Data Mining und Maschinelles Lernen

TECHNISCHE UNIVERSITÄT DARMSTADT

10. Übungsblatt

Aufgabe 1 Regressionsbäume

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

a) 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.

b) Zeichnen Sie den eben erzeugten Baum.

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

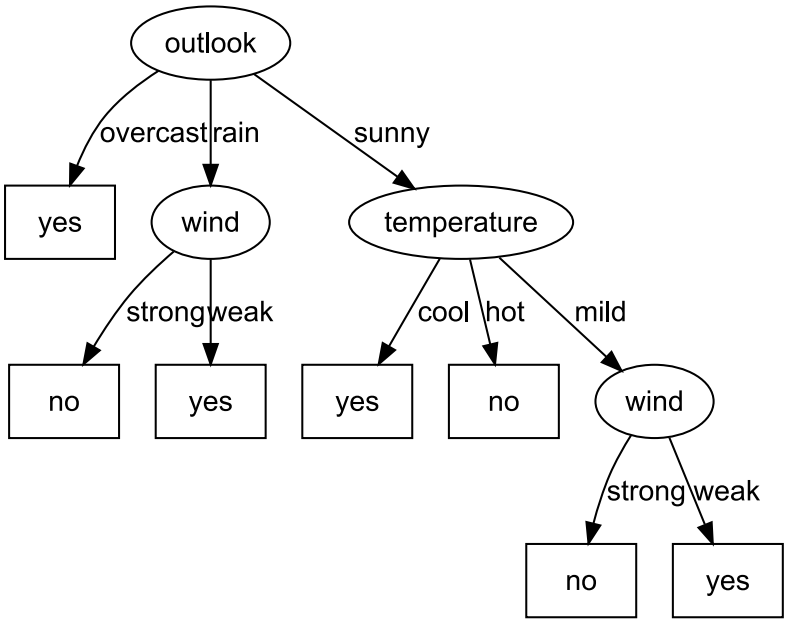
Beachten Sie bitte, daß sich der Mean-Squared-Error folgendermaßen berechnet:

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

Die Werte in der Formel sind genau wie im Skript definiert.

Aufgabe 2 Reduced Error Pruning

Gegeben sei folgender Entscheidungsbaum,



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, 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.