Data Mining und Maschinelles Lernen

Wintersemester 2015/2016

Lösungsvorschlag für das 10. Übungsblatt

TECHNISCHE UNIVERSITÄT DARMSTADT

Aufgabe 1: Regressionsbäume (1)

TECHNISCHE UNIVERSITÄT DARMSTADT

Gegeben sei folgende Beispielmenge:

A1	A2	A3	A4	Value
C	K	T	X	0.28
B	J	S	X	0.50
C	J	S	Z	0.35
B	I	R	Y	5.50
A	J	T	Z	0.35
A	K	S	Z	0.80
C	I	R	Y	5.10
A	I	R	Y	5.70
C	I	S	Y	0.76
B	I	S	X	1.03
B	K	R	Y	0.46
C	K	T	Z	0.39
B	K	S	X	0.28
A	K	R	X	1.10

Aufgabe 1: Regressionsbäume (2)

TECHNISCHE UNIVERSITÄT DARMSTADT

Erzeugen Sie einen Regressionsbaum mittels des wie folgt modifizierten Verfahrens ID3.
Verwenden Sie hierzu das Maß Standard Deviation Reduction (SDR) zur Auswahl der Tests und den Mittelwert der Instanzen eines Blattes als Vorhersagewert.
Hierbei soll ein Knoten, sobald er weniger als 3 Instanzen abdeckt, nicht weiter aufgeteilt und zu einem Blatt umgewandelt werden.
Sollte bei einem Test ein Testausgang keine Instanzen abdecken, fließt er nicht in die Berechnung des SDRs ein und soll, da keine Daten für ihn vorhanden sind, als Blatt verwendet werden, das den Mittelwert seines Elternknotens vorhersagt.
Im Falle zweier gleichwertiger Tests überlegen Sie sich, wie man diesen Konflikt lösen kann.

Aufgabe 1: Regressionsbäume (3)

TECHNISCHE UNIVERSITÄT DARMSTADT

Die Berechnung dieser Variante eines Regressionsbaumes erfolgt analog zur Berechnung eines mit ID3 erstellten Entscheidungbaumes. Wir bestimmen zuerst die Standardabweichung aller Trainingsdaten und die Standardabweichungen der möglichen Ausgänge eines Tests (analog zur Bestimmung der Entropien für die Berechnung des Maßes Information Gain). Zur Erinnerung: die Standardabweichung einer Menge von Werten $x_1, \cdots, x_k$ mit dem Mittelwert $\bar{x}$ berechnet sich wie folgt:

$$SD(x_1, \cdots, x_k) = \sqrt{\frac{1}{k-1} \sum_{i=1}^{k} (x_i - \bar{x})^2}$$

Die Werte in der Formel sind genau wie im Skript definiert.
Äquivalent zu Information Gain, wird SDR berechnet als:

$$SDR(S, A) = SD(S) - \sum_{i} \frac{|S_i|}{|S|} SD(S_i)$$

Die Berechung der Standard Deviation Reduction SDR werden wie in vorherigen Übungen nur für ausgewählte Knoten und jeweils nur für ein bestimmtes Testattribut ausführlich erläutern. Die übrigen Berechnungen werden wir nur tabellarisch angeben.

Wurzelknoten (1)

Die SDR des Attributes *A2* werden wir exemplarisch berechnen. Hierfür benötigen wir die SDs der Testausgänge *I*, *J* und *K*.
Mittelwert von *I*:

$$\bar{x}_I = \frac{5,5 + 5,1 + 5,7 + 0,76 + 1,03}{5} \approx 3,62.$$

entsprechend ist die Standardabweichung *SD*:

$$\begin{aligned} SD(S_I) &\approx \sqrt{\frac{1}{4}\left((5,5 - 3,62)^2 + (5,1 - 3,62)^2 + (5,7 - 3,62)^2 + (0,76 - 3,62)^2 + (1.03 - 3,62)^2\right)} \\ &\approx 2,5 \end{aligned}$$

Wurzelknoten (2)

Analog für die Testausgänge J und K

$$
\begin{aligned}
SD(S_J) &\approx 0,09 \\
SD(S_K) &\approx 0,33
\end{aligned}
$$

und für den gesamten Datensatz

$$SD(S) \approx 2,09$$

Somit haben wir alle zur Berechnung von $SDR(S, A2)$ benötigten Werte bestimmt:

$$
\begin{aligned}
SDR(S, A2) &= SD(S) - \frac{|S_I|}{|S|} \cdot SD(S_I) - \frac{|S_J|}{|S|} \cdot SD(S_J) - \frac{|S_K|}{|S|} \cdot SD(S_K) \\
&= 2,09 - \frac{5}{14} \cdot 2,5 - \frac{3}{14} \cdot 0,09 - \frac{6}{14} \cdot 0,33 \\
&\approx 1,04
\end{aligned}
$$

Wurzelknoten (3)

Entsprechend berechnen sich die SDRs der verbleibenden Tests:

Attribut	Wert	$\bar{x}_i$	$SD(S_i)$	$SDR(S,S_i)$
Verteilung		1,61	2,09	
A1	A	1,99	2,49	-0,163
	B	1,55	2,22	
	C	1,38	2,09	
A2	I	3,62	2,50	**1,038**
	J	0,40	0,09	
	K	0,55	0,33	
A3	R	3,57	2,57	1,036
	S	0,62	0,29	
	T	0,34	0,06	
A4	X	0,64	0,40	0,937
	Y	3,50	2,65	
	Z	0,47	0,22	

Der Tabelle entnehmen wir, daß der Test *A2* optimal ist. Da keiner seiner Testausgänge weniger als 3 Beispiele abdeckt, müssen alle drei Teilmengen weiter untersucht werden.

Zusatzaufgabe: Für Attribut A1 ist die SDR negativ, d.h. die durchschnittliche Standardabweichung ist in dem Kindknoten größer als in der Ursprungsmenge in der Wurzel. Kann dies auch für Informationgain für Klassifikationsaufgaben geschehen?

Wurzel $\rightarrow A2 = I$ (1)

TECHNISCHE UNIVERSITÄT DARMSTADT

Zuerst betrachten wir die Teilmenge für $A2 = I$:

Attribut	Wert	$\bar{x}_i$	$SD(S_i)$	$SDR(S, S_i)$
A2=I		3,62	2,50	
A1	A	5,7	0,00	0,005
	B	3,27	3,16	
	C	5,1	3,07	
A3	R	5,43	0,31	**2,237**
	S	0,9	0,19	
	T	n/a	n/a	
A4	X	1,03	0,00	0,617
	Y	4,27	2,35	
	Z	n/a	n/a	

Diesmal ist der Test $A3$ optimal.

$A3 = T$: Dabei treffen wir aber auf ein Problem: für den Testausgang T sind keine Werte bekannt. Dieses Problem lösen wir, indem wir dem Testausgang den Mittelwert der Instanzen des inneren Knoten zuweisen. Also wird für T ein Blatt mit Regressionswert 3,62 erzeugt.

$A3 = S$: Da für den Testausgang S weniger als drei Beispiele abgedeckt werden, können wir auch hier ein Blatt mit dem Wert 0,9 erstellen.

Wurzel $\rightarrow A2 = I \rightarrow A3 = R$: (1)

Für den Testausgang R müssen wir jedoch noch weitere Tests bestimmen:

Attribut	Wert	$\bar{x}_i$	$SD(S_i)$	$SDR(S, S_i)$
A2=I, A3=R		5,43	0,31	
A1	A	5,7	0,00	**0,306**
	B	5,5	0,00	
	C	5,1	0,00	
A4	X	n/a	n/a	0,000
	Y	5,43	0,00	
	Z	n/a	n/a	

Der optimale Test $A1$ stellt eine perfekte Trennung der verbleibenden Beispiele dar. Aus diesem Grund sind keine weiteren Tests nötig.

Wurzel $\rightarrow$ *A2* = *J* (1)

TECHNISCHE UNIVERSITÄT DARMSTADT

Zurück am Wurzelknoten betrachten wir Testausgang *J*:

Attribut	Wert	$\bar{x}_i$	$SD(S_i)$	$SDR(S, S_i)$
A2=J		0,4	0,09	
A1	A	0,35	0	**0,087**
	B	0,5	0	
	C	0,35	0	
A3	R	n/a	n/a	0,016
	S	0,43	0,11	
	T	0,35	0	
A4	X	0,5	0	0,087
	Y	n/a	n/a	
	Z	0,35	0	

Hierbei sind sowohl Test *A1* als auch *A4* mit 0,087 optimal.

Entscheidungskriterium: Als mögliches weiteres Entscheidungskriterium bietet sich hier die Anzahl der *besetzten* Knoten ab (also Knoten, die Beispiele abdecken), da die Vorhersage leerer Blätter
1) auf eine größere Anzahl Beispiele als die der anderen Blätter beruht und
2) keines dieser Beispiele auf den Test passt.

Beides erhöht die Varianz bzw. den erwarteten Fehler des Baumes.

Dementsprechend wählen wir Test *A1* aus. Da die drei abgedeckten Beispiele auf die 3 Testausgänge verteilt werden, werden keine weiteren Tests benötigt.

Wurzel $\rightarrow$ $A2 = K$ (1)

Berechnen wir nun den letzten Testausgang des Wurzelknotens K:

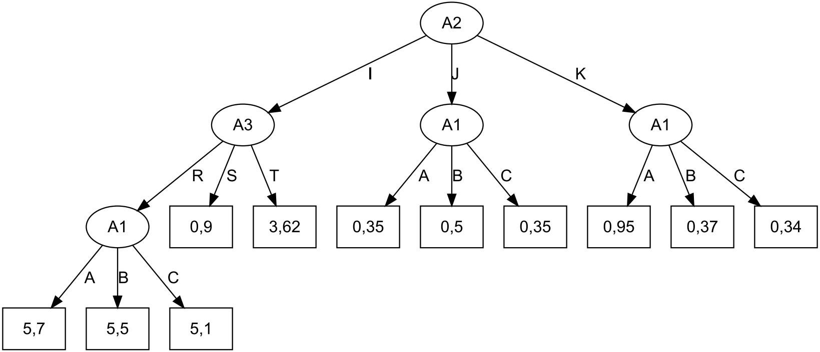
Attribut	Wert	$\bar{x}_i$	$SD(S_i)$	$SDR(S, S_i)$
A2=K		0,55	0,33	
A1	A	0,95	0,21	**0,191**
	B	0,37	0,13	
	C	0,34	0,08	
A3	R	0,78	0,45	0,030
	S	0,54	0,37	
	T	0,34	0,08	
A4	X	0,55	0,47	0,068
	Y	0,46	0	
	Z	0,6	0,08	

Wiederum ist der Test *A1* der optimale Test.
Diesmal erhält jeder Testausgang 2 Instanzen, demnach sind keine weiteren Tests nötig.

Abbildung Baum (1)

b) Zeichnen Sie den eben erzeugten Baum.

Lösung:



Fehler (1)

c) Bestimmen Sie den Mean-Squared-Error des Baumes

- auf den Trainingsdaten
- auf den folgenden Testdaten:

A1	A2	A3	A4	Value
B	J	T	Z	0.51
C	K	R	Y	1.90
B	J	R	X	0.90
A	J	S	Y	0.47
A	K	T	Z	0.54

Beachten Sie bitte, daß sich der Mean-Squared-Error folgendermaßen berechnet:

$$MSE = \frac{1}{n} \sum_{j=1}^{n} (y_j - r_j)^2$$

Fehler (2)

TECHNISCHE UNIVERSITÄT DARMSTADT

Lösung:

Auf den Trainingsdaten:

A1	A2	A3	A4	Value	Tree	$(Value - Reg)^2$
C	K	T	X	0,28	0,34	0,0036
B	J	S	X	0,50	0,5	0,0000
C	J	S	Z	0,35	0,35	0,0000
B	I	R	Y	5,50	5,5	0,0000
A	J	T	Z	0,35	0,35	0,0000
A	K	S	Z	0,80	0,95	0,0225
C	I	R	Y	5,10	5,1	0,0000
A	I	R	Y	5,70	5,7	0,0000
C	I	S	Y	0,76	0,9	0,0196
B	I	S	X	1,03	0,9	0,0169
B	K	R	Y	0,46	0,37	0,0081
C	K	T	Z	0,39	0,34	0,0025
B	K	S	X	0,28	0,37	0,0081
A	K	T	X	1,10	0,95	0,0225
$MSE = \sum (Value - Tree)^2 / 14$						0,0074

Fehler (3)

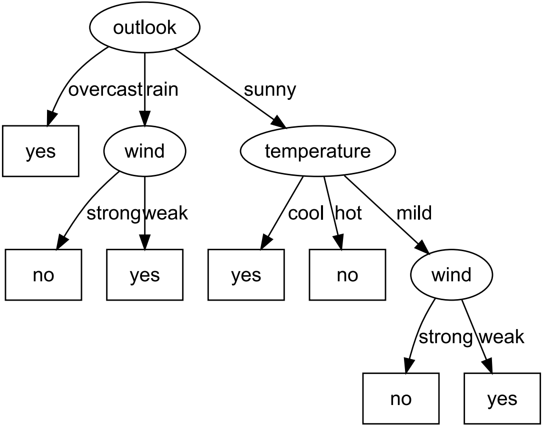
TECHNISCHE UNIVERSITÄT DARMSTADT

Auf den Testdaten:

A1	A2	A3	A4	Value	Tree	$(Value - Reg)^2$
B	J	T	Z	0,51	0,50	0,0001
C	K	R	Y	1,90	0,34	2,4336
B	J	R	X	0,90	0,50	0,1600
A	J	S	Y	0,47	0,35	0,0144
A	K	T	Z	0,54	0,95	0,1681
$MSE = \sum (Value - Tree)^2/5$						0,5552

Aufgabe 2: Reduced Error Pruning (1)

Gegeben sei folgender Entscheidungsbaum,



Aufgabe 2: Reduced Error Pruning (2)

TECHNISCHE UNIVERSITÄT DARMSTADT

der auf der Trainingsmenge

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No
D15	Sunny	Mild	Normal	Weak	No

gelernt wurde,

Aufgabe 2: Reduced Error Pruning (3)

und außerdem folgende Pruning-Menge (Validierungsmenge):

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D16	Sunny	Mild	High	Strong	No
D17	Rain	Hot	Normal	Weak	Yes
D18	Overcast	Cool	High	Strong	No
D19	Overcast	Mild	Normal	Strong	Yes
D20	Sunny	Cool	High	Strong	No

Wenden Sie Reduced-Error Pruning (Foliensatz *Entscheidungsbaum-Lernen*, Folie "*Reduced Error Pruning*") auf den Entscheidungsbaum an. Benutzen Sie als Evaluierungsmaß die Anzahl der korrekt klassifizierten Beispiele auf der Pruning-Menge.

Lösung: Beim Reduced-Error Pruning ersetzt man sukzessive Knoten durch Blätter, die dann die Majority-Klasse anhand der Trainingsmenge im jeweiligen Knoten vorhersagen.

Dieser Vorgang wiederholt sich so lange, bis keine Verbesserung auf dem Pruning-Set mehr erreicht wird, wobei mit Verbesserung auch Gleichheit der Genauigkeit gemeint ist (also $\geq$ ursprüngliche Genauigkeit), da sonst nicht die kleinste Version des Baumes erzeugt werden würde.

Die Genauigkeit des ursprünglichen Baumes liegt bei $^3/_5$.

Überprüfung der Knoten (1)

TECHNISCHE UNIVERSITÄT DARMSTADT

Wurzelknoten *outlook*:
Die Majority-Klasse ist “yes” (9 mal “yes” und 6 mal “no”). Sagt man also immer “yes” vorher, erreicht man auf der Pruning-Menge eine Genauigkeit von $2/5 < 3/5$.

outlook $\rightarrow$ *wind*:
Als nächstes testen wir das Ersetzen des Knotens *wind*, bei welchem “yes” die Majority-Klasse ist. Mit dieser Ersetzung würde der Baum eine Genauigkeit von $3/5$ erreichen.

outlook $\rightarrow$ *temperature*:
Der nachfolgende Test schneidet den Knoten *temperature* ab, in welchem “no” die Majority-Klasse ist. Es ergibt sich eine Genauigkeit von $4/5 \geq 3/5$.

outlook $\rightarrow$ *temperature* $\rightarrow$ *wind*:
Als letztes wird noch im rechten Teilbaum der Knoten *wind* getestet. Das Blatt sagt “no” vorher und der Baum erreicht eine Genauigkeit von $3/5$.

Prunen (1)

TECHNISCHE UNIVERSITÄT DARMSTADT

Nun suchen wir den kleinsten Baum, der mindestens eine Genauigkeit von $3/5$ aufweist (Zeile 'as long as the error on the pruning set does not increase'). Dafür gehen wir den Baum von den Blättern aus durch, z.B. durch eine Tiefensuche.

outlook $\rightarrow$ *wind*:
Wir können ohne einen Verlust an Genauigkeit den Knoten *wind* in ein Blatt verwandeln, das die Klasse "yes" vorhersagt.

outlook $\rightarrow$ *temperature* $\rightarrow$ *wind*:
Wir können den Knoten abschneiden und stattdessen "no" mit einer Genauigkeiten von $3/5$ vorhersagen.
outlook $\rightarrow$ *temperature*:
Es gilt auch nach dem Prunen von *outlook* $\rightarrow$ *temperature* $\rightarrow$ *wind* $4/5 \geq 3/5$. Also wird der Knoten zu einem Blatt, welches die Klasse "no" vorhersagt, umgewandelt.

Wurzelknoten *outlook*:
Es gibt keine Verbesserung der Genauigkeit, deshalb wird der Wurzelknoten nicht geprunt. Dies ist aber generell möglich.

Zusatzaufgabe: Ist auch eine andere Traversierung des Baumes möglich? Kommt dabei der gleiche Baum heraus?

Geprunter Baum (1)

Der geprunte Baum hat eine Genauigkeit von $^4/_5$ und sieht dann wie folgt aus: