Chair of Decision Sciences and Systems
TUM School of Computation, Information and Technology
Technical University of Munich

Business Analytics & Machine Learning
Tutorial sheet 5: Decision trees – Solution

Prof. Dr. Martin Bichler, Prof. Dr. Jalal Etesami
Julius Durmann, Markus Ewert, Johannes Knörr, Yutong Chao

Exercise T5.1 *Entropy and Information*

Compute the following:

a) entropy((0.1, 0.9))

b) entropy((0.8, 0.2))

c) entropy((0.5, 0.5))

d) entropy((0.8, 0.1, 0.1))

e) info([2, 3])

f) info([5, 4])

g) info([2, 3], [5, 4])

h) info([2, 3], [9, 0])

Solution

a) $\text{entropy}((0.1, 0.9)) = -0.1 \cdot \log_2(0.1) - 0.9 \cdot \log_2(0.9) \approx 0.469$

b) $\text{entropy}((0.8, 0.2)) = -0.8 \cdot \log_2(0.8) - 0.2 \cdot \log_2(0.2) \approx 0.722$

c) $\text{entropy}((0.5, 0.5)) = -0.5 \cdot \log_2(0.5) - 0.5 \cdot \log_2(0.5) \approx 1.0$

d) $\text{entropy}((0.8, 0.1, 0.1)) = -0.8 \cdot \log_2(0.8) - 0.1 \cdot \log_2(0.1) - 0.1 \cdot \log_2(0.1) \approx 0.922$

e) $\text{info}([2, 3]) = \text{entropy}((\frac{2}{2+3}, \frac{3}{2+3})) \approx 0.971$

f) $\text{info}([5, 4]) = \text{entropy}((\frac{5}{5+4}, \frac{4}{5+4})) \approx 0.991$

g) $\text{info}([2, 3], [5, 4]) = \frac{2+3}{(2+3)+(5+4)} \cdot \text{info}([2, 3]) + \frac{5+4}{(2+3)+(5+4)} \cdot \text{info}([5, 4]) \approx 0.984$

h) $\text{info}([2, 3], [9, 0]) = \frac{2+3}{(2+3)+(9+0)} \cdot \text{info}([2, 3]) + \frac{9+0}{(2+3)+(9+0)} \cdot \text{info}([9, 0]) \approx 0.347$

Exercise T5.2 *Optimal splits*

Compute the optimal splits for the following data using info gain.

a)

35	35	37	40	40	40
F	F	T	F	T	T

b)

36.8	36.8	37.2	38.3	38.3	39.7
T	F	F	T	F	F

Solution

Choose a split, which maximizes the gain, or equivalently, minimizes information after split:

a) Choosing the split point s yields two distinct cases:
   - Split point $s \in (35, 37) \rightarrow \text{info}([0, 2], [3, 1]) \approx 0.541$
   - Split point $s \in (37, 40) \rightarrow \text{info}([1, 2], [2, 1]) \approx 0.918$

   $\rightarrow$ w.l.o.g. optimal split at $s = 36$.

b) Choosing the split point s yields three distinct cases:
   - Split point $s \in (36.8, 37.2) \rightarrow \text{info}([1, 1], [1, 3]) \approx 0.874$
   - Split point $s \in (37.2, 38.3) \rightarrow \text{info}([1, 2], [1, 2]) \approx 0.918$
   - Split point $s \in (38.3, 39.7) \rightarrow \text{info}([2, 3], [0, 1]) \approx 0.809$

   $\rightarrow$ optimal split for e.g. $s = 39$.

Exercise T5.3 *Construct decision tree based on gini index*

Past Trend	Open Interest	Trading volume	Return
Positive	Low	High	Up
Negative	High	Low	Down
Positive	Low	High	Up
Positive	High	High	Up
Negative	Low	Low	Down
Positive	Low	High	Down
Negative	High	High	Down
Negative	Low	Low	Down
Positive	Low	High	Down
Positive	High	High	Up

In order to construct the decision tree for variable Return,

a) Compute the Gini index of Past Trend, the Gini index of Open Interest, and the Gini index of Trading Volume.

b) Choose the criterion of the decision tree at depth 1 by using Gini index

Solution

a) Calculating the Gini Index for past trend:

We denote $S_1 = \{x|\text{Past Trend}(x) = \text{Positive}\}$ and $S_2 = \{x|\text{Past Trend}(x) = \text{Negative}\}$.

$$\frac{|S_1|}{|S|} = \frac{3}{5}, \ \frac{|S_2|}{|S|} = \frac{2}{5}.$$

$$\frac{|\{x \in S_1|\text{Return}(x) = \text{Up}\}|}{|S_1|} = \frac{2}{3}, \ \frac{|\{x \in S_1|\text{Return}(x) = \text{Down}|\}}{|S_1|} = \frac{1}{3}.$$

$$\mathcal{G}(S_1) = 1 - (\frac{|\{x \in S_1|\text{Return}(x) = \text{Up}\}|}{|S_1|})^2 - (\frac{|\{x \in S_1|\text{Return}(x) = \text{Down}|\}}{|S_1|})^2 = \frac{4}{9}.$$

$$\frac{|\{x \in S_2|\text{Return}(x) = \text{Up}\}|}{|S_2|} = 0, \ \frac{|\{x \in S_2|\text{Return}(x) = \text{Down}|\}}{|S_2|} = 1.$$

$$\mathcal{G}(S_2) = 1 - (\frac{|\{x \in S_2|\text{Return}(x) = \text{Up}\}|}{|S_2|})^2 - (\frac{|\{x \in S_2|\text{Return}(x) = \text{Down}|\}}{|S_2|})^2 = 0.$$

$$\implies \mathcal{G}(S_1, S_2) = \frac{|S_1|}{S} * \mathcal{G}(S_1) + \frac{|S_2|}{S} * \mathcal{G}(S_2) = \frac{4}{15}.$$

Calculating the Gini Index for Open Interest:
We denote $S_1 = \{x|\text{Open Interest}(x) = \text{High}\}$ and $S_2 = \{x|\text{Open Interest}(x) = \text{Low}\}$.

$$\implies \mathcal{G}(S_1, S_2) = \frac{7}{15}.$$

Calculating the Gini Index for Trading Volume:
We denote $S_1 = \{x|\text{Trading Volume}(x) = \text{High}\}$ and $S_2 = \{x|\text{Trading Volume}(x) = \text{Low}\}$.

$$\implies \mathcal{G}(S_1, S_2) = \frac{12}{35}.$$

b) Since Past Trend has the lowest Gini Index, it is the criterion of the decision tree at depth 1