# Maschinelles Lernen: Symbolische Ansätze

Übungsblatt für den 16.1.2007

## Aufgabe 1

Gegeben sei folgende Beispielmenge:

| Day | Outlook | Temperature | Humidity | Wind | PlayTennis |
|---|---|---|---|---|---|
| D1 | Sunny | 26 | High | | No |
| D2 | Sunny | 28 | High | Strong | No |
| D3 | Overcast | 29 | High | Weak | Yes |
| D4 | Rain | 23 | High | Weak | Yes |
| D5 | Rain | | Normal | Weak | Yes |
| D6 | Rain | 12 | Normal | Strong | No |
| D7 | Overcast | 8 | | Strong | Yes |
| D8 | Sunny | 25 | High | Weak | No |
| D9 | Sunny | 18 | Normal | Weak | Yes |
| D10 | Rain | 20 | Normal | Weak | Yes |
| D11 | Sunny | 20 | Normal | Strong | |
| D12 | Overcast | 21 | High | Strong | Yes |
| D13 | | 26 | Normal | Weak | Yes |
| D14 | Rain | 24 | High | Strong | No |
| D15 | Sunny | 23 | Normal | Weak | No |
| D16 | Sunny | 21 | Normal | Weak | Yes |

a) Überlegen Sie sich eine gute Abstandsfunktion für die einzelnen Attribute.

b) Benutzen Sie 3-NN zum Ausfüllen der fehlenden Werte.

Beziehen Sie hier die Klassifikation mit ein oder nicht? Warum?

c) Welchen Klassifikationswert gibt $k$-NN für die folgende Instanz aus?

1. Outlook=Sunny, Temperature=23, Humidity=High, Wind=Strong

Testen Sie verschiedene $k$. Für welches $k$ ändert sich die Klassifikation gegenüber $k = 1$?

d) Berechnen Sie den Klassifikationswert obiger Instanz mittels abstandsgewichtetem NN (Shepards Methode).

# Aufgabe 2

Ein Datenset enthält $2 \times n$ Beispiele, wobei genau $n$ Beispiele positiv sind und $n$ Beispiele negativ sind. Der einfache Algorithmus `ZeroRule` betrachtet nur die Klassenverteilung der Trainings-Daten und sagt für alle Beispiele die Klasse $+$ voraus, wenn mehr positive als negative Beispiele in den Trainings-Daten enthalten sind, und die Klasse $-$ falls es umgekehrt ist. Bei Gleichverteilung entscheidet er sich zufällig für eine der beiden Klassen, die er dann immer vorhersagt.

- Wie groß ist die Genauigkeit dieses Klassifizierers, wenn die Verteilung der Trainings-Daten der Gesamt-Verteilung entspricht (d.h., wenn die Trainings-Daten repräsentativ sind)?
- Schätzen Sie die Genauigkeit von `ZeroRule` mittels Leave-One-Out Cross-Validation ab.

# Aufgabe 3

Sie vergleichen zwei Algorithmen A und B auf 20 Datensets und beobachten folgende Genauigkeitswerte:

| Datenset | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Algorithm A | 0,91 | 0,86 | 0,93 | 0,74 | 0,65 | 0,91 | 0,87 | 0,95 | 0,78 | 0,86 |
| Algorithm B | 0,94 | 0,80 | 0,96 | 0,88 | 0,84 | 0,94 | 0,97 | 0,67 | 0,86 | 0,89 |
| Datenset | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Algorithm A | 0,98 | 0,96 | 0,74 | 0,53 | 0,95 | 0,67 | 0,98 | 0,96 | 0,97 | 0,91 |
| Algorithm B | 0,87 | 0,90 | 0,79 | 0,51 | 0,96 | 0,69 | 0,79 | 0,98 | 0,98 | 0,76 |

Läßt sich mit Hilfe des Vorzeichentests nachweisen, ob einer der beiden Algorithmen A oder B signifikant besser ist als der andere? Folgt daraus, daß er nicht besser ist?

# Aufgabe 4

Gegeben sei ein Datensatz mit 300 Beispielen, davon $2/3$ positiv und $1/3$ negativ.

a) Ist die Steigung der Isometrien für Accuracy im Coverage Space für dieses Problem $< 1$, $= 1$ und $> 1$?

b) Ist die Steigung der Isometrien für Accuracy im ROC Space für dieses Problem $< 1$, $= 1$ und $> 1$?

c) Sie verwenden einen Entscheidungsbaum, um die Wahrscheinlichkeit für die positive Klasse zu schätzen. Sie evaluieren drei verschiedene Thresholds $t$ (alle Beispiele mit einer geschätzten Wahrscheinlichkeit $> t$ werden als positiv, alle anderen als negativ klassifiziert) und messen folgende absolute Anzahlen von False Positives und False Negatives:

| t | fn | fp |
|---|---|---|
| 0.7 | 40 | 20 |
| 0.5 | 30 | 60 |
| 0.3 | 10 | 80 |

Geben Sie für jeden Threshold an, für welchen Bereich des Kostenverhältnisses $\frac{c(+|-)}{c(-|+)}$ der Threshold optimal ist.

d) Wie hoch ist die maximale Genauigkeit (Accuracy), die Sie im Szenario von Punkt c bei einer False Positive Rate von maximal 30% erreichen können? Wie gehen Sie dabei vor?

e) Sie erfahren, daß in Ihrer Anwendung ein False Positive 2 Cents kostet und ein False Negative 5 Cents kostet. Mit Welchem Threshold können Sie die Kosten minimieren? Wie hoch sind die entstanden minimalen Kosten für diese 300 Beispiele?

f) Sie bekommen die Möglichkeit, zusätzlich zu den vorhandenen 300 Beispielen noch 400 selbst auszuwählen. Wie würden Sie die Auswahl treffen, damit ein Lerner, der Kosten nicht berücksichtigen kann, unter den in Punkt e angegebenen Kosten möglichst effektiv wird?