# Vorlesung Machine Learning und Data Mining

Übungsblatt für den 8.12.2005

## Aufgabe 1

Gegeben sei eine Beispielmenge mit folgenden Eigenschaften:

- Die Datenmenge enthält 1000 Beispiele.
- Jedes Beispiel ist durch 10 nominale Attribute $A_1, \ldots, A_{10}$ beschrieben.
- Jedes dieser Attribute hat 10 Werte.

a) Wie viele Entscheidungsbäume müßten bei vollständiger Suche untersucht werden (Abschätzung der Größenordnung)?
(Hinweis: Dies ist analog zu der Frage: Wie viele Entscheidungsbäume gibt es für diese Daten?)

b) Wie viele (partielle) Entscheidungsbäume müssen maximal beim Verfahren des TDIDT untersucht werden?

c) Wie oft wird jedes Beispiel im Worst-Case angefaßt?

d) Was würde sich bei a) und b) ändern, wenn die Attribute nicht nominal, sondern numerisch wären (die sonstigen Annahmen bleiben gleich)?

# Aufgabe 2

Gegeben sei folgende Beispielmenge:

| Day | Outlook | Temperature | Humidity | Wind | PlayTennis |
|---|---|---|---|---|---|
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overcast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overcast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overcast | Mild | High | Strong | Yes |
| D13 | Overcast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |
| D15 | Sunny | Mild | Normal | Weak | No |

a) Erzeugen Sie einen Entscheidungsbaum mittels des Verfahrens ID3 (TDIDT mit Maß Gain).

Anmerkung: Hier bietet es sich an, in Gruppen zu arbeiten.

b) Wiederholen Sie die Berechnungen für die Auswahl des Tests in der Wurzel mit den Maßen Information-Gain-Ratio und Gini-Index. Ändert sich etwas?

c) Ersetzen Sie das Beispiel D1 durch

| Day | Outlook | Temperature | Humidity | Wind | PlayTennis |
|---|---|---|---|---|---|
| D1 | ? | Hot | High | Weak | No |

(Attributwert für Outlook unbekannt) und berechnen Sie nun den Gain des Attributs Outlook, indem Anteile der Beispiele in die Teilbäume propagiert werden (siehe letzter Punkt auf Folie 34).