

# Maschinelles Lernen: Symbolische Ansätze



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

Wintersemester 2009/2010  
Musterlösung für das 8. Übungsblatt

## Aufgabe 1: RISE

Gegeben sei folgende Beispielmenge:

Outlook	Temperature	Humidity	Wind	PlayTennis
Overcast	Hot	High	False	Yes
Rainy	Cool	Normal	True	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	High	False	No
Overcast	Mild	High	True	Yes
Sunny	Cool	Normal	False	Yes
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	True	No
Sunny	Hot	High	False	No
Sunny	Mild	Normal	True	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Hot	High	True	No
Rainy	Mild	High	False	Yes
Overcast	Hot	Normal	False	Yes

- a) Wenden Sie den Algorithmus RISE (Foliensatz Instance-based Learning, Folie 42) auf den obigen Datensatz an. Berechnen Sie nur die erste Iteration und diese nur für die negativen Beispiele.

Bedenken Sie, dass die Accuracy berechnet wird, indem man für jede Regel alle Beispiele durchgeht und jeweils das nächste Beispiel nimmt, das nicht schon ohnehin abgedeckt wird. Dies ist notwendig, da sonst alle Beispiele immer mit den Regeln klassifiziert werden würden, die aus ihnen erstellt wurden und sich folglich die Accuracy nie verbessern könnte.

Benutzen Sie zur Berechnung der Distanz der Attribute die *Value Difference Metric* und nehmen Sie hierbei an, dass  $k = 1$  gilt. Als Vereinfachung müssen Sie die *VDM* nicht normieren.

Zur Bestimmung des Abstandes eines Beispiels zu einer Regel verwenden Sie die euklidische Distanz wobei Sie das Ziehen der Wurzel weglassen können.

Müssen Sie 2 Regeln zusammenfassen, so ist es Voraussetzung, dass diese die gleiche Klasse vorhersagen.

**Lösung:** Als erstes berechnen wir die *VDM*:

	outlook				temperature				humidity			wind	
	sunny	overcast	rainy		cool	hot	mild		high	normal		false	true
yes	2	4	3	yes	3	2	4	yes	3	6	yes	6	3
no	3	0	2	no	1	2	2	no	4	1	no	2	3

Daraus folgen dann die Abstände, die bereits hier quadriert werden:

- für das Attribut *outlook*

$$d(\text{sunny}, \text{overcast}) = \left| \frac{2}{5} - 1 \right| + \left| \frac{3}{5} - 0 \right| = 1,2 \rightarrow 1,2^2 = 1,44$$

$$d(\text{sunny}, \text{rainy}) = \left| \frac{2}{5} - \frac{3}{5} \right| + \left| \frac{3}{5} - \frac{2}{5} \right| = 0,4 \rightarrow 0,4^2 = 0,16$$

$$d(\text{overcast}, \text{rainy}) = \left| 1 - \frac{3}{5} \right| + \left| 0 - \frac{2}{5} \right| = 0,8 \rightarrow 0,8^2 = 0,64$$

- für das Attribut *temperature*

$$d(\text{cool}, \text{hot}) = \left| \frac{3}{4} - \frac{1}{2} \right| + \left| \frac{1}{4} - \frac{1}{2} \right| = 0,5 \rightarrow 0,5^2 = 0,25$$

$$d(\text{cool}, \text{mild}) = \left| \frac{3}{4} - \frac{2}{3} \right| + \left| \frac{1}{4} - \frac{1}{3} \right| = 0,5 \rightarrow 0,5^2 = 0,25$$

$$d(\text{hot}, \text{mild}) = \left| \frac{1}{2} - \frac{2}{3} \right| + \left| \frac{1}{2} - \frac{1}{3} \right| = \frac{1}{3} \rightarrow \frac{1}{3}^2 = \frac{1}{9}$$

- für das Attribut *humidity*

$$d(\text{high}, \text{normal}) = \left| \frac{3}{7} - \frac{6}{7} \right| + \left| \frac{4}{7} - \frac{1}{7} \right| = \frac{6}{7} \rightarrow \frac{6}{7}^2 = \frac{36}{49}$$

- für das Attribut *wind*

$$d(\text{false}, \text{true}) = \left| \frac{3}{4} - \frac{1}{2} \right| + \left| \frac{1}{4} - \frac{1}{2} \right| = 0,5 \rightarrow 0,5^2 = 0,25$$

Nun müssen wir die initiale Accuracy ausrechnen. Dazu müssen wir herausfinden von welcher Regel jedes Beispiel klassifiziert wird und wie es klassifiziert wird. Hierzu müssen wir die Distanzen berechnen:

Als erstes ordnen wir die Beispiele, indem wir zuerst alle negativen und dann alle positiven aufführen und benennen diese mit dem Buchstaben N für "No" sowie dem Index des Beispiels (Y für "Yes"):

Name	Outlook	Temperature	Humidity	Wind	PlayTennis
N1	Rainy	Cool	Normal	True	No
N2	Sunny	Mild	High	False	No
N3	Rainy	Mild	High	True	No
N4	Sunny	Hot	High	False	No
N5	Sunny	Hot	High	True	No
Y1	Rainy	Mild	High	False	Yes
Y2	Sunny	Mild	Normal	True	Yes
Y3	Rainy	Mild	Normal	False	Yes
Y4	Overcast	Hot	Normal	False	Yes
Y5	Overcast	Mild	High	True	Yes
Y6	Overcast	Cool	Normal	True	Yes
Y7	Overcast	Hot	High	False	Yes
Y8	Sunny	Cool	Normal	False	Yes
Y9	Rainy	Cool	Normal	False	Yes

Dann finden wir für jedes Beispiel die Regel, die es klassifiziert (also deren Manhattan Distanz am geringsten ist). Wir zeigen dies beispielhaft für das erste Beispiel:

zu klass. Beispiel	Regel	Outlook	Temperature	Humidity	Rainy	Distance
<b>N1</b>		<b>Rainy</b>	<b>Cool</b>	<b>Normal</b>	<b>True</b>	<b>0</b>
	N2	0,16	0,25	$\frac{36}{49}$	0,25	$0,16+0,25+\frac{36}{49}+0,25=1,39$
	N3	0	0,25	$\frac{36}{49}$	0	0,98
	N4	0,16	0,25	$\frac{36}{49}$	0,25	1,39
	N5	0,16	0,25	$\frac{36}{49}$	0	1,14
	Y1	0	0,25	$\frac{36}{49}$	0,25	1,23
	Y2	0,16	0,25	0	0	0,41
	Y3	0	0,25	0	0,25	0,5
	Y4	1,44	0,25	0	0,25	1,94
	Y5	1,44	0,25	$\frac{36}{49}$	0	2,42
	Y6	1,44	0	0	0	1,44
	Y7	1,44	0,25	$\frac{36}{49}$	0,25	2,67
	Y8	0,16	0	0	0,25	0,41
	<b>Y9</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0,25</b>	<b>0,25</b>

Wie man in der Tabelle sehen kann hat die Regel Y9 die geringste Distanz. Daher würde das Beispiel mit dieser Regel klassifiziert werden als Yes. Nachfolgend sind die weiteren Klassifizierungen aufgeführt:

zu klass. Beispiel	Regel	Distanz	Vorhersage
N2	N4	0,11	No
N3	Y1	0,25	Yes
N4	N2	0,11	No
N5	N4	0,25	No
Y1	N2	0,16	No
Y2	N1,Y3	0,41	Yes
Y3	Y9	0,25	Yes
Y4	Y6	0,5	Yes
Y5	Y7	0,36	Yes
Y6	Y4	0,5	Yes
Y7	Y5	0,36	Yes
Y8	Y9	0,16	Yes
Y9	Y8	0,16	Yes

Zu beachten ist hier, dass man bereits berechnete Distanzen wiederverwenden kann und so z.B. die Distanzen zu N1 bis N3 für das Beispiel N4 nicht berechnen braucht.

Bei Beispiel Y2 kommt es zu einem Tie. Hier nimmt man dann z.B. die Majority Class (also "Yes").

Nun sieht man, dass folgende Beispiel falsch klassifiziert werden: N1 (ist "No", klassifiziert als "Yes"), N3 (ist "No", klassifiziert als "Yes"), Y1 (ist "Yes", klassifiziert als "No"). Daher ist die initiale Accuracy  $\frac{11}{14} \approx 0,79$ .

Als nächstes beginnen wir mit dem ersten Schritt des Algorithmus. Dazu konvertieren wir alle Beispiele in Regeln. Dann gehen wir alle (negativen) Regeln durch und suchen jeweils das nicht abgedeckte Beispiel mit dem geringsten Abstand.

Für die Regel (das Beispiel) N1:

Regel	Beispiel	Distanz
N1	N2	1,39
N1	N3	0,98
N1	N4	1,39
N1	N5	1,14

Wir generalisieren damit N1 und N3 minimal, was zu folgender Regel N13 führt:

Outlook	Temperature	Humidity	Windy	PlayTennis
Rainy	?	?	True	No

Nun müssen wir überprüfen, ob sich die Accuracy nicht verschlechtert hat. Dazu müssen wir die Abstände zu N13 neu berechnen und überprüfen, ob sich die kürzeste Distanz verändert hat:

Beispiel	neue Distanz	momentan kürzester Abstand	beste Regel
N1	0	0,16	N13 (N5)
N2	0,41	0,11	N4
N3	0	0,25	N13 (N4)
N4	0,41	0,11	N2
N5	0,16	0,25	N13 (N4)
Y1	0,25	0,16	N2
Y2	0,16	0,41	N13 (Y3)
Y3	0,25	0,25	Y9
Y4	0,89	0,5	Y6
Y5	0,64	0,36	Y7
Y6	0,64	0,5	Y4
Y7	0,89	0,36	Y5
Y8	0,41	0,16	Y9
Y9	0,25	0,16	Y8

Wie man sehen kann ändern sich die Abstände zu den Beispielen N1, N3, N5 und Y2 (hier ist die vorherige beste Regel in Klammern dargestellt). N1 und N2 werden nun richtig klassifiziert, bei N5 ändert sich nichts, während Y2 nun falsch klassifiziert wird. Daher wird die Accuracy etwas besser und beträgt nun  $\frac{12}{14} \approx 0,86$ . Aus diesem Grund ist die Generalisierung zulässig und wird durchgeführt.

Für die Regel (das Beispiel) N2:

Als erstes müssen nun wir uns die Distanzen ansehen:

Regel	Beispiel	Distanz
N2	N1	1,39
N2	N3	0,41
N2	N4	0,11
N2	N5	0,36

Wir generalisieren damit N2 und N4 minimal, was zu folgender Regel N24 führt:

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	?	High	False	No

Nun müssen wir überprüfen, ob sich die Accuracy nicht verschlechtert hat. Dazu müssen wir die Abstände zu N24 neu berechnen und überprüfen, ob sich die kürzeste Distanz verändert hat:

Beispiel	neue Distanz	momentan kürzester Abstand	Regel
N1	1,14	0	N13
N2	0	0,11	N24 (N4)
N3	0,41	0	N13
N4	0	0,11	N24 (N2)
N5	0,16	0,16	N13
Y1	0,16	0,16	N24
Y2	0,98	0,41	N13
Y3	0,89	0,25	Y9
Y4	2,17	0,5	Y6
Y5	1,69	0,36	Y7
Y6	2,42	0,5	Y4
Y7	1,44	0,36	Y5
Y8	0,73	0,16	Y9
Y9	0,89	0,16	Y8

Es ändert sich nichts an der Accuracy, da kein Beispiel nun anders klassifiziert wird. Daher ist auch diese Generalisierung zulässig und wird hinzugefügt.

Für die Regel (das Beispiel) N3:

Wir sehen uns wieder die Distanzen an:

Regel	Beispiel	Distanz
N3	N1	0,98
N3	N2	0,41
N3	N4	0,52
N3	N5	0,27

Wir generalisieren damit N3 und N5 minimal, was zu folgender Regel N35 führt:

Outlook	Temperature	Humidity	Windy	PlayTennis
?	?	High	True	No

Nun müssen wir überprüfen, ob sich die Accuracy nicht verschlechtert hat. Dazu müssen wir die Abstände zu N35 neu berechnen und überprüfen, ob sich die kürzeste Distanz verändert hat:

Beispiel	neue Distanz	momentan kürzester Abstand	beste Regel
N1	0,73	0	N13
N2	0,25	0	N24
N3	0	0	N13
N4	0,25	0	N24
N5	0	0,25	N13
Y1	0,25	0,16	N24
Y2	0,73	0,16	N13
Y3	0,98	0,25	Y9
Y4	0,98	0,5	Y6
Y5	0	0,36	N35
Y6	0,73	0,5	Y4
Y7	0,25	0,36	N35
Y8	0,98	0,16	Y9
Y9	0,98	0,16	Y8

Bei N3 und N5 ändert sich nichts, aber Y5 und Y7 werden nun (fehlerhaft) negativ klassifiziert. Daher verschlechtert sich die Accuracy und wir fügen diese Generalisierung nicht hinzu.

Für die Regel (das Beispiel) N4:

Die Distanzen sind wie folgt:

Regel	Beispiel	Distanz
N4	N1	1,39
N4	N2	0,11
N4	N3	0,52
N4	N5	0,25

N4 und N2 werden minimal generalisiert, was zu folgender Regel N42 führt:

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	?	High	False	No

Da diese Regel bereits in der Liste enthalten ist, wird sie nicht hinzugefügt (im Algorithmus im Foliensatz Instance-based Learning, Folie 39 in Zeile 2. iv).

Für die Regel (das Beispiel) N5:

Die Distanzen sind wie folgt:

Regel	Beispiel	Distanz
N5	N1	1,14
N5	N2	0,36
N5	N3	0,27
N5	N4	0,25

N5 und N4 werden minimal generalisiert, was zu folgender Regel N54 führt:

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	?	No

Nun müssen wir überprüfen, ob sich die Accuracy nicht verschlechtert hat. Dazu müssen wir die Abstände zu N54 neu berechnen und überprüfen, ob sich die kürzeste Distanz verändert hat:

Beispiel	neue Distanz	momentan kürzester Abstand	beste Regel
N1	0,14	0	N13
N2	0,11	0	N24
N3	0,27	0	N13
N4	0	0	N24
N5	0	0,16	N13
Y1	0,27	0,16	N24
Y2	0,85	0,16	N13
Y3	1,01	0,25	Y9
Y4	2,17	0,5	Y6
Y5	1,55	0,36	Y7
Y6	2,42	0,5	Y4
Y7	1,44	0,36	Y5
Y8	0,98	0,16	Y9
Y9	1,14	0,16	Y8

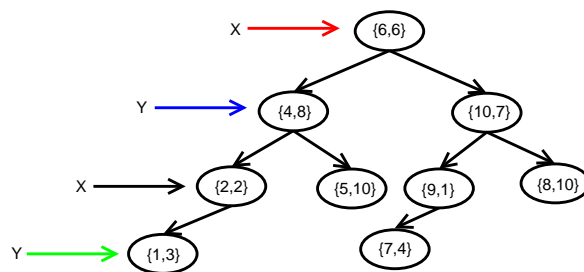
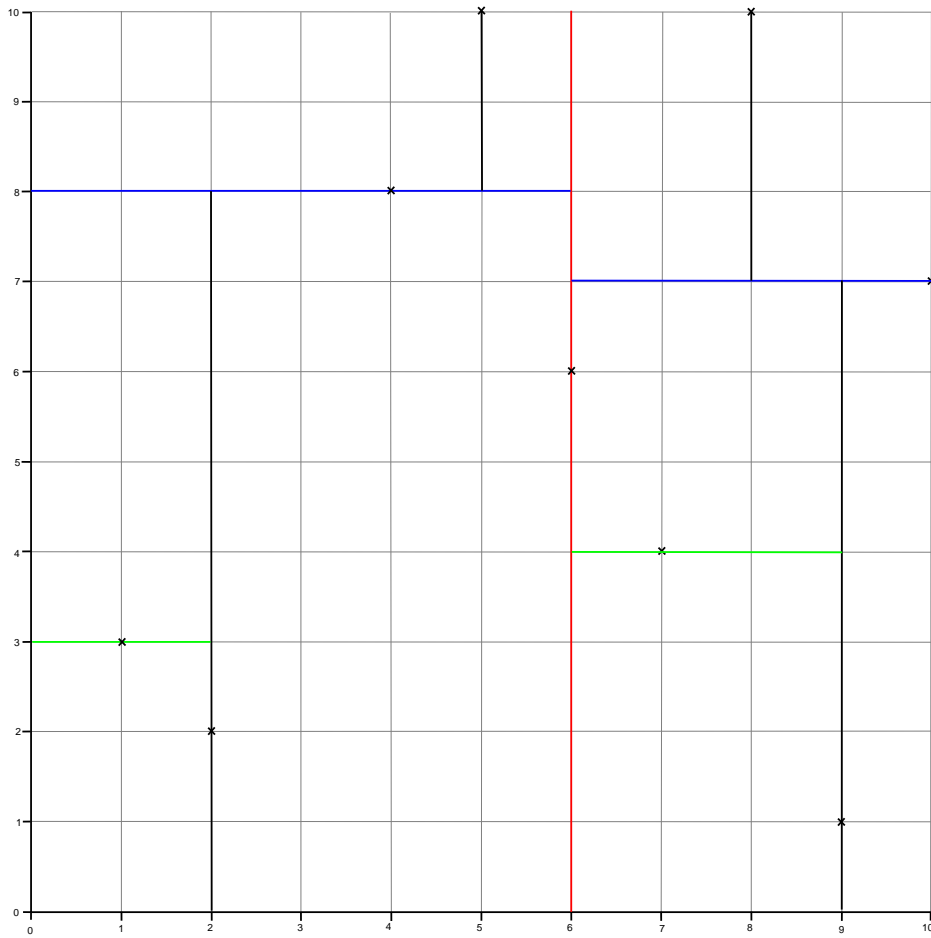
Nur bei N5 könnte sich etwas ändern, hier war aber die Klassifikation bereits richtig. Daher ändert sich an der Accuracy nichts, die Generalisierung wird hinzugefügt und wir erhalten folgende Regelmenge:

Name	Outlook	Temperature	Humidity	Windy	PlayTennis
N13	Rainy	?	?	True	No
N24	Sunny	?	High	False	No
N3	Rainy	Mild	High	True	No
N54	Sunny	Hot	High	?	No
Y1	Rainy	Mild	High	False	Yes
Y2	Sunny	Mild	Normal	True	Yes
Y3	Rainy	Mild	Normal	False	Yes
Y4	Overcast	Hot	Normal	False	Yes
Y5	Overcast	Mild	High	True	Yes
Y6	Overcast	Cool	Normal	True	Yes
Y7	Overcast	Hot	High	False	Yes
Y8	Sunny	Cool	Normal	False	Yes
Y9	Rainy	Cool	Normal	False	Yes

## Aufgabe 2: KD-Trees

- a) Bauen Sie einen KD-Tree aus der folgenden 2D Punktmenge auf und zeichnen Sie sowohl den Baum als auch die grafische Lösung im 2D-Raum:  
 $\{\{4, 8\}, \{7, 4\}, \{5, 10\}, \{1, 3\}, \{2, 2\}, \{9, 1\}, \{10, 7\}, \{8, 10\}, \{6, 6\}\}.$

**Lösung:** Zuerst wird nach der Variablen  $x$  geschaut. Daher müssen die Punkte aufsteigend nach ihrer  $x$ -Koordinate sortiert werden:  $X = \{\{1, 3\}, \{2, 2\}, \{4, 8\}, \{5, 10\}, \{6, 6\}, \{7, 4\}, \{8, 10\}, \{9, 1\}, \{10, 7\}\}.$  Dann macht man die erste Trennung beim Median, also beim Punkt  $\{6, 6\}$  (in der Grafik entspricht das der roten Linie). Als nächstes trennt man nach der Variablen  $y$ . Man sortiert wieder beide Mengen:  $Y_1 = \{\{2, 2\}, \{1, 3\}, \{4, 8\}, \{5, 10\}\},$   
 $Y_2 = \{\{9, 1\}, \{7, 4\}, \{10, 7\}, \{8, 10\}\}.$  Nun nimmt man wieder die Mediane der beiden Menge als Trennlinien (in der Grafik blau). Da es sich um eine gerade Anzahl handelt entschließen wir uns für den rechten Wert. Danach teilt man wieder nach  $x$  auf:  $X_1 = \{\{1, 3\}, \{2, 2\}\}, X_2 = \{\{5, 10\}\}, X_3 = \{\{7, 4\}, \{9, 1\}\}$  und  $X_4 = \{\{8, 10\}\}$  (schwarze Linien). Dann muss man nur noch die beiden verbleibenden Punkte  $\{1, 3\}$  und  $\{7, 4\}$  als Trennlinien für  $y$  nehmen (grüne Linien).



- b) Wenden Sie 1-NN für die folgenden beiden Queries  $\{7, 9\}$  und  $\{1, 1\}$  auf den Baum an und geben Sie die genaue Traversierung des Baumes an.

**Lösung:** Der Algorithmus verfährt so wie wenn er die neue Instanz in den Baum einordnen würde. In einem Blatt bildet er dann einen Kreis um die zu klassifizierende Instanz der den Radius des Abstands zwischen dieser und der aktuellen Instanz hat. Er setzt die aktuelle Instanz als beste Instanz und schaut nach, ob der Kreis andere Trennlinien schneidet. Ist dies so, so muss nochmal in dem anderen Blatt nachgeschaut werden (genau so wie beim Start des Algorithmus). Ist dies nicht so, dann geht man im Baum eine Ebene weiter hoch und braucht den kompletten anderen Teilbaum nicht mehr zu betrachten.

Instanz  $\{7, 9\}$ :

Man landet im Blatt  $\{8, 10\}$ . Der Kreis schneidet die Trennlinie der Instanz  $\{6, 6\}$ . Mögliche Kindknoten von  $\{8, 10\}$  müssen nicht durchsucht werden. Dann geht man eine Ebene weiter hoch im Baum zu  $\{10, 7\}$ . Diese Distanz ist größer, daher braucht man die Kinder ( $\{9, 1\}$  und  $\{7, 4\}$ ) nicht zu durchsuchen. Nun geht man wieder eine Ebene höher in die Wurzel und sucht den Quadranten in dem der Schnitt vorliegt ( $\{5, 10\}$ ). Die möglichen Kinder müssen nicht durchsucht werden, da die Distanz größer ist. Danach geht man zum Knoten  $\{4, 8\}$  dessen Distanz ebenfalls größer ist. Daher muss der linke Teilbaum nicht durchsucht werden. Nun geht man wieder eine Ebene höher und findet heraus, dass die Wurzel ebenfalls eine größere Distanz hat. Daher ist der Knoten mit der geringsten Distanz  $\{8, 10\}$ .

---

Instanz  $\{1, 1\}$ :

Man landet im Blatt  $\{1, 3\}$ . Der Kreis schneidet die Linie von der Instanz  $\{2, 2\}$  und die Distanz zu diesem Knoten ist kleiner. Daher wird dieser Knoten der neue beste Knoten. Man geht zu  $\{4, 8\}$  und merkt, dass die Distanz größer ist. Daher schneidet man die Kindknoten ab und geht zu  $\{6, 6\}$ , dessen Distanz auch größer ist. Der rechte Teilbaum wird abgeschnitten und der Knoten mit der geringsten Distanz ist  $\{2, 2\}$ .