Maschinelles Lernen: Symbolische Ansätze

TECHNISCHE UNIVERSITÄT DARMSTADT

Wintersemester 2009/2010
Musterlösung für das 7. Übungsblatt

Aufgabe 1: Nearest Neighbour

Gegeben sei folgende Beispielmenge:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	26	High		No
D2	Sunny	28	High	Strong	No
D3	Overcast	29	High	Weak	Yes
D4	Rain	23	High	Weak	Yes
D5	Rain		Normal	Weak	Yes
D6	Rain	12	Normal	Strong	No
D7	Overcast	8		Strong	Yes
D8	Sunny	25	High	Weak	No
D9	Sunny	18	Normal	Weak	Yes
D10	Rain	20	Normal	Weak	Yes
D11	Sunny	20	Normal	Strong	
D12	Overcast	21	High	Strong	Yes
D13		26	Normal	Weak	Yes
D14	Rain	24	High	Strong	No
D15	Sunny	23	Normal	Weak	No
D16	Sunny	21	Normal	Weak	Yes

a) Um fehlende Werte zu behandeln, kann man diese einfach auffüllen, indem man die am naheliegendsten Nachbarn zu diesem Beispiel verwendet. Benutzen Sie hierfür 3-NN zum Ausfüllen dieser Werte. Normieren Sie das numerische Attribut, wie im Skript beschrieben (Foliensatz Instance-based Learning, Folie 15), nehmen Sie für nominale Attribute die 0/1-Distanz (Foliensatz Instance-based Learning, Folie 16) und benutzen Sie als Distanzfunktion für 3-NN die Manhattan Distanz.

Beziehen Sie hier die Klassifikation mit ein oder nicht? Warum?

Lösung: Wir betrachten beim Ausfüllen der fehlenden Werte eines Beispiels dessen eigene Klasse. Das heißt, wir suchen die k nächsten Nachbarn des Beispieles, die zu dessen Klasse gehören. Damit ist gewährleistet, daß keine Eigenschaften einer anderen Klasse übernommen werden. Wie in der Übung bereits diskutiert geht es hier darum, dass jede Klasse eigene Eigenschaften (also Attribut-Wert-Paare) aufweist, die für die jeweilige Klasse charakterisierend sind. Würde man nun fehlende Werte auffüllen, indem man Beispiele, die zu einer anderen Klasse gehören, verwendet, würde man Eigenschaften dieser Klasse verwenden, die nicht für die Klasse charakterisierend sind, zu der das Beispiel gehört.

Für numerische Attribute verwenden wir die im Skript angebene normierte Abstandsfunktion, da deren Funktionswerte auch im Intervall $[0,1]$ liegen und deshalb besser mit den Abständen bei nominalen Attributen vergleichbar sind. Im Augenblick liegen die Werte des Attributes Temperatur zwischen 8 und 29, damit ergibt sich eine maximale Differenz von 21. Sollten später weitere Instanzen hinzukommen, deren Attributwerte nicht in dem Intervall $[8,21]$ liegen, müssen die Grenzen und die maximale Differenz angepaßt werden, da der normierte Wert ansonsten auch größer als 1 werden kann.

Nichtsdestotrotz ergibt sich für die vorliegenden Daten die folgende Abstandsfunktion:

$$d(a_1, a_2) = \frac{|a_1 - a_2|}{21}$$

Für nominale Attibute setzen wir die 0/1 Distanz ein. Zur Erinnerung, diese sieht für zwei nominale Attribute a_1 und a_2 wie folgt aus:

$$d(a_1, a_2) = \begin{cases} 0, & \text{falls } a_1 = a_2 \\ 1, & \text{sonst} \end{cases}$$

Die endgültige Distanzfunktion ergibt sich dann aus der Summe der Distanzen aller Attribute (Manhatten Distanz).

D1: Betrachten wir nun die Beispiele, die zum Ausfüllen der Werte des Beispiels D1 benötigt werden, und berechnen deren Abstand zu D1:

Day	Outlook	Temperature	Humidity	Wind	Abstand
D2	Sunny	28	High	**Strong**	**2/21**
D6	Rain	12	Normal	Strong	2 14/21
D8	Sunny	25	High	**Weak**	**1/21**
D14	Rain	24	High	**Strong**	**1 2/21**
D15	Sunny	23	Normal	Weak	1 3/21

Die Beispiele D2, D8 und D14 sind die nächsten Beispiele. Wir erhalten zweimal *Strong* und einmal *Weak* und setzen deshalb den fehlenden Wert von D1 auf *Strong*.

D5: Diesmal müssen wir einen numerischen Wert auffüllen. Hierfür bestimmen wir wiederum die k nächsten Nachbarn und berechnen die mittlere Distanz ihrer Attributwerte.

Day	Outlook	Temperature	Humidity	Wind	Abstand
D3	Overcast	29	High	Weak	2
D4	Rain	23	High	**Weak**	**1**
D7	Overcast	8		Strong	3
D9	Sunny	18	Normal	**Weak**	**1**
D10	Rain	20	Normal	**Weak**	**0**
D12	Overcast	21	High	Strong	3
D13		26	Normal	**Weak**	**1**
D16	Sunny	21	Normal	**Weak**	**1**

Wie man sieht treten bei D5 zwei Probleme auf. Zum einen fehlen den Instanzen D7 und D13 auch Attributwerte, wir treffen hier die Annahme, daß sich diese (fehlenden) Attributwerte von dem von D5 unterscheiden (Abstand 1). Zum anderen können wir nicht genau 3 nächste Nachbarn bestimmen, da 4 Beispiele den Abstand 1 haben. Wir können nun einfach den Mittelwert der Attributwerte der fünf nächsten Beispiele verwenden oder einfach 3 zufällig auswählen. Wir entscheiden uns für die erste Variante, bei der wir anschließend auf die nächste ganze Zahl abrunden:

$$\left\lfloor \frac{23 + 18 + 20 + 26 + 21}{5} \right\rfloor = \left\lfloor \frac{108}{5} \right\rfloor = 21$$

Wir füllen den fehlenden Wert also mit 21 auf.

D7, D13: Das Auffüllen der fehlende Werte erfolgt analog zu D1.

D11: Das Beispiel D11 entfernen wir komplett aus den Daten, da es uns keinen Nutzen für unsere Klassifikation bringt.

Damit erhalten wir den folgenden vollständigen Datensatz.

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	26	High	**Strong**	No
D2	Sunny	28	High	Strong	No
D3	Overcast	29	High	Weak	Yes
D4	Rain	23	High	Weak	Yes
D5	Rain	**21**	Normal	Weak	Yes
D6	Rain	12	Normal	Strong	No
D7	Overcast	8	**High**	Strong	Yes
D8	Sunny	25	High	Weak	No
D9	Sunny	18	Normal	Weak	Yes
D10	Rain	20	Normal	Weak	Yes
D12	Overcast	21	High	Strong	Yes
D13	**Rain**	26	Normal	Weak	Yes
D14	Rain	24	High	Strong	No
D15	Sunny	23	Normal	Weak	No
D16	Sunny	21	Normal	Weak	Yes

b) Benutzen Sie für die Berechnung von *k*-NN die gleichen Eckdaten wie in der vorherigen Aufgabe (Normierung für numerische Attribute, 0/1-Distanz für nominale Attribute und die Manhatten Distanz). Klassifizieren Sie so mit 1-NN das folgende Beispiel.

- D17: Outlook=Sunny, Temperature=23, Humidity=High, Wind=Strong

Lösung:

Wir berechnen zuerst die Abstände der einzelnen Trainingsbeispiele zum Klassifikationsbeispiel.

Day	Outlook	Temperature	Humidity	Wind	Klasse	Abstand
D1	Sunny	26	High	Strong	No	$^3/_{21}$
D2	Sunny	28	High	Strong	No	$^5/_{21}$
D3	Overcast	29	High	Weak	Yes	$2\ ^6/_{21}$
D4	Rain	23	High	Weak	Yes	2
D5	Rain	21	Normal	Weak	Yes	$3\ ^2/_{21}$
D6	Rain	12	Normal	Strong	No	$2\ ^{11}/_{21}$
D7	Overcast	8	High	Strong	Yes	$1\ ^{15}/_{21}$
D8	Sunny	25	High	Weak	No	$1\ ^2/_{21}$
D9	Sunny	18	Normal	Weak	Yes	$2\ ^5/_{21}$
D10	Rain	20	Normal	Weak	Yes	$3\ ^3/_{21}$
D12	Overcast	21	High	Strong	Yes	$1\ ^2/_{21}$
D13	Rain	26	Normal	Weak	Yes	$3\ ^3/_{21}$
D14	Rain	24	High	Strong	No	$1\ ^1/_{21}$
D15	Sunny	23	Normal	Weak	No	2
D16	Sunny	21	Normal	Weak	Yes	$2\ ^2/_{21}$

Wir sortieren die Beispiel aufsteigend nach ihrem Abstand zum Klassifikationsbeispiel.

Day	Outlook	Temperature	Humidity	Wind	Klasse	Abstand
D1	Sunny	26	High	Strong	No	$^3/_{21}$
D2	Sunny	28	High	Strong	No	$^5/_{21}$
D14	Rain	24	High	Strong	No	$1\ ^1/_{21}$
D8	Sunny	25	High	Weak	No	$1\ ^2/_{21}$
D12	Overcast	21	High	Strong	Yes	$1\ ^2/_{21}$
D7	Overcast	8	High	Strong	Yes	$1\ ^{15}/_{21}$
D4	Rain	23	High	Weak	Yes	2
D15	Sunny	23	Normal	Weak	No	2
D16	Sunny	21	Normal	Weak	Yes	$2\ ^2/_{21}$
D9	Sunny	18	Normal	Weak	Yes	$2\ ^5/_{21}$
D3	Overcast	29	High	Weak	Yes	$2\ ^6/_{21}$
D6	Rain	12	Normal	Strong	No	$2\ ^{11}/_{21}$
D5	Rain	21	Normal	Weak	Yes	$3\ ^2/_{21}$
D10	Rain	20	Normal	Weak	Yes	$3\ ^3/_{21}$
D13	Rain	26	Normal	Weak	Yes	$3\ ^3/_{21}$

Betrachten wir nun diese Tabelle sehen wir, daß das Beispiel für $k = 1$ (1 Beispiel negativ, 0 positiv) negativ klassifiziert wird.

c) Testen Sie nun verschiedene k. Für welches k ändert sich die Klassifikation gegenüber $k = 1$?

Lösung: Betrachten wir wieder die sortierte Tabelle aus der vorherigen Aufgabe sehen wir, dass sich erst beim Einbeziehen des elftnächsten Beispiels (D3, 6 positiv, 5 negativ, $k = 11$) eine Veränderung der Klassifikation ergibt.

d) Berechnen Sie den Klassifikationswert obiger Instanz mittels abstandsgewichtetem NN (Inverse Distance Weighting). Überlegen Sie sich hierzu, wie Sie diese Methode auf nominale Attribute anwenden können.

Lösung: Die im Skript angebene Methode (siehe Instanzenbasiertes Lernen, Folie 9 bezieht sich auf einen numerischen Klassenwert. Aus diesem Grund müssen wir uns überlegen wie wir diese auf eine nominale Klasse anwenden können. Hierfür gibt es mehrere Möglichkeiten. Wir entscheiden uns für die folgende: Wir berechnen für beide Klassen getrennt die Summe der Kehrwerte der Abstände zwischen dem Trainingsbeispiel der jeweiligen Klasse und des Klassifikationsbeispieles. Anschließend normieren wir diese Summen, indem wir sie durch ihre Summe teilen.

Fangen wir mit der positiven Klasse an:

$$\begin{aligned} sum_+ &= \left(\frac{21}{23}\right)^2 + \left(\frac{21}{36}\right)^2 + \left(\frac{21}{42}\right)^2 + \left(\frac{21}{44}\right)^2 + \left(\frac{21}{47}\right)^2 + \left(\frac{21}{48}\right)^2 \\ &+ \left(\frac{21}{65}\right)^2 + \left(\frac{21}{66}\right)^2 + \left(\frac{21}{66}\right)^2 \approx 2,337 \end{aligned}$$

Analog für die negative Klasse:

$$\begin{aligned} sum_- &= \left(\frac{21}{3}\right)^2 + \left(\frac{21}{5}\right)^2 + \left(\frac{21}{22}\right)^2 + \left(\frac{21}{23}\right)^2 + \left(\frac{21}{42}\right)^2 + \left(\frac{21}{53}\right)^2 \\ &\approx 68,792 \end{aligned}$$

Man sieht jetzt schon, daß die negative die bessere (höhere) Bewertung bekommt. Der Form halber normieren wir diese Werte noch. Für positiv gilt:

$$\frac{sum_+}{sum_+ + sum_-} \approx 0,033$$

Für negativ gilt:

$$\frac{sum_-}{sum_+ + sum_-} \approx 0,967$$

Wie bereits zuvor erwähnt, wird das Beispiel mittels abstandsgewichtetem NN als negativ klassifiziert.

e) Gehen Sie nun vom unveränderten Datensatz aus und benutzen Sie für die Berechnung von k-NN wieder für numerische Attribute die normierte Distanzfunktion und für nominale Attribute diesmal die *Value Difference Metric (VDM)* (Foliensatz Instance-based Learning, Folie 16). Nehmen Sie für die Berechnung der *VDM* einen Wert von $k = 1$ an und denken Sie daran die Distanzen mit der Anzahl der Klassen zu normieren. Klassifizieren Sie so das Beispiel aus Aufgabe a) (verwenden Sie 1-NN und die euklidische Distanz (Foliensatz Instance-based Learning, Folie 14)). Ändert sich die Klassifikation im Vergleich zur Aufgabe a)?

Lösung:

Für numerische Werte benutzen wir wieder die normierte Distanz wie in Aufgabe a). Für nominale Werte wird nun aber die *VDM* verwendet.

Die *VDM* war definiert als:

$$d_A(v_1, v_2) = \sum_c \left| \frac{n_{1,c}}{n_1} - \frac{n_{2,c}}{n_2} \right|^k$$

Da $k = 1$ gilt fällt diese Potenz also weg. Um die Distanzen auszurechnen gehen wir analog zum Skript vor und berechnen Tabellen für die jeweiligen Attribute:

	outlook		
	sunny	overcast	rain
yes	2	3	3
no	4	0	2

Diese Werte kommen zustande, indem man für jede Klasse die Häufigkeit der jeweiligen Ausprägungen des Attributs *outlook* zählt, wobei man unbekannte Attributwerte (egal ob im Attribut oder in der Klasse) einfach außer Acht lässt.

Nun vervollständigen wir die Tabellen für die beiden verbleibenden Attribute:

	humidity	
	high	normal
yes	3	5
no	4	2

	wind	
	strong	weak
yes	2	7
no	3	2

Wir berechnen nun alle nötigen Distanzen im Voraus und normieren diese gleich mit der Anzahl der Klassen (2):

- für das Attribut *outlook*

  $d(sunny, overcast) = |\frac{1}{3} - 1| + |\frac{2}{3} - 0| \approx 1,3 \rightarrow \frac{1,3}{2} = 0,65$

  $d(sunny, rain) = |\frac{1}{3} - \frac{3}{5}| + |\frac{2}{3} - \frac{2}{5}| \approx 0,53 \rightarrow \frac{0,53}{2} = 0,265$
- für das Attribut *humidity*

  $d(high, normal) = |\frac{3}{7} - \frac{5}{7}| + |\frac{4}{7} - \frac{2}{7}| \approx 0,57 \rightarrow \frac{0,57}{2} = 0,285$
- für das Attribut *wind*

  $d(strong, weak) = |\frac{2}{5} - \frac{7}{9}| + |\frac{3}{5} - \frac{2}{9}| \approx 0,76 \rightarrow \frac{0,76}{2} = 0,38$

Nun können wir die Distanzen mittels der euklidischen Distanz $d(x_1, x_2) = \sqrt{\sum_A d_A(v_{1,A}, v_{2,A})^2}$ ermitteln (wir gehen davon aus, dass die Distanz zwischen fehlenden Attributwerten und einem regulären Attributwert maximal ist, wie im Skript unter Foliensatz Instance Based Learning, Folie 19 beschrieben):

$$
\begin{aligned}
d(D17, D1) &= \sqrt{0 + \frac{3}{21}^2 + 0 + 1^2} \approx 1{,}0102 \\
d(D17, D2) &= \sqrt{0 + \frac{5}{21}^2 + 0 + 0} \approx \mathbf{0{,}2381} \\
d(D17, D3) &= \sqrt{0,65^2 + \frac{6}{21}^2 + 0 + 0,38^2} \approx 0{,}8053 \\
d(D17, D4) &= \sqrt{0,265^2 + 0 + 0 + 0,38^2} \approx 0{,}4633 \\
d(D17, D5) &= \sqrt{0,265^2 + 1 + 0,285^2 + 0,38^2} \approx 1{,}1384 \\
d(D17, D6) &= \sqrt{0,265^2 + \frac{11}{21}^2 + 0,285^2 + 0} \approx 0{,}8821 \\
d(D17, D7) &= \sqrt{0,65^2 + \frac{15}{21}^2 + 1 + 0} \approx 1{,}3902 \\
d(D17, D8) &= \sqrt{0 + \frac{2}{21}^2 + 0 + 0,38^2} \approx 0{,}3918 \\
d(D17, D9) &= \sqrt{0 + \frac{5}{21}^2 + 0,285^2 + 0,38^2} \approx 0.5313 \\
d(D17, D10) &= \sqrt{0,265^2 + \frac{3}{21}^2 + 0,285^2 + 0,38^2} \approx 0{,}5624 \\
d(D17, D11) &= \sqrt{0 + \frac{3}{21}^2 + 0,285^2 + 0} \approx 0{,}3188 \\
d(D17, D12) &= \sqrt{0,65^2 + \frac{2}{21}^2 + 0 + 0} \approx 0{,}6569 \\
d(D17, D13) &= \sqrt{1 + \frac{3}{21}^2 + 0,285^2 + 0,38^2} \approx 1{,}1163 \\
d(D17, D14) &= \sqrt{0,265^2 + \frac{1}{21}^2 + 0 + 0} \approx 0{,}2692 \\
d(D17, D15) &= \sqrt{0 + 0 + 0,285^2 + 0,38^2} \approx 0{,}475 \\
d(D17, D16) &= \sqrt{0 + \frac{2}{21}^2 + 0,285^2 + 0,38^2} \approx 0{,}4845
\end{aligned}
$$

Wie man in der Tabelle ablesen kann, hat das Beispiel D2 die geringste Distanz. Das gegebene Beispiel wird also als negativ klassifiziert. Da die Klassifikation in Aufgabe a) ebenfalls negativ war ändert sich nichts.