Data Mining und Maschinelles Lernen

TECHNISCHE UNIVERSITÄT DARMSTADT

Wintersemester 2015/16
1. Übungsblatt

Aufgabe 1 Anwendungsszenario

Überlegen Sie sich ein neues Szenario des klassifizierenden Lernens (kein aus der Vorlesung bekanntes).

a) Bestimmen Sie die zu verwendenden Trainings- und Testdaten Ihres Klassifikationsproblems.

b) Aus welchen Typen von Attributen (nominal, numerisch, ...) setzen sich die Beispiele zusammen?

c) Welche Kriterien würden Sie verwenden, um die Performanz des resultierenden Klassifizierers zu bewerten? Bedenken Sie bei Ihren Überlegungen, dass die Performanz abhängig von dem gewählten Problem ist (bei der Klassifizierung von Spam Mail ist es beispielsweise wichtig, echte Mails nicht als Spam einzuordnen).

Aufgabe 2 Praktische Anwendung

Gegeben sei das folgende 3-Klassenproblem, bei dem einer Person abhängig von ihrer Schulbildung, ihrem Familienstand (verheiratet/ledig mit (keinen) Kindern) und ihrem Geschlecht ein Wagentyp (Familien-, Klein- oder Sportwagen) zugeordnet werden soll.

Von einigen Personen sind uns folgende Daten bekannt (Tabelle 1):

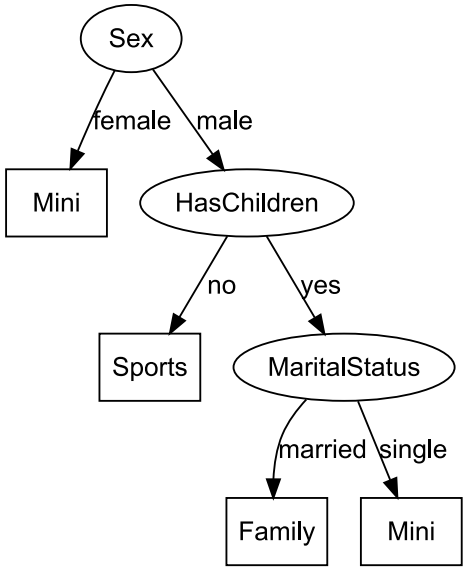
No.	Education	Marital Status	Sex	Has Children	Car
1	primary	married	female	no	mini
2	primary	married	male	no	sports
3	primary	married	female	yes	mini
4	primary	married	male	yes	family
5	primary	single	female	no	mini
6	primary	single	male	no	sports
7	secondary	married	female	no	mini
8	secondary	married	male	no	sports
9	secondary	married	male	yes	family
10	secondary	single	female	no	mini
11	secondary	single	female	yes	mini
12	secondary	single	male	yes	mini
13	university	married	male	no	mini
14	university	married	female	yes	mini
15	university	single	female	no	mini
16	university	single	male	no	sports
17	university	single	female	yes	mini
18	university	single	male	yes	mini

Tabelle 1: Trainingsdaten

No.	Education	Marital Status	Sex	Has Children	Car
19	primary	single	female	yes	?
20	primary	single	male	yes	?
21	secondary	married	female	yes	?
22	secondary	single	male	no	?
23	university	married	male	yes	?
24	university	married	female	no	?

Tabelle 2: Testdaten

a) Klassifizieren Sie die in Tabelle 2 angegebene Testmenge, deren Klassenlabel uns unbekannt sind, mit Hilfe des abgebildeten Entscheidungsbaums.



b) Der Baum klassifiziert nicht alle Trainings-Beispiele korrekt. Wie müßte man den vorhandenen Baum erweitern, damit er alle Trainings-Beispiele korrekt klassifiziert? Wie schätzen Sie die Qualität des resultierenden Baums ein?

c) Für das gegebene Klassifikationsproblem liefert Ihnen ein Regellerner folgende Regelmenge:

MaritalStatus = married	$\wedge$	HasChildren = yes	$\wedge$	Sex = male	$\Rightarrow$	Car = family
HasChildren = no	$\wedge$	Sex = male			$\Rightarrow$	Car = sports
Sex = female					$\Rightarrow$	Car = mini
MaritalStatus = single	$\wedge$	HasChildren = yes			$\Rightarrow$	Car = mini
Education = university	$\wedge$	MaritalStatus = married			$\Rightarrow$	Car = mini

Verwenden Sie diese Regelmenge zur Klassifikation der gegebenen Testmenge. Von welchen Regeln werden die Testbeispiele jeweils klassifiziert? Welches Verhalten und welche Eigenschaften von Regelmengen im Vergleich zu Entscheidungsbäumen fallen Ihnen hierbei auf?

d) Klassifizieren Sie nun dieselbe Testmenge mit dem Lernalgorithmus Nearest Neighbour aus der Vorlesung. Verwenden Sie als Distanzfunktion die Anzahl der Attributwerte, in denen sich die zu vergleichenden Beispiele unterscheiden. Bestimmen Sie alle Trainingsbeispiele mit minimaler Distanz zum jeweiligen Testbeispiel. Sagen Sie anhand der Klassenlabel dieser Trainingsbeispiele die Klasse des Testbeispiels voraus.