Data Mining und Maschinelles Lernen

TECHNISCHE UNIVERSITÄT DARMSTADT

10. Übungsblatt

Aufgabe 1 AdaBoost

Rechnen Sie das AdaBoost-Beispiel aus der Vorlesung (Ensemble-Methoden, Folien 14ff) nach. Verwenden Sie für die einzelnen Datenpunkte die folgenden Koordinaten (x, y, Klasse):

```
1,  5,  +          3,  1,  -
2,  2,  +          4,  6,  -
5,  8,  +          7,  4,  -
6, 10,  +          9,  3,  -
8,  7,  +         10,  9,  -
```

Als Basis-Lerner sollen Decision Stumps (also waagrechte bzw. senkrechte Splits, z.B. $x > 4 \rightarrow +$) verwendet werden. Der Basislerner wählt unter allen möglichen Splits jenen aus, bei dem die Gesamtsumme der Gewichte der falsch klassifizierten Beispiele minimiert wird. Wählen Sie bei Gleichstand den zuerst gefundenen Test beginnend mit vertikalen Splits mit aufsteigenden Thresholds.

a) Berechnen Sie die ersten 3 AdaBoost-Iterationen.

b) Generieren Sie aus den eben berechneten Decision Stumps einen Entscheidungsbaum.

Aufgabe 2 Stacking

In dieser Aufgabe sollen Sie unter Verwendung mehrerer Basislerner und der Ensemble-Methode Stacking einen Entscheidungsbaum lernen. Verwenden Sie hierfür den Datensatz und die drei Decision Stumps aus der vorherigen Aufgabe.

a) Konvertieren Sie diesen Datensatz, d.h. ersetzen sie die Attribute durch eine neue Attributsmenge, die jeweils ein Attribut für jeden Decision Stump beinhaltet. Als Attributwerte werden die Vorhersagen des entsprechenden Klassifzierers verwendet.

b) Bestimmen Sie nun auf dem konvertierten Datensatz einen Entscheidungsbaum mittels des Verfahrens ID3 und Maß Information Gains. Entscheiden Sie sich bei Gleichstand z.B. für den als erstes gefundenen Test.

c) Zeichnen Sie diesen Baum und vergleichen Sie ihn mit dem Entscheidungsbaum aus Aufgabe 1.

Aufgabe 3 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 (Foliensatz "*Instanzenbasiertes Lernen*", Folie "*RELIEF (2)*") 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 (Es kann aber natürlich mehrfach als Nearest-Neighbor auftauchen.). Als Distanz-Funktion nehmen Sie einfach die Anzahl der verschiedenen Attribute.

Aufgabe 4 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:

- equal-width
- equal-frequency
- chi-merge
- info-split

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).