Maschinelles Lernen: Symbolische Ansätze

TECHNISCHE UNIVERSITÄT DARMSTADT

Wintersemester 2009/2010
Musterlösung für das 12. Übungsblatt

Aufgabe 1: `RelieF`

Gegeben sind folgende 12 Beispiele der Wetter-Daten:

ID	outlook	temperature	humidity	windy	play
1	sunny	hot	high	FALSE	no
2	rainy	mild	high	FALSE	yes
3	rainy	cool	normal	FALSE	yes
4	rainy	cool	normal	TRUE	no
5	overcast	cool	normal	TRUE	yes
6	sunny	mild	high	FALSE	no
7	sunny	cool	normal	FALSE	yes
8	rainy	mild	normal	FALSE	yes
9	sunny	mild	normal	TRUE	yes
10	overcast	mild	high	TRUE	yes
11	overcast	hot	normal	FALSE	yes
12	rainy	mild	high	TRUE	no

Berechnen Sie die `RelieF` Feature-Gewichte für alle 4 Attribute (die ID ist nur zur leichten Identifizierung eines Beispiels vorhanden). Berechnen Sie den Nearest Hit und Nearest Miss für jedes Beispiel ($r = 12$). Gehen Sie davon aus, dass jedes Beispiel genau einmal gewählt wird. Als Distanz-Funktion nehmen Sie einfach die Anzahl der verschiedenen Attribute.

Lösung: Der Algorithmus geht folgendermaßen vor: Zuerst setzt der Benutzer den Wert für r (in der Aufgabe gilt $r = 12$). Als nächstes wählt der Algorithmus ein Beispiel zufällig aus (hier wird **nicht** sichergestellt, dass das Beispiel nicht schon einmal verwendet wurde). Dann findet er den *nearest hit* und den *nearest miss* und macht ein Update auf den Gewichtswert jedes Attributs.

In der Aufgabe gehen wir davon aus, dass ein bereits verwendetes Beispiel nicht noch einmal genommen wird. Wir machen uns die Eigenschaft zu Nutze, dass man auch eine Tabelle der Distanzen für die *Hits* und die *Misses* erstellen kann und dann einfach direkt für jedes Beispiel den *nearest hit* und *nearest miss* ausrechnen kann (indem man die Werte für jedes Attribut einfach addiert, was dem Update-Schritt des Algorithmus entspricht).

Wir erstellen also 2 Tabellen für die *Hits* und eine für die *Misses* (zu beachten ist, dass die Tabellen an der Diagonalen gespiegelt sind):

Hits für +								
ID	2	3	5	7	8	9	10	11
2	0	2	4	3	1	3	2	3
3	2	0	2	1	1	3	4	2
5	4	2	0	2	3	2	2	2
7	3	1	2	0	2	2	4	2
8	1	1	3	2	0	2	3	2
9	3	3	2	2	2	0	2	3
10	2	4	2	4	3	2	0	3
11	3	2	2	2	2	3	3	0

Hits für -				
ID	1	4	6	12
1	0	4	1	3
4	4	0	4	2
6	1	4	0	2
12	3	2	2	0

Misses				
ID	1	4	6	12
2	2	3	1	1
3	3	1	3	3
5	4	1	4	3
7	2	2	2	4
8	3	2	2	2
9	3	2	2	2
10	3	3	2	1
11	2	3	3	4

In den Tabellen sind die *nearest neighbors* der gleichen Klasse rot markiert (wobei bei gleicher Distanz immer das erste Beispiel in der Tabelle ausgewählt wurde). Nun erstellt man eine große Tabelle in der man für jedes Attribut die Unterscheidungen festhält (gleicher Wert = Distanz 0, unterschiedlicher Wert = Distanz 1). In dieser Tabelle repräsentiert eine Zeile jeweils einen Durchlauf des Algorithmus (wobei wir die Gewichte erst am Ende updaten, bzw. berechnen).

	Hits					Misses				
	N	outl.	temp.	hum.	wind	N	outl.	temp.	hum.	wind
2	8	0	0	1	0	6	1	0	0	0
3	7	1	0	0	0	4	0	0	0	1
5	3	1	0	0	1	4	1	0	0	0
7	3	1	0	0	0	1	0	1	1	0
8	2	0	0	1	0	4	0	1	0	1
9	5	1	1	0	0	4	1	1	0	0
10	2	1	0	0	1	12	1	0	0	0
11	3	1	1	0	0	1	1	0	1	0
1	6	0	1	0	0	2	1	1	0	0
4	12	0	1	1	0	3	0	0	0	1
6	1	0	1	0	0	2	1	0	0	0
12	4	0	1	1	0	2	0	0	0	1
		6	6	4	2		7	4	2	4

Da man nun alle Werte zu Verfügung hat, kann man die `RelieF` Feature-Gewichte errechnen:
$W(outlook) = -6/12 + 7/12 = 1/12$
$W(temperature) = -6/12 + 4/12 = -1/6$
$W(humidity) = -4/12 + 2/12 = -1/6$
$W(wind) = -2/12 + 4/12 = 1/6$

Aufgabe 2: Diskretisierungsmethoden

Gegeben sei folgende Version der Wetter-Daten mit 12 Trainings-Beispielen und 2 numerischen Attributen.

ID	outlook	temperature	humidity	windy	play
1	sunny	85	85	FALSE	no
2	rainy	70	96	FALSE	yes
3	rainy	68	80	FALSE	yes
4	rainy	65	70	TRUE	no
5	overcast	64	65	TRUE	yes
6	sunny	72	95	FALSE	no
7	sunny	69	70	FALSE	yes
8	rainy	75	80	FALSE	yes
9	sunny	75	70	TRUE	yes
10	overcast	72	90	TRUE	yes
11	overcast	81	75	FALSE	yes
12	rainy	71	91	TRUE	no

Diskretisieren Sie die beiden numerischen Attribute mit den Verfahren, die Sie in der Vorlesung kennen gelernt haben. Wählen Sie die Anzahl der Intervalle so, daß Sie die bekannten Daten erhalten könnten (drei Werte für Temperature, zwei für Humidity). Vergleichen Sie die Resultate miteinander und mit den bekannten Daten (die aus der vorherigen Aufgabe).

- equal-width

  Lösung: Bei equal-width wird versucht gleich große Intervalle zu bilden:

  - Attribut “temperature”: $85 - 64 = 21/3 = 7 \rightarrow$ Die Intervalle sollten also die Größe 7 haben. Es resultieren folgende Intervalle:

    $[64, 70], [71, 77], [78, 85]$ ($[64, 71], [72, 78], [79, 85]$ wäre auch zulässig, man muss sich hier eine geeignete Einteilung überlegen)

    Dies ergibt dann die Intervalle $(-\infty, 70], (70, 77], (77, \infty)$

  - Attribut “humidity”: $96-65 = \lfloor 31/2 \rfloor = 15 \rightarrow$ Die Intervalle sollten also die Größe 15 haben: $[65, 80], [81, 96]$

    Dies ergibt dann die Intervalle $(-\infty, 80], (80, \infty)$.

- equal-frequency

  Lösung: Bei equal-frequency wird vesucht möglichst gleich viele Beispiele pro Intervall zu haben:

  - Attribut “temperature”: $12/3 = 4$ Es sollten also jeweils 4 Beispiele in einem Intervall liegen.

    Daher liegen die Beispiele

    Nr.5 temp.=64, Nr.4 temp.=65, Nr.3 temp.=68, Nr.7 temp.=69

    im ersten Intervall, die Beispiele

    Nr.2 temp.=70, Nr.12 temp.=71, Nr.6 temp.=72, Nr.10 temp.=72

    im zweiten und die Beispiele

    Nr.8 temp.=75, Nr.9 temp.=75, Nr.11 temp.=81, Nr.1 temp.=85

    im dritten Intervall.

  - Attribut “humidity”: $12/2 = 6$ Es sollten jeweils 6 Beispiele in einem Intervall liegen.

    Die Beispiele

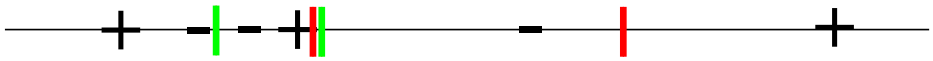
    Nr.5 hum.=65, Nr.4 hum.=70, Nr.7 hum.=70, Nr.9 hum.=70, Nr.11 hum.=75, Nr.3 hum.=80, Nr.8 hum.=80 (die beiden Beispiel mit humidity = 80 könnten auch im zweiten Intervall liegen)

    im ersten und die Beispiele

    Nr.1 hum.=85, Nr.10 hum.=90, Nr.12 hum.=91, Nr.6 hum.=95, Nr.2 hum.=96

    im zweiten Intervall.

Die folgende Grafik erklärt den Unterschied beider Methoden nochmals anschaulich, wobei es sich hier um ein allgemeines Beispiel handelt, was nicht den vorherigen Aufgaben entspricht (die roten Striche sind für equal-width und die grünen für equal-frequency):



- chi-merge

  Lösung: Wir werden nicht auf die Berechnung jedes einzelnen χ^2-Wertes eingehen, jedoch möchten wir kurz einen Trick zeigen und anschließend ein χ^2-Wert als Beispiel berechnen. Für zwei benachbarte Intervalle I_1 und I_2 berechnet sich χ^2 wie folgt:

$$\begin{aligned}\chi^2 &= \sum_{i\in\{I_1,I_2\}}\sum_{j\in\{+,-\}}\frac{(A_{ij}-E_{ij})^2}{E_{ij}}\\ &= \sum_{j\in\{+,-\}}\frac{(A_{I_1j}-E_{I_1j})^2}{E_{I_1j}}+\sum_{j\in\{+,-\}}\frac{(A_{I_2j}-E_{I_2j})^2}{E_{I_2j}}\end{aligned}$$

  Wie man sieht, kann jede dieser Summe auch für die Berechnung des χ^2-Wertes eines Paares von benachbarten Intervallen, zu denen eines der eben verwendeten Intervallen I_1 und I_2 gehört, verwendet werden. Das heißt, falls I_3 das andere benachbarte Intervall von I_2 ist, kann die obige Summe für I_2 zur Berechnung des χ^2-Wertes des Intervallpaares I_2 und I_3 benutzt werden. Aus diesem Grund berechnen wir zuerst diese Teilsumme und addieren sie danach für benachbarte Paare (=Summe).

  Die Erwartungswerte (E_{ij}) für die Anzahl von positiven bzw. negativen Beispielen innerhalb eines Intervalles erhalten wir, indem wir die Gesamtanzahl der Beispiele des Intervalls mit der jeweiligen a priori Wahrscheinlichkeiten multiplizieren. Da 8 positive und 4 negative Beispiele vorliegen, ist die a priori Wahrscheinlichkeit für ein positives Beispiel $2/3$ und dementsprechend für ein negatives $1/3$.

  Beispiel: Die Berechnung von χ^2 sieht für die benachbarten Intervalle I_1 und I_2 folgendermaßen aus. Angenommen I_1 beinhaltet 3 positive und 2 negative Beispiel, entsprechend I_2 2 positive und 4 negative. Der Erwartungswert für positive bzw. negative Beispiele in I_1 ist $(3+2)*2/3 = 10/3$ bzw. $(3+2)*1/3 = 5/3$. Entsprechend erwarten wir für I_2 $(2+4)*2/3 = 4$ positive und $(2+4)*1/3 = 2$ negative Beispiele. Jetzt haben wir alle benötigten Werte vorliegen. Berechnen wir nun zunächst die Teilsumme der beiden Intervalle

$$\frac{(3-\frac{10}{3})^2}{\frac{10}{3}}+\frac{(2-\frac{5}{3})^2}{\frac{5}{3}}=\frac{(\frac{1}{3})^2}{\frac{10}{3}}+\frac{(\frac{2}{3})^2}{\frac{5}{3}}=\frac{1}{9}\cdot\frac{3}{10}+\frac{4}{9}\cdot\frac{3}{5}=0,3 \tag{1}$$

$$\frac{(2-4)^2}{4}+\frac{(4-2)^2}{2}=\frac{4}{4}+\frac{4}{2}=3. \tag{2}$$

  Addieren wir nun die Teilsummen (1) und (2), erhalten wir den χ^2-Wert dieses Intervallpaares.

$$\chi^2 = 0,3+3=3,3$$

  Betrachten wir nun die eigentliche Berechnung:

  Temperature: Wir möchten mittels χ-Merge Temperature in 3 diskrete Werte aufteilen. Zuerst sortieren wir die Werte des Attributs *Temperature*. Je nach auf- bzw. absteigender Sortierung erhält man unterschiedliche Ergebnisse. Wir entscheiden uns für eine aufsteigende Sortierung.

Wert	positiv	negativ	Teilsumme	Summe
64	1	0	0,5	2,5
65	0	1	2	2,5
68	**1**	**0**	**0,5**	**1**
69	**1**	**0**	**0,5**	1
70	1	0	0,5	2,5
71	0	1	2	2,25
72	1	1	0,25	1,25
75	2	0	1	1,5
81	1	0	0,5	2,5
85	0	1	2	-

Die niedrigsten χ^2-Werte haben die Intervallpaare 68-69 und 69-70. Wir entscheiden uns für das Intervall 68-69. Für diese benötigen wir die Teilsumme (für 2 positive und 0 negative Beispiele), die wir bereits berechnet haben, und die Summen für die benachbarten Intervalle.

Wert	positiv	negativ	Teilsumme	Summe
64	1	0	0,5	2,5
65	0	1	2	3
68-69	2	0	1	1,5
70	1	0	0,5	2,5
71	0	1	2	2,25
72	**1**	**1**	**0,25**	**1,25**
75	**2**	**0**	**1**	1,5
81	1	0	0,5	2,5
85	0	1	2	-

Wir verschmelzen 72 und 75 und führen ein Update durch.

Wert	positiv	negativ	Teilsumme	Summe
64	1	0	0,5	2,5
65	0	1	2	3
68-69	2	0	1	1,5
70	1	0	0,5	2,5
71	0	1	2	2,125
72-75	**3**	**1**	**0,125**	**0,625**
81	**1**	**0**	**0,5**	2,5
85	0	1	2	-

Wir verschmelzen 72-75 und 81 und aktualisieren die Tabelle.

Wert	positiv	negativ	Teilsumme	Summe
64	1	0	0,5	2,5
65	0	1	2	3
68-69	**2**	**0**	**1**	**1,5**
70	**1**	**0**	**0,5**	2,5
71	0	1	2	2,4
72-81	4	1	0,4	2,4
85	0	1	2	-

Die Intervalle 68-69 und 70 werden zusammengelegt.

Wert	positiv	negativ	Teilsumme	Summe
64	1	0	0,5	2,5
65	0	1	2	3,5
68-70	3	0	1,5	3,5
71	**0**	**1**	**2**	**2,4**
72-81	**4**	**1**	**0,4**	2,4
85	0	1	2	-

Wir verschmelzen 71 und 72-81.

Wert	positiv	negativ	Teilsumme	Summe
64	1	0	0,5	2,5
65	0	1	2	3,5
68-70	**3**	**0**	**1,5**	**1,5**
71-81	**4**	**2**	**0**	2
85	0	1	2	-

Die Intervalle 68-70 und 71-81 werden zusammengelegt.

Wert	positiv	negativ	Teilsumme	Summe
64	**1**	**0**	**0,5**	**2,5**
65	**0**	**1**	**2**	2,5
68-81	7	2	0,5	2,5
85	0	1	2	-

Wir verschmelzen 64 und 65 und erhalten damit die folgenden Intervalle: 64-65, 68-81 und 85. Offensichtlich können nicht alle möglichen Temperaturen einem dieser Intervalle zugeordnet werden. Aus diesem Grund müssen die enstandenden Intervalle angepaßt werden. Hierfür gibt es verschiedene Möglichkeiten diese Intervalle zu bearbeiten, u.a.

- $(-\infty, 65]$, $(65, 81]$ und $(81, \infty)$ oder
- $(-\infty, 68)$, $[68, 85)$ und $[85, \infty)$ oder
- $(-\infty, 66.5)$, $[66.5, 83.5)$ und $[83.5, \infty)$ usw.

Humidity: Für Humidity gehen wir analog vor, wobei wir diesmal nur zwei diskrete Werte haben wollen.

Wert	positiv	negativ	Teilsumme	Summe
65	**1**	**0**	**0,5**	**0,5**
70	**2**	**1**	**0**	0,5
75	1	0	0,5	1,5
80	2	0	1	3
85	0	1	2	2,5
90	1	0	0,5	2,5
91	0	1	2	4
95	0	1	2	2,5
96	1	0	0,5	-

Wir verschmelzen 65 und 70.

Wert	positiv	negativ	Teilsumme	Summe
65-70	**3**	**1**	**0,125**	**0,625**
75	**1**	**0**	**0,5**	1,5
80	2	0	1	3
85	0	1	2	2,5
90	1	0	0,5	2,5
91	0	1	2	4
95	0	1	2	2,5
96	1	0	0,5	-

65-70 und 75 werden zusammengelegt.

Wert	positiv	negativ	Teilsumme	Summe
65-75	**4**	**1**	**0,4**	**1,4**
80	**2**	**0**	**1**	3
85	0	1	2	2,5
90	1	0	0,5	2,5
91	0	1	2	4
95	0	1	2	2,5
96	1	0	0,5	-

65-75 und 80 werden verschmolzen.

Wert	positiv	negativ	Teilsumme	Summe
65-80	6	1	0,571	2,571
85	**0**	**1**	**2**	**2,5**
90	**1**	**0**	**0,5**	2,5
91	0	1	2	4
95	0	1	2	2,5
96	1	0	0,5	-

Wir legen 85 und 90 zusammen.

Wert	positiv	negativ	Teilsumme	Summe
65-80	**6**	**1**	**0,571**	**0,821**
85-90	**1**	**1**	**0,25**	2,25
91	0	1	2	4
95	0	1	2	2,5
96	1	0	0,5	-

Wir verschmelzen 65-80 und 95-90.

Wert	positiv	negativ	Teilsumme	Summe
65-90	**7**	**2**	**0,5**	**2,5**
91	**0**	**1**	**2**	4
95	0	1	2	2,5
96	1	0	0,5	-

65-90 und 91 werden zusammengefaßt.

Wert	positiv	negativ	Teilsumme	Summe
65-91	**7**	**3**	**0,167**	**2,167**
95	**0**	**1**	**2**	2,5
96	1	0	0,5	-

Wir fassen 65-91 und 95 zusammen und erhalten damit die folgenden Intervalle: [65,91] und [96, 96]. Diese Intervalle müssen wie bereits oben erwähnt noch so bearbeitet werden, daß jeder Zahlenwert zu einem der bearbeiteten Intervallen zugeteilt werden kann.

- info-split (entropy-split)

  Lösung: Bei Entropy-split zerlegen wir eines der bestehenden Intervalle in jeweils zwei Teilintervalle, falls das zu zerlegende Intervall den niedristen E-Score (Berechnung von Entropy-split ist äquivalent zu diesem Maß) aufweist. E-Score ist wiederum äquivalent zu Information-Gain. Beide Maße und deren Berechnung wurden bereits ausführlich in den Übungen über Entscheidungsbäume und Inkrementelles Lernen (ID5r) besprochen. Aus diesem Grund wollen wir nicht weiter auf deren Berechnung eingehen.

  Temperature: Wir sortieren zunächst die Attributwerte (aufsteigend).

	A<Wert		A≥Wert		
Wert	positiv	negativ	positiv	negativ	E-Score
64	0	0	8	4	0,918
65	1	0	7	4	0,867
68	1	1	7	3	0,901
69	2	1	6	3	0,918
70	3	1	5	3	0,907
71	4	1	4	3	0,876
72	4	2	4	2	0,918
75	5	3	3	1	0,907
81	7	3	1	1	0,901
85	8	3	0	1	**0,775**

Der Teilungspunkt 85 hat den besten (niedrigsten) E-Score. Wir teilen das Intervall $(-\infty, \infty)$ in $(-\infty, 85)$ und $[85, \infty)$ auf.

Wert	A<Wert positiv	A<Wert negativ	A≥Wert positiv	A≥Wert negativ	E-Score
64	0	0	8	3	0,845
65	1	0	7	3	0,801
68	1	1	7	3	0,807
69	2	1	6	2	0,840
70	3	1	5	2	0,844
71	4	1	4	2	0,829
72	4	2	4	1	0,829
75	5	3	3	0	**0,694**
81	7	3	1	0	0,801
85	0	0	0	1	-

Jetzt ist 75 der beste Teilungspunkt. Wir zerlegen das Intervall $(-\infty, 85)$ in $(-\infty, 75)$ und $[75, 85)$. Demnach sieht die Diskretisierung von Temperature wie folgt aus: $(-\infty, 75)$, $[75, 85)$ und $[85, \infty)$.

Humidity:

Wert	A<Wert positiv	A<Wert negativ	A≥Wert positiv	A≥Wert negativ	E-Score
65	0	0	8	3	0,918
70	1	0	7	4	0,867
75	3	1	5	3	0,907
80	4	1	4	3	0,876
86	6	1	2	3	**0,750**
90	6	2	2	2	0,874
91	7	2	1	2	0,803
95	7	3	1	1	0,867
96	7	4	1	0	0,801

Wir teilen das Intervall $(-\infty, \infty)$ in $(-\infty, 86)$ und $[86, \infty)$ auf.

Nun gehen wir noch kurz auf den Vergleich der Ergebnisse ein: Da man nicht weiß, wie die numerischen Daten des originalen Datensatzes diskretisiert wurden, kann man hier keine genaue Aussage treffen. Daher kann man sich nur die Unterschiede zwischen den einzelnen Methoden und den Originaldaten anschauen (diese sind rot markiert):

						equal-freq.		equal-width		$\chi - merge$		info-split	
Bsp.Nr.	outl.	temp.	hum.	windy	play?	temp.	hum.	temp.	hum.	temp.	hum.	temp.	hum.
1	sunny	hot	high	false	no	hot	high	hot	high	hot	normal	hot	normal
2	rainy	mild	high	false	yes	mild	high	cool	high	mild	high	cool	high
3	rainy	cool	normal	false	yes	cool	normal	cool	normal	mild	normal	cool	normal
4	rainy	cool	normal	true	no	cool	normal	cool	normal	cool	normal	cool	normal
5	overcast	cool	normal	true	yes	cool	normal	cool	normal	cool	normal	cool	normal
6	sunny	mild	high	false	no	mild	high	mild	high	mild	normal	cool	high
7	sunny	cool	normal	false	yes	cool	normal	cool	normal	mild	normal	cool	normal
8	rainy	mild	normal	false	yes	hot	normal	mild	normal	mild	normal	mild	normal
9	sunny	mild	normal	true	yes	hot	normal	mild	normal	mild	normal	mild	normal
10	overcast	mild	high	true	yes	mild	high	mild	high	mild	normal	cool	high
11	overcast	hot	normal	false	yes	hot	normal	hot	normal	mild	normal	mild	normal
12	rainy	mild	high	true	no	mild	high	cool	high	mild	normal	cool	high