Gottfried Wilhelm Leibniz Universität Hannover
Institut für Praktische Informatik
Fachgebiet Datenbanken und Informationssysteme
Prof. Dr. Udo Lipeck/M.Sc. Oliver Pabst

03.06.–10.06.2015

Übung zur Vorlesung „Data Mining“ im Sommersemester 2015

# Übungsblatt 5

**Aufgabe 1** *(Klassifikation mit Entscheidungsbäumen)*

Gegeben sei folgende Datenbank aus der Vorlesung:

| Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|
| Yes | Single | 125000 | No |
| No | Married | 100000 | No |
| No | Single | 70000 | No |
| Yes | Married | 120000 | No |
| No | Divorced | 95000 | Yes |
| No | Married | 60000 | No |
| Yes | Divorced | 220000 | No |
| No | Single | 85000 | Yes |
| No | Married | 75000 | No |
| No | Single | 90000 | Yes |

Wie in der Vorlesung soll für diese Daten ein Entscheidungsbaum zum Klassen-Attribut „Cheat“ kontruiert werden.

Für die nominalen Attribute „Refund“ und „Marital Status“ sollen alle möglichen binären Aufteilungen (Splits) der Form *Attribut* = *Wert* bzw. *Attribut* $\neq$ *Wert* in Betracht gezogen werden. Für das kontinuierliche Attribut „Taxable Income“ soll jede binäre Aufteilung *Attribut* $>$ *Wert* bzw. *Attribut* $\leq$ *Wert* möglich sein.

Benutzen Sie jeweils folgende Kriterien um die besten Splits zu bestimmen:

a) das Gini-Kriterium

b) den Missklassifikationsfehler

Ein Split soll innvollerweise nur dann durchgeführt werden, wenn sich das entsprechende Kriterium verbessert.

*bitte wenden*

**Aufgabe 2** *(Split-Kriterien für Entscheidungsbäume)*

Gegeben sei folgende Datenbank mit den binären Attributen A, B und C.

| A | B | C |
|---|---|---|
| ja | nein | 1 |
| ja | ja | 1 |
| ja | ja | 1 |
| ja | nein | 2 |
| ja | ja | 1 |
| nein | nein | 2 |
| nein | nein | 2 |
| nein | nein | 2 |
| ja | ja | 2 |
| ja | nein | 2 |

Zur Klassifikation nach dem Attribut C soll für diese Datenbank ein Entscheidungsbaum aufgebaut werden. Eine erste Aufteilung kann entweder nach Attribut A oder B geschehen. Verwenden Sie jeweils folgende Kriterien um zu entscheiden welche die sinnvollere ist:

a) Gini-Kriterium
b) Information Gain
c) Gain Ratio

Zusatzfrage: Die Werte des Gini- und Entropie-Maßes für eine Wahrscheinlichkeit $p$ wachsen jeweils monoton für $0 < p < 0.5$ und fallen für $0.5 < p < 1$ (vgl. Bild auf Folie 43 aus der Vorlesung). Ist es trotzdem möglich, dass Information Gain und Gini-Kriterium zur Wahl *verschiedener* Aufteilungen führen?

Abgaben zu Beginn nächsten Vorlesung am Mittwoch, 10.06.2015