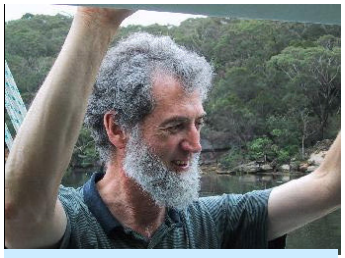
Entscheidungsbaum-Lernen (Ein Mini-Beispiel)



J. Ross Quinlan, australischer Informatiker; entwickelte die ID3- und C4.5-Algorithmen sowie weitere Data Mining Verfahren. Gründer von Rulequest Research, Austr.

Dichotomie bedeutet die Aufteilung von etwas in zwei Strukturen oder zwei Begriffe. Im mathematisch-logischen Sinne schließen sich die neuen Begriffe aus.
Bsp.: Die Aufteilung der Zahlen in rationale und irrationale Zahlen.

- Damit haben wir alle notwendigen Vorbereitungen für das Berechnen von Entscheidungsbäumen getroffen.
- Man betrachtet z.B. Y als Klassifikationsattribut und X als Informationsattribut.
- Für jedes Informationsattribut wird nun der Informationsgewinn IG berechnet. Dann wird das Attribut mit dem höchsten Informationsgewinn als Knoten im Baum ausgewählt. Der Restbaum wird mit den noch verfügbaren Informationsattributen in rekursiver Weise bestimmt.
- Das ist das Grundprinzip des ID3-Algorithmus von J. Ross Quinlan. ID steht dabei für *Interactive Dichotomizer*.
- Der Entscheidungsbaum kann verbessert werden, wenn man anstelle des IG das Gain-Ratio-Maß verwendet. In Nachverarbeitungsschritten kann der entstandene Entscheidungsbaum dann noch „geprunt" (abgeschnitten/beschnitten) werden.
- Darauf basiert der Algorithmus C4.5, der eine Verbesserung des ID3-Algorithmus darstellt.

Entscheidungsbaum-Lernen

- Ein Beispiel:

  Die Spalten S1 und S2 beinhalten z.B. den Anteil zweier Substanzen in einer Blutprobe. Wir berechnen den IG der beiden Variablen bzgl. der Klassenbestimmung.

Blutproben-ID	S1	S2	Klasse
1	hoch	niedrig	krank
2	hoch	niedrig	krank
3	niedrig	niedrig	gesund
4	hoch	hoch	krank
5	hoch	hoch	gesund
6	hoch	niedrig	krank
7	niedrig	niedrig	gesund
8	niedrig	hoch	krank
9	niedrig	niedrig	gesund
10	niedrig	niedrig	gesund

- IG(Klasse| S1)
  = H(Klasse) – H(Klasse| S1)

g ≡ gesund
k ≡ krank

- H(Klasse)
  $= -p(g)\cdot\log_2 p(g) - p(k)\cdot\log_2 p(k)$
  $= -5/10\cdot\log_2(5/10) - 5/10\cdot\log_2(5/10)$
  $= -½\cdot(-1) - ½\cdot(-1) = 1$ [Sh]

	niedrig	hoch
S1	4 g 1 k	1 g 4 k
S2	4 g 3 k	1 g 2 k

- H(Klasse| S1) = 0,72 [Sh]
- H(Klasse| S2) = 0,95 [Sh]

(Berechnung s. näch. Seite)

- Insgesamt erhalten wir

  IG(Klasse|S1) = H(Klasse) – H(Klasse| S1) = 1 – 0,72 = 0,28 [Sh]
  IG(Klasse|S2) = H(Klasse) – H(Klasse| S2) = 1 – 0,96 = 0,04 [Sh]

- Wir wählen daher das Attribut S1 als ersten Entscheidungsknoten aus.

Entscheidungsbaum-Lernen

$\log_2 x = \log_{10} x / \log_{10} 2$

Bsp.: **3** = $\log_2 8 = \log_{10} 8 / \log_{10} 2$
= 0,90309 /0,30103 = 3

Blutproben-ID	S1	S2	Klasse
1	hoch	niedrig	krank
2	hoch	niedrig	krank
3	niedrig	niedrig	gesund
4	hoch	hoch	krank
5	hoch	hoch	gesund
6	hoch	niedrig	krank
7	niedrig	niedrig	gesund
8	niedrig	hoch	krank
9	niedrig	niedrig	gesund
10	niedrig	niedrig	gesund

g ≡ gesund
k ≡ krank

	nied-rig	hoch
S1	4 g 1 k	1 g 4 k
S2	4 g 3 k	1 g 2 k

- Berechnung von H(Klasse|S1):

- H(Klasse| S1=hoch) = $-1/5 \cdot \log_2(1/5) - 4/5 \cdot \log_2(4/5)$
  = $-0.2 \cdot (-2{,}32) - 0.8 \cdot (-0{,}32)$
  = 0,46 + 0,26 = 0,72
- H(Klasse| S1=nied.) = $-1/5 \cdot \log_2(1/5) - 4/5 \cdot \log_2(4/5)$ = 0,72
- H(Klasse| S1) = p(S1=„hoch“) · H(K|S1=„hoch“)
  +p(S1=„nied.“) · H(K|S1=„nied“)
  = 5/10 · 0,72 + 5/10 · 0,72 = 0,72

Entscheidungsbaum-Lernen

Blutproben-ID	S1	S2	Klasse
1	hoch	niedrig	krank
2	hoch	niedrig	krank
3	niedrig	niedrig	gesund
4	hoch	hoch	krank
5	hoch	hoch	gesund
6	hoch	niedrig	krank
7	niedrig	niedrig	gesund
8	niedrig	hoch	krank
9	niedrig	niedrig	gesund
10	niedrig	niedrig	gesund

g ≡ gesund
k ≡ krank

	nied-rig	hoch
S1	4 g 1 k	1 g 4 k
S2	4 g 3 k	1 g 2 k

- Berechnung von H(Klasse|S2):

- H(Klasse| S2=hoch) $= -1/3 \cdot \log_2(1/3) - 2/3 \cdot \log_2(2/3)$
  $= -0.33 \cdot (-1{,}58) - 0.67 \cdot (-0{,}58)$
  $= 0{,}52 + 0{,}39 = 0{,}91$
- H(Klasse| S2=nied.) $= -4/7 \cdot \log_2(4/7) - 3/7 \cdot \log_2(3/7)$
  $= 0{,}57 \cdot 0{,}81 + 0{,}43 \cdot 1{,}22$
  $= 0{,}46 + 0{,}52 = 0{,}98$
- H(Klasse| S2) $= p(S2 = \text{„hoch“}) \cdot H(K|\ S2 = \text{„hoch“})$
  $+ p(S2 = \text{„nied.“}) \cdot H(K|\ S2 = \text{„nied“})$
  $= 3/10 \cdot 0{,}91 + 7/10 \cdot 0{,}98$
  $= 0{,}27 + 0{,}69 = 0.96$

Berechnung IG:

- IG(Klasse|S1) $= 1 - 0{,}72 = 0{,}28$
- IG(Klasse|S2) $= 1 - 0.96 = 0{,}04$

➔ Wähle S1