Lösung zur Aufgabe 23

Patient	Fieber (X_1)	Husten (X_2)	Kurzatmigkeit (X_3)	infiziert (T)
1	Nein	Nein	Nein	Nein
2	Ja	Ja	Ja	Ja
3	Ja	Ja	Nein	Nein
4	Ja	Nein	Ja	Ja
5	Ja	Ja	Ja	Ja
6	Nein	Ja	Nein	Nein
7	Ja	Nein	Ja	Ja
8	Ja	Nein	Ja	Ja
9	Nein	Ja	Ja	Ja
10	Ja	Ja	Nein	Ja
11	Nein	Ja	Nein	Nein
12	Nein	Ja	Ja	Nein
13	Nein	Ja	Ja	Nein
14	Ja	Ja	Nein	Nein

a) Es gibt drei binäre Inputvariablen x_1, x_2, x_3 und die qualitative binäre Outputvariable T besitzt die beiden Ausprägungen $\mathscr{C}_0$ (Patient ist nicht infiziert) und $\mathscr{C}_1$ (Patient ist infiziert). Der Inputraum $\mathscr{D}$ besteht aus insgesamt 14 Beobachtungen (Daten von Patienten). Der Trainingsdatensatz ist somit gegeben durch

$$\mathscr{T} = \left\{ (\mathbf{x}_n, t_n) \in \mathbb{R}^3 \times \{\mathscr{C}_0, \mathscr{C}_1\} : n = 1, \ldots, 14 \right\}.$$

Die Entropie einer Teilmenge $R \subseteq \mathscr{D}$ ist

$$\varepsilon_3(R) = -\sum_{k=0}^{1} r_k(R) \log_2 r_k(R) \in \left[0, \frac{1}{2}\right],$$

$\log_2(K)$ $\log_2(2) = 1$

wobei

$$r_k(R) = \frac{\sum_{(\mathbf{x}_n, t_n) \in \mathscr{T}} 1_{\{\mathscr{C}_k\}}(t_n) \times 1_R(\mathbf{x}_n)}{\sum_{(\mathbf{x}_n, t_n) \in \mathscr{T}} 1_R(\mathbf{x}_n)} \quad \text{für } \mathscr{C}_k \in \{\mathscr{C}_0, \mathscr{C}_1\}$$

die relative Häufigkeiten der Klasse $\mathscr{C}_k$ ist. Umso näher der Wert von $\varepsilon_3(R)$ bei 0 liegt, desto homogener (reiner) ist die Teilmenge R.

1. Ebene

Zerlegung des Inputraums $R = \mathscr{D}$ mittels der Inputvariablen x_1 liefert die beiden Teilmengen

$$R_0(1) = \{\mathbf{x}_1, \mathbf{x}_6, \mathbf{x}_9, \mathbf{x}_{11}, \mathbf{x}_{12}, \mathbf{x}_{13}\} \quad \text{und} \quad R_1(1) = \{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_7, \mathbf{x}_8, \mathbf{x}_{10}, \mathbf{x}_{14}\}.$$

Mit den relativen Häufigkeiten

$$r_0(R_0(1)) = \frac{5}{6},\ r_1(R_0(1)) = \frac{1}{6},\ r_0(R_1(1)) = \frac{2}{8} \quad \text{und} \quad r_1(R_1(1)) = \frac{6}{8}$$

erhält man für die Entropie der beiden Teilmengen:

$-5/6 \log_2(5/6) - 1/6 \log_2(1/6) = 0{,}6500$

$$\varepsilon_3(R_0(1)) = -\sum_{k=0}^{1} r_k(R_0(1)) \log_2 r_k(R_0(1)) = 0{,}6500224$$

$$\varepsilon_3(R_1(1)) = -\sum_{k=0}^{1} r_k(R_1(1)) \log_2 r_k(R_1(1)) = 0{,}8112781$$

Der Wert der Fehlerfunktion für die Inputvariable x_1 beträgt somit

$$\frac{|R_0(1)|}{|R|} \cdot \varepsilon_3(R_0(1)) + \frac{|R_1(1)|}{|R|} \cdot \varepsilon_3(R_1(1)) = \frac{6}{14} \cdot 0{,}6500224 + \frac{8}{14} \cdot 0{,}8112781$$
$$= 0{,}7421685.$$

Zerlegung des Inputraums $R = \mathscr{D}$ mittels der Inputvariablen x_2 liefert die beiden Teilmengen

$$R_0(2) = \{\mathbf{x}_1, \mathbf{x}_4, \mathbf{x}_7, \mathbf{x}_8\} \quad \text{und} \quad R_1(2) = \{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_9, \mathbf{x}_{10}, \mathbf{x}_{11}, \mathbf{x}_{12}, \mathbf{x}_{13}, \mathbf{x}_{14}\}.$$

Mit den relativen Häufigkeiten

$$r_0(R_0(2)) = \frac{1}{4},\ r_1(R_0(2)) = \frac{3}{4},\ r_0(R_1(2)) = \frac{6}{10} \quad \text{und} \quad r_1(R_1(2)) = \frac{4}{10}$$

erhält man für die Entropie der beiden Teilmengen:

$$\varepsilon_3(R_0(2)) = -\sum_{k=0}^{1} r_k(R_0(2)) \log_2 r_k(R_0(2)) = 0{,}8112781$$

$$\varepsilon_3(R_1(2)) = -\sum_{k=0}^{1} r_k(R_1(2)) \log_2 r_k(R_1(2)) = 0{,}9709506$$

Der Wert der Fehlerfunktion für die Inputvariable x_2 beträgt somit

$$\frac{|R_0(2)|}{|R|} \cdot \varepsilon_3(R_0(2)) + \frac{|R_1(2)|}{|R|} \cdot \varepsilon_3(R_1(2)) = \frac{4}{14} \cdot 0{,}8112781 + \frac{10}{14} \cdot 0{,}9709506$$
$$= 0{,}9253299.$$

Zerlegung des Inputraums $R = \mathscr{D}$ mittels der Inputvariablen x_3 liefert die beiden Teilmengen

$$R_0(3) = \{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_6, \mathbf{x}_{10}, \mathbf{x}_{11}, \mathbf{x}_{14}\} \quad \text{und} \quad R_1(3) = \{\mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_7, \mathbf{x}_8, \mathbf{x}_9, \mathbf{x}_{12}, \mathbf{x}_{13}\}.$$

Mit den relativen Häufigkeiten

$$r_0(R_0(3)) = \frac{5}{6},\ r_1(R_0(3)) = \frac{1}{6},\ r_0(R_1(3)) = \frac{2}{8} \quad \text{und} \quad r_1(R_1(3)) = \frac{6}{8}$$

erhält man für die Entropie der beiden Teilmengen:

$$\varepsilon_3(R_0(3)) = -\sum_{k=0}^{1} r_k(R_0(3)) \log_2 r_k(R_0(3)) = 0{,}6500224$$

$$\varepsilon_3(R_1(3)) = -\sum_{k=0}^{1} r_k(R_1(3)) \log_2 r_k(R_1(3)) = 0{,}8112781$$

Der Wert der Fehlerfunktion für die Inputvariable x_3 beträgt somit

$$\frac{|R_0(3)|}{|R|} \cdot \varepsilon_3(R_0(3)) + \frac{|R_1(3)|}{|R|} \cdot \varepsilon_3(R_1(3)) = \frac{6}{14} \cdot 0{,}6500224 + \frac{8}{14} \cdot 0{,}8112781$$
$$= 0{,}7421685.$$

Der Wert der Fehlerfunktion wird also bei einer Zerlegung bzgl. der Inputvariablen x_1 und x_3 minimiert, so dass für die Verzweigung x_1 oder x_3 gewählt werden kann. Im Folgenden wird die Inputvariable x_1 gewählt, d.h. es werden nun die „linke“ Teilmenge

$$R_0(1) = \{\mathbf{x}_1, \mathbf{x}_6, \mathbf{x}_9, \mathbf{x}_{11}, \mathbf{x}_{12}, \mathbf{x}_{13}\}$$

und anschließend die „rechte“ Teilmenge

$$R_1(1) = \{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_7, \mathbf{x}_8, \mathbf{x}_{10}, \mathbf{x}_{14}\}.$$

betrachtet.

2. Ebene

Um die Notation möglichst einfach zu halten, wird im Folgenden $R = R_0(1)$ gesetzt. Eine Zerlegung dieser Teilmenge mittels x_2 liefert die Teilmengen

$$R_0(2) = \{\mathbf{x}_1\} \quad \text{und} \quad R_1(2) = \{\mathbf{x}_6, \mathbf{x}_9, \mathbf{x}_{11}, \mathbf{x}_{12}, \mathbf{x}_{13}\}.$$

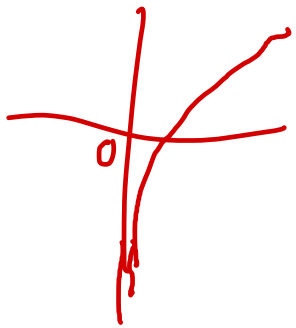
Mit den relativen Häufigkeiten

$$r_0(R_0(2)) = \frac{1}{1},\ r_1(R_0(2)) = \frac{0}{1},\ r_0(R_1(2)) = \frac{4}{5} \quad \text{und} \quad r_1(R_1(2)) = \frac{1}{5}$$

erhält man für die Entropie der beiden Teilmengen:

$$\varepsilon_3(R_0(2)) = -\sum_{k=0}^{1} r_k(R_0(2)) \log_2 r_k(R_0(2)) = 0$$

$$\varepsilon_3(R_1(2)) = -\sum_{k=0}^{1} r_k(R_1(2)) \log_2 r_k(R_1(2)) = 0{,}7219281$$



Der Wert der Fehlerfunktion für die Inputvariable x_2 beträgt somit

$$\frac{|R_0(2)|}{|R|} \cdot \varepsilon_3(R_0(2)) + \frac{|R_1(2)|}{|R|} \cdot \varepsilon_3(R_1(2)) = \frac{1}{6} \cdot 0 + \frac{5}{6} \cdot 0{,}7219281$$
$$= 0{,}6016067.$$

Zerlegung der „linken“ Teilmenge $R = R_0(1)$ mittels der Inputvariablen x_3 liefert die beiden Teilmengen

$$R_0(3) = \{\mathbf{x}_1, \mathbf{x}_6, \mathbf{x}_{11}\} \quad \text{und} \quad R_1(3) = \{\mathbf{x}_9, \mathbf{x}_{12}, \mathbf{x}_{13}\}.$$

Mit den relativen Häufigkeiten

$$r_0(R_0(3)) = \frac{3}{3},\ r_1(R_0(3)) = \frac{0}{3},\ r_0(R_1(3)) = \frac{2}{3} \quad \text{und} \quad r_1(R_1(3)) = \frac{1}{3}$$

erhält man für die Entropie der beiden Teilmengen:

$$\varepsilon_3(R_0(3)) = -\sum_{k=0}^{1} r_k(R_0(3)) \log_2 r_k(R_0(3)) = 0$$

$$\varepsilon_3(R_1(3)) = -\sum_{k=0}^{1} r_k(R_1(3)) \log_2 r_k(R_1(3)) = 0{,}9182958$$

Der Wert der Fehlerfunktion für die Inputvariable x_3 beträgt somit

$$\begin{aligned}\frac{|R_0(3)|}{|R|} \cdot \varepsilon_3(R_0(3)) + \frac{|R_1(3)|}{|R|} \cdot \varepsilon_3(R_1(3)) &= \frac{3}{6} \cdot 0 + \frac{3}{6} \cdot 0{,}7219281 \\ &= 0{,}4591479.\end{aligned}$$

Der Wert der Fehlerfunktion wird also bei einer Zerlegung bzgl. der Inputvariablen x_3 minimiert. Für die Verzweigung wird daher x_3 gewählt und die Teilmenge $R_0(3)$ wird so zu einem Blatt (Endknoten) mit dem Prognosewert $\mathscr{C}_0$ (Patient ist nicht infiziert).

Es wird nun die „rechte“ Teilmenge $R = R_1(1)$ mittels der Inputvariablen x_2 zerlegt. Um die Notation möglichst einfach zu halten, wird im Folgenden wieder $R = R_1(1)$ gesetzt. Eine Zerlegung dieser Teilmenge liefert die beiden Teilmengen

$$R_0(2) = \{\mathbf{x}_4, \mathbf{x}_7, \mathbf{x}_8\} \quad \text{und} \quad R_1(2) = \{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_5, \mathbf{x}_{10}, \mathbf{x}_{14}\}.$$

Mit den relativen Häufigkeiten

$$r_0(R_0(2)) = \frac{0}{3},\ r_1(R_0(2)) = \frac{3}{3},\ r_0(R_1(2)) = \frac{2}{5} \quad \text{und} \quad r_1(R_1(2)) = \frac{3}{5}$$

erhält man für die Entropie der beiden Teilmengen:

$$\varepsilon_3(R_0(2)) = -\sum_{k=0}^{1} r_k(R_0(2)) \log_2 r_k(R_0(2)) = 0$$

$$\varepsilon_3(R_1(2)) = -\sum_{k=0}^{1} r_k(R_1(2)) \log_2 r_k(R_1(2)) = 0{,}8352666$$

Der Wert der Fehlerfunktion für die Inputvariable x_2 beträgt somit

$$\frac{|R_0(2)|}{|R|} \cdot \varepsilon_3(R_0(2)) + \frac{|R_1(2)|}{|R|} \cdot \varepsilon_3(R_1(2)) = \frac{3}{8} \cdot 0 + \frac{5}{8} \cdot 0{,}8352666$$
$$= 0{,}5220416.$$

Zerlegung der „rechten“ Teilmenge $R = R_1(1)$ mittels der Inputvariablen x_3 liefert die beiden Teilmengen

$$R_0(3) = \{\mathbf{x}_3, \mathbf{x}_{10}, \mathbf{x}_{14}\} \quad \text{und} \quad R_1(3) = \{\mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_7, \mathbf{x}_8\}.$$

Mit den relativen Häufigkeiten

$$r_0(R_0(3)) = \frac{2}{3},\ r_1(R_0(3)) = \frac{1}{3},\ r_0(R_1(3)) = \frac{0}{5} \quad \text{und} \quad r_1(R_1(3)) = \frac{5}{5}$$

erhält man für die Entropie der beiden Teilmengen:

$$\varepsilon_3(R_0(3)) = -\sum_{k=0}^{1} r_k(R_0(3)) \log_2 r_k(R_0(3)) = 0{,}9182958$$

$$\varepsilon_3(R_1(3)) = -\sum_{k=0}^{1} r_k(R_1(3)) \log_2 r_k(R_1(3)) = 0$$

Der Wert der Fehlerfunktion für die Inputvariable x_3 beträgt somit

$$\frac{|R_0(3)|}{|R|} \cdot \varepsilon_3(R_0(3)) + \frac{|R_1(3)|}{|R|} \cdot \varepsilon_3(R_1(3)) = \frac{3}{8} \cdot 0{,}9182958 + \frac{5}{8} \cdot 0$$
$$= 0{,}3443609.$$

Der Wert der Fehlerfunktion wird also bei einer Zerlegung bzgl. der Inputvariablen x_3 minimiert. Für die zweite Verzweigung wird daher x_3 gewählt und die Teilmenge $R_1(3)$ wird so zu einem Blatt (Endknoten) mit dem Prognosewert $\mathscr{C}_1$ (Patient ist infiziert).

3. Ebene

Die in der 2. Ebene resultierende „rechte“ Teilmenge

$$R_1(3) = \{\mathbf{x}_9, \mathbf{x}_{12}, \mathbf{x}_{13}\}$$

kann mittels der einzigen noch verbleibenden Inputvariablen x_2 in die beiden Teilmengen

$$R_0(2) = \{\} \quad \text{und} \quad R_1(2) = \{\mathbf{x}_9, \mathbf{x}_{12}, \mathbf{x}_{13}\}$$

zerlegt werden. Mittels majority vote wird der „linken“ Teilmenge $R_0(2)$ der Prognosewert $\mathscr{C}_0 \vee \mathscr{C}_1$ (Patient ist nicht infiziert oder infiziert) und der „rechten“ Teilmenge $R_1(2)$ der Prognosewert $\mathscr{C}_0$ (Patient ist nicht infiziert) zugeordnet.

Analog kann die in der 2. Ebene resultierende „rechte“ Teilmenge

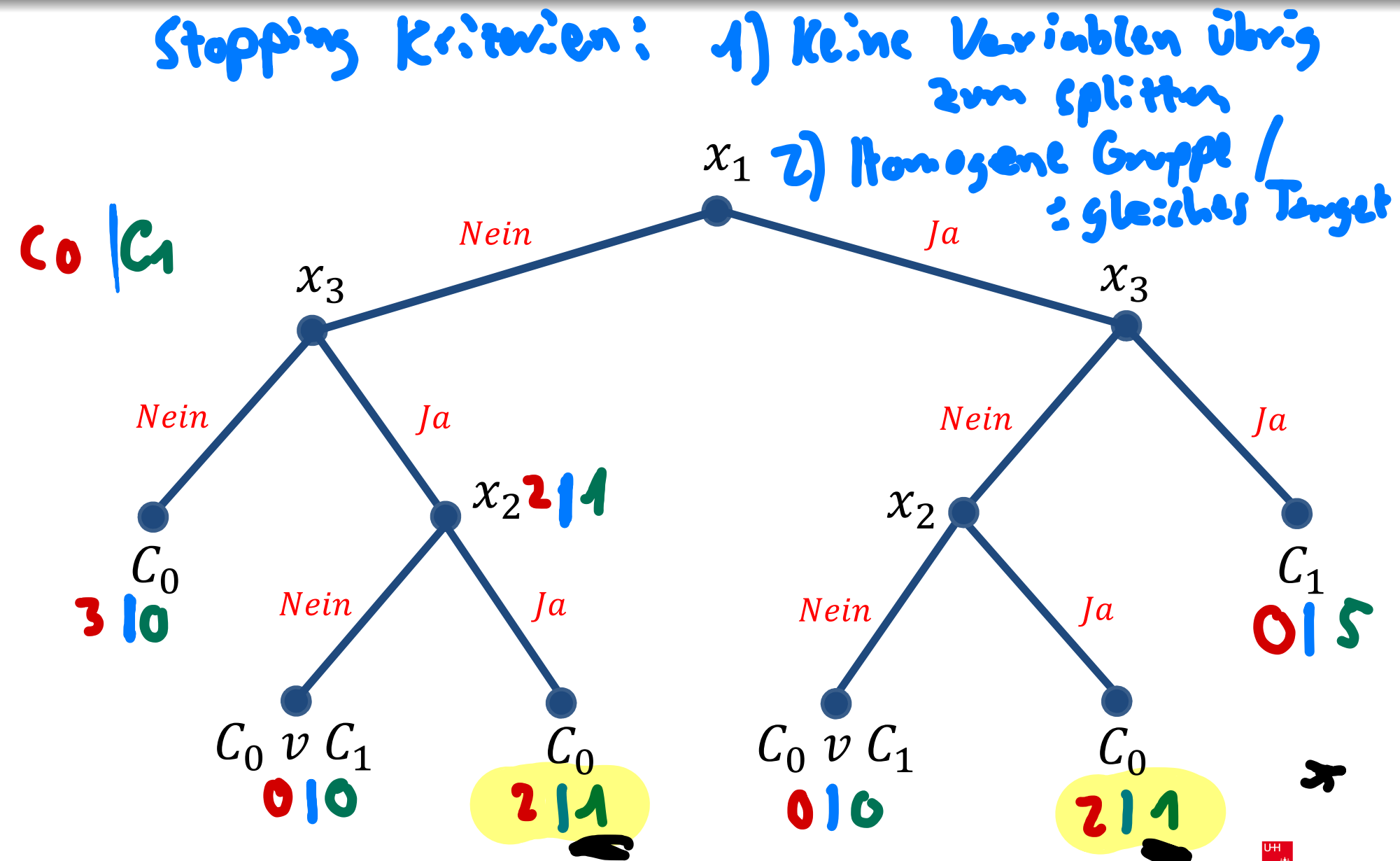
$$R_0(3) = \{\mathbf{x}_3, \mathbf{x}_{10}, \mathbf{x}_{14}\}$$

mittels der einzigen noch verbleibenden Inputvariablen x_2 in die beiden Teilmengen

$$R_0(2) = \{\} \quad \text{und} \quad R_1(2) = \{\mathbf{x}_3, \mathbf{x}_{10}, \mathbf{x}_{14}\}$$

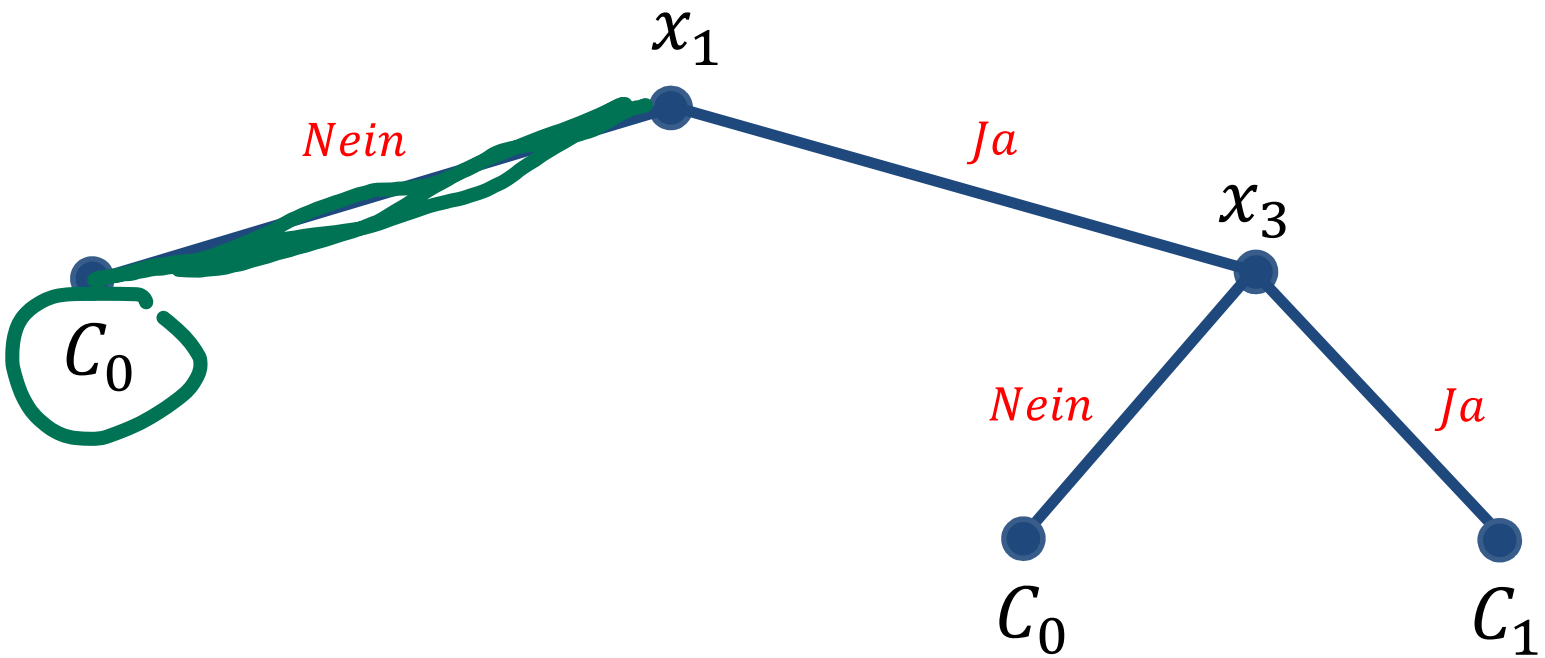
zerlegt werden. Mittels majority vote wird der „linken“ Teilmenge $R_0(2)$ der Prognosewert $\mathscr{C}_0 \vee \mathscr{C}_1$ (Patient ist nicht infiziert oder infiziert) und der „rechten“ Teilmenge $R_1(2)$ der Prognosewert $\mathscr{C}_0$ (Patient ist nicht infiziert) zugeordnet.

Man erhält somit den untenstehenden Klassifikationsbaum:


Stopping Kriterien: 1) Keine Variablen übrig zum splitten
2) Homogene Gruppe / gleiches Target
C0 | C1
x_1
Nein
Ja
x_3
x_3
Nein
Ja
Nein
Ja
x_2 2 | 1
x_2
C_0
3 | 0
Nein
Ja
Nein
Ja
C_1
0 | 5
C_0 v C_1
0 | 0
C_0
2 | 1
C_0 v C_1
0 | 0
C_0
2 | 1

b) Der in Aufgabenteil a) resultierende Klassifikationsbaum weist eine Fehlklassifikationsrate von $\frac{2}{14}$ auf. D.h. bei den 14 Patienten kommt es zu insgesamt 2 Fehlklassifikationen (Patienten 9 und 10).

c) Das Symptom „Husten (x_2)“ liefert offensichtlich keine Informationen für eine richtige Klassifikation der Patienten. D.h. wenn die beiden Verzweigungen mittels der Inputvariablen x_2 jeweils durch einen Blatt (Endknoten) mit dem Progosewert $\mathscr{C}_0$ ersetzt werden, verschlechtert sich die Fehlklassifikationsrate dadurch nicht. Da dann auf der linken Seite bei der Verzweigung mittels der Inputvariablen x_3 in beiden Blättern (Endknoten) der Progosewert $\mathscr{C}_0$ resultiert, kann auch die linke Verzweigung mittels der Inputvariablen x_3 durch einen Blatt (Endknoten) mit dem Progosewert $\mathscr{C}_0$ ersetzt werden. Ein weiteres Beschneiden des Baumes würde jedoch zu einer Erhöhung der Fehlklassifikationsrate führen. Nach dieser Beschneidung erhält man den untenstehenden Klassifikationsbaum.



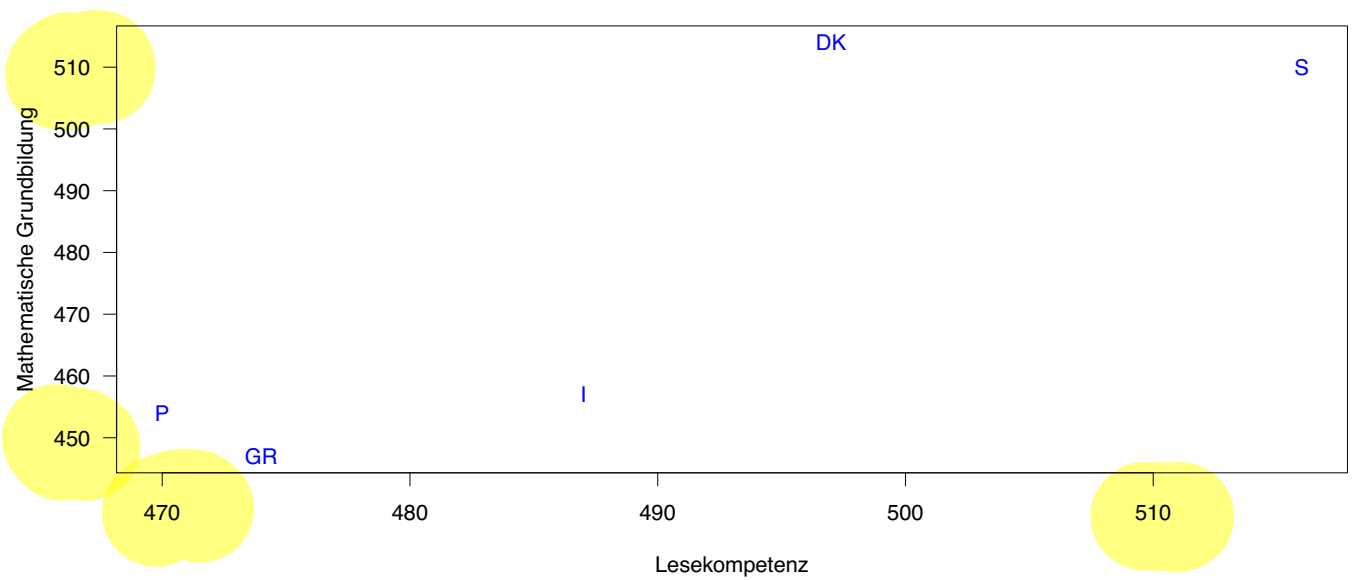
d) Bei einem Patienten ohne Fieber und ohne Husten, aber mit Kurzatmigkeit lautet die Prognose „nicht infiziert“.

Lösung zur Aufgabe 24

Land	Lesekompetenz (X_1)	Mathematische Grundbildung (X_2)
Dänemark (DK)	497	514
Griechenland (GR)	474	447
Italien (I)	487	457
Portugal (P)	470	454
Schweden (S)	516	510

a) Aus dem untenstehenden Streudiagramm (Letterdiagramm) ist zu erkennen:

- das Merkmal "Mathematische Grundbildung X_2" streut viel stärker als das Merkmal „Lesekompetenz X_1" und
- mit $\{\mathrm{P}, \mathrm{GR}, \mathrm{I}\}$ (südliche Länder) und $\{\mathrm{DK}, \mathrm{S}\}$ (skandinavische Länder) gibt es zwei Gruppen von Ländern, die sich bzgl. dieser beiden Merkmake relativ stark unterscheiden.



b) Mit dem Mittelwertvektor

$$\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2) = \frac{1}{5}\sum_{n=1}^{5} \mathbf{x}_n = (488{,}8; 476{,}4)$$

erhält man die zentrierte Datenmatrix

$$\mathbf{D} = \begin{pmatrix} \mathbf{x}_1 - \bar{\mathbf{x}} \\ \mathbf{x}_2 - \bar{\mathbf{x}} \\ \mathbf{x}_3 - \bar{\mathbf{x}} \\ \mathbf{x}_4 - \bar{\mathbf{x}} \\ \mathbf{x}_5 - \bar{\mathbf{x}} \end{pmatrix} = \begin{pmatrix} 8{,}2 & 37{,}6 \\ -14{,}8 & -29{,}4 \\ -1{,}8 & -19{,}4 \\ -18{,}8 & -22{,}4 \\ 27{,}2 & 33{,}6 \end{pmatrix}.$$

Die Matrix **D** gibt an, wie stark sich die 5 Länder bei den beiden Merkmalen vom jeweiligen Mittelwert unterscheiden. Die zugehörige empirische Varianz-Kovarianz-Matrix ist gegeben durch

$$\mathbf{S} = \frac{1}{5}\sum_{n=1}^{5} (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T = \frac{1}{5}\mathbf{D}^T\mathbf{D} = \begin{pmatrix} 276{,}56 & 422{,}68 \\ 422{,}68 & 857{,}04 \end{pmatrix}.$$

c) Die empirische Varianz-Kovarianz-Matrix $\mathbf{S}$ besitzt das charakteristische Polynom

$$p_{\mathbf{S}}(\lambda) = \det \begin{pmatrix} 276{,}56 - \lambda & 422{,}68 \\ 422{,}68 & 857{,}04 - \lambda \end{pmatrix}$$
$$= (276{,}56 - \lambda)(857{,}04 - \lambda) - 422{,}68^2.$$

Durch Nullsetzen und eine kurze Umformung erhält man daraus die quadratische Gleichung

$$\lambda^2 - 1133{,}6\lambda + 58364{,}6 = 0.$$

Lösen dieser Gleichung liefert die beiden auf zwei Nachkommastellen gerundete Eigenwerte $\lambda_1 = 1079{,}54$ und $\lambda_2 = 54{,}06$.

größter Eigenwert

Der Eigenvektor $\mathbf{u}_1 \in \mathbb{R}^2$ zum Eigenwert $\lambda_1 = 1079{,}54$ resultiert als Lösung des linearen Gleichungssystems $\mathbf{S}\mathbf{u}_1 = \lambda_1 \mathbf{u}_1$ bzw. $(\mathbf{S} - \lambda_1 \mathbf{E})\mathbf{u}_1 = \mathbf{0}$. D.h. es ist das lineare Gleichungssystem

$$-802{,}98u_{11} + 422{,}68u_{12} = 0$$
$$422{,}68u_{11} - 222{,}50u_{12} = 0$$

$$\begin{pmatrix} -802{,}98 & 422{,}68 \\ 422{,}68 & -222{,}50 \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

zu lösen.

Zusammen mit $\|\mathbf{u}_1\|^2 = u_{11}^2 + u_{12}^2 = 1$ liefert dies den normierten Eigenvektor

$$\mathbf{u}_1 = \begin{pmatrix} 0{,}466 \\ 0{,}885 \end{pmatrix}.$$

$u_{12} = \pm \left(\left(\frac{422{,}68}{802{,}98} \right)^2 + 1 \right)^{-1/2}$

Der Eigenvektor $\mathbf{u}_2 \in \mathbb{R}^2$ zum Eigenwert $\lambda_2 = 54{,}06$ löst das lineare Gleichungssystem $\mathbf{S}\mathbf{u}_2 = \lambda_2 \mathbf{u}_2$ bzw. $(\mathbf{S} - \lambda_2 \mathbf{E})\mathbf{u}_2 = \mathbf{0}$. D.h. zu lösen ist das lineare Gleichungssystem

$u_{11} = \frac{422{,}68}{802{,}98} u_{12}$

$$\begin{aligned} 222{,}50 u_{21} + \ 422{,}68 u_{22} &= 0 \\ 422{,}68 u_{21} + 802{,}98 u_{22} &= 0. \end{aligned}$$

Zusammen mit $\|\mathbf{u}_2\|^2 = u_{21}^2 + u_{22}^2 = 1$ liefert dies den normierten Eigenvektor

$$\mathbf{u}_2 = \begin{pmatrix} 0{,}885 \\ -0{,}466 \end{pmatrix}.$$

Die beiden Hauptkomponenten sind somit gegeben durch:

$$\begin{aligned} Z_1 &= 0{,}466(X_1 - \overline{X}_1) + 0{,}885(X_2 - \overline{X}_2) \\ Z_2 &= 0{,}885(X_1 - \overline{X}_1) - 0{,}466(X_2 - \overline{X}_2) \end{aligned}$$

Interpretation: Z_1: gewichteter Mittel beider Merkmale (X_1, X_2); X_2 ist fast doppelt so stark gewichtet

Z_2: Kontrast (Differenz) aus beiden Merkmalen