesen lässt:

$\{HD\}$	0,42
$\{HD, VD\}$	0,12
$\{HD, D, P\}$	0,28
Θ	0,08

Im allgemeinen kann es vorkommen, dass eine Menge mehrfach in der Tabelle auftritt. In diesem Fall müssen alle zugehörigen Werte addiert werden.
Außerdem gibt es Kombinationen, so dass die leere Menge auftritt. Dieser muss aber der Wert Null zugeordnet werden, da ein Ereignis aus Θ immer eintritt. Damit die Summe aller Werte von Teilmengen eins ergibt, müssen deshalb alle Werte skaliert werden. Der Faktor ist dabei der Kehrwert der Wahrscheinlichkeit für die leere Menge[6].

3.5 Fuzzy Sets & Fuzzy Logic

3.5.1 Fuzzy Sets

In der klassischen Mengetheorie ist für jedes Element der Grundmenge G klar definiert, ob es Element einer Menge ist oder nicht. Betrachtet man die Menge der Studenten der Uni Karlsruhe, so kann für jede Person festgestellt werden, ob sie zu dieser Menge gehört. Dabei gibt es verschiedene Möglichkeiten diese Menge zu beschreiben:

- Aufzählung aller Elemente
- Beschreibung der Auswahlkriterien (z.B. eine Person ist Student an der Uni Karlsruhe, wenn sie immatrikuliert ist)
- Charakteristische Funktion $G \rightarrow \{0, 1\}$

In der Realität gibt es aber auch Mengen, die nicht klar abgegrenzt sind. Betrachtet man die Menge der Studenten, die regelmäßig die Vorlesung besuchen, so lässt sich nicht klar definieren, ob ein Student zu dieser Menge gehört. Bei der klassischen Mengenlehre muss es klar definierte Auswahlkriterien geben. So würde ein Student, der genau die Mindestanzahl an Stunden besucht hat, zur Menge gehören. Ein Student, der nur eine Stunde weniger die Vorlesung besucht hat hingegen nicht. Gründe, die zu einem solchen Problem führen, sind zum einen ungenaue Definitionen (Wann ist ein Sandhaufen ein Sandhaufen?) und zum anderen zu viele Bedingungen für eine Definition, so dass Überprüfung nur schwer möglich ist.
Betracht man nun Fuzzy Sets, so wird jedem Element mittels einer Funktion μ ein Wert zugeordnet, der den Grad der Mitgliedschaft des Elements zu dieser Menge angibt. Es gibt mehrere Möglichkeiten ein solches Fuzzy Set dazustellen:

- Aufzählung aller Elemente mit jeweiligem Grad der Zugehörigkeit
- Angabe einer Funktion, die den Elementen den Grad der Mitgliedschaft zuordnet (*membership function*)

[6] entspricht dem Faktor K, aus 3.15.

Den Elementen müssen nur Werte größer oder gleich 0 zugeordnet werden, nicht unbedingt Werte zwischen 0 und 1. Falls die Funktion den Elementen Werte zwischen 0 und 1 zuordnet, so nennt man dieses Fuzzy Set normalisiert. Es kann allerdings jedes Fuzzy Set normalisiert werden, indem man jeden Wert durch das Supremum der Funktion teilt.

3.5.2 Mengenoperationen auf Fuzzy Sets

Um mit diesen Mengen arbeiten zu können, müssen Durchschnitt, Vereinigung und Komplement auf Fuzzy Sets definiert werden. Dazu betrachtet man erneut die *membership function*.
Der Durchschnitt zweier Mengen $C = A \cap B$ ist elementweise definiert durch die *membership function* μ_C:

$$\mu_C(X) = min(\mu_A(x), \mu_B(x)) \tag{3.17}$$

Die Vereinigung zweier Mengen $D = A \cup B$ ist elementweise definiert durch die membership function μ_D:

$$\mu_D(X) = max(\mu_A(x), \mu_B(x)) \tag{3.18}$$

Das Komplement einer Menge ist elementweise definiert durch die membership function $\mu_{\neg A}$:

$$\mu_{\neg A}(x) = 1 - \mu_A(x) \tag{3.19}$$

Beispiel:

$$\begin{aligned} B &= besucht\_Vorlesung = \{(Fritz, 0.2), (Klaus, 0.9), (Ute, 0.8)\} \\ F &= sind\_Fleißig = \{(Fritz, 0.5), (Klaus, 0.7), (Ute, 0.6), (Kim, 0.1)\} \end{aligned}$$

$$\begin{aligned} &\Rightarrow C = B \cap F = \{(Fritz, 0.2), (Klaus, 0.7), (Ute, 0.6)\} \\ &\Rightarrow D = B \cup F = \{(Fritz, 0.5), (Klaus, 0.9), (Ute, 0.8), (Kim, 0.1)\} \\ &\Rightarrow E = \neg B = \{(Fritz, 0.8), (Klaus, 0.1), (Ute, 0.2), (Kim, 0.1)\} \end{aligned}$$

3.5.3 Fuzzy Logic

Die Fuzzy Logic bildet eine Erweiterung zu der Booleschen Logik. In der Booleschen Logik werden Aussagen nur die Werte *true* oder *false* zugeordnet. Diese Aussage kann auch als Zugehörigkeit zu einer Menge interpretiert werden.
Beispiel: Die Aussage "Ute ist Studentin" ist genau dann wahr, wenn Ute in der Menge der Studenten enthalten ist. In der Fuzzy Logic wird die Zugehörigkeit über Fuzzy Sets dargestellt. So kann der Aussage "Ute ist fleißig" nicht ein Wert aus der Menge $\{true, false\}$ zugeordnet werden, sondern ein Wert aus dem Intervall $[0, 1]$. Dieser Wert entspricht dabei der Mitgliedschaft im Fuzzy Set. 0 heißt dabei, dass die Aussage falsch ist und 1, dass die Aussage wahr ist. Dabei ordnet die Funktion T jeder Aussage ihre Wahrscheinlichkeit zu.
Daraus ergeben sich die folgenden Definitionen für die logischen Funktionen "*und*", "*oder*" und "$Negation$":

Seien A und B Aussagen, so gilt:

$$\begin{aligned} T(A \wedge B) &= min(T(A), T(B)) \\ T(A \vee B) &= max(T(A), T(B)) \\ T(\neg A) &= 1 - T(A) \end{aligned}$$

Daraus ergibt sich für die Implikation:

$$\begin{aligned} T(A \Rightarrow B) &= T(\neg A \vee B) \\ &= max(T(\neg A), T(B)) \\ &= max(1 - T(A), T(B)) \end{aligned}$$

Beispielsweise seien folgende Prädikate definiert (wobei μ_B und μ_F die membership Funktionen der zugehörigen Fuzzy Sets aus dem Beispiel oben sind):

$B(X) := X$ besucht die Vorlesung regelmäßig, mit
$T(B(X)) = \mu_B(X)$ und
$F(X) := X$ ist ein fleißiger Student, mit
$T(F(X)) = \mu_F(X)$

Somit ist der Aussage "Fritz und Ute besuchen die Vorlesung regelmäßig" folgender Wert zugeordnet:

$$\begin{aligned} T(B(Fritz) \wedge B(Ute)) &= min(T(B(Fritz)), T(B(Ute))) \\ &= min(0.2, 0.8) = 0.2 \end{aligned}$$

Zu beachten ist noch, dass aus den oben definierten Regeln auch folgendes abgeleitet werden kann:

$$\begin{aligned} T(B(Fritz) \wedge \neg B(Fritz)) &= min(B(Fritz), 1 - B(Fritz)) \\ &= min(0.2, 0.8) = 0.2 \end{aligned}$$

Also wird der offensichtlich falschen Aussage ein Wert größer 0 zugeordnet.

3.6 Zusammenfassung

Diese Ausarbeitung beschäftigte sich zunächst mit den Grundlagen der Wahrscheinlichkeitsrechnung. Dabei wurde die Problematik aufgezeigt, aus der *full joint distribution* Wahrscheinlichkeiten effizient zu berechnen. Abhilfe schafft das Bayes Netzwerk, welches eine Möglichkeit gibt, Wahrscheinlichkeiten, die man aus der *full joint distribution* erhält, auf eine effizientere Weise zu speichern. Außerdem wurden verschiedene Wege aufgezeigt, wie man mittels approximativer Inferenz schneller als bei der Exakten die Wahrscheinlichkeiten für beliebige Ereignisse berechnen kann. Dies ist jedoch mit Einschränkungen bei der Genauigkeit verbunden, die nur mit höherem Zeitaufwand verringert werden konnten.
Der Abschnitt über die Dempster Shafer Theorie lieferte mittels *Dempsters rule* eine Möglichkeit, Aussagen über die Wahrscheinlichkeit für Ereignisse zu geben und im besonderen auch Hinweise über die Genauigkeit, die mit Hilfe des zu dem Ereignis gehörigen Intervalls beschrieben wird. Als letztes ging die Ausarbeitung

auf Fuzzy Sets und Fuzzy Logik ein. Fuzzy Sets bieten eine Möglichkeit, Mengen die nicht genau beschrieben sind darzustellen. Fuzzy Logik lässt Schlüsse und Folgerungen zu, die auf Fuzzy Sets “arbeiten”.

Literaturverzeichnis

[1] Russel S., Norvig P.: *Artificial Intelligence – A Modern Approach*, Second Edition, Prentice Hall, 2003.

[2] Ginsberg M.: *Essentials of Artificial Intelligence*, Morgan Kaufmann Publishers, 1996.

[3] Zimmermann H.-J.: *Fuzzy Set Theory – And Its Applications*, Second Edition, Kluwer Academic Publishers, 1991.

[4] Rich E., Knight K.: *Artificial Intelligence*, Second Edition, McGraw-Hill 1991.

[5] Shafer,G. *Dempster-Shafer Theory*, www.glennshafer.com/assets/downloads/article48.pdf

[6] O'Neill, A. *Dempster-Shafer Theory*, http://www.quiver.freeserve.co.uk/binaries/dempster.pdf