

Maschinelles Lernen: Symbolische Ansätze



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Wintersemester 2009/2010
Musterlösung für das 1. Übungsblatt

Aufgabe 1: Anwendungsszenario

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

Lösung: Es sollen medizinische Daten von Patienten benutzt werden, um festzustellen, ob die Anwendung einer Chemotherapie Erfolge zeigen würde.

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

Lösung:

Trainingsdaten: Daten vorhandener Patientenakten, bei denen die Patienten mit Chemotherapie behandelt wurden. Als Klassenlabel wird ausgegeben, ob die Chemotherapie angeschlagen hat oder nicht

Testdaten: dieselben Daten wie die zum Trainieren verwendeten. Hier gibt es 3 verschiedene Möglichkeiten (siehe Einführung, Folie 16):

1. Expertenwissen: Ein Experte in der Domäne bewertet die Ausgaben des Klassifizierers, bzw. gibt eine Klassifizierung der Beispiele vor
2. Bewertung der Güte über bereits gelabelte Daten; hier wird üblicherweise ein Teil der Trainingsdaten, der die gleiche Klassenverteilung wie die gesamten Trainingsdaten aufweist, als Testdaten verwendet
3. On-Line Überprüfung: der Klassifizierer gibt die Vorhersage aus und diese wird direkt überprüft

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

Lösung:

nominal: Geschlecht, Alter als kategoriale Werte (<25, 25–60, >60), ...

numerisch: Blutdruck, Alter, ...

- 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).

Lösung: Zur Bewertung der Performanz wird der Fehler auf den Testdaten berechnet (siehe Aufgabe 1a)).

In dem gewählten Beispiel kommt es auf die Sichtweise an. Das Krankenhaus ist bestrebt die Kosten für eine Chemotherapie bei einem Patienten, wo diese nicht anschlagen würde, einzusparen. Der Patient hingegen würde nichts unversucht lassen, um den Krebs zu behandeln.

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:

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

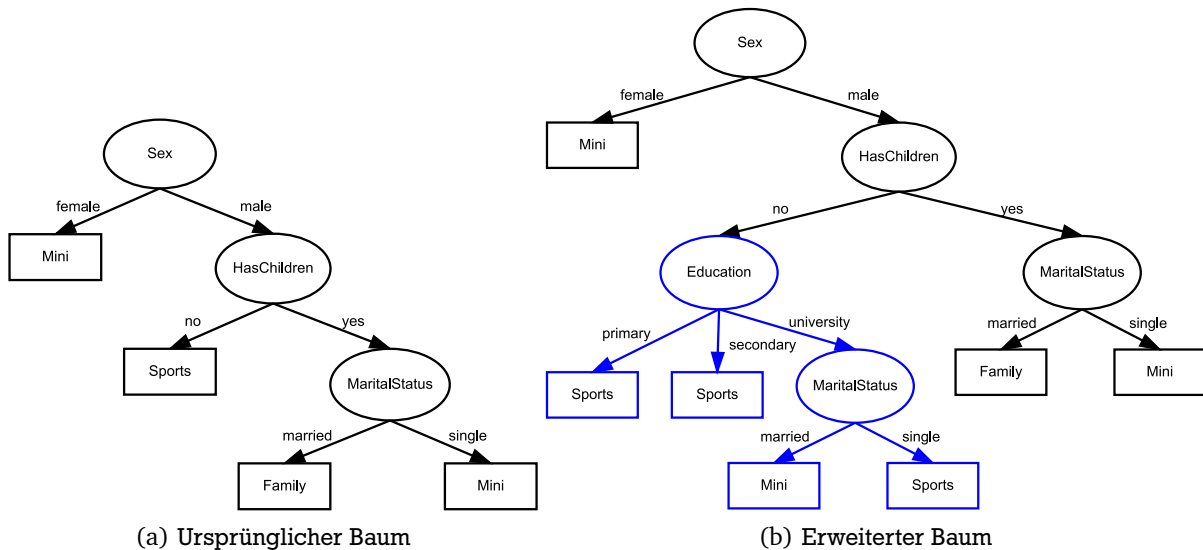
- a) Klassifizieren Sie die folgende Testmenge, deren Klassenlabel uns unbekannt sind, mit Hilfe des abgebildeten Entscheidungsbaums.

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

Lösung:

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



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?

Lösung: Nur das Beispiel

No.	Education	Marital Status	Sex	Has Children	Car
13	university	married	male	no	mini

wird von dem zweiten Blatt (von links) falsch klassifiziert. Um den Baum zu erweitern, betrachten wir zunächst alle Beispiele, die durch dieses Blatt klassifiziert werden:

No.	Education	Marital Status	Sex	Has Children	Car
2	primary	married	male	no	sports
6	primary	single	male	no	sports
8	secondary	married	male	no	sports
13	university	married	male	no	mini
16	university	single	male	no	sports

Wie man sieht ist keiner der verbleibenden Tests, *Education* und *Marital Status*, alleine ausreichend, um die Beispiele sauber zu trennen. Also müssen wir uns für einen der beiden Tests entscheiden. Wir haben uns für den Test *Education* entschieden. Somit erhalten wir 3 Testausgänge. Da die Testausgänge *Primary* und *Secondary* beide nur Beispiele der Klasse *Sports* abdecken, können wir bei diesen jeweils ein Blatt mit der Vorhersage *Sports* anhängen. Für den Testausgang *University* benötigen wir noch einen weiteren Test auf *Marital Status*, wobei bei *Married* ein Blatt mit der Vorhersage *Mini* und bei *Single* eines mit *Sports* angehängt wird.

Der ursprüngliche und eben erweiterte Baum ist oben abgebildet.

Die Qualität des Baums hat sich nun verändert:

- Einerseits klassifiziert der Baum nun alle Trainingsbeispiele korrekt aber
- andererseits ist er um einiges angewachsen (von 4 Blättern auf 7 Blätter).

Da man *Overfitting*, also eine Überanpassung an die Trainingsmenge, vermeiden will, sollte man generell kleinere Bäume bevorzugen, selbst wenn sie einige wenige Trainingsbeispiele falsch klassifizieren.

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

- R1: MaritalStatus = married \wedge HasChildren = yes \wedge Sex = male \Rightarrow Car = family
 R2: HasChildren = no \wedge Sex = male \Rightarrow Car = sports
 R3: Sex = female \Rightarrow Car = mini
 R4: MaritalStatus = single \wedge HasChildren = yes \Rightarrow Car = mini
 R5: Education = university \wedge MaritalStatus = married \Rightarrow Car = mini

Verwenden Sie diese Regelmenge zu 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?

Lösung:

No.	Education	Marital Status	Sex	Has Children	Car	Rule
19	primary	single	female	yes	Mini	R3, R4
20	primary	single	male	yes	Mini	R4
21	secondary	married	female	yes	Mini	R3
22	secondary	single	male	no	Sports	R2
23	university	married	male	yes	?	R1, R5
24	university	married	female	no	Mini	R3, R5

Regelmengen und Entscheidungsbäume unterscheiden sich in folgenden wesentlichen Punkten:

- Regeln können überlappen wie bei den Beispielen 19, 23 und 24. Da bei den Beispielen 19 und 24 beide Regeln die gleiche Klasse vorhersagen, wählt man diese zur Klassifikation (indem man einfaches *Voting* macht, also die Klasse nimmt die am häufigsten von den Regeln vorhergesagt wird). Bei dem Beispiel 23 unterscheidet sich jedoch die Klassifikation und es kommt zu einem Gleichstand. Um dennoch eine Klassifikation vornehmen zu können, gibt es verschiedene Möglichkeiten:
 - man nimmt die *Majority Class* der Klassen aus der Vorhersage (also die Klasse die in den Trainingsdaten am häufigsten vorkommt)
 - man nimmt die Klasse der kürzeren Regel
 - falls die Regeln nach ihrer Qualität sortiert sind, verwendet man die mit der höheren Qualität
- Die obige Regelmenge kann bestimmte Beispiele nicht klassifizieren, da diese Beispiele von keiner Regel abgedeckt werden. Daher verwendet man eine sog. *Default-Rule* welche im Regelkörper den Wert `TRUE` hat, also alle Beispiele abdeckt und die Majority Class vorhersagt.

- 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.

Lösung: Wir berechnen zunächst alle Abstände aller Testbeispiele zu allen Trainingsbeispielen. Diese sind in der folgenden Tabelle jeweils in einer der letzten sechs Spalten aufgelistet.

No.	Education	Marital Status	Sex	Has Children	Car	19	20	21	22	23	24
1	primary	married	female	no	mini	2	3	2	3	3	1
2	primary	married	male	no	sports	3	2	3	2	2	2
3	primary	married	female	yes	mini	1	2	1	4	2	2
4	primary	married	male	yes	family	2	1	2	3	1	3
5	primary	single	female	no	mini	1	2	3	2	4	2
6	primary	single	male	no	sports	2	1	4	1	3	3
7	secondary	married	female	no	mini	3	4	1	2	3	1
8	secondary	married	male	no	sports	4	3	2	1	2	2
9	secondary	married	male	yes	family	3	2	1	2	1	3
10	secondary	single	female	no	mini	2	3	2	1	4	2
11	secondary	single	female	yes	mini	1	2	1	2	3	3
12	secondary	single	male	yes	mini	2	1	2	2	2	4
13	university	married	male	no	mini	4	3	3	2	1	1
14	university	married	female	yes	mini	2	3	1	4	1	1
15	university	single	female	no	mini	2	3	3	2	3	1
16	university	single	male	no	sports	3	2	4	1	2	2
17	university	single	female	yes	mini	1	2	2	3	2	2
18	university	single	male	yes	mini	2	1	3	2	1	3

Tabelle 3: Distanzen zum jeweiligen Beispiel

Da wir die Nachbarn mit minimalen Abstand zur Klassifikation verwenden, selektieren wir diejenigen mit einem Abstand von 1 aus der Tabelle (**rot markiert**). Da diese Nachbarn als Votes für ihre jeweilige Klasse benutzt werden, erhalten wir die folgende Abstimmung, bei der jeweils die Klasse vorhergesagt wird, die am meisten Stimmen erhält.

No.	Family	Mini	Sports	Prediction
19	0	4	0	Mini
20	1	2	1	Mini
21	1	4	0	Mini
22	0	1	3	Sports
23	2	3	0	Mini
24	0	5	0	Mini

Tabelle 4: Voting der Nearest Neighbours