Decision Trees

Implementierungstechniken für
Hauptspeicherdatenbanksysteme

Dominik Vinan

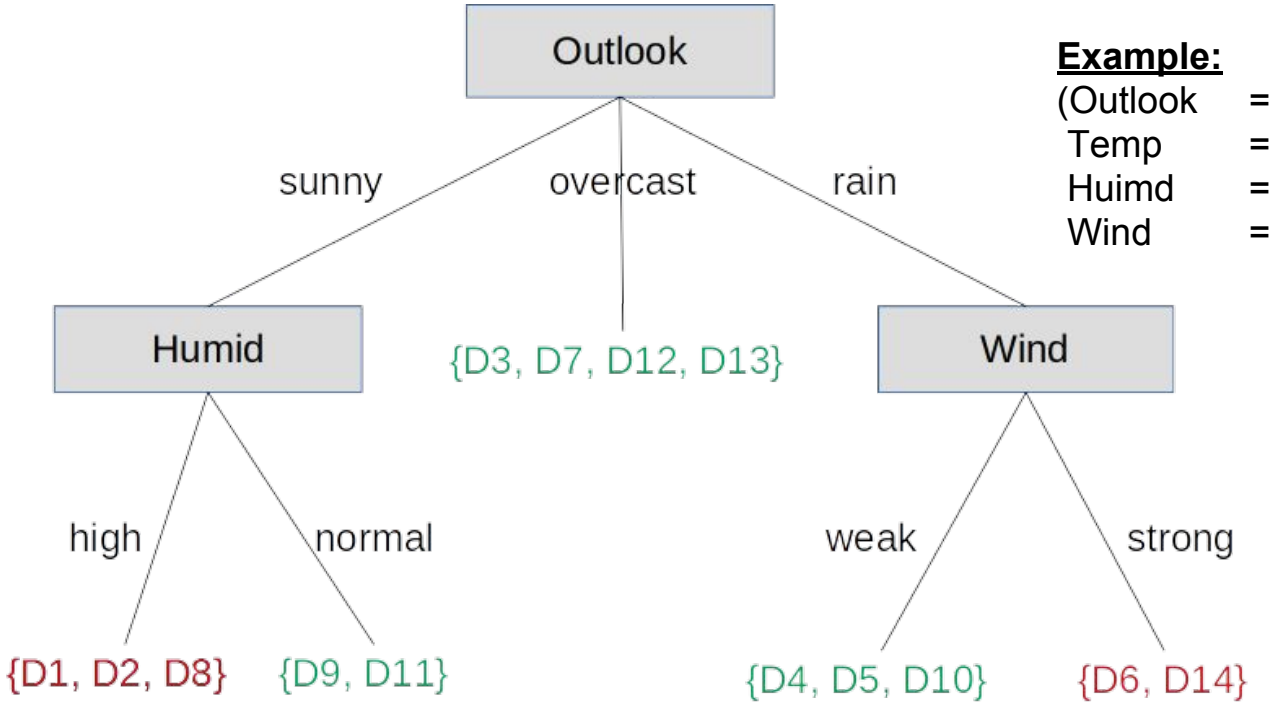
Outline

- Introduction
- CART/TDIDT
  - ID3
  - C4.5
- Advantages & Disadvantages
- Implementation ID3

Example Database

Day	Outlook	Temp	Humid	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

Introduction



Example:

(Outlook = sunny,
Temp = mild,
Huimd = normal,
Wind = weak)

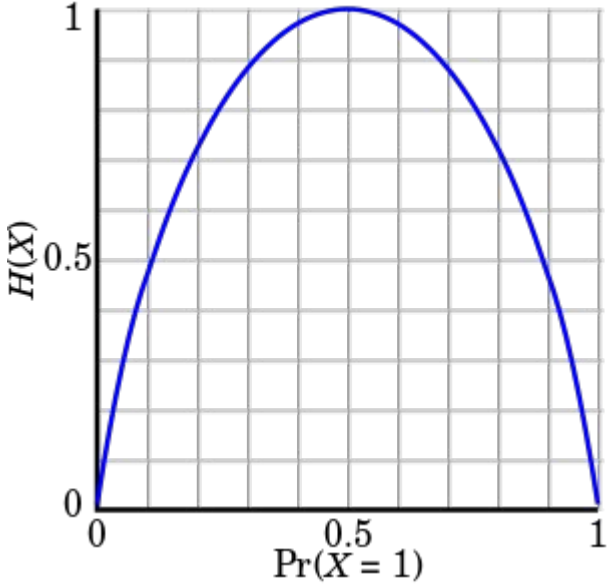
CHAID & CART/TDIDT

- Top-Down Induction of Decision Trees (CART/TDIDT)
  - Non-Incremental approach i.e. needs to start over after change of training data.
  - Needs Pruning as Trees can become overly complex.
  - Examples: CART(Algorithm), ID3, C4.5, C5.0.
- Chi-square Automatic Interaction Detectors (CHAID)
  - Main Difference to CART:
    - Tree growth is limited => avoids pruning.

Information

- Im/purity (Entropy)
- Information Gain
  - Expected gain of information after splittng.

gain(A) = I(p, n) - H(A)



https://en.wikipedia.org/wiki/Entropy_(information_theory)#/media/File:Binary_entropy_plot.svg

Classification- vs. Regression Trees

Classification Tree:

- Classifies categorial target values.
- E.g.: Response {‘Yes’, ‘No’}.

Regression Tree:

- Finds splitting value for continuous target values.
- E.g.: ExpectedTemp { 20.5, 10.7, 30.0, 17.7, ... }.

=> Predictors can be either numeric or categorial.

Iterative Dichotomiser 3 (ID3)

- By J. Ross Quinlan
- Selection criterion: **Information Gain** or **Gain Ratio.**
- Information Gain
  - Bits of gained information.
  - Has a bias towards variables with multiple values.
- Gain Ratio
  - Takes number and size of branches into account.
  - Is not always defined.

Example Database

Day	Outlook	Temp	Humid	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

ID3 Pseudocode

1. Select primary key, target attribute, and the dataset
2. **If not (pure or stopping criterion met)**: Call **ID3**
    a. Calculate **Entropy** & **Information Gain** for every attribute.
    b. Select attribute **X** with **MAX(Information Gain)**.
    c. **Make a tree node** using attribute **X**.
    d. **Split dataset into subsets** for every value of **X**.
    e. **Recursive** call of ID3 for every subset.

Example calculation ID3 (Root)

$$i(Sunny) = \frac{5}{14} \cdot \left(-\frac{2}{5} \cdot log(\frac{2}{5}) - \frac{3}{5} \cdot log(\frac{3}{5}) \right) = 0.346768$$

$$i(Rain) = \frac{5}{14} \cdot \left(-\frac{3}{5} \cdot log(\frac{3}{5}) - \frac{2}{5} \cdot log(\frac{2}{5}) \right) = 0.346768$$

$$i(Overcast) = \frac{4 \cdot (-1 \cdot log(1))}{14} = 0$$

$$E(Outlook, D) = (i(Sunny) + i(Overcast) + i(Rain)) = 0.693486$$

$$gain(Outlook) = I(D) - E(Outlook, D) = 0.940286 - 0.693486 = 0.2468$$

$$gain(Humid) = I(D) - E(Humid, 0) = 0.1518$$

$$gain(Wind) = I(D) - E(Wind, 0) = 0.0481$$

$$gain(Temp) = I(D) - E(Temp, 0) = 0.0292$$

C4.5/C5.0

- Improvements over ID3
  - Handles continuous and categorial variables.
  - Handles missing values.
  - More efficient pruning.

  => C4.5 makes ID3 applicable in practice.

- C5.0 - Commercial implementation
  - Increased performance.
  - Less memory.
  - More precise.

Advantages & Disadvantages

+ Human readable rules.
+ Limited computation power (for application).
+ Handles continuous and categorial values.

- Not optimal for predicting specific values.
- Growing and pruning trees is computationally complex.
- Inefficient for non-rectangular regions.

Implementation

Day	Overcast	Temp	Humid	Wind	Day
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
...	...	...	...	...	...

Thank You!