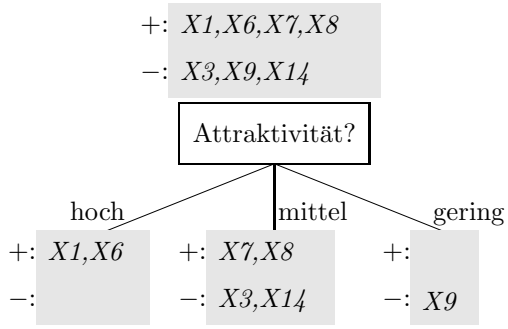
Lösungshinweise zu Kapitel 5

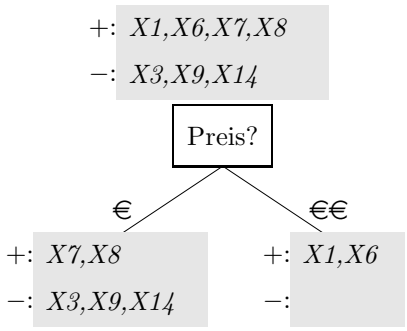
zu Selbsttestaufgabe 5.4 (Attributauswahl)

1. Für jedes der neun übriggebliebenen Attribute geben wir die Aufteilung analog zu Abbildung 5.4 an:

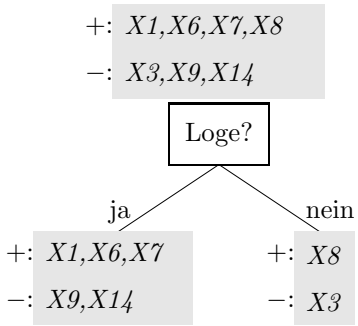
(a)



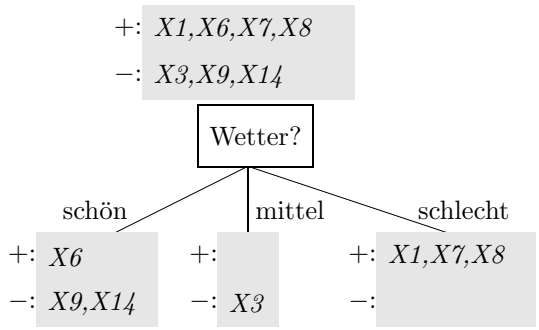
(b)



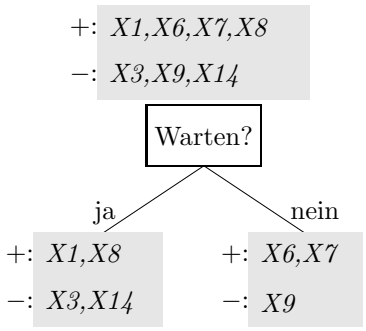
(c)



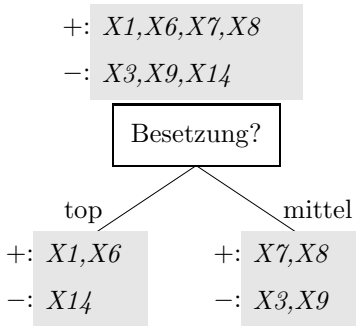
(d)


+: X1,X6,X7,X8
−: X3,X9,X14
Wetter?
schön
mittel
schlecht
+: X6
−: X9,X14
+:
−: X3
+: X1,X7,X8
−:

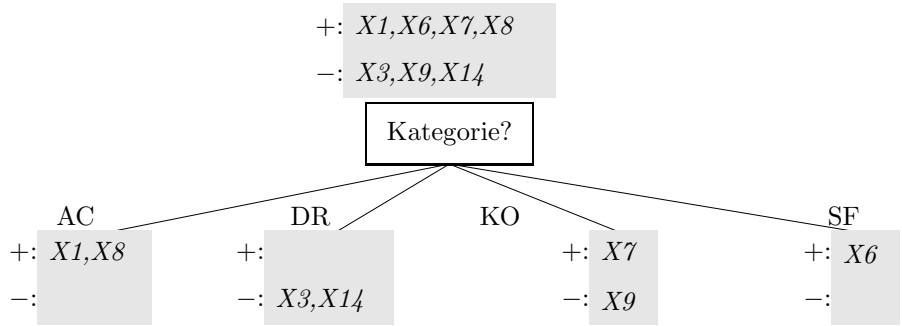
(e)


+: X1,X6,X7,X8
−: X3,X9,X14
Warten?
ja
nein
+: X1,X8
−: X3,X14
+: X6,X7
−: X9

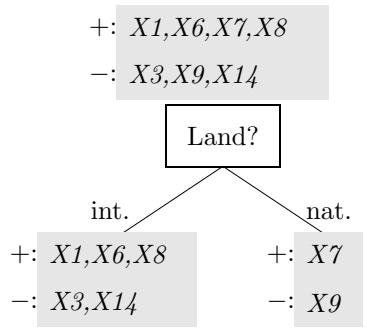
(f)


+: X1,X6,X7,X8
−: X3,X9,X14
Besetzung?
top
mittel
+: X1,X6
−: X14
+: X7,X8
−: X3,X9

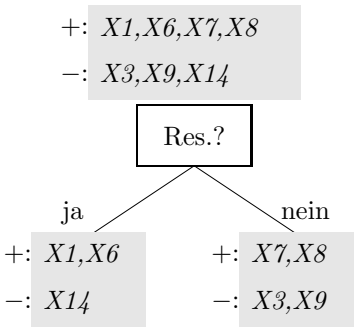
(g)


+: X1,X6,X7,X8
−: X3,X9,X14
Kategorie?
AC
DR
KO
SF
+: X1,X8
−:
+:
−: X3,X14
+: X7
−: X9
+: X6
−:

(h)



(i)



2. Mit dem Attribut *Kategorie* können also 5 der verbliebenen Beispiele $(X_1, X_3, X_6, X_8, X_{14})$ endgültig klassifiziert werden. Bei allen anderen Attributen sind es weniger Beispiele. Bezüglich dieses Kriteriums ist daher *Kategorie* hier das wichtigste Attribut.

zu Selbsttestaufgabe 5.5 (Lernen von Entscheidungsbäumen) Der Teil des von DT erzeugten Entscheidungsbaumes mit den Attributen *Gruppe* und *Wetter* ist in Abbildung 5.4(c) angegeben. Gemäß Selbsttestaufgabe 5.4 ist für die verbleibende, noch nicht klassifizierte Beispielmenge *Kategorie* das wichtigste Attribut. Nach dem Test von *Kategorie* sind nur noch die Fälle $\{X_7, X_9\}$ zu klassifizieren. Die beiden einzigen Attribute, die diese beiden Fälle unterscheiden, sind die Attribute *Attraktivität* und *Wetter*. Da beide Attribute beide Fälle endgültig klassifizieren, sind beide Attribute nach dem gegebenen Kriterium gleich wichtig. Zur weiteren Bearbeitung wählen wir das Attribut *Wetter* aus. (Vergewissern Sie sich bitte, dass die folgenden Überlegungen aber auch gelten, wenn wir stattdessen das Attribut *Attraktivität* auswählen.)

1. Der von DT erzeugte Entscheidungsbaum ist in Abbildung L.6 zu sehen.

2. Der von DT erzeugte Baum ist offensichtlich einfacher und kompakter. Es werden aber alle Beispiele aus der Trainingsmenge von beiden Entscheidungsbäumen gleich klassifiziert. (Andernfalls würde DT nicht korrekt arbeiten.)

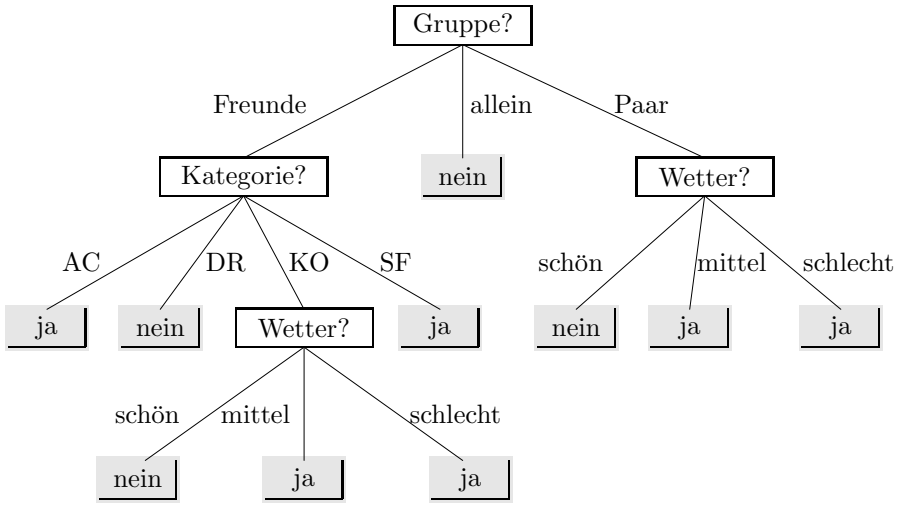


Abbildung L.6 Der von DT erzeugte Entscheidungsbaum zu Aufgabe 5.5

3. Die Trainingsmenge enthält keine Beispiele für den Fall, dass die Attraktivität des Filmes mittelmäßig ist, man alleine ins Kino geht und nicht warten muss. In dem Entscheidungsbaum aus Abbildung 5.2 werden alle Fälle mit diesen Attributwerten positiv klassifiziert. In dem von DT erzeugten Baum würde aber ein solcher Fall mit dem Attributwert $Gruppe = allein$ negativ klassifiziert. Jedes Beispiel mit den Attributwerten

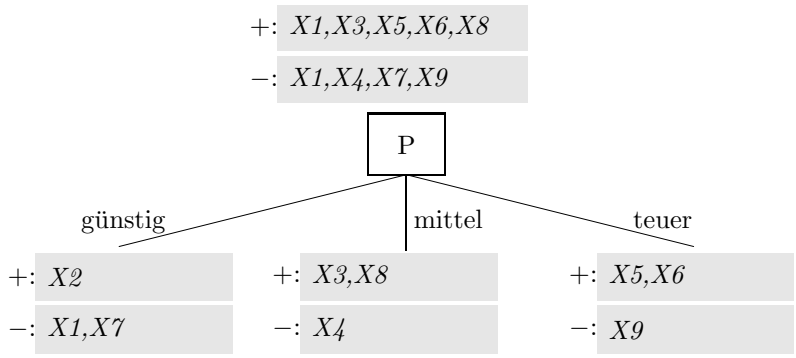
$$\begin{array}{lcl} \mathit{Attraktivität} & = & \text{mittel} \\ \mathit{Warten} & = & \text{nein} \\ \mathit{Gruppe} & = & \text{allein} \end{array}$$

würde also von den beiden Bäumen unterschiedlich klassifiziert werden.

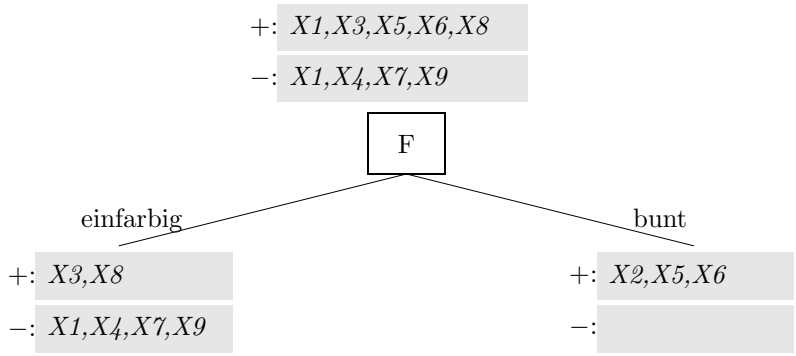
zu Selbsttestaufgabe 5.6 (Feuerwerk)

1. Wahl des Attributs für die erste Ebene des Entscheidungsbaums:

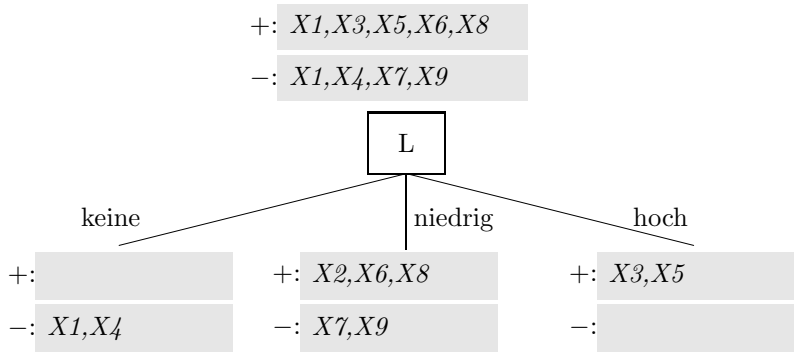
   (a) Auswahl von Attribut P:



   (b) Auswahl von Attribut F:



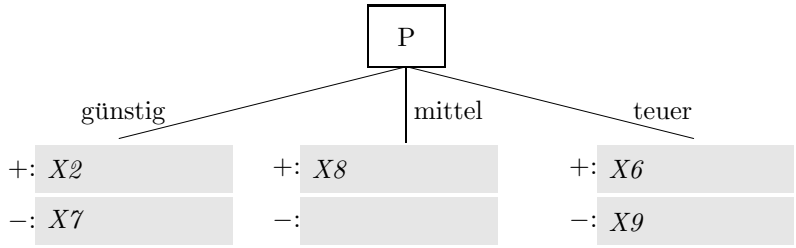
   (c) Auswahl von Attribut L:



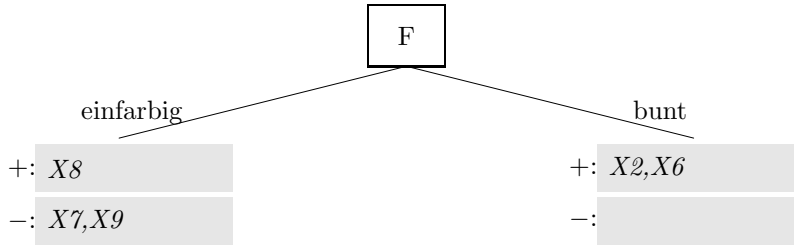
Offensichtlich können durch Attribut L auf der ersten Ebene des Entscheidungsbaums 4 Beispiele der Trainingsmenge vollständig klassifiziert werden, während durch Attribut F nur 3 Beispiele vollständig klassifiziert werden können und durch Attribut P keines der Beispiele vollständig klassifiziert werden kann. Damit wird gemäß dem Kardinalitätskriterium L als Attribut auf der ersten Ebene des Entscheidungsbaums gewählt.

2. Wahl des Attributs für die zweite Ebene des Entscheidungsbaums für den Fall $L =$ niedrig zur weiteren Klassifikation der Beispiele $X2, X6, X7, X8, X9$:

   (a) Auswahl von Attribut P:

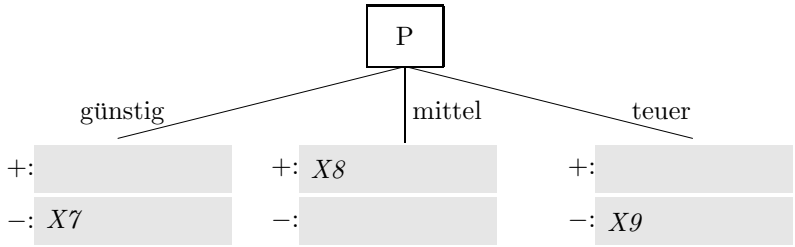


   (b) Auswahl von Attribut F:



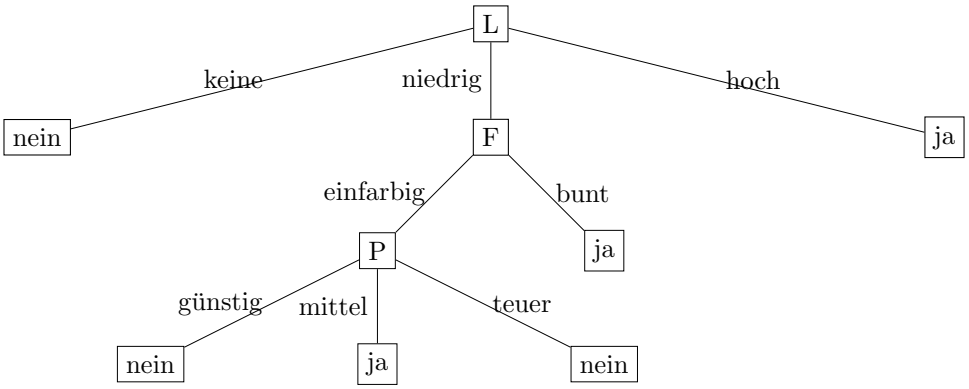
   Offensichtlich können durch Attribut F auf der zweiten Ebene des Entscheidungsbaums 2 Beispiele der Trainingsmenge vollständig klassifiziert werden, während durch Attribut P nur 1 Beispiel vollständig klassifiziert werden kann. Damit wird gemäß dem Kardinalitätskriterium F als Attribut auf der zweiten Ebene des Entscheidungsbaums gewählt.

3. Da für den Fall $F =$ *einfarbig* noch nicht alle Beispiele vollständig klassifiziert sind, wird im Entscheidungsbaum noch eine dritte Ebene benötigt. Für diese steht nur noch das Attribut P zur Verfügung:



Offensichtlich werden auf der dritten Ebene durch Attribut P alle verbliebenen Beispiele vollständig klassifiziert.

4. Insgesamt ergibt sich damit der folgende Entscheidungsbaum:



zu Selbsttestaufgabe 5.7 (Entlassung)

1. Wir untersuchen zunächst, wie die Beispiele von den Attributen jeweils klassifiziert werden:

   (a) Abteilung

EDV	
+	1,3
-	2,4

Kundenbetreuung	
+	5,6
-	7

Marketing	
+	10
-	8,9

   (b) Firmenzugehörigkeit

kurz	
+	1,3,5
-	2,7,8

lang	
+	6,10
-	4,9

   (c) Alter

jung	
+	1,5
-	2,7,8,9

älter	
+	3,6,10
-	4

   (d) Tätigkeit

Sachbearbeiter	
+	1,10
-	4,7,8

Führungsposition	
+	3,5,6
-	2,9

   Keins der Beispiele wird durch eines der Attribute eindeutig klassifiziert. Als erstes Attribut wählen wir daher das erste: “Abteilung”. Dies ist die Wurzel des aufzubauenden Entscheidungsbaumes.

Für den mit "EDV" markierten Ast suchen wir ein Attribut, das möglichst viele der Beispiele $\{1, 2, 3, 4\}$ eindeutig klassifiziert. Jedoch separiert keines der Attribute die negativen Beispiele $\{2, 4\}$ vollständig von den positiven $\{1, 3\}$. Das Attribut "Firmenzugehörigkeit" klassifiziert als einziges Attribut eines der Beispiele vollständig, nämlich Beispiel 4. Damit sind nur noch die Beispiele $\{1, 2, 3\}$ zu klassifizieren. Auch hierfür gibt es noch kein Attribut, das alle Beispiele vollständig einordnet. Vom Attribut "Alter" wird jedoch Beispiel 3 vollständig klassifiziert; vom Attribut "Tätigkeit" wird ebenfalls ein Beispiel eindeutig klassifiziert. Es gilt also die vorgegebene Attributreihenfolge. Die Beispiele 1 und 2 werden über das Attribut "Tätigkeit" separiert. Damit endet der "EDV- Ast" des Entscheidungsbaumes.

In dem mit "Kundenbetreuung" markierten Ast des Entscheidungsbaumes sind die positiven Beispiele 5, 6 und das negative Beispiel 7 zu untersuchen. Das Attribut "Tätigkeit" separiert die negativen und positiven Beispiele, da die Beispiele 5 und 6 unter die Kategorie "Führungsposition" fallen und das Beispiel 7 unter "Sachbearbeiter". Es müssen also keine weiteren Attribute betrachtet werden.

Im Zweig "Marketing" ist gleichfalls nur noch ein weiteres Attribut heranzuziehen: Da das positive Beispiel 10 unter "älter" fällt und die beiden negativen Beispiele 8 und 9 unter "jung" eingeordnet sind, ergibt sich hier eine vollständige Klassifizierung (s. Abbildung L.7).

2. Die aus diesem Entscheidungsbaum abzulesenden Entscheidungsregeln lauten:

   (R1) **if** EDV **and** kurz_in_Firma **and** jung **and** Sachbearbeiter **then** entlassen

   (R2) **if** EDV **and** kurz_in_Firma **and** jung **and** Führungsposition **then** nicht_entlassen

   (R3) **if** EDV **and** kurz_in_Firma **and** älter **then** entlassen

   (R4) **if** EDV **and** lang_in_Firma **then** nicht_entlassen

   (R5) **if** Kundenbetreuung **and** Sachbearbeiter **then** nicht_entlassen

   (R6) **if** Kundenbetreuung **and** Führungsposition **then** entlassen

   (R7) **if** Marketing **and** jung **then** nicht_entlassen

   (R8) **if** Marketing **and** älter **then** entlassen

3. Die Unternehmensberatung hätte also in etwa sagen können: "Entlassen Sie die Leute wieder, die Sie erst kürzlich für die EDV-Abteilung eingestellt haben (sie finden schnell was Neues), außer den jungen Leuten, die bereits in Führungspositionen tätig sind. Die Sachbearbeiter in Ihrer Kundenbetreuung arbeiten gut, werden jedoch schlecht geführt, versetzen Sie die Führungskräfte. Ihr Marketing ist altmodisch; verjüngen Sie die Abteilung, indem Sie den Älteren andere Aufgaben zuweisen."

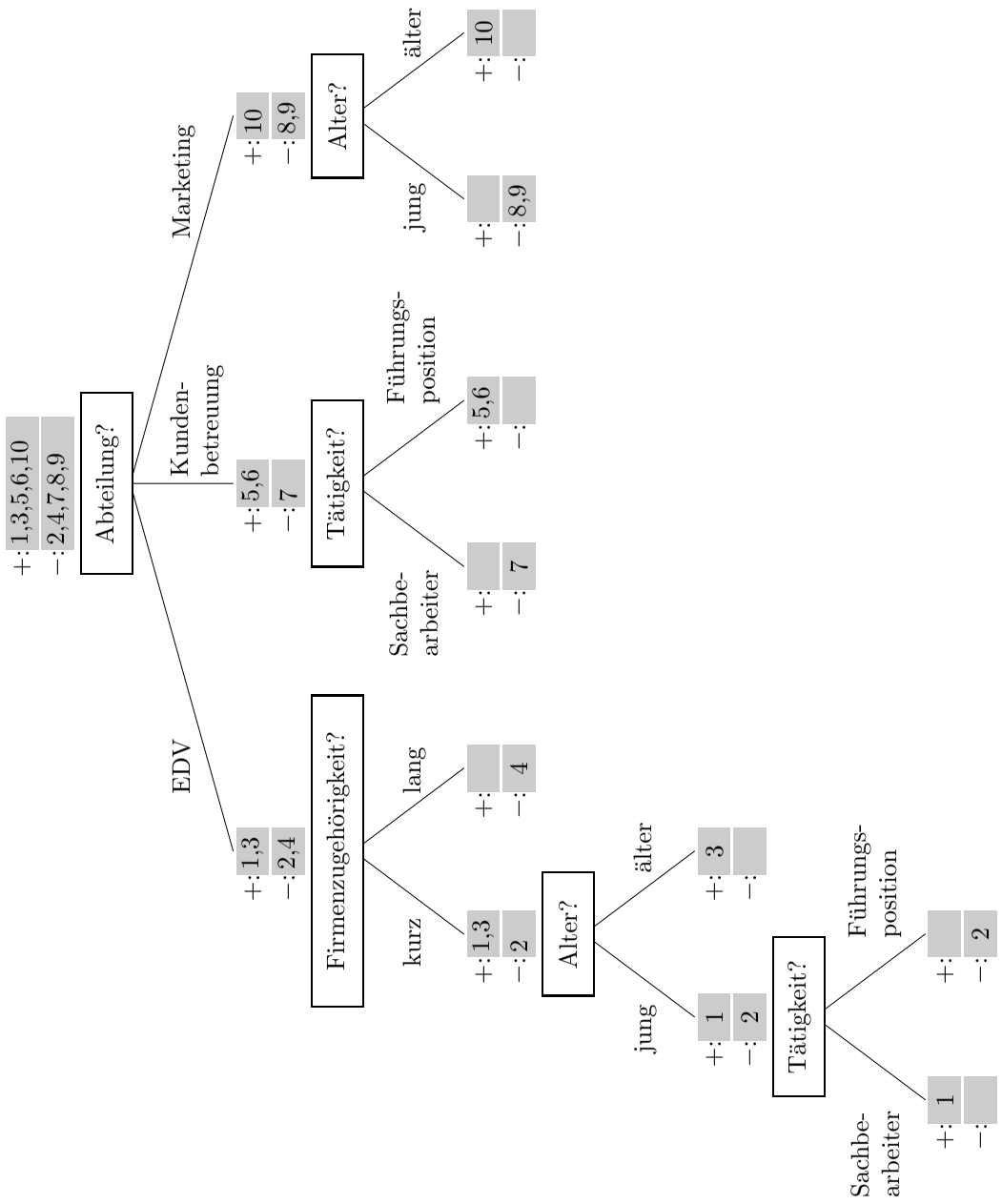


Abbildung L.7 Entscheidungsbaum zu Selbsttestaufgabe 5.7, Teil 1

4. Die Behauptung ist in dieser plakativen Form unzutreffend. Beispielsweise lassen sich die Beispiele 2 und 5 ohne Betrachten der Abteilung nicht unterschiedlich klassifizieren; es handelt sich beide Male um junge, der Firma erst kurz angehörende Führungskräfte, die eine wird jedoch entlassen, die andere nicht.

5. Es ergibt sich der in Abbildung L.8 gezeigte Entscheidungsbaum.

   Die Interpretation ist hier ein wenig anders; das erste Augenmerk gilt der Unterscheidung in Sachbearbeiter und Führungskräfte. Der Rat hätte also auch lauten können: "Behalten Sie Ihre Sachbearbeiter bis auf die jungen Leute in der EDV und die Älteren im Marketing. Behalten Sie die jungen Führungskräfte in der EDV und im Marketing und entlassen Sie die anderen." Oder anders ausgedrückt: "Entlassen Sie Sachbearbeiter, wenn sie jung sind und in der EDV-Abteilung arbeiten oder wenn sie älter sind und in der Abteilung 'Marketing' angestellt sind. Entlassen Sie die jungen Führungskräfte in der Kundenbetreuung und alle Führungskräfte über 35." An dieser Konstruktion des Entscheidungsbaums zeigt sich auch, dass das Attribut "Firmenzugehörigkeit" für die Klassifizierung entbehrlich ist und dass die Reihenfolge der Attributauswahl wesentlich ist: Ein junger Sachbearbeiter, der schon lange in der EDV arbeitet, wird nach den Entscheidungsregeln in Teil 2 nicht entlassen – nach dem Entscheidungsbaum der Zusatzaufgabe schon.

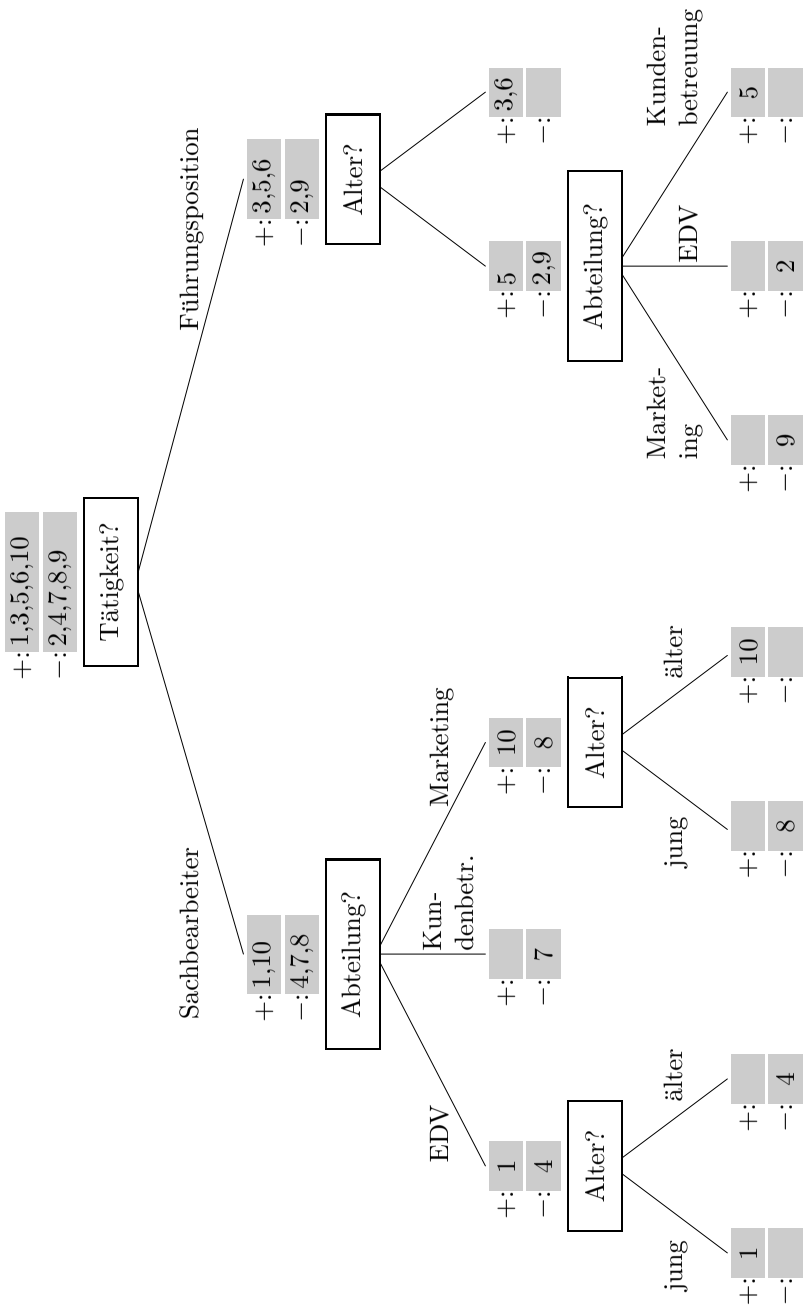


Abbildung L.8 Entscheidungsbaum zu Selbsttestaufgabe 5.7, Teil 5

zu Selbsttestaufgabe 5.8 (Informationsgehalt, Entropie) In der Trainingsmenge des Kinoproblems gibt es $p = 8$ positive und $n = 7$ negative Beispiele. Der Informationsgehalt der Antwort beträgt somit

$$H(\frac{p}{p+n}; \frac{n}{p+n}) = H(\frac{8}{15}; \frac{7}{15}) \approx 0.9968\, bit$$

zu Selbsttestaufgabe 5.10 (Informationsgewinn) Wir bestimmen nacheinander alle bedingten mittleren Informationsgehalte (bedingte Entropien) für das Kinoproblem:

$I(E \mid \textit{Attraktivität}$ bekannt)	≈ 0.8109,	$I(E \mid \textit{Preis}$ bekannt)	≈ 0.9837
$I(E \mid \textit{Loge}$ bekannt)	≈ 0.9675,	$I(E \mid \textit{Wetter}$ bekannt)	≈ 0.8569
$I(E \mid \textit{Warten}$ bekannt)	≈ 0.9118,	$I(E \mid \textit{Besetzung}$ bekannt)	≈ 0.9453
$I(E \mid \textit{Kategorie}$ bekannt)	≈ 0.9334,	$I(E \mid \textit{Land}$ bekannt)	≈ 0.9675
$I(E \mid \textit{Reservierung}$ bekannt)	≈ 0.9837,	$I(E \mid \textit{Gruppe}$ bekannt)	≈ 0.7004

Da $I(E)$ für alle Attribute a gleich ist, ist $gain(a)$ genau dann maximal, wenn $I(E \mid a$ bekannt) minimal ist. Dies ist für $a = \textit{Gruppe}$ der Fall.

zu Selbsttestaufgabe 5.11 (Informationsgewinn) In $E_{Rest} = \{X_1, X_3, X_6, X_7, X_8, X_9, X_{14}\}$ sind vier positive und drei negative Beispiele. Daher ist

$$I(E_{Rest}) = H(\frac{4}{7}; \frac{3}{7}) \approx 0.9852\, bit$$

1. Die Lösungshinweise zu Selbsttestaufgabe 5.4(a) zeigen, wie die drei möglichen Attributwerte zu *Attraktivität* die Menge E_{Rest} aufteilen. Analog zu Beispiel 5.9 erhalten wir:

$$\begin{aligned} gain(\textit{Attraktivität}) &= I(E_{Rest}) - I(E_{Rest} \mid \textit{Attraktivität} \text{ bekannt}) \\ &\approx 0.9852 - [\tfrac{2}{7}H(1;0) + \tfrac{4}{7}H(\tfrac{2}{4};\tfrac{2}{4}) + \tfrac{1}{7}H(0;1)] \\ &\approx 0.9852 - \tfrac{4}{7} \cdot 1 \\ &\approx 0.9852 - 0.5714 \\ &= 0.4138\, bit \end{aligned}$$

2. Die Lösungshinweise zu Selbsttestaufgabe 5.4(g) zeigen, wie die vier möglichen Attributwerte zu *Kategorie* die Menge E_{Rest} aufteilen. Damit gilt:

$$\begin{aligned} gain(Kategorie) &= I(E_{Rest}) - I(E_{Rest} \mid \textit{Kategorie} \text{ bekannt}) \\ &\approx 0.9852 - [\tfrac{2}{7}H(1;0) + \tfrac{2}{7}H(0;1) + \tfrac{2}{7}H(\tfrac{1}{2};\tfrac{1}{2}) + \tfrac{1}{7}H(1;0)] \\ &\approx 0.9852 - \tfrac{2}{7} \cdot 1 \\ &\approx 0.9852 - 0.2857 \\ &= 0.6995\, bit \end{aligned}$$

zu Selbsttestaufgabe 5.12 (gain ratio) Wir berechnen auf der Grundlage von Abbildung 5.3

$$\begin{aligned} \textit{split info}(\textit{Attraktivität}) &= H(\tfrac{4}{15};\tfrac{9}{15};\tfrac{2}{15}) \approx 1.3383 \\ \textit{split info}(Wetter) &= H(\tfrac{4}{15};\tfrac{6}{15};\tfrac{5}{15}) \approx 1.5656 \\ \textit{split info}(Gruppe) &= H(\tfrac{7}{15};\tfrac{5}{15};\tfrac{3}{15}) \approx 1.5058 \end{aligned}$$

Aus den Ergebnissen der Selbsttestaufgabe 5.10 erhalten wir dann

$$\begin{aligned} \textit{gain ratio}(\textit{Attr.}) &\approx \tfrac{0.9968-0.8109}{1.3383} \approx 0.1389 \\ \textit{gain ratio}(Wetter) &\approx \tfrac{0.9968-0.8569}{1.5656} \approx 0.0894 \\ \textit{gain ratio}(Gruppe) &\approx \tfrac{0.9968-0.7004}{1.5058} \approx 0.1968 \end{aligned}$$

Also ist auch nach dem *gain ratio*-Kriterium das Attribut *Gruppe* das beste (dieser drei Attribute).

zu Selbsttestaufgabe 5.13 (Auftragsmanagement)

1. Wir untersuchen zunächst, wie die Beispiele von den Attributen jeweils klassifiziert werden; dies ist in Abbildung L.9 dargestellt.

   Wir haben $E = \{1,\ldots,12\}$ mit den positiven Beispielen 3,5,7,11,12 und den negativen Beispielen 1,2,4,6,9,10. Daher ist die Anzahl der positiven Beispiele $p = 6$ und die Anzahl der negativen Beispiele $n = 6$. Also ist

   $$I(E) = H\big(\frac{p}{p+n}\big) + H\big(\frac{n}{p+n}\big) = -2 \cdot (0.5 \cdot log_2 0.5) = 1$$

   Wir berechnen für die Attribute *Bereich, Aufwand, Attraktivität* und *Bauchgefühl* den Informationsgewinn (*gain*).

   Für *Bereich* haben wir:

   $p_1 = 1$ positive Beispiele für die Ausprägung „Handwerker“
   $n_1 = 3$ negative Beispiele für die Ausprägung „Handwerker“
   $p_2 = 3$ positive Beispiele für die Ausprägung „Beratungsnetz“
   $n_2 = 1$ negative Beispiele für die Ausprägung „Beratungsnetz“
   $p_3 = 2$ positive Beispiele für die Ausprägung „Online-Shop“
   $n_3 = 2$ positive Beispiele für die Ausprägung „Online-Shop“

   Daraus berechnet sich

   $$\begin{aligned} \textit{gain}(\textit{Bereich}) &= I(E) - I(E|\textit{Bereich bekannt}) \\ &= 1 - \textstyle\sum_{i=1}^{3} \frac{p_i+n_i}{12} H(\frac{p_i}{p_i+n_i};\frac{n_i}{p_i+n_i}) \\ &= 1- \\ &\quad \big((\tfrac{4}{12}\cdot(-1)\cdot(\tfrac{1}{4}log_2\tfrac{1}{4}+\tfrac{3}{4}log_2\tfrac{3}{4})) \\ &\quad +(\tfrac{4}{12}\cdot(-1)\cdot(\tfrac{3}{4}log_2\tfrac{3}{4}+\tfrac{1}{4}log_2\tfrac{1}{4})) \\ &\quad +(\tfrac{4}{12}\cdot(-1)\cdot(\tfrac{2}{4}log_2\tfrac{2}{4}+\tfrac{2}{4}log_2\tfrac{2}{4}))\big) \\ &= 0.126 \end{aligned}$$

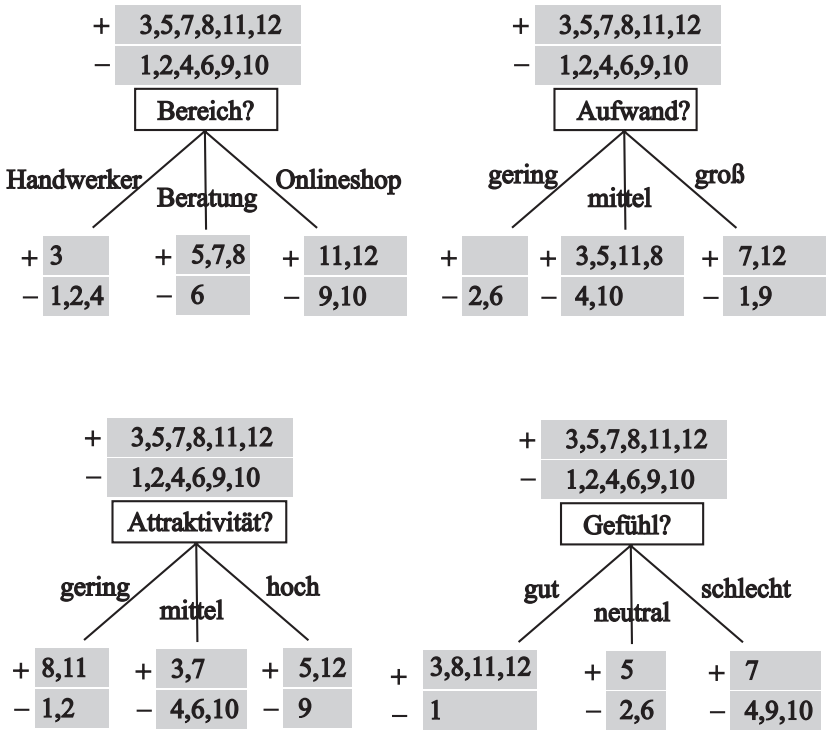


Abbildung L.9 Aufteilung nach Attributen (Selbsttestaufgabe 5.13)

Als nächstes berechnen wir *split info* von *Bereich*:

$$H(Bereich) = -\frac{4}{12} \cdot log_2 \frac{4}{12} - \frac{4}{12} \cdot log_2 \frac{4}{12} - \frac{4}{12} \cdot log_2 \frac{4}{12} = 1.58$$

Daraus ergibt sich

$$gain\ ratio(Bereich) = \frac{gain(Bereich)}{splitinfo(Bereich)} = \frac{0.126}{1.58} = 0.08$$

Auf demselben Rechenweg ergibt sich

$$\begin{aligned} gainratio(Aufwand) &= 0.14 \\ gain\ ratio\ (Attraktivität) &= 0.02 \\ gain\ ratio\ (Bauchgefühl) &= 0.13 \end{aligned}$$

Im Entscheidungsbaum wird also zuerst nach dem Attribut „Aufwand“ differenziert. Für die Ausprägung „gering“ des Attributs „Aufwand“ sind bereits

beide Beispiele vollständig klassifiziert. Für die beiden anderen Ausprägungen betrachten wir wiederum die Aufteilung nach den anderen Atributen und berechnen *gain ratio* (vgl. Abbildung L.10).

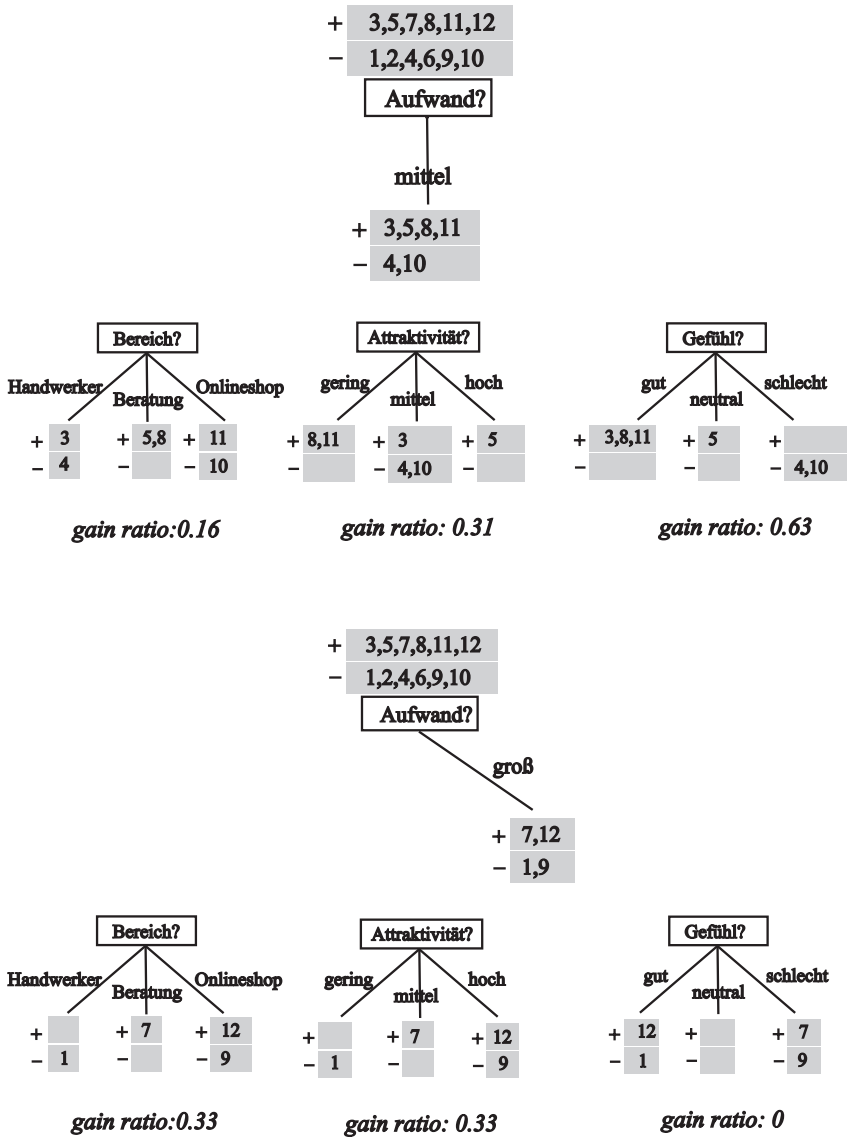


Abbildung L.10 Aufteilung nach „Aufwand = mittel“ und „Aufwand = groß“ (Selbsttestaufgabe 5.13)

Im Ast „Aufwand = mittel“ wird, wenn man als nächstes das Attribut „Bauchgefühl“ wählt, die Beispielmenge vollständig aufgeteilt. Dies spiegelt sich auch im maximalen Wert von *gain ratio* wieder.

Im „Aufwand: groß“-Ast fällt die Wahl auf das Attribut „Bereich“.

Nun ist nur noch für die Ausprägung „Online-Shop“ zu entwickeln. Nur das Attribut „Bauchgefühl“ bewirkt eine vollständige Klassifizierung der verbliebenen Beispiele. Der daraus resultierende Entscheidungsbaum ist in Abbildung L.11 dargestellt.

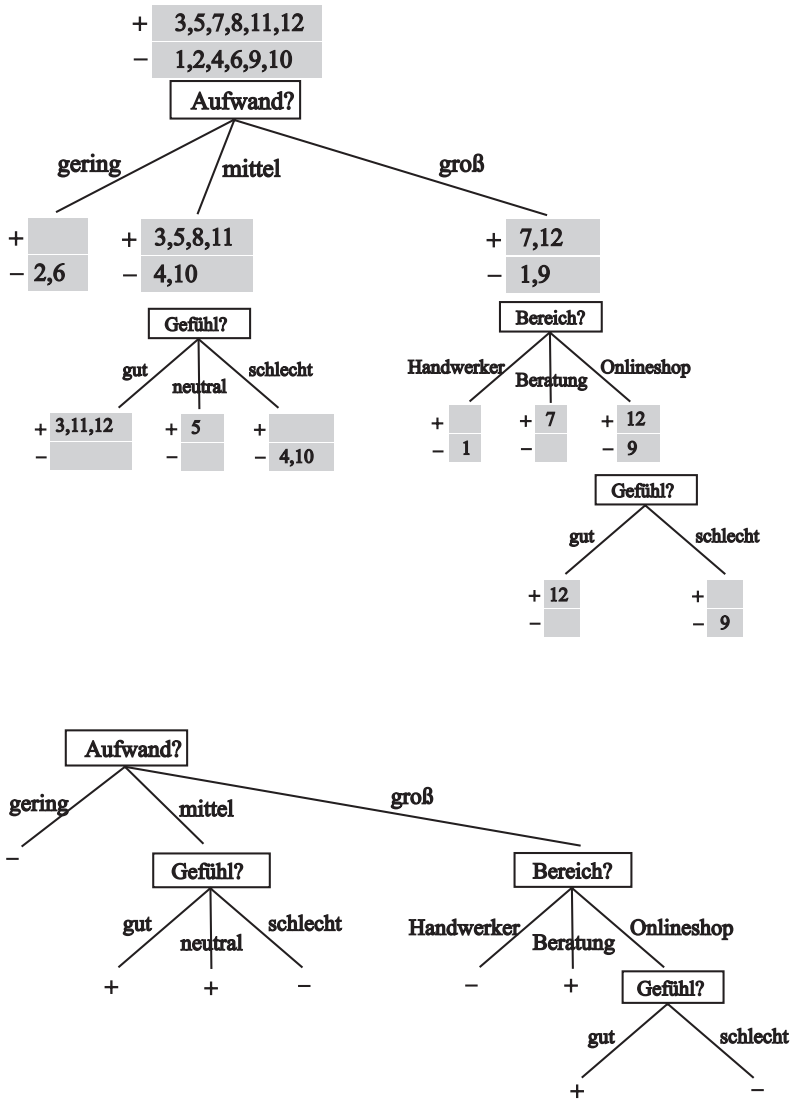


Abbildung L.11 Alle Beispiele sind klassifiziert; es ergibt sich der Entscheidungsbaum (Selbsttestaufgabe 5.13)

2. Die dem Entscheidungsbaum entsprechenden Entscheidungsregeln lauten:

   `if` Aufwand = gering
   `then` Entscheidung = nein

   `if` Aufwand = mittel `and` Gefühl = gut
   `then` Entscheidung = ja

   `if` Aufwand = mittel `and` Gefühl = neutral
   `then` Entscheidung = ja

   `if` Aufwand = mittel `and` Gefühl = schlecht
   `then` Entscheidung = nein

   `if` Aufwand = groß `and` Bereich = Handwerker
   `then` Entscheidung = nein

   `if` Aufwand = groß `and` Bereich = Beratung
   `then` Entscheidung = ja

   `if` Aufwand = groß `and` Bereich = Online-Shop `and` Gefühl=gut
   then Entscheidung = ja

   `if` Aufwand = groß `and` Bereich = Online-Shop `and` Gefühl = schlecht
   `then` Entscheidung = nein

3. Eine Formel E (in disjunktiver Normalform) lässt sich direkt aus den Entscheidungsregeln angeben, indem die Bedingungen für die positive Entscheidung verodert werden:

$$E := (am \wedge gg) \vee (am \wedge gn) \vee (ao \wedge bb) \vee (ao \wedge bo \wedge gg)$$

   Zusätzlich muss zum Ausdruck gebracht werden, dass eine Ausprägung eines Attributes die anderen Ausprägungen ausschließt, woraus sich eine Menge von Zusatzbedingungen ergibt:

$$\begin{array}{lll} Z & := & \neg(bh \wedge (bb \vee bo)) \wedge \neg(bb \wedge (bh \vee bo)) \wedge \neg(bo \wedge (bh \vee bb)) \\ & \wedge & \neg(ao \wedge (am \vee ag)) \wedge \neg(am \wedge (ao \vee ag)) \wedge \neg(ag \wedge (ao \vee am)) \\ & \wedge & \neg(atg \wedge (atm \vee ath)) \wedge \neg(atm \wedge (atg \vee ath)) \wedge \neg(ath \wedge (atg \vee atm)) \\ & \wedge & \neg(gg \wedge (gn \vee gs)) \wedge \neg(gn \wedge (gg \vee gs)) \wedge \neg(gs \wedge (gg \vee gn)) \end{array}$$

   Daraus ergibt sich die gesuchte Formel als $Z \wedge E$.

   Betrachtet man die ursprüngliche Tabelle, so kann man sie auch als Wahrheitstabelle auffassen, wobei die Klassifizierung dem Wahrheitswert entspricht (+ für 1 und - für 0). Daraus erhält man die folgende (längere) disjunktive Normalform

$$\begin{array}{ll} & (bh \wedge am \wedge atm \wedge gg) \vee (bb \wedge am \wedge ath \wedge gn) \\ \vee & (bb \wedge ag \wedge atm \wedge bs) \vee (bb \wedge am \wedge atg \wedge gg) \\ \vee & (bb \wedge ao \wedge atm \wedge gs) \vee (bb \wedge am \wedge atg \wedge gg) \\ \vee & (bo \wedge am \wedge atg \wedge gg) \vee (bo \wedge ao \wedge ath \wedge gg)) \end{array}$$

zu Selbsttestaufgabe 5.23 (Korrektheit und Vollständigkeit) Die Hypothese in S deckt die 3 positiven Beispiele aus der Trainingsmenge ab, während die negativen Beispiele X_3 und X_5 die Hypothese nicht erfüllen. Die Hypothese in S ist daher bzgl. der Menge aller Trainingsbeispiele korrekt und vollständig.

zu Selbsttestaufgabe 5.26 (*VS* und Begrenzungsmengen) Wir zeigen allgemein, dass nach jeder Bearbeitung eines Beispiels durch *VS* für die beiden Begrenzungsmengen S und G die Invariante

$$Inv(S,G) = Inv_1(S,G) \wedge Inv_2(S,G)$$

mit

$$\begin{aligned} Inv_1(S,G): &\quad \text{Für jedes } s \in S \text{ gibt es ein } g_s \in G \text{ mit } s \leq g_s. \\ Inv_2(S,G): &\quad \text{Für jedes } g \in G \text{ gibt es ein } s_g \in S \text{ mit } s_g \leq g. \end{aligned}$$

gilt. Für die initialen Mengen S_{init} und G_{init} der speziellsten und der allgemeinsten Hypothesen der Konzeptsprache gilt offensichtlich $Inv(S_{init}, G_{init})$. Wir nehmen nun an, dass S_{pre} und G_{pre} die aktuellen Begrenzungsmengen vor der Bearbeitung von e sind, dass $Inv(S_{pre}, G_{pre})$ gilt, dass e das nächste zu bearbeitende Beispiel ist und dass die Mengen S_{post} und G_{post} die Begrenzungsmengen nach der Bearbeitung von e gemäß Algorithmus *VS* in Abbildung 5.13 sind.

Bei einem positiven Beispiel werden in der Begrenzungsmenge der allgemeinsten Hypothesen gegebenenfalls Hypothesen entfernt, aber keine neuen hinzugefügt; entsprechendes gilt bei einem negativen Beispiel für die Begrenzungsmenge der speziellsten Hypothesen. In unserem Beweis werden wir daher die beiden folgenden Eigenschaften (*properties*) verwenden:

- Wenn e ein *positives* Beispiel ist, gilt:

$$\textbf{Prop}[G_{post}, e \text{ pos}]: \quad G_{post} = G_{pre} \setminus \{h \in G_{pre} \mid h(e) = 0\}$$

- Wenn e ein *negatives* Beispiel ist, gilt:

$$\textbf{Prop}[S_{post}, e \text{ neg}]: \quad S_{post} = S_{pre} \setminus \{h \in S_{pre} \mid h(e) = 1\}$$

Wir zeigen im Folgenden, dass $Inv_1(S_{post}, G_{post})$ gilt. Ganz analog lässt sich zeigen, dass auch $Inv_2(S_{post}, G_{post})$ gilt.

Sei $s \in S_{post}$. Für den Nachweis von $Inv_1(S_{post}, G_{post})$ müssen wir also zeigen, dass es ein $g_s \in G_{post}$ mit $s \leq g_s$ gibt:

- Falls $s \in S_{pre}$, dann gibt es wegen $Inv(S_{pre}, G_{pre})$ ein $g_{pre} \in G_{pre}$ mit $s \leq g_{pre}$. Es sind nun prinzipiell drei Fälle zu unterscheiden:

  1. Falls g_{pre} das Beispiel e korrekt klassifiziert, so gilt $g_{pre} \in G_{post}$ und wir setzen $g_s = g_{pre}$.

  2. Falls g_{pre} das Beispiel e fälschlicherweise negativ klassifiziert, muss wegen $s \leq g_{pre}$ auch s das Beispiel e fälschlicherweise negativ klassifizieren. Da dies aber im Widerspruch zu der Annahme steht, dass $s \in S_{post}$ gilt, kann dieser Fall nicht auftreten.

3. Falls g_{pre} das Beispiel e fälschlicherweise positiv klassifiziert, muss e ein negatives Beispiel sein, und wegen $g_{pre}(e) = 1$ ist $g_{pre} \in G_{pre}$ bei der Bearbeitung von e gemäß Algorithmus *VS* in Abbildung 5.13 durch die Hypothesen in der Menge $H \subseteq G_{post}$ mit

$$H = \{h' \mid h'(e) = 0,\ h' \leq g_{pre},\ \text{es gibt ein } s' \in S_{post} \text{ mit } s' \leq h' \\ \text{und es gibt kein } h'' \text{ mit } h''(e) = 0 \wedge h' < h'' \wedge h'' \leq g_{pre}\}$$

ersetzt worden, wobei S_{post} wie in **Prop**$[S_{post}, e \text{ neg}]$ angegeben ist. Wegen $s \in S_{post}$ muss gemäß **Prop**$[S_{post}, e \text{ neg}]$ daher $s(e) = 0$ gelten. Wegen $s(e) = 0$ und $g_{pre}(e) = 1$ muss es wegen $s \leq g_{pre}$ ein $h' \in H$ mit $s \leq h' < g_{pre}$ geben: damit können wir $g_s = h'$ setzen.

- Falls $s \notin S_{pre}$, dann muss e ein positives Beispiel sein und es muss eine Hypothese $s_{pre} \in S_{pre}$ geben, die e fälschlicherweise negativ klassifiziert und aus der s entstanden ist. D.h., wegen $s_{pre}(e) = 0$ ist $s_{pre} \in S_{pre}$ bei der Bearbeitung von e gemäß Algorithmus *VS* in Abbildung 5.13 durch die Hypothesen in der Menge $H \subseteq S_{post}$ mit

$$H = \{h' \mid h'(e) = 1,\ s_{pre} \leq h',\ \text{es gibt ein } g' \in G_{post} \text{ mit } h' \leq g' \\ \text{und es gibt kein } h'' \text{ mit } h''(e) = 1 \wedge s_{pre} < h'' \wedge h'' \leq h'\}$$

ersetzt worden, wobei G_{post} wie in **Prop**$[G_{post}, e \text{ pos}]$ angegeben ist. Da $s \in H$ gilt, gibt es daher nach Konstruktion von H auch ein $g_s \in G_{post}$ mit $s \leq g_s$.

zu Selbsttestaufgabe 5.27 (Versionenraumlernen) Der Grund, dass h nicht in die Menge G aufgenommen werden darf, ist, dass h unvollständig bzgl. der bisher aufgetretenen positiven Beispiele ist. Anstatt alle bisherigen Beispiele erneut bzgl. jeder neuen Hypothese in G zu testen, reicht es aus, die aktuelle Begrenzungsmenge S zu überprüfen. Da in S die speziellsten noch möglichen Hypothesen enthalten sind, muss jede andere zulässige Hypothese durch zumindest eine Hypothese in S begrenzt werden.

Im Algorithmus *VS* wird dies wie folgt erreicht: Im Fall eines negativen Beispiels wird eine neue Hypothese h' nur dann in G aufgenommen, wenn es ein $s \in S$ gibt mit $s \leq h'$ (vgl. drittletzte Zeile in Abbildung 5.13). Zu der Hypothese $h = \langle$ *Fußball, ?, ?, ?, ?* $\rangle$ gibt es jedoch kein solches s in der aktuellen Begrenzungsmenge S_3.

zu Selbsttestaufgabe 5.28 (Versionenraumlernen) Das Beispiel X_4 ist für die Hypothese $h = \langle$ *?, ?, draußen, ?, ?* $\rangle$ fälschlicherweise negativ. Da G aber bereits die allgemeinsten noch möglichen Hypothesen enthält, kann h nicht mehr verallgemeinert werden, ohne die Korrektheit zu verlieren. h muss daher aus G entfernt werden. Im Algorithmus *VS* (Abbildung 5.13) geschieht dies als erste Anweisung in der **if**-Klausel für positive Beispiele.

zu Selbsttestaufgabe 5.29 (CD-ROM-Produktion)

1. Wir starten das Verfahren mit den Mengen

   $\boxed{S_0 := \{< \emptyset, \emptyset, \emptyset, \emptyset, \emptyset >\}}$ und

   $\boxed{G_0 := \{<?, ?, ?, ?, ? >\}}$

   (a) X_1 ist ein positives Beispiel, das von der einzigen Hypothese in S_0 nicht abgedeckt wird. Daher wird S_0 minimal generalisiert zu

   $\boxed{S_1 := \{<$ Kinder, Spiel,***, Tom, Karin $>\}}$

   G_0 deckt X_1 wunschgemäß ab und bleibt daher unverändert:

   $\boxed{G_1 := G_0}$

   (b) X_2 ist ebenfalls positiv. Dieses Beispiel wird von G_1, aber nicht von S_1 erfasst, daher wird S_1 minimal generalisiert zu

   $\boxed{S_2 := \{<$?, Spiel, ***, Tom, Karin $>\}}$

   und wir setzen

   $\boxed{G_2 := G_1}$

   (c) X_3 ist ein negatives Beispiel, das von S_2 nicht abgedeckt wird, jedoch fälschlicherweise von G_2. Daher bleibt S_2 unverändert; G_2 muss minimal spezialisiert werden:

   $\boxed{S_3 := S_2}$

   $\boxed{G_3 := \{<$?,?, ***, ?, ? $>, <$?, ?, ?, ?, Karin $>\}}$

   (d) X_4 ist wieder ein positives Beispiel. Es wird von S_3 nicht erfasst, daher wird S_3 minimal generalisiert:

   $\boxed{S_4 := \{<$?, Spiel, ?, Tom, Karin $>\}}$

   X_4 wird von der Hypothese $<$?,?, ***, ?, ? $>$ in G_3 fälschlich als negativ bewertet, daher muss diese Hypothese aus G_3 entfernt werden. Die zweite Hypothesen in G_3 stuft X_4 zu Recht als positiv ein. Daher haben wir

   $\boxed{G_4 := \{<$?, ?, ?, ?, Karin $>\}}$

   (e) X_5 ist positiv und wird von der Hypothese in G_4 auch so bewertet. S_4 muss jedoch minimal generalisiert werden, da X_5 fälschlich als negativ bewertet wird. Wir erhalten also

   $\boxed{S_5 := \{<$?, ?, ?, Tom, Karin $>\}}$ und

   $\boxed{G_5 := G_4}$

   (f) X_6 ist negativ und wird von der Hypothese in S_5 auch so bewertet, jedoch von der Hypothese in G_5 fälschlich positiv eingestuft. Daher muss letztere spezialisiert werden; das Ergebnis ist $<$?, ?, ?, Tom, Karin $>$:

   $\boxed{S_6 := S_5 = \{<$?, ?, ?, Tom, Karin $>\} =: G_6}$

Damit endet das Verfahren mit der einzigen Hypothese

$$< ?, ?, ?, \text{Tom}, \text{Karin} >$$

Tom und Karin geben also offenbar das Erfolgsduo ab.

2. Die Mengen $S_1, \ldots, S_5$ bestimmen sich jeweils wie im ersten Teil.

   (a) X_6' ist negativ, wird jedoch von der Hypothese $< ?, ?, ?, \text{Tom}, \text{Karin} >$ in S_5 fälschlich positiv bewertet, daher muss diese Hypothese aus S_5 entfernt werden. Damit ist S_6' jedoch leer; der Versionenraum ist zur leeren Menge kollabiert.

   (b) X_6'' ist positiv, wird jedoch von der einzigen Hypothese in G_5 als negativ bewertet. Diese muss also aus G_5 entfernt werden, woraufhin der Versionenraum ebenfalls zur leeren Menge kollabiert.

zu Selbsttestaufgabe 5.33 (Assoziationsregeln) Seien X, Y, Y' Itemmengen, $Y' \subseteq Y \subset X$. Es ist

$$\begin{aligned} \mathit{confidence}((X - Y') \to Y') &= \frac{\mathit{support}(X)}{\mathit{support}(X - Y')} \\ \mathit{confidence}((X - Y) \to Y) &= \frac{\mathit{support}(X)}{\mathit{support}(X - Y)} \end{aligned}$$

da $(X - Y) \cup Y = X = (X - Y') \cup Y'$. $\mathit{confidence}((X - Y') \to Y')$ und $\mathit{confidence}((X - Y) \to Y)$ unterscheiden sich also nur in ihren Nennern. Weiter folgt aus $Y' \subseteq Y \subset X$ auch $(X - Y') \supseteq (X - Y)$ und daher $\mathit{support}(X - Y') \leq \mathit{support}(X - Y)$. Damit gilt die Behauptung.

zu Selbsttestaufgabe 5.34 (`AprioriGen`-Algorithmus) Dass $\{A, B, C, E\}$ in C_4 nicht enthalten ist, liegt an der Bedingung, dass nur Mengen aufgenommen werden, die sich aus zwei Mengen ergeben, deren $k-2$ erste Elemente (nach der definierten Ordnung) gleich sind. Um $\{A, B, C, E\}$ in C_4 aufzunehmen, müssten also $\{A, B, C\}$ und $\{A, B, E\}$ in L_3 enthalten sein, $\{A, B, E\}$ ist jedoch nicht in L_3 .

Allgemein gilt: Alle $k-1$-elementigen Teilmengen einer Menge in L_k müssen häufig sein, also Elemente von L_{k-1}. Jede Menge in L_k lässt sich geordnet darstellen:

$$\{e_1, \ldots, e_k\} \text{ mit } \; e_1 \leq e_2 \leq \cdots \leq e_k$$

Um den "Teilmengencheck" bestehen zu können, müssen insbesondere die Mengen

$$\{e_1, \ldots, e_{k-1}\} \text{ und } \; \{e_1, \ldots, e_{k-2}, e_k\}$$

in L_{k-1} enthalten sein. Lässt sich die Menge nicht in dieser Weise erzeugen, dann kann sie auch nicht häufig sein. Daher erübrigen sich alle anderen Möglichkeiten, Mengen für C_k zu erzeugen.

zu Selbsttestaufgabe 5.35 (Werbeagentur)

1. In der folgenden Tabelle sind die Sendungen zusammen mit ihrem Support angegeben:

L.	*Sendung*	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	z_{10}	z_{11}	z_{12}	*supp.*
s_1	Wer will Millionen?	•	•					•		•	•			$\frac{5}{12}$
s_2	Quiz mit Dirk Plauer			•		•	•		•			•	•	$\frac{6}{12}$
s_3	Quiz-Express		•			•		•			•			$\frac{4}{12}$
s_4	Birkenallee	•							•			•	•	$\frac{4}{12}$
s_5	Gut Josepha			•		•	•		•			•	•	$\frac{6}{12}$
s_6	Gutes Bier, schlechtes B.	•	•	•				•		•	•	•		$\frac{7}{12}$
s_7	Bettina Schläfer				•	•					•			$\frac{3}{12}$
s_8	Nora am Abend		•		•	•		•	•		•			$\frac{6}{12}$
s_9	Olli-G.-Show		•		•	•			•		•			$\frac{5}{12}$

Daraus ergibt sich die Menge der häufigen 1-Itemmengen

$$I_1 := \{\{s_1\}, \{s_2\}, \{s_5\}, \{s_6\}, \{s_8\}, \{s_9\}\}$$

Der *AprioriGen*-Algorithmus untersucht nun die aus den Mengen aus I_1 gebildeten 2-elementigen Itemmengen auf ihren Support.[1] Der Teilmengencheck erübrigt sich in der ersten Stufe. Das Ergebnis ist

$\{s_1, s_2\}[0]$ $\{s_1, s_5\}[0]$ $\{s_1, s_6\}[\frac{5}{12}]$ $\{s_1, s_8\}[\frac{3}{12}]$ $\{s_1, s_9\}[\frac{2}{12}]$
$\{s_2, s_5\}[\frac{6}{12}]$ $\{s_2, s_6\}[\frac{2}{12}]$ $\{s_2, s_8\}[\frac{2}{12}]$ $\{s_2, s_9\}[\frac{2}{12}]$
$\{s_5, s_6\}[\frac{2}{12}]$ $\{s_5, s_8\}[\frac{2}{12}]$ $\{s_5, s_9\}[\frac{2}{12}]$
$\{s_6, s_8\}[\frac{3}{12}]$ $\{s_6, s_9\}[\frac{2}{12}]$
$\{s_8, s_9\}[\frac{5}{12}]$

Wir erhalten als Menge der häufigen 2-Itemmengen

$$L_2 = \{\{s_1, s_6\}, \{s_2, s_5\}, \{s_8, s_9\}\}$$

[1] Das heißt, er zählt die Transaktionen, die beide Elemente der Menge beinhalten, Beispiel: $support(s_1, s_6)$ ist die Anzahl der Zuschauer, die sowohl "Wer will Millionen" als auch "Gutes Bier, schlechtes Bier" schauen, geteilt durch 12.

Aus $\{s_1, s_6\}$ ergeben sich zwei Assoziationsregeln:
Für die Regel $s_1 \rightarrow s_6$ gilt

$$\mathit{confidence}(s_1 \rightarrow s_6) = \frac{\mathit{support}(s_1 \rightarrow s_6)}{\mathit{support}(s_1)} = \frac{\frac{5}{12}}{\frac{5}{12}} = 1$$

und für $s_6 \rightarrow s_1$ erhalten wir

$$\mathit{confidence}(s_6 \rightarrow s_1) = \frac{\mathit{support}(s_6 \rightarrow s_1)}{\mathit{support}(s_6)} = \frac{\frac{5}{12}}{\frac{7}{12}} \approx 0.71$$

Aus $\{s_2, s_5\}$ ergeben sich ebenfalls zwei Assoziationsregeln:
Für die Regel $s_2 \rightarrow s_5$ gilt

$$\mathit{confidence}(s_2 \rightarrow s_5) = \frac{\mathit{support}(s_2 \rightarrow s_5)}{\mathit{support}(s_2)} = \frac{\frac{6}{12}}{\frac{6}{12}} = 1$$

und für die Regel $s_5 \rightarrow s_2$ gilt

$$\mathit{confidence}(s_5 \rightarrow s_2) = \frac{\mathit{support}(s_5 \rightarrow s_2)}{\mathit{support}(s_5)} = \frac{\frac{6}{12}}{\frac{6}{12}} = 1$$

Betrachten wir nun die 2-Itemmenge $\{s_8, s_9\}$. Wir erhalten

$$\mathit{confidence}(s_8 \rightarrow s_9) = \frac{\frac{5}{12}}{\frac{6}{12}} = \frac{5}{6} \approx 0.83$$

und

$$\mathit{confidence}(s_9 \rightarrow s_8) = \frac{\frac{5}{12}}{\frac{5}{12}} = 1$$

Häufige 3-Itemmengen gibt es nicht, denn der paarweise Schnitt der häufigen 2-Itemmengen ist leer und deshalb ist auch C_3 leer (vgl. Aufgabe 5.7).

Als Assoziationsregeln, deren Konfidenz oberhalb der 0.8 - Schwelle liegt, ergeben sich also

$s_1 \rightarrow s_6$ Konfidenz : 1

$s_2 \rightarrow s_5$ Konfidenz : 1

$s_5 \rightarrow s_2$ Konfidenz : 1

$s_8 \rightarrow s_9$ Konfidenz : 0.83

$s_9 \rightarrow s_8$ Konfidenz : 1

2. In der folgenden Tabelle werden die Transaktionen für die zusammengefassten Produktgruppen mit ihrem Support dargestellt:

Kürzel	Art der Send.	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	z_{10}	z_{11}	z_{12}	*support*
Q	Quiz	•	•	•		•	•	•	•	•	•	•	•	$\frac{11}{12}$
S	Serie	•	•	•		•	•	•	•	•	•	•	•	$\frac{11}{12}$
T	Talk		•		•	•		•	•		•			$\frac{6}{12}$

Alle drei 1-Itemmengen sind häufig. Wir berechnen den Support der 2-Itemmengen:

$$\{Q,S\}[\tfrac{11}{12}] \quad \{S,T\}[\tfrac{5}{12}] \quad \{Q,T\}[\tfrac{5}{12}]$$

Alle 2-Itemmengen sind ebenfalls häufig, d.h.

$$L_2 = \{\{Q,S\},\{S,T\},\{Q,T\}\}$$

Daraus ergeben sich sechs Assoziationsregeln:

$$\begin{aligned}
&\mathit{confidence}(Q \to S) = \tfrac{11}{12} \cdot \tfrac{12}{11} = 1\\
&\mathit{confidence}(S \to Q) = \tfrac{11}{12} \cdot \tfrac{12}{11} = 1\\
&\mathit{confidence}(Q \to T) = \tfrac{5}{12} \cdot \tfrac{12}{11} = \tfrac{5}{11} \approx 0.45\\
&\mathit{confidence}(T \to Q) = \tfrac{5}{12} \cdot \tfrac{12}{6} = \tfrac{5}{6} \approx 0.83\\
&\mathit{confidence}(S \to T) = \tfrac{5}{12} \cdot \tfrac{12}{11} = \tfrac{5}{11} \approx 0.45\\
&\mathit{confidence}(T \to S) = \tfrac{5}{12} \cdot \tfrac{12}{6} = \tfrac{5}{6} \approx 0.83
\end{aligned}$$

Die folgenden Assoziationsregeln liegen über der angegebenen Schwelle:

$$\begin{aligned}
&Q \to S \quad \text{Konfidenz} : 1\\
&S \to Q \quad \text{Konfidenz} : 1\\
&T \to Q \quad \text{Konfidenz} : 0.83\\
&T \to S \quad \text{Konfidenz} : 0.83
\end{aligned}$$

Die Menge $\{Q,S,T\}$ besteht den Teilmengencheck, da alle 2-Itemmengen häufig sind. Der Support von $\{Q,S,T\}$ ist $\frac{5}{12}$.

Es werden zunächst drei Assoziationsregeln auf ihre Konfidenz untersucht:

$$\begin{aligned}
&\mathit{confidence}(QS \to T) = \tfrac{5}{12} \cdot \tfrac{12}{11} = \tfrac{5}{11} \approx 0.45\\
&\mathit{confidence}(QT \to S) = \tfrac{5}{12} \cdot \tfrac{12}{5} = 1\\
&\mathit{confidence}(ST \to Q) = \tfrac{5}{12} \cdot \tfrac{12}{5} = 1
\end{aligned}$$

Da die erste Regel nicht über der geforderten Konfidenzschwelle liegt, wird T aus der Menge H_3 der möglichen Konklusionen der 3-Item-Mengen entfernt. Es bleibt nun nur noch die Regel $T \rightarrow QS$ zu untersuchen. Diese hat die Konfidenz $\frac{5}{12} \cdot \frac{12}{6} = 0.83$. Insgesamt haben wir also die folgenden, der Konfidenzgrenze genügenden Assoziationsregeln:

$$T \rightarrow QS \quad \text{Konfidenz} : 0.83$$
$$QT \rightarrow S \quad \text{Konfidenz} : 1$$
$$ST \rightarrow Q \quad \text{Konfidenz} : 1$$

3. Wir erhalten folgende Tabelle mit Support-Angaben:

Kürzel	Sendeanstalt	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	z_{10}	z_{11}	z_{12}	*support*
fun	FUN-TV	•	•	•	•	•		•	•	•	•	•		$\frac{10}{12}$
zarf	ZARF	•		•		•	•		•			•	•	$\frac{7}{12}$
sat	SAT		•		•	•		•	•		•			$\frac{6}{12}$

Alle 1-Itemmengen sind häufig. Für die 2-Itemmengen erhalten wir

$$\{zarf, fun\}[\tfrac{5}{12}] \quad \{fun, sat\}[\tfrac{6}{12}] \quad \{zarf, sat\}[\tfrac{2}{12}]$$

Zwei der 2-Itemmengen sind häufig, nämlich $\{zarf, fun\}$ und $\{fun, sat\}$. Wir erhalten die folgenden Konfidenzen:

$$\begin{aligned}
&\mathit{confidence}(fun \rightarrow zarf) = \tfrac{5}{12} \cdot \tfrac{12}{10} = 0.5 \\
&\mathit{confidence}(zarf \rightarrow fun) = \tfrac{5}{12} \cdot \tfrac{12}{7} \approx 0.71 \\
&\mathit{confidence}(fun \rightarrow sat) = \tfrac{6}{12} \cdot \tfrac{12}{10} = 0.6 \\
&\mathit{confidence}(sat \rightarrow fun) = \tfrac{6}{12} \cdot \tfrac{12}{6} = 1
\end{aligned}$$

Nur eine Assoziationsregel liegt über der angegebenen Schwelle:

$$sat \rightarrow fun \quad \text{Konfidenz} : 1$$

Häufige 3-Itemmengen gibt es nicht; $C_3 = \emptyset$.

4. Betrachten wir alle Assoziationsregeln, deren Konfidenz über der angegebenen Schwelle liegen:

Wer will Millionen? → Gutes Bier, schlechtes Bier	Konfidenz : 1
Quiz mit Dirk Plauer → Gut Josepha	Konfidenz : 1
Gut Josepha → Quiz mit Dirk Plauer	Konfidenz : 1
Olli-G.-Show → Nora am Abend	Konfidenz : 1
Nora am Abend → Olli-G.-Show	Konfidenz : 0.83
Quiz → Serie	Konfidenz : 1
Serie → Quiz	Konfidenz : 1
Talk → Quiz	Konfidenz : 0.83
Talk → Serie	Konfidenz : 0.83
Talk → Quiz und Serie	Konfidenz : 0.83
Quiz und Talk → Serie	Konfidenz : 1
Serie und Talk → Quiz	Konfidenz : 1
sat → *fun*	Konfidenz : 1

   Das ließe sich wie folgt kommentieren:

   "Die *Wer-will-Millionen?* - Zuschauer sind auch Fans von *Gutes-Bier-schlechtes-Bier*. Die Zuschauer von Dirk Plauers Quiz sind identisch mit den *Gut-Josepha*-Fans. Wer *Olli G.* mag, schaut auch *Nora am Abend*, und viele Fans von Nora mögen auch Olli G.s Show.

   Wer Quizsendungen schaut, mag auch Serien und umgekehrt. Wer Talkshows sieht, schaut auch gerne Serien und Quizsendungen. Die SAT-Zuschauer schauen auch FUN-TV, ansonsten bleiben die Zuschauer ihrem Sender offenbar eher treu."

5. Dass die Zuschauer, die sich die vorabendlichen Krimi-Wiederholungen im ZARF anschauen, öfter ein Gebiss tragen als die, die sich nachmittags durch die Talkshows zappen, liegt natürlich nicht an der Wirkung des Sendungsformats auf die Zahngesundheit, sondern an der unterschiedlichen Altersstruktur dieser Zuschauergruppen. Die angegebene Assoziationsregel macht eine strukturelle Aussage, kann jedoch keinen Kausalzusammenhang herstellen. Bei diesem Beispiel ist das unmittelbar einleuchtend, andere Datenauffälligkeiten verführen jedoch leicht dazu, aus strukturellen Gegebenheiten zweifelhafte oder falsche Kausalzusammenhänge herzustellen. Hier ist also Vorsicht geboten (Fernsehen macht dick/dumm/gewalttätig ...) Vergleichen Sie auch mit der Interpretation der Absichten der Unternehmensberatung in Selbsttestaufgabe 5.7: Die Absichten könnten auch ganz anders gelautet haben, angegeben wurde nur eine plausible Erklärung.

zu Selbsttestaufgabe 5.36 (Data Mining in Wörtern)

	ent-spannend	ver-reisen	da-durch	ge-dehnt	er-neut	wie-der	end-lich	Rest	zu-letzt	nicht	
E	•	•		•	•	•	•	•	•		0.8
D	•		•	•		•	•				0.5
R		•	•		•	•		•			0.5
N	•	•		•	•		•			•	0.6
EuR		•			•	•		•			0.4
EuN	•	•		•	•		•				0.5
ER		•			•	•					0.3
EN	•	•					•				0.3

Alle einelementigen Itemmengen bis auf $\{EN\}$ und $\{ER\}$ sind häufig.

Wir bilden die 2-elementigen Itemmengen über den häufigen einelementigen Itemmengen und berechnen ihren Support:

$\boxed{\{E, D\}[0.4]}$	$\boxed{\{E, R\}[0.4]}$	$\boxed{\{E, N\}[0.5]}$	$\boxed{\{E, EuR\}[0.4]}$	$\boxed{\{E, EuN\}[0.5]}$
$\{D, R\}[0.2]$	$\{D, N\}[0.3]$	$\{D, EuR\}[0.1]$	$\{D, EuN\}[0.3]$	$\{R, N\}[0.2]$
$\boxed{\{R, EuR\}[0.4]}$	$\{R, EuN\}[0.2]$	$\{N, EuR\}[0.2]$	$\boxed{\{N, EuN\}\}[0.5]}$	$\{EuR, EuN\}[0.2]$

Als Menge der häufigen 2-Itemmengen erhalten wir somit

$$L_2 = \{\, \{E, D\}, \{E, R\}, \{E, N\}, \{E, EuR\}, \{E, EuN\}, \{R, EuR\}, \{N, EuN\} \,\}$$

Für jede dieser Itemmengen $\{X, Y\}$ sind jetzt die Werte

$$\mathit{confidence}(X \to Y) = \frac{\mathit{support}(X \to Y)}{\mathit{support}(X)}$$

und

$$\mathit{confidence}(Y \to X) = \frac{\mathit{support}(Y \to X)}{\mathit{support}(Y)}$$

zu berechnen.

Für die Menge $\{E, D\}$ ergibt sich

$$\mathit{confidence}(E \to D) = \frac{\mathit{support}(E \to D)}{\mathit{support}(E)} = \frac{0.4}{0.8} = 0.5$$

und

$$\mathit{confidence}(D \to E) = \frac{\mathit{support}(D \to E)}{\mathit{support}(D)} = \frac{0.4}{0.5} = 0.8$$

Auf diese Weise ergeben sich die folgenden Werte:

Regel	Konfidenz	Regel	Konfidenz
$E \to D$	0.5	$D \to E$	0.8
$E \to R$	0.5	$R \to E$	0.8
$E \to N$	0.625	$N \to E$	0.833
$E \to EuR$	0.5	$EuR \to E$	1
$E \to EuN$	0.625	$EuN \to E$	1
$R \to EuR$	0.8	$EuR \to R$	1
$N \to EuN$	0.833	$EuN \to N$	1

Über der Konfidenzschwelle liegen die folgenden Regeln:

Regel	Konfidenz	Regel	Konfidenz
$D \to E$	0.8	$R \to E$	0.8
$N \to E$	0.833	$EuR \to E$	1
$EuN \to E$	1	$R \to EuR$	0.8
$EuR \to R$	1	$N \to EuN$	0.833
$EuN \to N$	1		

Als Ordnung für die Itemmengen zur Bestimmung von C_3 wählen wir Reihenfolge des Auftretens in der Tabelle. Die als nächstes von AprioriGen erzeugten dreielementigen Kandidatenmengen sind dann

$$\begin{aligned} C_3 \quad = \quad & \{\{E,D,R\},\{E,D,N\},\{E,D,EuR\},\{E,D,EuN\},\{E,D,ER\},\{E,R,N\}, \\ & \{E,R,EuR\},\{E,R,EuN\},\{E,R,ER\},\{E,N,EuR\},\{E,N,EuN\}, \\ & \{E,N,ER\},\{E,EuR,EuN\},\{E,EuR,ER\},\{E,EuN,ER\}\}\} \end{aligned}$$

Beachten Sie, dass beispielsweise die Menge $\{N, EuN, EuR\}$ nicht in C_3 enthalten ist.

Als nächstes wird für jedes Element aus C_3 überprüft, ob jede zweielementige Teilmenge ebenfalls häufig sind. Nur die Mengen $\{E, R, EuR\}$ und $\{E, N, EuN\}$ bestehen den Teilmengencheck, d. h.,

$$L_3 = \{\{E,R,EuR\},\{E,N,EuN\}\}$$

.

Für die Menge $\{E, R, EuR\}$ (support: 0.4) ergibt sich

$$confidence(E \to R, EuR) = \frac{support(E \to R, EuR)}{support(E)} = \frac{0.4}{0.8} = 0.5$$

$$confidence(R \to E, EuR) = \frac{support(R \to E, EuR)}{support(R)} = 0.4/0.5 = 0.8$$

$$confidence(EuR \to E, R) = \frac{support(EuR \to E, R)}{support(EuR)} = 0.4/0.4 = 1$$

$$confidence(E, R \rightarrow EuR) = \frac{support(E, R \rightarrow EuR)}{support(E, R)} = 0.4/0.4 = 1$$

$$confidence(E, EuR \rightarrow R) = \frac{support(E, EuR \rightarrow R)}{support(E, EuR)} = 0.4/0.4 = 1$$

$$confidence(R, EuR \rightarrow E) = \frac{support(R, EuR \rightarrow E)}{support(R, EuR)} = 0.4/0.4 = 1$$

Für die Menge $\{E, N, EuN\}$ (support: 0.5) ergibt sich

$$confidence(E \rightarrow N, EuN) = \frac{support(E \rightarrow N, EuN)}{support(E)} = \frac{0.5}{0.8} = 0.625$$

$$confidence(N \rightarrow E, EuN) = \frac{support(N \rightarrow E, EuN)}{support(N)} = 0.5/0.6 = 0.833$$

$$confidence(EuN \rightarrow N, E) = \frac{support(EuN \rightarrow N, E)}{support(EuN)} = 0.5/0.5 = 1$$

$$confidence(EuN, N \rightarrow E) = \frac{support(EuN, N \rightarrow E)}{support(EuN, N)} = 0.5/0.5 = 1$$

$$confidence(E, N \rightarrow EuN) = \frac{support(E, N \rightarrow EuN)}{support(E, N)} = 0.5/0.5 = 1$$

$$confidence(EuN, E \rightarrow N) = \frac{support(EuN, E \rightarrow N)}{support(EuN, E)} = 0.5/0.5 = 0.1$$

Als Assoziationsregeln über der geforderten Konfidenzschwelle ergeben sich also

Regel	Konfidenz
$N \rightarrow E, EuN$	0.833
$EuN \rightarrow N, E$	1
$EuN, N \rightarrow E$	1
$E, N \rightarrow EuN$	1
$EuN, E \rightarrow N$	1
$R \rightarrow E, EuR$	0.8
$EuR \rightarrow E, R$	1
$E, R \rightarrow EuR$	1
$E, EuR \rightarrow R$	1
$R, EuR \rightarrow E$	1

Die Menge C_4 ist leer: Keine vierelementige Teilmenge der Itemmenge ist häufig. Somit sind alle Assoziationsregeln berechnet.

zu Selbsttestaufgabe 5.37 (Data Mining - Restaurant)

1. Die folgende Tabelle enthält die Datenbasis aus der Aufgabenstellung mit zusätzlicher Angabe des Supports der jeweiligen 1-Itemmengen.

Label	*Speise*	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	
A	marinierte Sardinen	•			•			•			•	0.4
B	geräucherter Schwertfisch					•				•		0.2
C	schwarze Kräuter-Oliven			•						•	•	0.3
D	mit Mandeln gefüllte Oliven	•				•						0.2
E	Manzanilla Oliven				•		•				•	0.3
F	gratinierte Miesmuscheln	•	•		•			•		•		0.5
G	Parmaschinken mit Melone					•			•			0.2
H	Vitello tonnato	•				•						0.2
I	Carpaccio	•				•						0.2
J	gefüllte Peperoni	•		•	•					•	•	0.5

Somit liegt also der Support der Artikel $\{A\}$, $\{F\}$ und $\{J\}$ oberhalb der gewählten Schwelle. Die Kombinationen von $\{A\}$, $\{F\}$ und $\{J\}$ sind die Mengen $\{A, F\}$, $\{A, J\}$ und $\{J, F\}$, deren Support allerdings jeweils 0.3 beträgt. Somit gibt es in der gegebenen Datenbasis nur die häufigen 1-Itemmengen $\{A\}$, $\{F\}$ und $\{J\}$ und keine (nicht-trivialen) Assoziationsregeln.

2. In der folgenden Tabelle enthält jede Bestellung bzw. Transaktion einen Eintrag für die Produktgruppen, aus denen (mindestens) eine Speise ausgewählt wurde.

Label	*Speise*	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	
K	Fisch	•			•	•		•		•	•	0.6
L	Oliven	•		•	•	•	•			•	•	0.7
M	gratinierte Miesmuscheln	•	•		•			•		•		0.5
N	Fleisch	•				•			•			0.3
O	gefüllte Peperoni	•		•	•					•	•	0.5

Somit sind also $\{K\}$, $\{L\}$, $\{M\}$ und $\{O\}$ häufige 1-Itemmengen. Mittels *AprioriGen* werden jeweils zuerst schrittweise Kandidaten für häufige k-Itemmengen aus den häufigen $k-1$-Itemmengen generiert:

(a) Je zwei häufige $(k-1)$-Itemmengen mit denselben ersten $(k-2)$ Elementen werden zu einer k-elementigen Menge vereinigt (im Fall $k = 2$ werden alle 1-Itemmengen paarweise vereinigt),

(b) aus der Ergebnismenge des ersten Schritts werden diejenigen Mengen entfernt, deren $k-1$-elementige Teilmengen nicht alle häufig sind (Teilmengencheck, im Fall $k = 2$ ist er nicht nötig).

Anschließend wird jeweils der Support dieser Kandidaten überprüft und die Kandidaten ggf. in die Ausgabe aufgenommen.

Häufige 2-Itemmengen (Eingabe ist $\{K\}$, $\{L\}$, $\{M\}$ und $\{O\}$):
Der Algorithmus *AprioriGen* generiert aus den häufigen 1-Itemmengen die Kandidatenmengen $\{K, L\}$ [0.5], $\{K, O\}$ [0.4], $\{L, O\}$ [0.5], $\{K, M\}$ [0.4], $\{M, O\}$ [0.3] und $\{L, M\}$ [0.3]. Mit dem jeweils in Klammern angegebenen Support sind $\{K, L\}$, $\{K, O\}$, $\{L, O\}$ und $\{K, M\}$ häufige 2-Itemmengen und werden deshalb vom *Apriori*-Algorithmus übernommen.

Häufige 3-Itemmengen (Eingabe ist $\{K, L\}$, $\{K, O\}$, $\{L, O\}$, $\{K, M\}$):
Im ersten Schritt von *AprioriGen* werden $\{K, L, O\}$, $\{K, L, M\}$ und $\{K, M, O\}$ erzeugt, im zweiten Schritt wird ein Teilmengencheck durchgeführt: alle 2-elementigen Teilmengen von $\{K, L, O\}$ sind häufig, die Menge $\{K, L, M\}$ enthält die nicht-häufige 2-elementige Teilmenge $\{L, M\}$ und die Menge $\{K, M, O\}$ enthält die nicht-häufige 2-elementige Teilmenge $\{M, O\}$. Es bleibt also lediglich die Menge $\{K, L, O\}$ als Kandidat übrig. Diese hat einen Support von 0.4 und wird deshalb von *Apriori* als häufige 3-Itemmenge übernommen.

Häufige 4-Itemmengen (Eingabe ist $\{K, L, O\}$):
Häufige 4-Itemmengen können nicht gebildet werden, da es nur eine häufige 3-Itemmenge gibt.

Die Ausgabe des *Apriori*-Algorithmus ist daher:

$$\{\{K\}, \{L\}, \{M\}, \{O\}, \{K, M\}, \{K, L\}, \{K, O\}, \{L, O\}, \{K, L, O\}\}$$

Nun werden zu jeder häufigen Itemmenge X die gesuchten Assoziationsregeln erzeugt.

Aus den häufigen 2-Itemmengen werden die gesuchten Assoziationsregeln direkt bestimmt. Folgende Regeln werden untersucht, jeweils mit Konfidenz:

Regel	*Konfidenz*	*Regel*	*Konfidenz*
K $\rightarrow$ M	$\frac{0.4}{0.6} < 0.82$	M $\rightarrow$ K	$\frac{0.4}{0.5} < 0.82$
K $\rightarrow$ L	$\frac{0.5}{0.6} > 0.82$	L $\rightarrow$ K	$\frac{0.5}{0.7} < 0.82$
K $\rightarrow$ O	$\frac{0.4}{0.6} < 0.82$	O $\rightarrow$ K	$\frac{0.4}{0.5} < 0.82$
L $\rightarrow$ O	$\frac{0.5}{0.7} < 0.82$	O $\rightarrow$ L	$\frac{0.5}{0.5} > 0.82$

Davon erfüllen also die Regeln $K \rightarrow L$ und $O \rightarrow L$ die Konfidenzbedingung und werden in die Ausgabeliste aufgenommen.

Wir betrachten nun die 3-Itemmenge $X = \{K, L, O\}$. Es bezeichne H_m die Menge der m-Item-Konklusionen dieser Menge. Es lassen sich aus X mit $H_1 = \{\{K\}, \{L\}, \{O\}\}$ die folgenden drei Regeln (jeweils mit Konfidenz in Klammern) generieren:

$$KL \to O \left[\frac{0.4}{0.5}\right], \qquad KO \to L \left[\frac{0.4}{0.4}\right], \qquad OL \to K \left[\frac{0.4}{0.5}\right]$$

Die Regel $KO \to L$ erfüllt die Konfidenzbedingung.

Es wurden also zunächst die Assoziationsregeln mit möglichst kurzer Konklusion (1-elementig) erzeugt. Der Algorithmus untersucht dann schrittweise, ob es auch passende Regeln mit erweiterten Konklusionen gibt:

- Setze $H_{m+1} := \mathit{AprioriGen}(H_m)$ (bei $m = 0$ ist in diesem Schritt nichts zu tun, H_1 sind zunächst alle einelementigen Teilmengen).
- Für alle Konklusionen $h_{m+1} \in H_{m+1}$ überprüft man die Konfidenz der Regel $(X - h_{m+1}) \to h_{m+1}$; liegt sie über der Schwelle *minconf*, so wird die Regel ausgegeben, andernfalls wird h_{m+1} aus H_{m+1} entfernt.

$\underline{m = 0:}$
Da nur die Regel $KO \to L$ die Konfidenzbedingung erfüllt, werden $\{O\}$ und $\{K\}$ aus H_1 entfernt: $H_1 := H_1 \setminus \{\{O\}, \{K\}\} = \{L\}$.

$\underline{m = 1:}$
Aus $H_2 = \mathit{AprioriGen}(H_1) = \emptyset$ lässt sich keine weitere Regel mit einer 2-elementigen Konklusion konstruieren.

Insgesamt erhalten wir folgende Assoziationsregeln:

Regel		*Support*	*Konfidenz*
Fisch	→ Oliven	0.4	0.833
gefüllte Peperoni	→ Oliven	0.5	1
Fisch, gefüllte Peperoni	→ Oliven	0.5	1

zu Selbsttestaufgabe 5.38 (Data Mining - Makler)

1. Berechnung des Supports der aufgelisteten Merkmale:

Label	Merkmal	t_1	t_2	t_3	t_4	t_5	Support
A	Balkon	•	•	•	•		$\frac{4}{5} = 0.8$
B	Loft			•			$\frac{1}{5} = 0.2$
C	Altbau	•	•		•	•	$\frac{4}{5} = 0.8$
D	Einbauküche		•		•	•	$\frac{3}{5} = 0.6$
E	Innenstadtnähe	•				•	$\frac{2}{5} = 0.4$
F	Garage / Stellplatz	•		•	•	•	$\frac{4}{5} = 0.8$
G	Keller	•			•		$\frac{2}{5} = 0.4$

2. Bestimmung der häufigen Itemmengen zu $minsupp = 0.5$ mit Hilfe des Apriori-Algorithmus:

   Die Menge der häufigen 1-Itemmengen enthält alle 1-Itemmengen, mit einem Support von mindestens 0.5:

   $$L_1 = \{\{A\}, \{C\}, \{D\}, \{F\}\}$$

   Bilde aus den Mengen von L_1 die Kandidatenmenge C_2:

   $$C_2 = \{\{A, C\}, \{A, D\}, \{A, F\}, \{C, D\}, \{C, F\}, \{D, F\}\}$$

   Bei C_2 entfällt der Teilmengencheck.
   Berechne den Support der 2-Itemmengen in C_2:

$C_2 =$	$\{\{A, C\},$	$\{A, D\},$	$\{A, F\},$	$\{C, D\},$	$\{C, F\},$	$\{D, F\}\}$
Support:	$\frac{3}{5} = 0.6$	$\frac{2}{5} = 0.4$	$\frac{3}{5} = 0.6$	$\frac{3}{5} = 0.6$	$\frac{3}{5} = 0.6$	$\frac{2}{5} = 0.4$

   Die Menge der häufigen 2-Itemmengen enthält alle Mengen aus C_2, mit einem Support von mindestens 0.5:

   $$L_2 = \{\{A, C\}, \{A, F\}, \{C, D\}, \{C, F\}\}$$

   Bilde aus den Mengen von L_2 die Kandidatenmenge C_3 indem je zwei Mengen, deren erstes Element jeweils gleich ist, zu einer Menge kombiniert werden:

   $$C_3 = \{\{A, C, F\}, \{C, D, F\}\}$$

   Teilmengencheck:
   $\{C, D, F\}$ wird aus C_3 entfernt, da die Teilmenge $\{D, F\}$ nicht häufig ist. Hingegen sind alle Teilmengen von $\{A, C, F\}$ häufig, so dass diese Menge erhalten bleibt. Berechne den Support der 3-Itemmengen in C_3':

$C_3' =$	$\{\{A, C, F\}\}$
Support:	$\frac{2}{5} = 0.4$

   Die Menge der häufigen 3-Itemmengen enthält alle Mengen aus C_3, mit einem Support von mindestens 0.5:

   $$L_3 = \emptyset$$

   Damit ergibt sich als Menge aller häufigen Itemmengen:

   $$L = L_1 \cup L_2$$

3. Ermittlung aller Assoziationsregeln mit 2 Items und $minconf = 0.6$:

   Aus allen 2-häufigen Itemmengen, werden Assoziationsregeln der Form $(X - Y) \rightarrow Y$ gebildet. Dabei gibt X eine 2-Itemmenge und Y eine 1-Itemmenge an, die Teilmenge von X ist. Alle möglichen Kombinationen müssen gebildet werden. Danach berechnet man $confidence((X-Y) \rightarrow Y)$ und übernimmt alle Assoziationsregeln mit einem confidence-Wert $\geq minconf$.

   Es kommen folgende Regeln zustande:

$\{A, C\}$	*confidence*	$\{A, F\}$	*confidence*	$\{C, D\}$	*confidence*	$\{C, F\}$	*confidence*
$A \rightarrow C$	$\frac{0.6}{0.8} = \frac{3}{4}$	$A \rightarrow F$	$\frac{3}{4}$	$C \rightarrow D$	$\frac{3}{4}$	$C \rightarrow F$	$\frac{3}{4}$
$C \rightarrow A$	$\frac{3}{4}$	$F \rightarrow A$	$\frac{3}{4}$	$D \rightarrow C$	$\frac{0.6}{0.6} = 1$	$F \rightarrow C$	$\frac{3}{4}$

   Die *confidence*-Werte wurden dabei nach folgender Regel errechnet (allgemein für k-Itemmengen):

$$confidence((X - Y) \rightarrow Y) = \frac{sup(X)}{sup(X-Y)}$$

   Die vereinfachte Formel für 2-Itemmengen lautet:

$$confidence(X \rightarrow Y) = \frac{sup(\{X,Y\})}{sup(X)}$$

   Da $minconf = 0.6$ ist, und jede der oben aufgestellten Regeln einen confidence-Wert von $\frac{6}{8} = \frac{3}{4}$ oder größer hat, wird jede der Regeln vom Algorithmus als gültig gewertet.