Tutorial Business Analytics

Tutorial 6 - Solution

Exercise 6.1

Calculate:

a) entropy(0.1, 0.9)
b) entropy(0.8, 0.2)
c) entropy(0.3, 0.7)
d) entropy(0.5, 0.5)
e) entropy(0.8, 0.1, 0.1)

Solution 6.1

a) entropy(0.1, 0.9) = $-0.1 \cdot \log_2 0.1 - 0.9 \cdot \log_2 0.9$ = 0.469.

b) entropy(0.8, 0.2) = 0.722.

c) entropy(0.3, 0.7) = 0.881.

d) entropy(0.5, 0.5) = 1.

e) entropy(0.8, 0.1, 0.1) = 0.922.

Exercise 6.2

Calculate:

a) info([2, 3])

b) info([5, 4])

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

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

Solution 6.2

a) info([2, 3]) = entropy($\frac{2}{2+3}$, $\frac{3}{2+3}$) = 0.971.

b) info([5, 4]) = 0.991.

c) info([2, 3], [5, 4]) = $\frac{2+3}{(2+3)+(5+4)}$ · info([2, 3]) + $\frac{5+4}{(2+3)+(5+4)}$ · info([5, 4]) = 0.984.

d) info([2, 3], [9, 0]) = 0.347.

Exercise 6.3

Construct a tree:

Temperature	Visibility	Snow depth	Sport
< -5	Clear	≥ 50	Skiing
< -5	Fog	≥ 50	Swimming
< -5	Fog	< 50	Swimming
< -5	Rain	≥ 50	Skiing
< -5	Rain	< 50	Swimming
≥ -5	Clear	≥ 50	Skiing
≥ -5	Clear	< 50	Skiing
≥ -5	Fog	< 50	Swimming
≥ -5	Rain	≥ 50	Skiing

Solution 6.3

Constructing the tree:

	< -5	≥ -5	Clear	Fog	Rain	< 50	≥ 50	∑
Skiing	2	3	3	0	2	1	4	**5**
Swimming	3	1	0	3	1	3	1	**4**

Attribute temperature

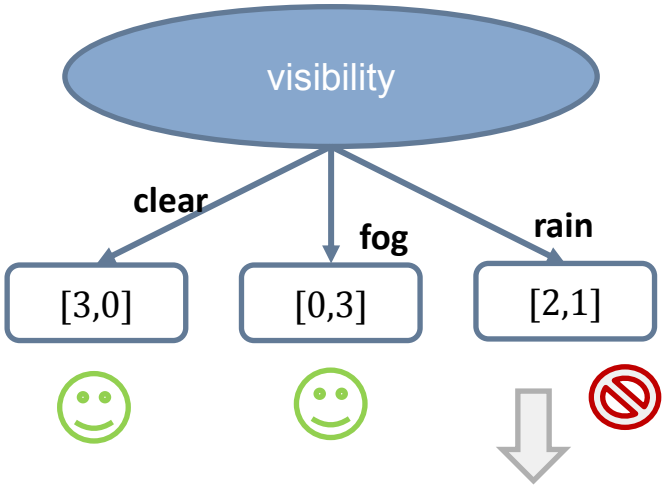
gain(temperature) = info(root) – info(temperature)
= info([5, 4]) – info([2, 3], [3, 1])
= 0.991 – 0.900 = 0.091

Attribute visibility

gain(visibility) = info(root) – info(visibility)
= info([5, 4]) – info([3, 0], [0, 3], [2, 1])
= 0.991 – 0.306 = 0.685

Attribute snow depth

gain(snow depth) = info(root) – info(snow depth)
= info([5, 4]) – info([1, 3], [4, 1])
= 0.991 – 0.762 = 0.229



Temperature	Visibility	Snow depth	Sport
< -5	Rain	≥ 50	**Skiing**
< -5	Rain	< 50	**Swimming**
≥ -5	Rain	≥ 50	**Skiing**

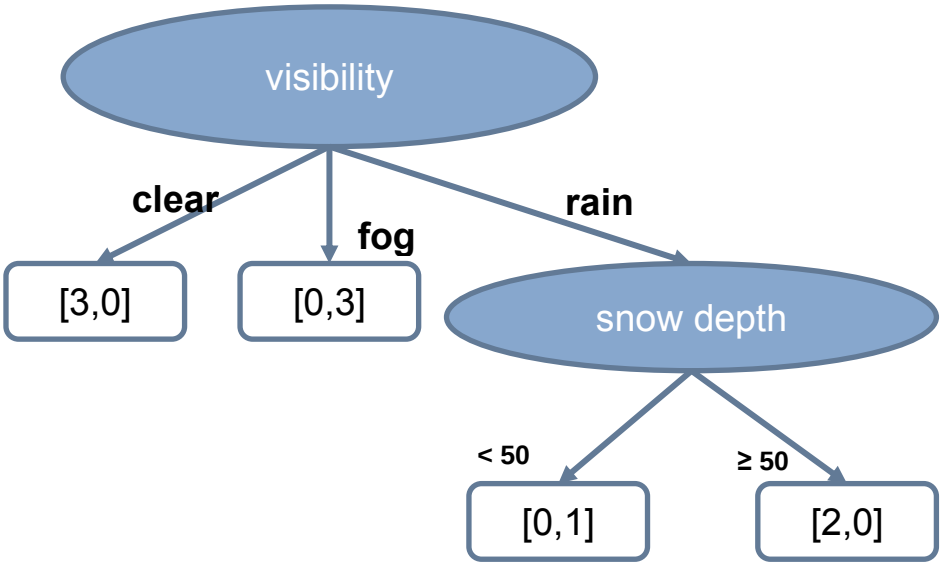
Attribute temperature

gain(temperature) = info([2, 1]) – info([1, 1], [1, 0])
= 0.918 – 0.667
= 0.251

Attribute snow depth

gain(snow depth) = info([2, 1]) – info([0, 1], [2, 0])
= 0.918 – 0
= 0.918

Result:



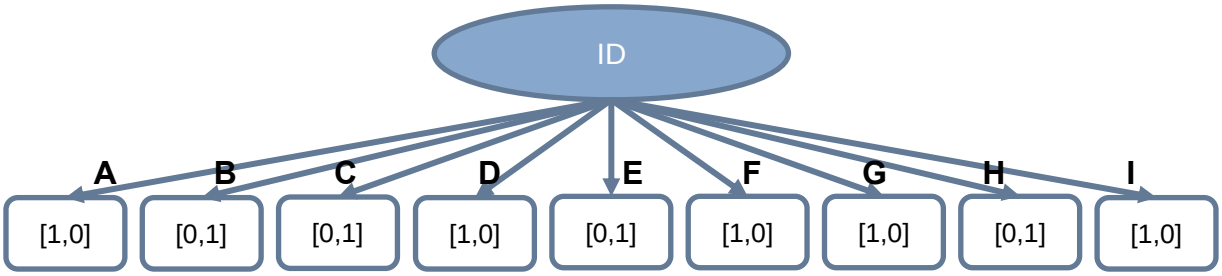
Exercise 6.4

Construct a tree for the following table, that has an additional ID attribute now.

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

Solution 6.4

Construct a tree:



Exercise 6.5

Construct the tree from exercise 6.4 a second time using gain ratio:

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

Solution 6.5

Construct the tree from exercise 6.4 a second time using gain ratio:

	< -5	≥ -5	Clear	Fog	Rain	< 50	≥ 50	∑
Skiing	2	3	3	0	2	1	4	**5**
Swimming	3	1	0	3	1	3	1	**4**

Attribute temperature

- gain(temperature) = info([5,4]) – info([2,3], [3,1]) = 0.991 – 0.900 = 0.091
- gainRatio(temperature) = 0.091 / info([5,4]) = 0.091 / 0.991 = 0.092

Attribute visibility

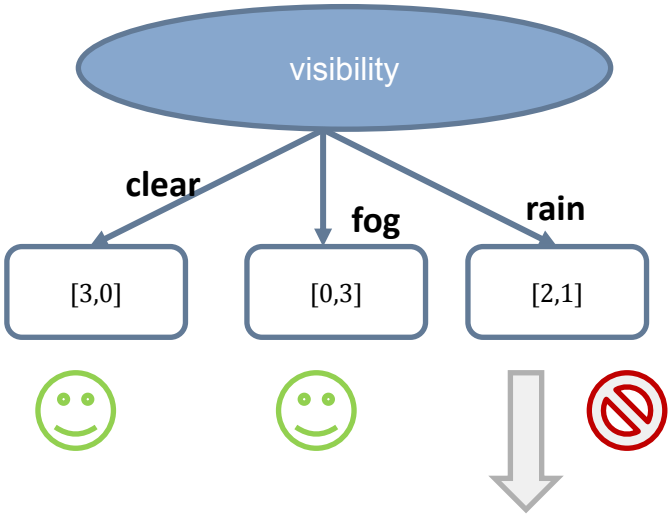
- gain(visibility) = info([5,4]) – info([3,0], [0,3], [2,1]) = 0.991 – 0.306 = 0.685
- gainRatio(visibility) = 0.685 / info([3,3,3]) = 0.685 / 1.585 = **0.432**

Attribute snow depth

- gain(snow depth) = info([5,4]) – info([1,3], [4,1]) = 0.991 – 0.762 = 0.229
- gainRatio(snow depth) = 0.229 / info([4,5]) = 0.229 / 0.991 = 0.231

Attribute ID

- gain(ID) = info([5,4]) – info([1,0], [1,0], [1,0], [1,0], [1,0], [1,0], [1,0], [1,0], [1,0]) = 0.991
- gainRatio(ID) = 0.991 / info([1,1,1,1,1,1,1,1,1]) = 0.991 / 3.17 = 0.313



ID	Temperature	Visibility	Snow depth	Sport
D	< -5	Rain	≥ 50	Skiing
E	< -5	Rain	< 50	Swimming
I	≥ -5	Rain	≥ 50	Skiing

Attribute temperature

- gain(temperature) = info([2, 1]) – info([1, 1], [1, 0]) = 0.918 – 0.667 = 0.251
- gainRatio(temperature) = 0.251 / info([2,1]) = 0.251 / 0.918 = 0.273

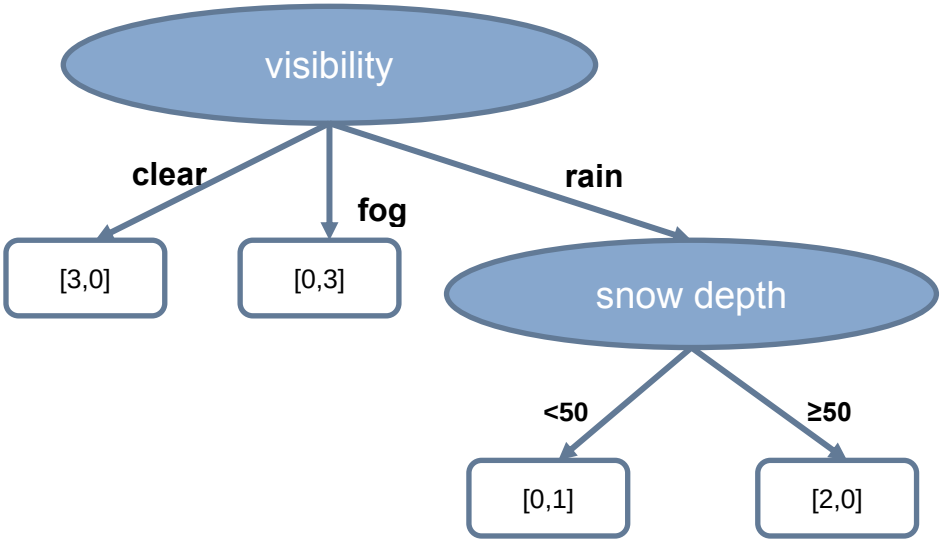
Attribute snow depth

- gain(snow depth) = info([2, 1]) – info([0, 1], [2, 0]) = 0.918 – 0 = 0.918
- gainRatio(snow depth) = 0.981 / info([1,2]) = 0.918 / 0.918 = 1

Attribute ID

- gain(ID) = info([2,1]) – info([1,0], [0,1], [1,0]) = 0.918 – 0 = 0.918
- gainRatio(ID) = 0.918 / info([1,1,1]) = 0.918 / 1.585 = 0.579

Result:



Exercise 6.6

Find the optimal binary splits.

a)

60	60	120	120	180	180	180
F	F	T	F	F	T	T

b)

5	5	7	7	7	8	9	9
T	T	T	T	F	T	F	F

Solution 6.6

Find the optimal binary splits.

a)

60	60	120	120	180	180	180
F	F	T	F	F	T	T

All possible splits: [0,2], [1,1], [2,1]

- info([0,2],[3,2]) = **0.694 split at 90**
- info([1,3],[2,1]) = 0.857

b)

5	5	7	7	7	8	9	9
T	T	T	T	F	T	F	F

All possible splits: [2,0],[2,1],[1,0],[0,2]

- info([2,0],[3,3]) = 0.75
- info([4,1],[1,2]) = 0.796
- info([5,1],[0,2]) = **0.488 split at 8.5**