
Data Mining und Maschinelles Lernen

Wintersemester 2015/2016

Lösungsvorschlag für das 9. Übungsblatt



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Aufgabe 1: Decision Trees (1)

Gegeben sei folgende Beispielmenge:

Age	Education	Married	Income	Credit?
old	secondary	no	medium	yes
young	college	yes	low	yes
old	secondary	yes	medium	yes
young	college	no	high	yes
young	secondary	no	high	yes
old	secondary	no	high	yes
old	college	no	high	yes
old	college	yes	low	yes
young	primary	yes	medium	no
old	primary	yes	low	no
young	secondary	yes	medium	no
young	college	no	medium	no

a) Erzeugen Sie einen Entscheidungsbaum mittels des Verfahrens ID3 (TDIDT mit Maß Gain) und zeichnen Sie diesen.

Aufgabe 1: Decision Trees (2)



a) Erzeugen Sie einen Entscheidungsbaum mittels des Verfahrens ID3 (TDIDT mit Maß Gain) und zeichnen Sie diesen.

Lösung:

In dieser Aufgabe verwenden wir die Klasse yes als positive Klasse (+) bzw. no als negative Klasse (-). Für die Berechnung des (Information) Gains für ein Attribut benötigen wir die Entropien der ursprünglichen Beispielmenge und der Beispielmengen, die sich durch dessen Attributwerte ergeben. Die Entropie aller Beispiele (Menge S , davon sind 8 positive und 4 negative Beispiele) berechnet sich wie folgt:

$$\begin{aligned} Entropy(S) &= -\frac{8}{8+4} \log_2 \left(\frac{8}{8+4} \right) - \frac{4}{8+4} \log_2 \left(\frac{4}{8+4} \right) \\ &= 0,918 \end{aligned}$$

Aufgabe 1: Decision Trees (3)

Jetzt benötigen wir noch die Entropien aller Tests (Attribute). Mögliche Attribute sind *Age*, *Education*, *Married* und *Income*. Wir verwenden *Age* exemplarisch zur Berechnung der Entropie eines Tests und dessen Information Gains.

Die Abdeckung von *Age = old* sind 5 positive und 1 negatives Beispiel ($p_+ = 5/6, p_- = 1/6$) und für *Age = young* sind jeweils 3 Beispiele ($p_+ = 1/2, p_- = 1/2$) abgedeckt.

Weiterhin ergeben sich $S_{Age=old}/S = 6/(6+6) = 0,5$ und $S_{Age=young}/S = 0,5$.

Berechnen wir nun die Entropien dieser beiden Testausgänge:

$$Entropy(S_{Age=old}) = -\frac{5}{6} \log_2 \left(\frac{5}{6} \right) - \frac{1}{6} \log_2 \left(\frac{1}{6} \right) \approx 0,65$$

$$Entropy(S_{Age=young}) = -\frac{1}{2} \log_2 \left(\frac{1}{2} \right) - \frac{1}{2} \log_2 \left(\frac{1}{2} \right) = 1$$

Subtrahieren wir nun diese Entropien mit $S_{Age=old}$ und $S_{Age=young}$ gewichtet von der Entropie der ursprünglichen Beispielmenge ab:

$$\begin{aligned} Gain(S, Age) &= Entropy(S) - \frac{S_{Age=old}}{S} \cdot Entropy(S_{Age=old}) - \frac{S_{Age=young}}{S} \cdot Entropy(S_{Age=young}) \\ &= 0,918 - 0,5 \cdot 0,65 - 0,5 \cdot 1 = 0,093 \end{aligned}$$

Aufgabe 1: Decision Trees (4)

Die Berechnung der Information Gains der verbleibenden Tests erfolgt analog. Die hierfür benötigten Werte und die erhaltenen Information Gains können Sie der folgenden Tabelle entnehmen:

Attribut	Wert	+	-	p_+	p_-	Entropie	S_i/s	Gain
Verteilung		8	4	0,67	0,33	0,918	-	
Age	old	5	1	0,83	0,17	0,650	1/2	0,093
	young	3	3	0,50	0,50	1,000	1/2	
Education	college	4	1	0,80	0,20	0,722	5/12	0,317
	primary	0	2	0,00	1,00	0,000	1/6	
	secondary	4	1	0,80	0,20	0,722	5/12	
Married	no	5	1	0,83	0,17	0,650	1/2	0,093
	yes	3	3	0,50	0,50	1,000	1/2	
Income	high	4	0	1,00	0,00	0,000	1/3	0,284
	low	2	1	0,67	0,33	0,918	1/4	
	medium	2	3	0,40	0,60	0,971	5/12	

Da das Maß Information Gain maximiert werden soll, entscheiden wir uns für den Test *Education*. Die ursprüngliche Beispielmenge wird auf die 3 Testausgänge von *Education* (*college*, *primary* und *secondary*) aufgeteilt.

Die Beispielmenge von *primary* besteht nur aus negativen Beispielen, hier werden keine weiteren Tests benötigt (*Education=primary* ist somit das erste Blatt des Baums mit der Vorhersage *no*).

Für *college* und *secondary* müssen wir noch weitere Tests einführen, da deren Beispielmengen noch positive und negative Beispiele enthalten.

Betrachten wir zuerst die Beispielmenge von *Education = college*. Auf die Berechnung werden wir nicht im Detail eingehen, die benötigten Werte und die Ergebnisse sind wieder in der folgenden Tabelle zusammengefaßt:

Attribut	Wert	+	-	p_+	p_-	Entropie	S_i/s	Gain
Verteilung		4	1	0,80	0,20	0,722	-	
Age	old	2	0	1,00	0,00	0,000	$\frac{2}{5}$	0,171
	young	2	1	0,67	0,33	0,918	$\frac{3}{5}$	
Married	no	2	1	0,67	0,33	0,918	$\frac{3}{5}$	0,171
	yes	2	0	1,00	0,00	0,000	$\frac{2}{5}$	
Income	high	2	0	1,00	0,00	0,000	$\frac{2}{5}$	0,722
	low	2	0	1,00	0,00	0,000	$\frac{2}{5}$	
	medium	0	1	0,00	1,00	0,000	$\frac{1}{5}$	

Der Test *income* trennt die verbleibenden Beispiele perfekt.

Test Education (2)

Keine weiteren Tests sind nötig, wir können also mit der Beispielmenge von *Education = secondary* weitermachen.

Attribut	Wert	+	-	p_+	p_-	Entropie	S_i/S	Gain
Verteilung		4	1	0,80	0,20	0,722	-	
Age	old	3	0	1,00	0,00	0,000	$\frac{3}{5}$	0,322
	young	1	1	0,50	0,50	1,000	$\frac{2}{5}$	
Married	no	3	0	1,00	0,00	0,000	$\frac{3}{5}$	0,322
	yes	1	1	0,50	0,50	1,000	$\frac{2}{5}$	
Income	high	2	0	1,00	0,00	0,000	$\frac{2}{5}$	0,171
	low	0	0	0,00	0,00	0,000	-	
	medium	2	1	0,67	0,33	0,918	$\frac{3}{5}$	

Die Tests *Age* und *Married* weisen den gleichen Information Gain auf. Wir entscheiden uns für den ersten Test, also *Age*, um diesen Konflikt aufzulösen.

Die Beispielmenge wird also auf zwei Beispielmengen, für *Age = old* und *Age = young*, aufgeteilt. Da die Beispielmenge von *Age = old* nur positive Beispiele enthält, müssen wir nur noch die Beispielmenge von *Age = young* betrachten:

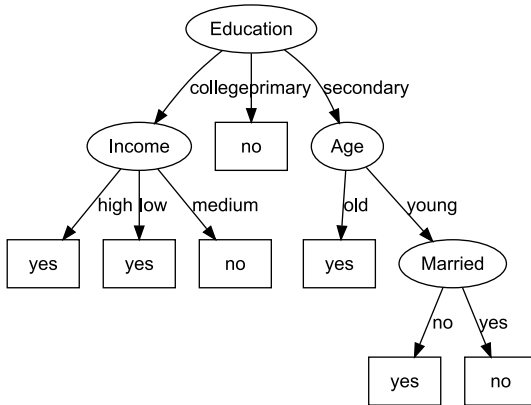
Test Age (1)

Attribut	Wert	+	-	p_+	p_-	Entropie	S_i/s	Gain
Verteilung		1	1	0,5	0,5	1,000	-	
Married	no	1	0	1,00	0,00	0,000	1/2	1,000
	yes	0	1	0,00	1,00	0,000	1/2	
Income	high	1	0	1,00	0,00	0,000	1/2	1,000
	low	0	0	0,00	0,00	0,000	-	
	medium	0	1	0,00	1,00	0,000	1/2	

Die verbleibenden Tests erzielen wiederum den gleichen Information Gain (beide trennen die Beispiele perfekt), wir entscheiden uns abermals für den ersten Test *Married*.

Entscheidungsbaum (1)

Da keinen weiteren Tests nötig sind, erhalten wir den folgenden Baum:



Aufgabe 2b) (1)



TECHNISCHE
UNIVERSITÄT
DARMSTADT

b) Wiederholen Sie die Berechnungen für die Auswahl des Tests in der Wurzel mit den Maßen Information-Gain-Ratio und Gini-Index. Ändert sich etwas?

Gain Ratio (1)

Zur Berechnung wird zusätzlich zu den Berechnungen des Information Gains die intrinsische Information von jedem Test benötigt. Das heißt, für jeden Test muß die Verteilung der Einheiten auf die möglichen Testausgänge bestimmt werden.

Betrachten wir beispielsweise den Test *Income*, die Beispiele teilen sich auf die Testausgänge wie folgt auf: 4, 3 und 5 Beispiele (in relativen Häufigkeiten $4/12$, $3/12$ und $5/12$).

$$Intl(S, Income) = -\frac{4}{12} \cdot \log\left(\frac{4}{12}\right) - \frac{3}{12} \cdot \log\left(\frac{3}{12}\right) - \frac{5}{12} \cdot \log\left(\frac{5}{12}\right) \approx 1,555$$

Teilen wir nun den Information Gain (siehe Teilaufgabe a)) durch den eben berechneten Wert, erhalten wir den Gain Ratio von *Income*.

$$GR(S, Income) = \frac{Gain(S, Income)}{Intl(S, Income)} = \frac{0,284}{1,555} \approx 0,183$$

Gain Ratio (2)

Entsprechend berechnen sich die Gain Ratios der anderen Attribute.

Attribut	Gain	S_i/s	Intrinsic	GainRatio
Age	0,093	$\{1/2, 1/2\}$	1,00	0,093
Education	0,317	$\{5/12, 1/6, 5/12\}$	1,483	0,214
Married	0,093	$\{1/2, 1/2\}$	1,00	0,093
Income	0,284	$\{1/3, 1/4, 5/12\}$	1,555	0,183

Da GainRatio maximiert werden soll, ist auch hier der Test *Education* optimal.

Die Berechnung des GiniIndex ist ähnlich zur Berechnung des Information Gains. Wobei hier anstatt der Entropie das Unreinheitsmaß Gini verwendet wird.

Die so erhaltenen Werte werden aufaddiert, aber nicht von der ursprünglichen Unreinheit abgezogen. Aus diesem Grund wird der GiniIndex nicht maximiert, sondern minimiert.

Berechnen wir beispielhaft den GiniIndex von *Age*. Die Tabelle aus Teilaufgabe a) können wir verwenden.

Attribut	Wert	+	-	p_+	p_-	Entropie	s_i/s	Gain
Verteilung		8	4	0,67	0,33	0,918	-	
Age	old	5	1	0,83	0,17	0,650	1/2	0,093
	young	3	3	0,50	0,50	1,000	1/2	

Berechnen wir nun zunächst die Gini-Werte für die beiden Testausgänge old und young.

$$Gini(S_{Age=old}) = 1 - p_+^2 - p_-^2 \approx 1 - 0,83^2 - 0,17^2 \approx 0,278$$

$$Gini(S_{Age=young}) = 1 - p_+^2 - p_-^2 \approx 1 - 0,5^2 - 0,5^2 \approx 0,5$$

Addieren wir diese Werte gewichtet durch die relative Häufigkeiten der beiden Testausgänge auf, erhalten wir den GiniIndex von *Age*.

$$Gini(S, Age) \approx 0,5 \cdot 0,278 + 0,5 \cdot 0,5 \approx 0,389$$

GiniIndex (2)

Die GiniIndizes der anderen Tests werden analog berechnet.

Attribut	Wert	+	-	p_+	p_-	Gini(S)	s_i/s	Gini(S,A)
Verteilung		8	4	0,67	0,33	0,918	-	
Age	old	5	1	0,83	0,17	0,278	1/2	0,389
	young	3	3	0,50	0,50	0,500	1/2	
Education	college	4	1	0,80	0,20	0,320	5/12	0,267
	primary	0	2	0,00	1,00	0,000	1/6	
	secondary	4	1	0,80	0,20	0,320	5/12	
Married	no	5	1	0,83	0,17	0,278	1/2	0,389
	yes	3	3	0,50	0,50	0,500	1/2	
Income	high	4	0	1,00	0,00	0,000	1/3	0,311
	low	2	1	0,67	0,33	0,444	1/4	
	medium	2	3	0,40	0,60	0,480	5/12	

Wie bereits erwähnt wird ein minimaler GiniIndex gesucht. Wiederum ist *Education* der optimale Test.

Aufgabe 1c) (1)

c) Klassifizieren Sie die folgenden Beispiele mit dem Baum aus der Teilaufgabe a):
'?' steht hier für einen unbekanntem/fehlenden Attributwert.

Age	Education	Married	Income	Credit?
?	secondary	no	medium	?
young	?	yes	low	?

Aufgabe 1c) (2)

Die zu klassifizierenden Beispiele sind unvollständig, d.h. manche Attributwerte sind unbekannt. Aus diesem Grund müssen wir uns überlegen, wie fehlende Attributwerte während der Klassifikation zu behandeln sind.

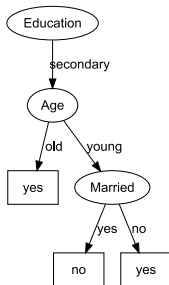
Fall 1 – Attribut wird auf dem Pfad zu einem Blatt nicht getestet: Uns stört der fehlende Wert nicht und die Klassifikation kann ohne Probleme durchgeführt werden.

Fall 2 – Das betreffende Attribut wird auf dem Klassifikation-Pfad als Test verwendet: Wir teilen die Beispiele in mehrere gewichtete Beispiele auf. Insgesamt werden genau so viele Beispiele erzeugt wie das Attribut Attributwerte besitzt. Die Gewichte entsprechen der relativen Häufigkeiten der Trainingsbeispiele, die sich auf die verschiedenen Attributwerte verteilt haben.

1. Beispiel (1)

Verdeutlichen wir diese Vorgehensweise an dem ersten zu klassifizierenden Beispiel.

Age	Education	Married	Income	Credit?
?	secondary	no	medium	?



Der erste Test ist *Education*, der Attributwert *secondary* liegt vor. Wir können also zum nächsten Knoten weitergehen.

1. Beispiel (1)

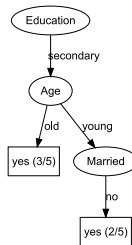
Wie man sieht, wird als nächstes das Attribut *Age* getestet.

Diesmal liegt uns der Attributwert nicht vor, das Beispiel muß nun aufgeteilt werden. Aus der Teilaufgabe

a) wissen wir, daß sich $\frac{3}{5}$ bzw. $\frac{2}{5}$ der Beispiele auf den Attributwert *old* bzw. *young* verteilen.

Demnach wird ein Beispiel auf der Kante *Age = old* mit dem Gewicht $\frac{3}{5}$ durchgeleitet, entsprechend für *Age = young* ein Beispiel mit dem Gewicht $\frac{2}{5}$.

Verfolgen wir nun den Weg dieser Beispiele in dem folgenden Baum.



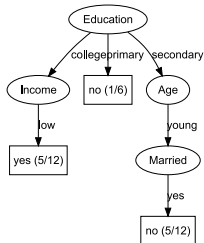
Beide Pfade enden in einem Blatt, das jeweils die Klasse *yes* voraussagt. Demnach wird das Beispiel als *yes* klassifiziert.

2. Beispiel (1)

Age	Education	Married	Income	Credit?
young	?	yes	low	?

Analog gehen wir bei dem zweiten zu klassifizierenden Beispiel vor, bei dem schon das erste Testattribut (der Wurzelknoten) fehlt.

Das Beispiel wird in dem Verhältnis $5/12, 2/12$ und $5/12$ aufgeteilt.
Es gibt keine weiteren fehlenden Attribute.



Insgesamt erhalten die Klassen yes $5/12$ und no $2/12 + 5/12 = 7/12$. Demnach wird die Klasse no vorhergesagt.