Blatt: DTL

Michel Bünger

DTL.01: Entscheidungsbäume mit CAL3 und ID3

CAL3

```
erster Durchlauf:
- - - - - - - - - - - - - - - - - - - - - - - - - - -

Alter(
    ≥35(O:1)
    <35: *
)

- - - - - - - - - - - - - - - - - - - - - - - - - - -

Alter(
    ≥35(O:1)
    <35: (O:1)
)

- - - - - - - - - - - - - - - - - - - - - - - - - - -

Alter(
    ≥35(O:1, M:1)
    <35: (O:1)
)

- - - - - - - - - - - - - - - - - - - - - - - - - - -

Alter(
    ≥35(O:1, M:2)
    <35: (O:1)
)

- - - - - - - - - - - - - - - - - - - - - - - - - - -
```

```
Alter(
    $\geq$35(O:2, M:2) $\rightarrow$ $S_1$ erreicht, $\frac{2}{4} < S_2$;
       Differenzierung
    <35: (O:1)
)

- - - - - - - - - - - - - - - - - - - - - - - - - - -

Alter(
    $\geq$35(
        hoch: (M:1 O:2)
        niedrig:(M:1)
    )

    <35: (O:2)
)

- - - - - - - - - - - - - - - - - - - - - - - - - - -

Alter(
    $\geq$35(
        hoch: (M:1 O:2)
        niedrig:(M:1)
    )

    <35: (O:2, M:1)
)

- - - - - - - - - - - - - - - - - - - - - - - - - - -
```

Erster Durchlauf abgeschlossen, Blätter nicht klar definiert.

zweiter Durchlauf:

```
- - - - - - - - - - - - - - - - - - - - - - - - - - -

Alter(
    $\geq$35(
        hoch: (M:1 O:3) $\rightarrow$ $S_1$ erreicht, $\frac{3}{4} > S_2$; Dominanz
            von O
        niedrig:(M:1)
    )

    <35: (O:2, M:1)
)

- - - - - - - - - - - - - - - - - - - - - - - - - - -
```

```
Alter(
    ≥35(
        hoch: O
        niedrig:(M:1)
    )

    <35: (O:3, M:1) → S_1 erreicht, 3/4 > S_2; Dominanz von O
)

- - - - - - - - - - - - - - - - - - - - - - - - - - -


Alter(
    ≥35(
        hoch: O ← Klasse M wird als Fehler klassifiziert
        niedrig:(M:1)
    )

    <35: O
)

- - - - - - - - - - - - - - - - - - - - - - - - - - -

Alter(
    ≥35(
        hoch: O
        niedrig:(M:2)
    )

    <35: O
)

- - - - - - - - - - - - - - - - - - - - - - - - - - -

...

Für Alter(≥ 35(hoch:O)) und Alter(<35: O) kann keine Differenzierung mehr geschehen.
Für Alter(≥35(niedrig: (M: 2))), wird im viertem Durchlauf M: 4 gelten und somit S_1 erreicht, 4/4 > S_2; Dominanz von M.
```

```
...

finaler Ergebnisbaum:

- - - - - - - - - - - - - - - - - - - - - - - - - - - -

Alter(
    ≥35(
        hoch: O
        niedrig:M
    )

    <35: O
)

- - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

ID3

Berechnen der Entropie der Trainingsmenge:

$H(S) = \frac{4}{7}log_2(\frac{4}{7}) - \frac{3}{7}log_2(\frac{3}{7})$

$H(S) = 0.985$ Bit

Berechnung der Informationsgewinne aller Attribute:

Attribut: Alter

- ≥ 35: O:2 M:2. $H(\geq 35) = -\frac{2}{4}log_2(\frac{2}{4}) - \frac{2}{4}log_2(\frac{2}{4}) = 1$
- < 35: O:2 M:1. $H(< 35) = -\frac{2}{3}log_2(\frac{2}{3}) - \frac{1}{3}log_2(\frac{1}{3}) \approx 0.918$

Gewichtete Entropie von Alter:
$H(S, Alter) = \frac{4}{7}H(\geq 35) + \frac{3}{7}H(< 35) \approx 0.965$

Informationsgewinn für Alter:
$Gain(S, Alter) = H(S) - H(S, Alter) = 0.02$

Attribut: Einkommen

- *hoch*: O:3 M:1. $H(hoch) = -\frac{3}{4}log_2(\frac{3}{4}) - \frac{1}{4}log_2(\frac{1}{4}) \approx 0.811$
- *niedrig*: O:1 M:2. $H(niedrig) = -\frac{1}{3}log_2(\frac{1}{3}) - \frac{2}{3}log_2(\frac{2}{3}) \approx 0.918$

Gewichtete Entropie von Einkommen:
$H(S, Einkommen) = \frac{4}{7}H(hoch) + \frac{3}{7}H(niedrig) \approx 0.857$

Informationsgewinn für Alter:
$Gain(S, Einkommen) = H(S) - H(S, Einkommen) \approx 0.128$

Attribut: Bildung

- *Abitur*: O:1 M:2. $H(Abitur) = -\frac{1}{3}log_2(\frac{1}{3}) - \frac{2}{3}log_2(\frac{2}{3}) \approx 0.918$
- *Bachelor*: O:1 M:1. $H(Bachelor) = -\frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2}) = 1$
- *Master*: O:2 M:0. Rein! $H(Master) = 0$

Gewichtete Entropie von Bildung:
$H(S, Bildung) = \frac{3}{7}H(Abitur) + \frac{2}{7}H(Bachelor) + \frac{2}{7}H(Master) \approx 0.679$

Informationsgewinn für Alter:
$Gain(S, Bildung) = H(S) - H(S, Bildung) \approx 0.306$
Der Größte Informationsgewinn wird von Bildung versprochen, weshalb dies das erste Attribut sein wird, das behandelt wird.

Untermenge: Master

Für die Untermenge 'Master' gehören alle Instanzen zu der Klasse O, was Bedeutet, dass 'Master' rein ist und mit dem Blattknoten 'O' abschließt.

0.0.1 Untermenge: Abitur

Die Berechnungen der neuen Entropien zeigt hier, dass als nächstes Attribut, Einkommen den höchsten Informationsgewinn besitzt.

Zudem ist dann zu sehen, dass für Abitur(Einkommen(hoch(O:1 M:0))) und Abitur(Einkommen(niedrig(O:0 M:2))) gilt, d.h. beide sind rein und Blattknoten von jeweils 'O' und M'

Untermenge: Bachelor

Die Berechnungen der neuen Entropien zeigt hier, dass als nächstes Attribut, Alter den höchsten Informationsgewinn besitzt.

Zudem ist dann zu sehen, dass für Bachelor(Alter($\geq$ 35(O:0 M:1))) und Bachelor(Alter(¡35(O:1 M:0))) gilt, d.h. beide sind rein und Blattknoten von jeweils 'M' und 'O'

fertiger Ergebnisbaum:

```
- - - - - - - - - - - - - - - - - - - - - - - - - - -

Bildung(

    Abitur(
        Einkommen(
            hoch: O
            niedrig: M
        )
    )
    Bachelor(
        Alter(
            ≥35: M
            <35: O
        )
    )

    Master: O
)

- - - - - - - - - - - - - - - - - - - - - - - - - - -
```

DTL.02: Pruning

Der gegebene Baum:

- -

x_3(

    x_2(
        x_1(C, A),
        x_1(B, A)
    ),

    x_1(
        x_2(C, B),
        A
)

)

- -

Lässt sich Kürzen, indem man die Allgemeine Transformationsregel:

$x_1(x_2(a, b), x_2(c, d)) \leftrightarrow x_2(x_1(a, c), x_1(b, d))$

Auf den ersten $x_2(...)$ in $x_3(x_2(...)...)$ anwendet.
Daraus entsteht:

- -

x_3(

    x_1(
        x_2(C, B),
        x_2(A, A) $\rightarrow$ A
    ),

    x_1(
        x_2(C, B),
        A
)

)

- -

Nun sind beide $x_1(...)$ in $x_3(x_1(...), x_1(...))$ gleich, weshalb sich der Baum letztlich ein letztes mal kürzen lässt:

- -

x_3(

x_1(
x_2(C, B),
x_2(A, A) $\rightarrow$ A
)

)

- -

Also ist der gekürzte Baum: $x_3(x_1(x_2(C, B), A))$