Kapitel 3

Entscheidungsbäume

3.1 Schlimmster und durchschnittlicher Fall

Gegeben sei eine endliche Menge S, der sogenannte *Suchraum*, sowie Tests mit einer festen Menge A möglicher Ausgänge. Ziel ist es, mit Hilfe dieser Tests in Kenntnis von S ein unbekanntes Objekt x aus S zu identifizieren.

Beispielsweise mag $S = \{1, \ldots, 7\}$ sein und ein Test von der Form „Ist $x \leq k$?" für ein $k \in \mathbb{N}$, mit $A = \{ja, nein\}$ als Menge der möglichen Ausgänge. Es zeigt sich, daß man mit höchstens drei derartigen Tests ein zunächst unbekanntes x identifizieren kann, mit weniger jedoch nicht.

Dabei denkt man sich die Menge C der zur Identifikation führenden Antwortfolgen als präfixfreien Code über A. Den von diesen Folgen und ihren Präfixen im Baum aller Wörter aus A^* induzierte Teilbaum T denkt man sich beschriftet: An den Blättern steht das durch die Antwortfolge vollständig identifizierte Objekt, an den anderen Ecken kann man sich die konkreten Tests (und auch den aktuellen Suchraum) vorstellen. Der Suchraum sei in dem Sinn irredundant, daß es zu jedem Objekt auch (wenigstens) eine Antwortfolge gibt, die es identifiziert (jedes Objekt „kann der Fall sein"). Daher besitzt T wenigstens $|S|$ Blätter, im allgemeinen aber mehr; also: $|C| \geq |S|$. Im Fall der Gleichheit $|C| = |S|$ stehen die Antwortfolgen in Bijektion zum Suchraum.

Schlimmstenfalls benötigt man $m := \max\{\ell(w) : w \in C\}$ viele Fragen bis zur Identifikation eines Objektes, und man ist natürlich daran interessiert, diese Größe klein zu halten. Der folgende Satz impliziert, daß es stets ein Objekt geben muß, zu dessen Identifikation wenigstens $\lceil \log_{|A|} |S| \rceil$ viele Fragen nötig sind.

Satz 1 *Für jeden präfixfreien Code C über A gilt* $\max\{\ell(w) : w \in C\} \geq \lceil \log_{|A|} |C| \rceil$.

Beweis. Sei $m := \max\{\ell(w) : w \in C\}$. Verlängert man jedes Wort von C auf ein Wort der Länge m, so ensteht ein präfixfreier m-Code mit $|C|$ Wörtern. Folglich ist $|C| \leq |A|^m$, woraus die Behauptung folgt. □

Wir denken uns jetzt den Suchraum mit einer Verteilung p versehen, die angibt, mit welcher Wahrscheinlichkeit $p(s)$ das zu identifizierende Objekt gleich s ist. Es ist dann sinnvoll, nach einem Entscheidungsbaum zu fragen, bei der die durchschnittliche Fragenlänge minimiert wird. Wenn dabei die Antwortfolgen bijektiv den Elementen aus A entsprechen, fragen wir nach einer Codierung f von S mit möglichst kleiner durchschnittlicher Wortlänge $\sum_{x\in S} p(x)\ell(f(x))$, also mit möglichst kleinem $L_A(S, f)$. Diese kann immer als Huffmann-Codierung realisiert werden. Ob sich die minimale Codierung *mittels des Fragenkatalogs* realisieren läßt, ist zunächst nachgeordnet; läßt man die Frage „Ist $x \in K$?" mit Antwortmenge $A = \{ja, nein\}$ für jedes $K \subseteq S$ zu, gelingt das immer (wie der folgende Satz zeigt), ansonsten gibt es durchaus Situationen, in denen ein optimaler Entscheidungsbaum nicht durch den Fragenvorrat dargestellt werden kann.

Satz 2 *Gegeben sei ein Suchraum S mit Verteilung p und für jedes $K \subseteq S$ sei die Frage „Ist $x \in K$?" mit Antwortmenge $A := \{J, N\}$ zulässig. Dann lassen sich die Objekte aus S mit einer durchschnittlichen Anzahl von $L_A(S)$ vielen Fragen identifizieren.*

Beweis. Sei f eine optimale Codierung von S über A und T der von C im Baum aller Wörter über A induzierte Baum. Für jedes Nicht-Blatt $w \in V(T) \setminus C$ sei $K_w := \{v \in C : v \text{ hat das Präfix } wJ\}$ die Menge der Blätter im J-Zweig von w. An die Ecke w schreibt man dann die Frage „Ist x in K_w?". □

Als Beispiel betrachten wir den Suchraum $S := \{1, 2, 3, 4, 5, 6, 7, 8\}$. Durch Fragen der Form „Ist $x \in K$?" mit $K \subseteq S$ soll ein unbekanntes Element aus S identifiziert werden. Es ist bekannt, daß mit folgenden Wahrscheinlichkeiten s das gesuchte Element ist (Angaben in Prozent):

s	1	2	3	4	5	6	7	8
$p(s)$	5	5	20	50	5	5	5	5

Gefragt ist nach einem Entscheidungsbaum, der dies mit einer minimalen durchschnittlichen Anzahl Fragen kann. Hierzu konstruieren wir zunächst den Huffman-Code für S mit p und daraus den Entscheidungsbaum wie in Abbildung 3.1. Die durchschnittliche Anzahl Fragen ist gleich $L_A(S)$, also gleich

$$1 \cdot 50\% + 3 \cdot 20\% + 4 \cdot 30\% = 2.3,$$

was gegenüber der durchschnittlichen Anzahl Fragen in einem balancierten Fragebaum mit acht Blättern zu Antwortfolgen der Länge 3 eine deutliche Verbesserung bedeutet. Der ausbalancierte Baum hat dagegen ein besseres worstcase-Verhalten: Man kommt dort immer mit drei Fragen zum Ziel, während

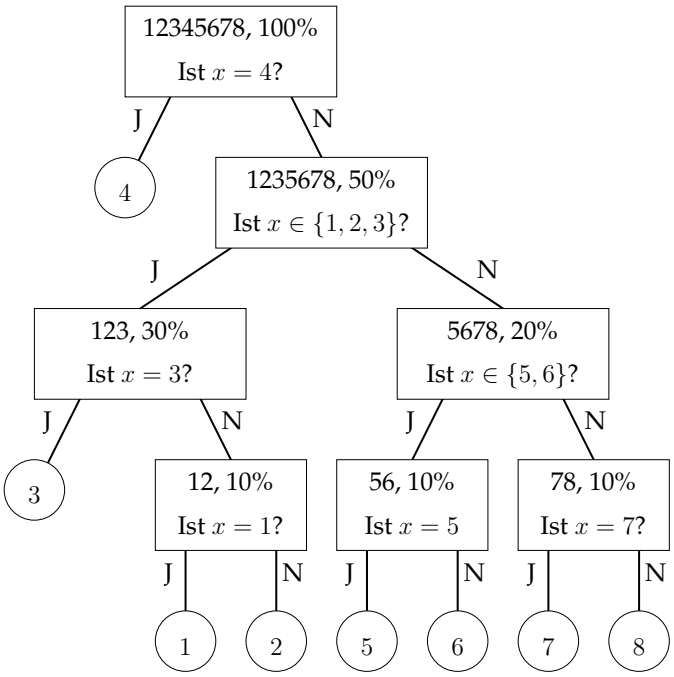

Abbildung 3.1: Entscheidungsbaum. In den Blättern (rund) steht das Suchergebnis, an den übrigen Ecken steht in der oberen Zeile der (verbleibende) Suchraum nebst seiner Wahrscheinlichkeitsmasse und in der unteren Zeile die nächste Frage. Die Frage „Ist $x = a$?" ist natürlich äquivalent zur formal zulässigen Frage „Ist $x \in \{a\}$?", wir verwenden sie der besseren Lesbarkeit wegen.

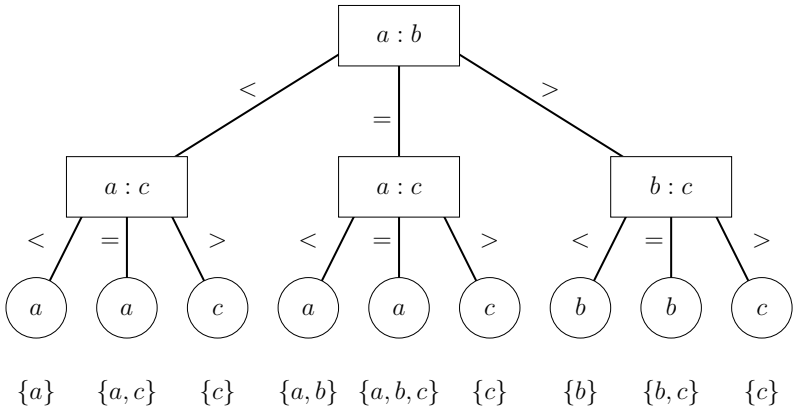

Abbildung 3.2: Entscheidungsbaum für ein Wiegeproblem. An den Blättern findet man einen Gegenstand minimalen Gewichtes. Unter den Blättern kann man die Menge aller Gegenstände minimalen Gewichtes ablesen, wie sie sich aus der Antwortfolge ergibt. Statt *eines* Gegenstandes minimalen Gewichtes findet man mit demselben Entscheidungsbaum *alle* Gegenstände minimalen Gewichts.

in immerhin 40% der Fälle man mit vier Fragen im Entscheidungsbaum aus Abbildung 3.1 dabei ist.

Eine weitere einfach zu beschreibende Problemklasse sind sogenannte Wiegeprobleme. Das Meßinstrument ist eine skalenlose Balkenwaage mit zwei Schalen, die eine Aussage darüber erlaubt, ob die Masse auf der linken Waagschale kleiner, gleich oder größer als die auf der rechten ist. Gegeben ist eine Menge M von Gegenständen nebst einer Gewichtsfunktion $w : M \to \mathbb{R}$, und Fragen sind durch zwei disjunkte Teilmengen L und R von M gegeben, über die per Wiegung eine der drei Aussagen $w(A) < w(B)$, $w(A) = w(B)$, $w(A) > w(B)$ getroffen werden kann. (Dabei ist natürlich $w(X) = \sum_{x \in X} w(x)$.) Die Frage wird durch $A : B$ abgekürzt, die Antwort durch $<, =, >$. Die Antwortmenge ist $A = \{<, =, >\}$, der Suchraum hängt von der konkreten Problemstellung ab. Betrachten wir als Beispiel $M = \{a, b, c\}$. Wenn wir auf einen Gegenstand minimalen Gewichts aus sind, dann ist der Suchraum $S = M$ und wir kommen nach nur *zwei* Suchabfragen zu einem Gegenstand minimalen Gewichts (siehe Abbildung 3.2). Zwar ist die „informationstheoretische untere Schranke" aus Satz 1 gleich $\lceil \log_{|A|} |C| \rceil = 1$; man kann sich aber leicht überlegen, daß mit einer einzigen Wiegung keinen Gegenstand minimalen Gewichts finden kann. Tatsächlich läßt sich mit demselben Entscheidungsbaum schon die Menge *aller* Gegenstände minimalen Gewichts ermitteln. Hier besteht der Suchraum aus allen sieben nichtleeren Teilmengen von $\{a, b, c\}$. Die informationstheoretische Schranke aus Satz 1 ist gleich $\lceil \log_3 7 \rceil = 2$, wird also angenommen.