Intermezzo: Entscheidungsbäume und Informationsgehalt

In einer Fernsehsendung soll ein Rateteam durch gezielte Fragen den Beruf eines Kandidaten in Erfahrung bringen (Abb. 1). Der Kandidat kann auf die Fragen nur mit Ja oder Nein antworten. Trotzdem schafft es das Rateteam häufig, mit wenigen Fragen den Beruf zu ermitteln, obwohl es Tausende von verschiedenen Berufen gibt. Dazu stellen sie erst sehr allgemeine Fragen und verfeinern sie dann schrittweise.

Abb. 1: Beruferaten – Was bin ich?

Zuerst wird häufig gefragt, ob zur Ausübung des Berufes eine Lehre erforderlich ist. Sollte diese Frage mit Ja beantwortet werden, könnte man fragen, ob es sich bei dem Beruf um eine Handwerk handelt. Das Frage-Antwort-Spiel lässt sich also als eine fortschreitende Verzweigung ansehen, welche man in einem **Baumdiagramm** verdeutlichen kann (Abb. 2).

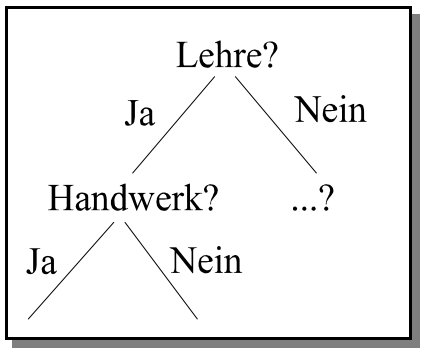

Abb. 2: Baumdiagramm

Ganz ähnlich kann man vorgehen, wenn man einen Gast in einem Hotel sucht und der Portier nur die Antworten Ja und Nein geben will. Man fragt zuerst, ob sich der Gast in der ersten Hälfte aller Zimmer befindet; dies entspricht einer sehr allgemeinen Frage. Antwortet unser Portier mit Nein, muss der gesuchte Gast sich in der zweiten Hälfte befinden. Dann teilen wir diese Hälfte ihrerseits in zwei gleiche Teile und befragen den Portier erneut. So verfahren wir weiter, bis wir das Zimmer unseres Gastes gefunden haben. Wie beim Beruferaten können wir unseren Suchvorgang auf Ja-Nein-Fragen zurückführen.

Wie viele Fragen dabei insgesamt benötigt werden, hängt von der Anzahl der Zimmer ab. Wenn das Hotel 8 Zimmer besitzt, brauchen wir 3 solcher Fragen, um den Gast zu finden, wenn es 128 Zimmer besitzt, brauchen wir 7 Fragen. Allgemein findet man durch m Fragen einen Gast in einer Gesamtheit von 2^m Zimmern. Wenn man umgekehrt aus n Zimmern ein bestimmtes herausfinden möchte, sind dazu $\log_2(n)$ Fragen erforderlich. Dabei bezeichnet $\log_2$ den Logarithmus zur Basis 2.

Durch jede Frage und jede Antwort erhöht sich der Kenntnisstand des Fragenden. Die Anzahl der Ja-Nein-Fragen, welche erforderlich ist, um vom aktuellen Kenntnisstand (hier: Gast befindet sich in einem bestimmten Hotel.) bis zum gewünschten Kenntnisstand (hier: In welchem Zimmer befindet sich der Gast?) zu gelangen, bezeichnet man als **Informationsgehalt *h***. Die Einheit des Informationsgehalts ist **1 bit**. 1 bit entspricht also dem Informationsgehalt, der in der Antwort auf eine Ja-Nein-Frage enthalten ist.

Beachte, dass der Informationsgehalt nicht nur vom gewünschten Kenntnisstand abhängt, sondern auch vom aktuellen. Wenn in unserem Beispiel nicht der Hotelname bekannt wäre, sondern nur der Name des Ortes, in welchem sich (neben anderen) unser Hotel mit dem gesuchten Gast befindet, wäre der Informationsgehalt größer.

Aufgaben

1. Überprüfe die Beziehung $\log_2(x) = \frac{\lg(x)}{\lg(2)}$ für $x = 3$ und $x = 7$. Leite die Formel her.
2. Ein Zahlenschloss habe 3 bzw. 4 Ringe mit den Ziffern von 0 bis 9. Wie groß ist jeweils der Informationsgehalt der Geheimzahl?

Wahrscheinlichkeit von Nachrichten

Die Nachricht „Grundschüler schlägt Großmeister in Schachpartie“ wird von einem Zeitungsredakteur sicherlich aufgegriffen und in einem Artikel veröffentlicht werden. Dagegen wird die gegenteilige Nachricht „Großmeister schlägt Grundschüler in Schachpartie“ kaum Aufmerksamkeit erregen. Offensichtlich wird die erste Nachricht vom Redakteur für wertvoller gehalten als die zweite. Woran liegt das? Ganz offensichtlich tritt die erste Situation im Gegensatz zur zweiten äußerst selten auf; sie ist sehr unwahrscheinlich.

Tatsächlich kann man einen Zusammenhang zwischen der **Wahrscheinlichkeit *p*** einer Nachricht und ihrem Informationsgehalt *h* herstellen. Für den Fall unseres Hotelgastes gelingt dies folgendermaßen: Wenn wir davon ausgehen, dass alle Zimmer gleich wahrscheinlich sind, ist bei einer Anzahl von *n* Zimmern die Wahrscheinlichkeit $p = 1/n$. Damit gilt für den Informationsgehalt der Nachricht, wo sich unser Gast aufhält:

$$h = \log_2(\frac{1}{p}) = -\log_2(p) = -\frac{\lg(p)}{\lg(2)}$$

Diese Formel gilt natürlich nicht nur für den Kenntnisstand über unseren Hotelgast, sondern für jede ähnliche Situation.

Bei kleinen Wahrscheinlichkeiten – wie im Fall unseres Grundschülers, der gegen den Großmeister gewinnt – ist der Logarithmus von *p* eine negative Zahl mit großem Betrag, der Informationsgehalt *h* ist demnach sehr groß. Dies erklärt auch die Reaktion des Redakteurs: Die unwahrscheinliche Nachricht hat eben einen großen Informationsgehalt; und deswegen lohnt es sich, sie zu veröffentlichen.

Aufgaben

1. Eine Nachricht hat die Wahrscheinlichkeit 0,1 bzw. 0,001. Wie groß ist ihr Informationsgehalt?

2. Ein Tresor hat 3 Räder mit je 24 Buchstaben. Wie groß ist die Wahrscheinlichkeit, auf gut Glück den richtigen Code einzustellen? Berechne auch den Informationsgehalt des Zugangscodes.

Suchbäume und Codes

Ähnlich wie beim Beruferaten lassen sich viele Nachrichten als das Ergebnis eines Suchbaums wie in Abb. 2 ansehen. Wenn wir statt der Antworten Ja und Nein eine 1 bzw. eine 0 schreiben, kann die Nachricht auch durch eine Dualzahl angegeben werden. In Abb. 3 führt z. B. der markierte Pfad zu der Information „F“. Dieser Pfad lässt sich auch durch die Dualzahl 101_2 beschreiben. Da es bei einem solchen Suchbaum keinen Unterschied ausmacht, ob man das Ziel oder den Pfad angibt, kann man 101_2 auch als ein Kodierung für die Information „F“ ansehen. Weil diese Verschlüsselung mithilfe von Dualzahlen geschieht, spricht man hier von einem **binären Code**.

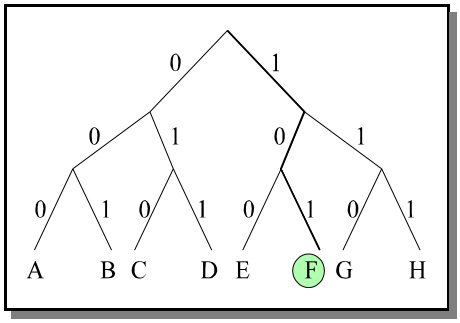

Abb. 3: Der binäre Code von F ist 101_2

Derartige Verzweigungsstrukturen findet man an vielen Automaten, z.B. auch bei Fahrkartenautomaten (Abb. 4). Einen vereinfachten Fahrkartenautomaten wollen wir hier etwas genauer unter die Lupe nehmen. Er soll in Mönchengladbach stehen und nur die Ziele Düsseldorf, Aachen, Viersen und Köln anbieten. Wenn man eine Karte kaufen möchte, stellt der Automat zuerst die Frage:

Abb. 4: Am Fahrkartenautomaten

Wollen Sie nach Aachen bzw. Köln?

Wir können die Antwort Ja oder Nein eingaben. Diese und die weiteren möglichen Reaktionen des Automaten siehst du in Abb. 5. Sie zeigt: Jeweils nach genau zwei Eingaben ist der Zielort erfasst und die entsprechende Fahrkarte kann (nach Bezahlung) ausgegeben werden. Wir haben es hier mit einer binären **Kodierung mit fester Codelänge** zu tun.

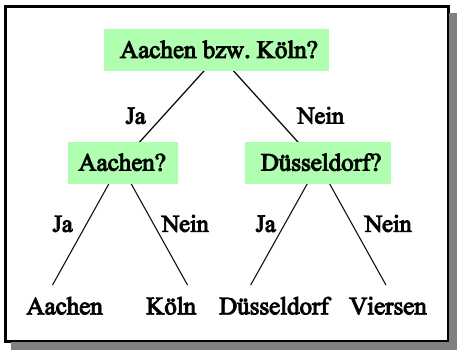

Abb. 5: Durch Verzweigung zum Ziel

Möglich sind aber auch **Kodierungen mit variabler Codelänge**, z. B.:

Zielort	Code	Wahrscheinlichkeit
Aachen	001	0,05
Köln	01	0,3
Düsseldorf	1	0,6
Viersen	000	0,05

Ein grundlegendes Problem bei Codes mit variabler Länge ist: Woher weiß die Maschine, dass das Ende des Codewortes erreicht ist? In diesem Fall reicht es aus, der Maschine folgende zwei Regeln einzubauen:

1. Die Codelänge ist höchstens 3.
2. Die Ziffer 1 beendet ein Codewort.

Auch wenn die Codes mit variabler Länge komplizierter aussehen, sind sie günstiger, wenn die Zielorte nicht gleich häufig benutzt werden. In unsere Tabelle hat der Zielort Düsseldorf die Wahrscheinlichkeit 0,6; sein Code ist 1. In diesem sehr häufigen Fall ist also nur eine einzige Tastenbetätigung erforderlich. Dafür müssen für die Ziele Viersen und Aachen, die seltener gewählt werden, jeweils 3 Tastendrücke erfolgen.

Im Mittel ist dies tatsächlich günstiger als der Code mit fester Länge 2. Die mittlere Codelänge berechnet man, indem man die einzelnen Codelängen mit den zugehörigen Wahrscheinlichkeiten multipliziert und die so erhaltenen Produkte addiert:

$$\text{mittlere Codelänge} = 3 \cdot 0{,}05 + 2 \cdot 0{,}3 + 1 \cdot 0{,}6 + 3 \cdot 0{,}05 = 1{,}5$$

Damit liegt die mittlere Codelänge nun deutlich unter der festen Codelänge 2, die wir zunächst benutzt haben.

Für die Hersteller von Automaten sind derartige Überlegungen wichtig, entscheiden sie doch darüber, wie schnell ein Kunde im Mittel(!) an sein gewünschtes Produkt kommt.

Aufgaben

1. Warum ist ein variabler Code für das Alphabet sinnvoll? Nenne Buchstaben, welche einen möglichst kurzen Code erhalten sollten.

2. Besorge dir ein Morsealphabet. Warum kann es neben den Zeichen Punkt und Strich nicht auf das Pausezeichen verzichten? Welche Buchstaben haben einen langen, welche einen kurzen Code?