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

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

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

Exercise H5.1 *Soccer results*

Host	Better Form	Referee's Preference	Tradition	Result
A	B	B	4	X
A	A	None	4	A
B	A	B	1	B
B	A	None	3	X
A	B	None	1	B
A	B	None	2	X
B	A	B	2	B
B	Same	None	1	B
A	Same	None	5	A
A	B	None	5	A
B	Same	None	4	A
B	Same	A	3	A
A	Same	A	3	A
A	B	None	3	A

Construct the first two levels of the decision tree using gain ratio.

Note: Tradition is a numerical attribute. You need to split it using a binary split. In order to construct the root, use 2.5 as the value for the split point. If necessary, find the optimal split point for the second level. The attribute Tradition indicates how many games team A won, out of the last six games.

Note: A and B are teams. The value "Same" indicates that both teams are in equally good form. X means that the game resulted in a draw.

Solution

Notation: [A, X, B] Entropy for complete dataset: info([7, 3, 4]) $\approx$ 1.493

depth = 1:

- gainRatio(Host) = $\frac{\text{gain(Host)}}{\text{intrinsic\_info(Host)}} = \frac{\text{info}([7,3,4]) - \text{info}([5,2,1],[2,1,3])}{\text{info}([8,6])} \approx 0.127$
- gainRatio(Better Form) = $\frac{\text{gain(Better Form)}}{\text{intrinsic\_info(Better Form)}} = \frac{\text{info}([7,3,4]) - \text{info}([1,1,2],[4,0,1],[2,2,1])}{\text{info}([4,5,5])} \approx 0.167$
- gainRatio(Referee's Preference) = $\frac{\text{gain(Ref's Preference))}}{\text{intrinsic\_info(Ref's Preference))}} = \frac{\text{info}([7,3,4]) - \text{info}([2,0,0],[5,2,2],[0,1,2])}{\text{info}([2,9,3])} \approx 0.290$
- gainRatio(Tradition, split 2.5) = $\frac{\text{gain(Tradition, split 2.5)}}{\text{intrinsic\_info(Tradition, split 2.5)}} = \frac{\text{info}([7,3,4]) - \text{info}([0,1,4],[7,2,0])}{\text{info}([5,9])} \approx 0.791$

$\longrightarrow$ first split (Tradition ≤ 2.5) yielding following tree:

depth = 2, Tradition ≤ 2.5, (left subtree):

Host	Better Form	Referee's Preference	Tradition	Result
B	A	B	1	B
A	B	None	1	B
A	B	None	2	X
B	A	B	2	B
B	Same	None	1	B

- gainRatio(Host) = $\frac{\text{gain(Host)}}{\text{intrinsic\_info(Host)}} = \frac{\text{info}([0,1,4]) - \text{info}([0,1,1],[0,0,3])}{\text{info}([2,3])} \approx 0.322$
- gainRatio(Better Form) = $\frac{\text{gain(Better Form)}}{\text{intrinsic\_info(Better Form)}} = \frac{\text{info}([0,1,4]) - \text{info}([0,0,2],[0,0,1],[0,1,1])}{\text{info}([2,1,2])} \approx 0.212$
- gainRatio(Referee's Preference) = $\frac{\text{gain(Ref's Preference))}}{\text{intrinsic\_info(Ref's Preference))}} = \frac{\text{info}([0,1,4]) - \text{info}([0,1,2],[0,0,2])}{\text{info}([3,2])} \approx 0.176$
- gainRatio(Tradition, split 1.5) = $\frac{\text{gain(Tradition, split 1.5)}}{\text{intrinsic\_info(Tradition, split 1.5)}} = \frac{\text{info}([0,1,4]) - \text{info}([0,0,3],[0,1,1])}{\text{info}([3,2])} \approx 0.322$

$\longrightarrow$ two equally good choices for left sub-tree split (Host, or Tradition ≤ 1.5).
W.l.o.g. choose (Tradition ≤ 1.5), yielding following tree: depth = 2, Tradition > 2.5, (right subtree):

Host	Better Form	Referee's Preference	Tradition	Result
A	B	B	4	X
A	A	None	4	A
B	A	None	3	X
A	Same	None	5	A
A	B	None	5	A
B	Same	None	4	A
B	Same	A	3	A
A	Same	A	3	A
A	B	None	3	A

- gainRatio(Host) = $\frac{\text{gain(Host)}}{\text{intrinsic\_info(Host)}} = \frac{\text{info}([7,2,0]) - \text{info}([5,1,0],[2,1,0])}{\text{info}([6,3])} \approx 0.027$
- gainRatio(Better Form) = $\frac{\text{gain(Better Form)}}{\text{intrinsic\_info(Better Form)}} = \frac{\text{info}([7,2,0]) - \text{info}([1,1,0],[4,0,0],[2,1,0])}{\text{info}([2,4,3])} \approx 0.154$
- gainRatio(Referee's Preference) = $\frac{\text{gain(Ref's Preference))}}{\text{intrinsic\_info(Ref's Preference))}} = \frac{\text{info}([7,2,0]) - \text{info}([2,0,0],[5,1,0],[0,1,0])}{\text{info}([2,6,1])} \approx 0.270$
- gainRatio(Tradition, split 3.5) = $\frac{\text{gain(Tradition, split 3.5)}}{\text{intrinsic\_info(Tradition, split 3.5)}} = \frac{\text{info}([7,2,0]) - \text{info}([3,1,0],[4,1,0])}{\text{info}([4,5])} \approx 0.002$
- gainRatio(Tradition, split 4.5) = $\frac{\text{gain(Tradition, split 4.5)}}{\text{intrinsic\_info(Tradition, split 4.5)}} = \frac{\text{info}([7,2,0]) - \text{info}([5,2,0],[2,0,0])}{\text{info}([7,2])} \approx 0.122$

$\longrightarrow$ right sub-tree split (Referee's Preference) yielding following tree:

Exercise H5.2 *Winter sports*

ID	Temperature	Visibility	Snow Depth	Sport
A	< -5	Clear	≥ 50	Skiing
B	< -5	Fog	≥ 50	Swimming
C	< -5	Fog	< 50	Swimming
D	< -5	Rain	≥ 50	Skiing
E	< -5	Rain	< 50	Swimming
F	≥ -5	Clear	≥ 50	Skiing
G	≥ -5	Clear	< 50	Skiing
H	≥ -5	Fog	< 50	Swimming
I	≥ -5	Rain	≥ 50	Skiing

a) Construct a decision tree for the variable Sport using gain ratio

b) Classify following data points:

- (Temperature = -3, Visibility = Fog, Snow Depth = 12)
- (Temperature = 10, Visibility = Clear, Snow Depth = 0)
- (Temperature = 5, Visibility = Rain, Snow Depth = 27)

Solution

a) Compute the gainRatio for each variable. For convenience, refer to the following concatenated frequency table:

	Temperature		Visibility			Snow Depth		
	< -5	≥ -5	clear	fog	rain	< 50	≥ 50	$\sum$
Skiing	2	3	3	0	2	1	4	**5**
Swimming	3	1	0	3	1	3	1	**4**

- gainRatio(Temperature) = $\frac{\text{gain(Temperature)}}{\text{intrinsic\_info(Temperature)}} = \frac{\text{info}([5,4])-\text{info}([2,3],[3,1])}{\text{intrinsic\_info}([5,4])} \approx \frac{0.991-0.900}{0.991} \approx 0.092$
- gainRatio(Visibility) = $\frac{\text{gain(Visibility)}}{\text{intrinsic\_info(Visibility)}} = \frac{\text{info}([5,4])-\text{info}([3,0],[0,3],[2,1])}{\text{info}([3,3,3])} \approx \frac{0.991-0.306}{1.585} \approx 0.432$
- gainRatio(Snow Depth) = $\frac{\text{gain(Snow Depth)}}{\text{intrinsic\_info(Snow Depth)}} = \frac{\text{info}([5,4])-\text{info}([1,3],[4,1])}{\text{info}([4,5])} \approx \frac{0.991-0.762}{0.991} \approx 0.231$
- gainRatio(ID) = $\frac{\text{gain(ID)}}{\text{intrinsic\_info(ID)}} = \frac{\text{info}([5,4])-\text{info}([1,0],[0,1],[0,1],[1,0],[0,1],[1,0],[1,0],[0,1],[1,0])}{\text{info}([1,1,1,1,1,1,1,1,1])} \approx \frac{0.991-0.0}{3.17} \approx 0.313$

As the variable Visibility maximizes gainRatio, the tree is split into three subtrees, with decision branches for "clear", "fog", and "rain".
As the paths "clear" and "fog" are pure, only "rain" has to be considered further. Again, compute the gainRatio for each variable. Conditioning the data to "Visibility = rain" yields the following concatenated frequency table:

	Temperature		Snow Depth		
	< -5	≥ -5	< 50	≥ 50	$\sum$
Skiing	1	1	0	2	**2**
Swimming	1	0	1	0	**1**

- gainRatio(Temperature) = $\frac{\text{gain(Temperature)}}{\text{intrinsic\_info(Temperature)}} = \frac{\text{info}([2,1])-\text{info}([1,1],[1,0])}{\text{intrinsic\_info}([2,1])} \approx \frac{0.918-0.667}{0.918} \approx 0.273$
- gainRatio(Snow Depth) = $\frac{\text{gain(Snow Depth)}}{\text{intrinsic\_info(Snow Depth)}} = \frac{\text{info}([2,1])-\text{info}([0,1],[2,0])}{\text{intrinsic\_info}([1,2])} \approx \frac{0.918-0.0}{0.918} \approx 1.0$
- gainRatio(ID) = $\frac{\text{gain(ID)}}{\text{intrinsic\_info(ID)}} = \frac{\text{info}([2,1])-\text{info}([0,1],[0,1],[1,0])}{\text{intrinsic\_info}([1,1,1])} \approx \frac{0.918-0.0}{1.585} \approx 0.579$

The next best split is via Snow Depth, as it maximizes gainRatio. This leads to the following (final) tree:

b) Using the built tree as a classifier yields:

- (Temperature = -3, Visibility = Fog, Snow Depth = 12) $\longrightarrow$ swimming
- (Temperature = 10, Visibility = Clear, Snow Depth = 0) $\longrightarrow$ skiing
- (Temperature = 5, Visibility = Rain, Snow Depth = 27) $\longrightarrow$ swimming