Maschinelles Lernen: Symbolische Ansätze

TECHNISCHE UNIVERSITÄT DARMSTADT

Wintersemester 2008/2009
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 Lernen einzelner Regeln, Folie 16):*

1. *Expertenwissen: Ein Experte in der Domäne bewertet die Ausgaben des Klassifizieres, 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 kategorische 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

Betrachten wir nochmals den Beispieldatensatz aus der Vorlesung.

Temperature	Outlook	Humidity	Windy	PlayGolf?
hot	sunny	high	false	no
hot	sunny	high	true	no
hot	overcast	high	false	yes
cool	rain	normal	false	yes
cool	overcast	normal	true	yes
mild	sunny	high	false	no
cool	sunny	normal	false	yes
mild	rain	normal	false	yes
mild	sunny	normal	true	yes
mild	overcast	high	true	yes
hot	overcast	normal	false	yes
mild	rain	high	true	no
cool	rain	normal	true	no
mild	rain	high	false	yes

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

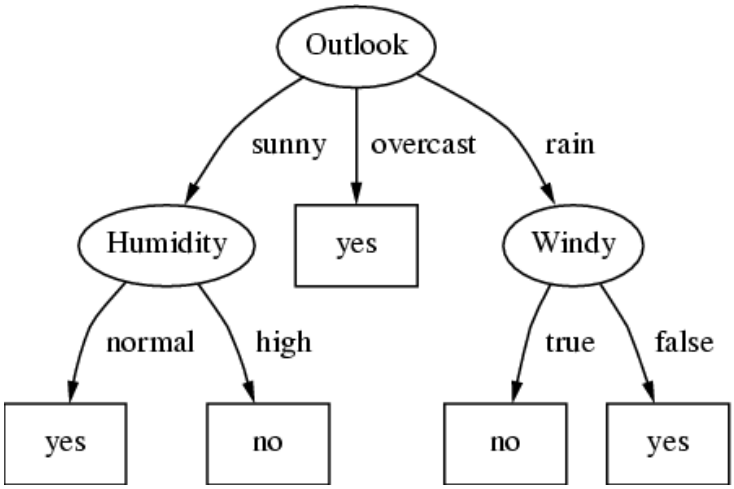
Temperature	Outlook	Humidity	Windy	PlayGolf?
hot	rain	high	true	yes
mild	sunny	normal	false	no
hot	rain	normal	false	yes
cool	overcast	high	true	yes
mild	rain	normal	true	no

Lösung:

Beispiel	Klassifizierung
1	no
2	yes
3	yes
4	yes
5	no

Errechnen Sie die Genauigkeit des Klassifizierers auf der Testmenge (die korrekt klassifizierten Beispiele geteilt durch die Gesamtzahl der Beispiele). Wie schätzen Sie die Güte des Klassifizierers ein? Begründen Sie Ihre Aussage.

Lösung: $Genauigkeit = \frac{3}{5} = 60\%$
Die Verteilung der positiven Beispiele in der Trainingsmenge ist: $\frac{9}{14} = 64,29\%$. *Hätte man also jedes Beispiel als positiv (yes) vorhergesagt, wäre man bereits besser gewesen.*



b) 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: *Die Tabelle der Trainingsdaten mit der jeweiligen Distanz zu den verschiedenen Beispielen:*

Temperature	Outlook	Humidity	Windy	PlayGolf?	Bsp.1	Bsp.2	Bsp.3	Bsp.4	Bsp.5
hot	sunny	high	false	no	2	2	2	3	4
hot	sunny	high	true	no	1	3	3	2	3
hot	overcast	high	false	yes	2	3	2	2	4
cool	rain	normal	false	yes	3	2	1	3	2
cool	overcast	normal	true	yes	3	3	3	1	2
mild	sunny	high	false	no	3	1	3	3	3
cool	sunny	normal	false	yes	4	1	2	3	3
mild	rain	normal	false	yes	3	1	1	4	1
mild	sunny	normal	true	yes	3	1	3	3	1
mild	overcast	high	true	yes	2	3	4	1	2
hot	overcast	normal	false	yes	3	2	1	3	3
mild	rain	high	true	no	1	3	3	2	1
cool	rain	normal	true	no	2	3	2	2	1
mild	rain	high	false	yes	2	2	2	3	2

Klassifikation mit der Nearest Neighbour Methode:

Beispiel	pos.Bsps.	neg.Bsps.	Klassifizierung
1	0	2	no
2	3	1	yes
3	3	0	yes
4	2	0	yes
5	2	2	yes

Wenn gleich viele positive und negative Beispiele die gleiche Distanz erhalten, wählt man die Klasse aus, die in der Trainingsmenge am häufigsten vorkommt.