Einführung in Web und Data Science

Übungsblatt 10 - Lösungsvorschlag

Tanya Braun, Ralf Möller

Übung 1. Entscheidungsbäume 12 P.

Gegeben sei die folgende Menge an Trainingsdaten:

Name	Geschlecht	Winter	Intrige	Leben	Familie
Arya	w	1	1	1	Stark
Cersei	w	0	0	1	Lannister
Jaime	m	0	0	1	Lannister
Joanna	w	0	0	0	Lannister
Jon	m	1	1	1	Stark
Margaery	w	0	1	0	Tyrell
Olenna	w	0	1	0	Tyrell
Robb	m	1	1	0	Stark
Sansa	w	1	1	1	Stark
Tyrion	m	1	1	1	Lannister
Tywin	m	0	0	0	Lannister

Die erste Spalte „Name“ identifiziert eine Instanz aus den Trainingsdaten. Die letzte Spalte „Familie“ stellt den Klassennamen dar. Die Spalten „Geschlecht“, „Winter“, „Intrige“ und „Leben“ stellen binäre Attribute mit der folgenden Bedeutung (sekundär für die Bearbeitung der Aufgabe) dar:

Geschlecht Beschreibt das Geschlecht der Figur mit den Ausprägungen weiblich und männlich, abgekürzt w und m.

Winter Beschreibt, ob die Figur daran glaubt, dass der Winter naht, mit den Ausprägungen ja und nein, codiert mit 1 und 0.

Intrige Beschreibt, ob die Figur gegen die Lannisters intrigiert hat, mit den Ausprägungen ja und nein, codiert mit 1 und 0.

Leben Beschreibt, ob die Figur noch am Leben ist, mit den Ausprägungen ja und nein, codiert mit 1 und 0.

Bauen Sie einen Entscheidungsbaum. Nutzen Sie den (maximalen) Informationsgewinn, um das nächste Attribut zum Aufteilen auszuwählen. Schreiben Sie Ihre Rechnungen und Entscheidungen so auf, dass sie nachvollziehbar sind. Zeichnen Sie am Ende Ihren gebauten Entscheidungsbaum.

Lösung 1.
Hinweis: Die gezeigte Lösung ist ausführlicher als für die Punkte nötig.

Für die Rechnungen benutze ich folgende Abkürzungen: G = Geschlecht, W = Winter, I = Intrige, L = Leben. Die Reihenfolge der Häufigkeiten der Klassen in der Informationsfunktion ist $[n_{Lannister}, n_{Stark}, n_{Tyrell}]$. Wir beginnen mit dem Wurzelknoten mit allen Trainingsdaten in diesem Knoten.
Vorherinformationsgehalt

$$\begin{aligned} N &= 11 \\ info([5,4,2]) &= -\frac{5}{11}\log_2\left(\frac{5}{11}\right) - \frac{4}{11}\log_2\left(\frac{4}{11}\right) - \frac{2}{11}\log_2\left(\frac{2}{11}\right) \approx 1,495 \end{aligned}$$

Informationsgewinn für G:

$$\begin{aligned} G = w, N_{G=w} &= 6 \\ info([2,2,2]) &= -\frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) && \approx 1,585 \\ G = m, N_{G=m} &= 5 \\ info([3,2,0]) &= -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{0}{5}\log_2\left(\frac{0}{5}\right) && \approx 0,971 \\ gain(G) &= info([5,4,2]) - info([2,2,2],[3,2,0]) \\ &= 1,495 - \left(\frac{6}{11}info([2,2,2]) + \frac{5}{11}info([3,2,0])\right) \approx 0,189 \end{aligned}$$

Informationsgewinn für W:

$$\begin{aligned} W = 1, N_{W=1} &= 5 \\ info([1,4,0]) &= -\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\log_2\left(\frac{4}{5}\right) - \frac{0}{5}\log_2\left(\frac{0}{5}\right) && \approx 0,722 \\ W = 0, N_{W=0} &= 6 \\ info([4,0,2]) &= -\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{0}{6}\log_2\left(\frac{0}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) && \approx 0,918 \\ gain(W) &= info([5,4,2]) - info([1,4,0],[4,0,2]) \\ &= 1,495 - \left(\frac{5}{11}info([1,4,0]) + \frac{6}{11}info([4,0,2])\right) \approx 0,666 \end{aligned}$$

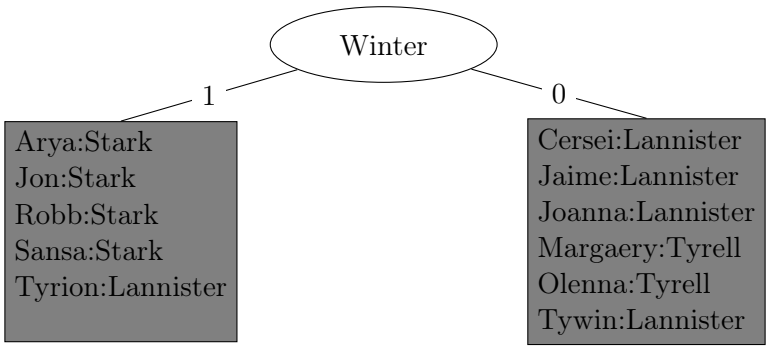
Informationsgewinn für I:

$$
\begin{aligned}
&I = 1, N_{I=1} = 7 \\
&\quad info([1,4,2]) = -\frac{1}{7}\log_2\left(\frac{1}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right) - \frac{2}{7}\log_2\left(\frac{2}{7}\right) && \approx 1,379 \\
&I = 0, N_{I=0} = 4 \\
&\quad info([4,0,0]) = -\frac{4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right) && = 0,000 \\
&\qquad gain(I) = info([5,4,2]) - info([1,4,2],[4,0,0]) \\
&\qquad\quad = 1,495 - \left(\frac{5}{11} info([1,4,2]) + \frac{6}{11} info([4,0,0])\right) && \approx 0,618
\end{aligned}
$$

Informationsgewinn für L:

$$
\begin{aligned}
&L = 1, N_{L=1} = 6 \\
&\quad info([3,3,0]) = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{0}{6}\log_2\left(\frac{0}{6}\right) && \approx 1,000 \\
&L = 0, N_{L=0} = 5 \\
&\quad info([2,1,2]) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) && \approx 1,522 \\
&\qquad gain(L) = info([5,4,2]) - info([3,3,1],[2,1,1]) \\
&\qquad\quad = 1,495 - \left(\frac{6}{11} info([3,3,0]) + \frac{5}{11} info([2,1,2])\right) && \approx 0,258
\end{aligned}
$$

Der größte Informationsgewinn liegt bei W. Wir generieren also zwei Kindknoten und teilen die Trainingsdaten basierend auf ihrem Wert für W auf die beiden Knoten auf. Die Kanten zu den Knoten kennzeichnen wir entsprechend der W-Werte in dem jeweiligen Kind. Der Baum sieht momentan so aus:



Weiter geht es mit dem Knoten, bei dem wir mit $W = 1$ landen. Die Trainingsdaten dort sehen, wie folgt, aus:

Name	Geschlecht	Winter	Intrige	Leben	Familie
Arya	w	1	1	1	Stark
Jon	m	1	1	1	Stark
Robb	m	1	1	0	Stark
Sansa	w	1	1	1	Stark
Tyrion	m	1	1	1	Lannister

Jetzt beginnen wir von vorn den Informationsgewinn für diese Daten und verbliebenen Attribute auszurechnen. Da der Klassenname „Tyrell“ nicht mehr vorkommt, lassen wir den Teil aus der Berechnung weg, der sich auf die Tyrells bezieht, da alle Werte 0 ergeben würden.
Vorherinformationsgehalt

$$N = 5$$
$$info\left([1,4]\right) = -\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\log_2\left(\frac{4}{5}\right) \approx 0,722$$

Informationsgewinn für G:

$$\begin{aligned}
G = w, N_{G=w} = 2 & \\
info\left([0,2]\right) &= -\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right) && = 0,000 \\
G = m, N_{G=m} = 3 & \\
info\left([1,2]\right) &= -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) && \approx 0,918 \\
gain(G) &= info\left([1,4]\right) - info\left([0,2],[1,2]\right) & & \\
&= 0,722 - \left(\frac{2}{5}info\left([0,2]\right) + \frac{3}{5}info\left([1,2]\right)\right) && \approx 0,171
\end{aligned}$$

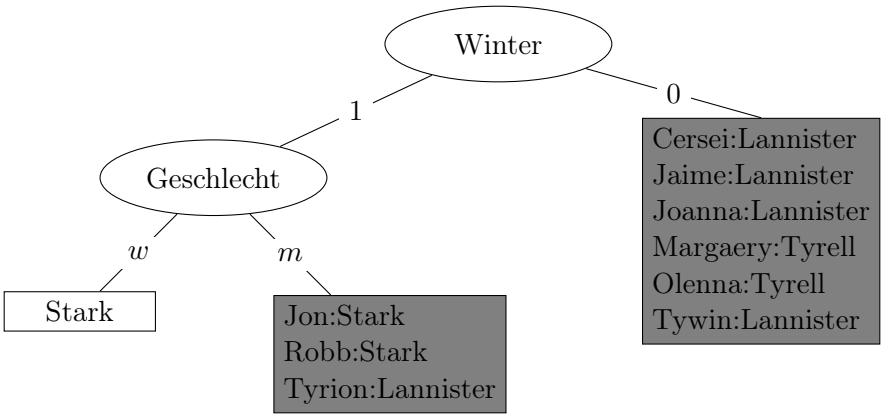
Informationsgewinn für I (alle Werte sind gleich, daher kann es keinen Informationsgewinn geben, bräuchte man also nicht rechnen):

$$\begin{aligned}
I = 1, N_{I=1} = 5 & \\
info\left([1,4]\right) &= -\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\log_2\left(\frac{4}{5}\right) && \approx 0,722 \\
I = 0, N_{I=0} = 0 & \\
info\left([0,0]\right) &= -\frac{0}{5}\log_2\left(\frac{0}{5}\right) - \frac{0}{5}\log_2\left(\frac{0}{5}\right) && = 0,000 \\
gain(I) &= info\left([1,4]\right) - info\left([1,4],[0,0]\right) & & \\
&= 0,722 - \left(\frac{5}{5}info\left([1,4]\right) + \frac{0}{5}info\left([0,0]\right)\right) && = 0,000
\end{aligned}$$

Informationsgewinn für L:

$$
\begin{aligned}
&L = 1, N_{L=1} = 4 \\
&\qquad info\left([1,3]\right) = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right) && \approx 0,811 \\
&L = 0, N_{L=0} = 1 \\
&\qquad info\left([0,1]\right) = -\frac{0}{1}\log_2\left(\frac{0}{1}\right) - \frac{1}{1}\log_2\left(\frac{1}{1}\right) && = 0,000 \\
&\qquad\quad gain(L) = info\left([1,4]\right) - info\left([1,3],[0,1]\right) \\
&\qquad\qquad\quad = 0,722 - \left(\frac{4}{5} info\left([1,3]\right) + \frac{1}{5} info\left([0,1]\right)\right) \approx 0,073
\end{aligned}
$$

Wir wählen G zum Aufteilen der Daten. Wir generieren also wieder zwei Kindknoten und teilen die Trainingsdaten dieses Mal basierend auf ihrem Wert für G auf die beiden Knoten auf. Die Kanten zu den Knoten kennzeichnen wir entsprechend der G-Werte in dem jeweiligen Kind. Die Trainingsdaten mit dem Wert w haben alle den gleichen Klassennamen. Wir hören also in dem Knoten auf zu splitten und fügen hier den Klassennamen „Stark“ hinzu.



Die anderen Daten haben noch keine eindeutige Klasse und wir haben noch zwei mögliche Attribute zum Teilen. Wir berechnen also weiter Informationsgewinne. Die übrig gebliebenen Trainingsdaten sind

Name	Geschlecht	Winter	Intrige	Leben	Familie
Jon	m	1	1	1	Stark
Robb	m	1	1	0	Stark
Tyrion	m	1	1	1	Lannister

Vorherinformationsgehalt

$$
\begin{aligned}
&N = 3 \\
&info\left([1,2]\right) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) \approx 0,918
\end{aligned}
$$

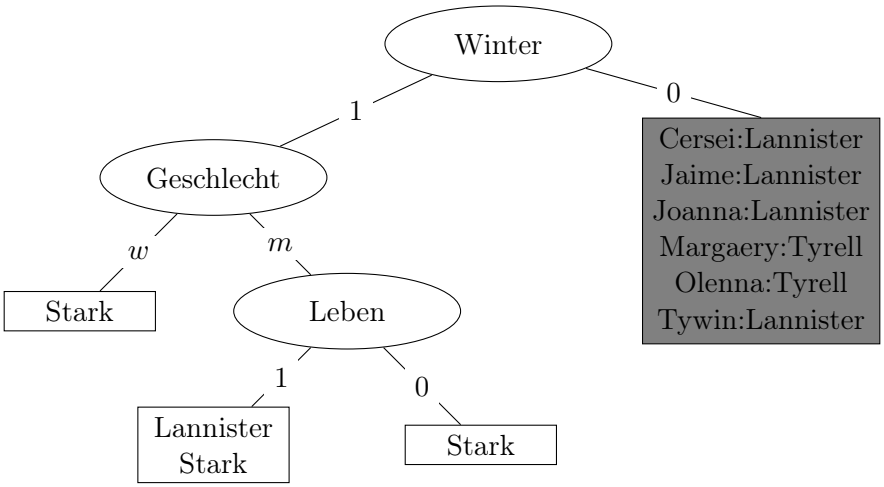
Informationsgewinn für I (ergab beim letzten Mal schon 0, bräuchten wir nicht noch mal neu ausrechnen):

$$
\begin{aligned}
&I = 1, N_{I=1} = 3 \\
&\quad info([1,2]) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) && \approx 0,918 \\
&I = 0, N_{I=0} = 0 \\
&\quad info([0,0]) = -\frac{0}{3}\log_2\left(\frac{0}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right) && = 0,000 \\
&\quad gain(I) = info([1,2]) - info([1,2],[0,0]) \\
&\qquad = 0,918 - \left(\frac{3}{3}info([1,2]) + \frac{0}{3}info([0,0])\right) = 0,000
\end{aligned}
$$

Informationsgewinn für L:

$$
\begin{aligned}
&L = 1, N_{L=1} = 2 \\
&\quad info([1,1]) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) && = 1,000 \\
&L = 0, N_{L=0} = 1 \\
&\quad info([0,1]) = -\frac{0}{1}\log_2\left(\frac{0}{1}\right) - \frac{1}{1}\log_2\left(\frac{1}{1}\right) && = 0,000 \\
&\quad gain(L) = info([1,2]) - info([1,1],[0,1]) \\
&\qquad = 0,918 - \left(\frac{2}{3}info([1,1]) + \frac{1}{3}info([0,1])\right) \approx 0,252
\end{aligned}
$$

Wir splitten basierend auf L. Der Kindknoten für $L = 0$ enthält ein Trainingsdatum mit dem Klassennamen „Stark“, bekommt also diese Klasse zugeordnet. Der andere Kindknoten enthält die Daten für Jon und Tyrion, die nicht weiter gesplittet werden können. Hier erhalten wir also eine Verteilung von $(\frac{1}{2}, \frac{1}{2})$ für die beiden Klassennamen „Lannister“ und „Stark“. Für den fertigen Baum müssen wir uns noch für eine Klasse entscheiden.



Jetzt gehen wir zurück zu dem Kindknoten mit dem Wert $W = 0$. Die Trainingsdaten dort sind die folgenden:

Name	Geschlecht	Winter	Intrige	Leben	Familie
Cersei	w	0	0	1	Lannister
Jaime	m	0	0	1	Lannister
Joanna	w	0	0	0	Lannister
Margaery	w	0	1	0	Tyrell
Olenna	w	0	1	0	Tyrell
Tywin	m	0	0	0	Lannister

Jetzt berechnen wir wieder die Informationsgewinne.
Vorherinformationsgehalt:

$$N = 6$$
$$info\left([4,2]\right) = -\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) \approx 0,918$$

Informationsgewinn für G:

$$\begin{aligned}
G = w, N_{G=w} = 4 & \\
info\left([2,2]\right) &= -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) && = 1,000 \\
G = m, N_{G=m} = 2 & \\
info\left([2,0]\right) &= -\frac{2}{2}\log_2\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right) && = 0,000 \\
gain(G) &= info\left([4,2]\right) - info\left([2,2],[2,0]\right) \\
&= 0,918 - \left(\frac{4}{6}info\left([2,2]\right) + \frac{2}{6}info\left([2,0]\right)\right) \approx 0,252
\end{aligned}$$

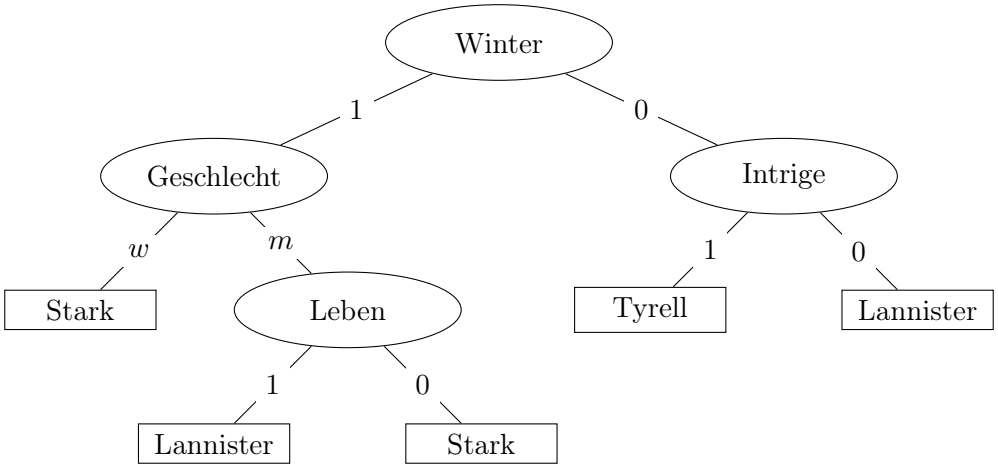
Informationsgewinn für I:

$$\begin{aligned}
I = 1, N_{I=1} = 2 & \\
info\left([0,2]\right) &= -\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right) && = 0,000 \\
I = 0, N_{I=0} = 4 & \\
info\left([4,0]\right) &= -\frac{4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right) && = 0,000 \\
gain(I) &= info\left([4,2]\right) - info\left([0,2],[4,0]\right) \\
&= 0,918 - \left(\frac{2}{6}info\left([0,2]\right) + \frac{4}{6}info\left([4,0]\right)\right) \approx 0,918
\end{aligned}$$

Informationsgewinn für L:

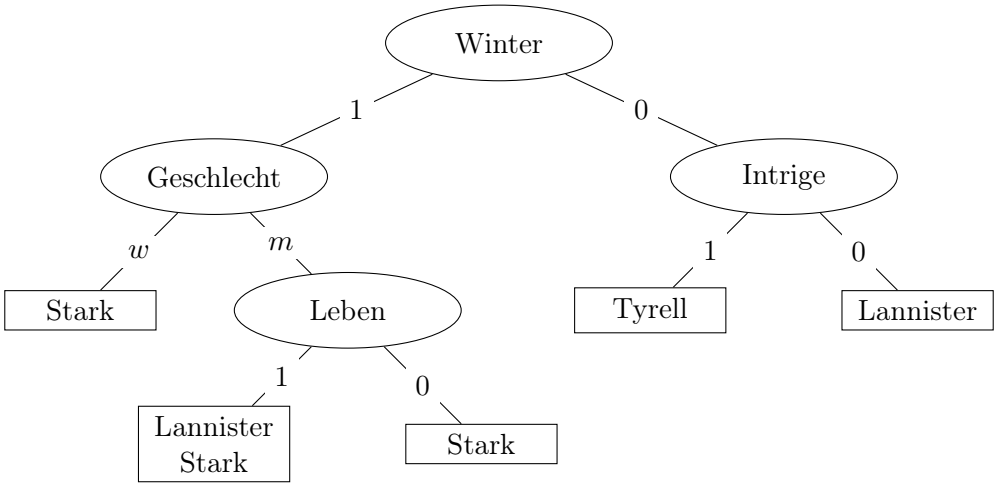
$$
\begin{aligned}
&L = 1, N_{L=1} = 2 \\
&\qquad info\left([2,0]\right) = -\frac{2}{2}\log_2\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right) && \approx 0,000 \\
&L = 0, N_{L=0} = 4 \\
&\qquad info\left([2,2]\right) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) && \approx 1,000 \\
&\qquad\quad gain(L) = info\left([4,2]\right) - info\left([2,0],[2,2]\right) \\
&\qquad\qquad\quad = 0,918 - \left(\frac{2}{6}info\left([2,0]\right) + \frac{4}{6}info\left([2,2]\right)\right) = 0,251
\end{aligned}
$$

Wir wählen I und haben eine perfekte Aufteilung der Daten. Wir generieren zwei Kindknoten für die beiden I-Werte und fügen bei dem $(I = 1)$-Kind den Klassennamen „Tyrell“ und bei dem anderen Kind den Klassennamen „Lannister“ hinzu. Der fertige Baum sieht, wie folgt, aus, wenn man sich entscheidet die Klasse Lannister im untersten Level zu verwenden:



Übung 2. Pruning bei Entscheidungsbäumen 8 P.

Gegeben sei der folgende Entscheidungsbaum:



Von den Ihnen zur Verfügung stehenden Daten haben Sie die folgenden während des Lernens nicht benutzt („hold-out set“).

Name	Geschlecht	Winter	Intrige	Leben	Familie
Joffrey	m	0	0	0	Lannister
Myrcella	w	0	0	0	Lannister
Tommen	m	0	0	0	Lannister
Tytos	m	0	0	0	Lannister
Jeyne	w	0	0	0	Lannister
Lancel	m	1	1	1	Lannister
Lyanna	w	1	0	0	Stark
Eddard	m	1	1	0	Stark
Rickard	m	1	0	0	Stark
Loras	m	0	1	0	Tyrell

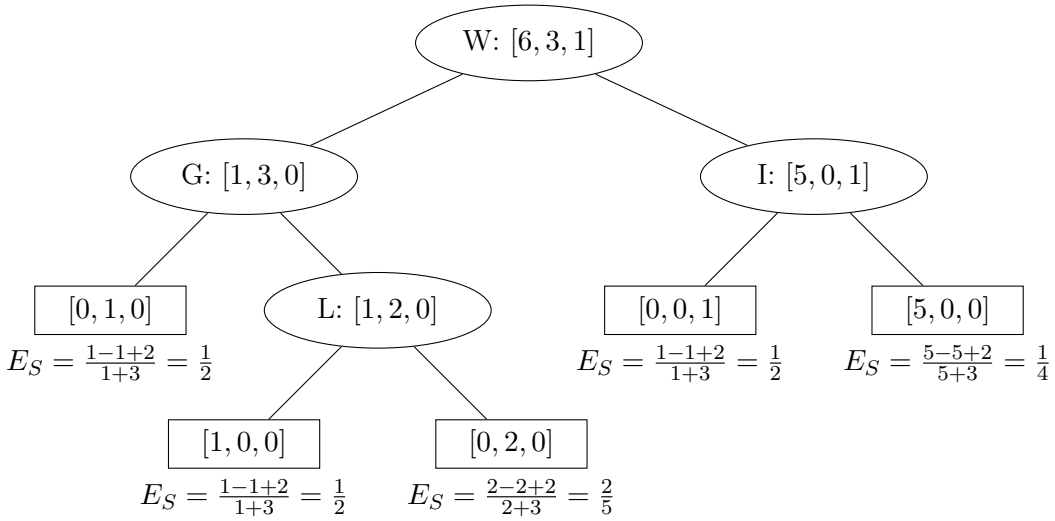
Nutzen Sie die Daten nun um den Baum mittels „Expected Error Pruning“ zu beschneiden. Schreiben Sie Ihre Rechnungen und Entscheidungen nachvollziehbar auf. Geben Sie am Ende Ihren geprunten Entscheidungsbaum an.

Lösung 2.
In den Knoten steht die Klassenverteilung. Die Reihenfolge der Klassennamen ist dabei Lannister, Stark, Tyrell ($[L, S, T]$); $k = 3$. Die folgenden Gleichungen werden benutzt.

$$E_S(Node) = \frac{N - n + k - 1}{N + k} = \frac{N - n + 2}{N + 3} \qquad \text{(Static Expected Error)}$$

$$E_B(Node) = \sum_i P_i \cdot E(Node_i) \qquad \text{(Backed-up Error)}$$

$$E(Node) = \min(E_S(Node), E_B(Node))$$



Rechnungen für die inneren Knoten
Knoten L

$$E_S([1,2,0]) = \frac{3-2+2}{3+3} = \frac{1}{2} = \frac{15}{30}$$
$$E_B = \frac{1}{3} \cdot \frac{1}{2} + \frac{2}{3} \cdot \frac{2}{5} = \frac{1}{6} + \frac{4}{15} = \frac{13}{30}$$

$E_S > E_B \rightarrow$ Nicht prunen
Knoten G

$$E_S([1,3,0]) = \frac{4-3+2}{4+3} = \frac{3}{7}$$
$$E_B = \frac{1}{4} \cdot \frac{1}{2} + \frac{3}{4} \cdot \frac{13}{30} = \frac{1}{8} + \frac{13}{40} = \frac{9}{20}$$

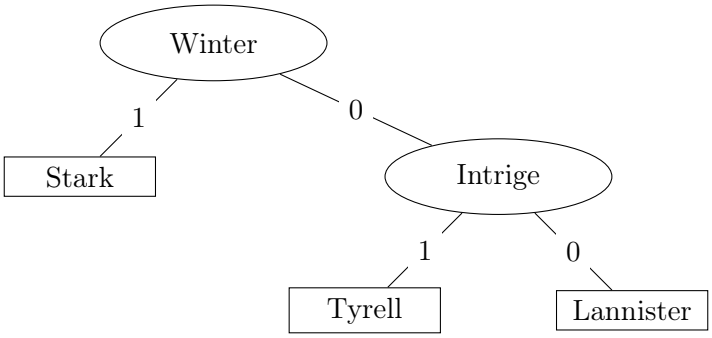
$E_S < E_B \rightarrow$ Prunen
Knoten I

$$E_S([5,0,1]) = \frac{6-5+2}{6+3} = \frac{1}{3} = \frac{8}{24}$$
$$E_B = \frac{1}{6} \cdot \frac{1}{2} + \frac{5}{6} \cdot \frac{1}{4} = \frac{1}{12} + \frac{5}{24} = \frac{7}{24}$$

$E_S > E_B \rightarrow$ Nicht prunen
Knoten W

$$E_S([6,3,1]) = \frac{10-6+2}{10+3} = \frac{6}{13}$$
$$E_B = \frac{4}{10} \cdot \frac{3}{7} + \frac{6}{10} \cdot \frac{7}{24} = \frac{6}{35} + \frac{7}{40} = \frac{97}{280}$$

$E_S > E_B \rightarrow$ Nicht prunen
Der geprunte Entscheidungsbaum:


Winter
1
0
Stark
Intrige
1
0
Tyrell
Lannister