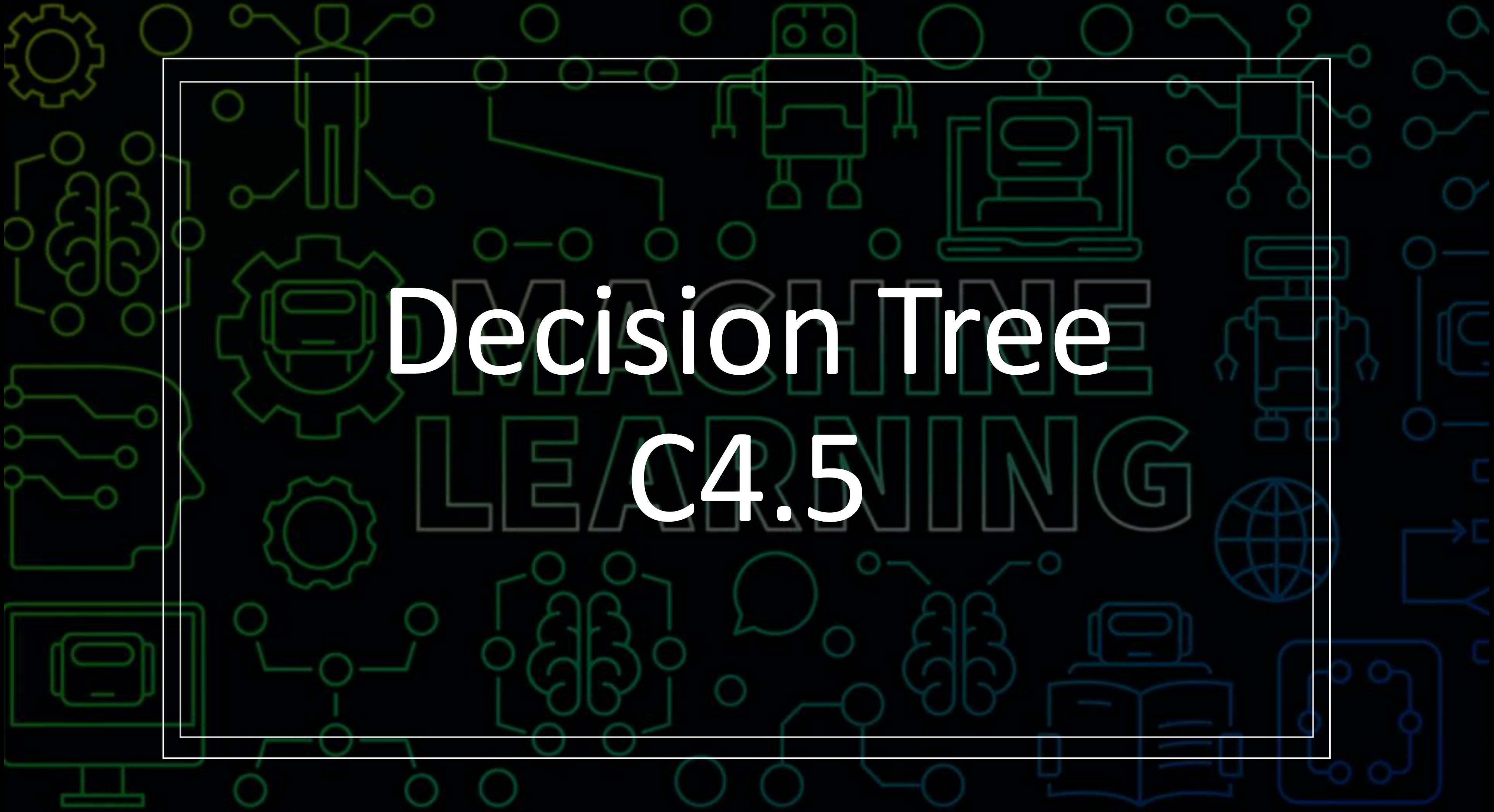
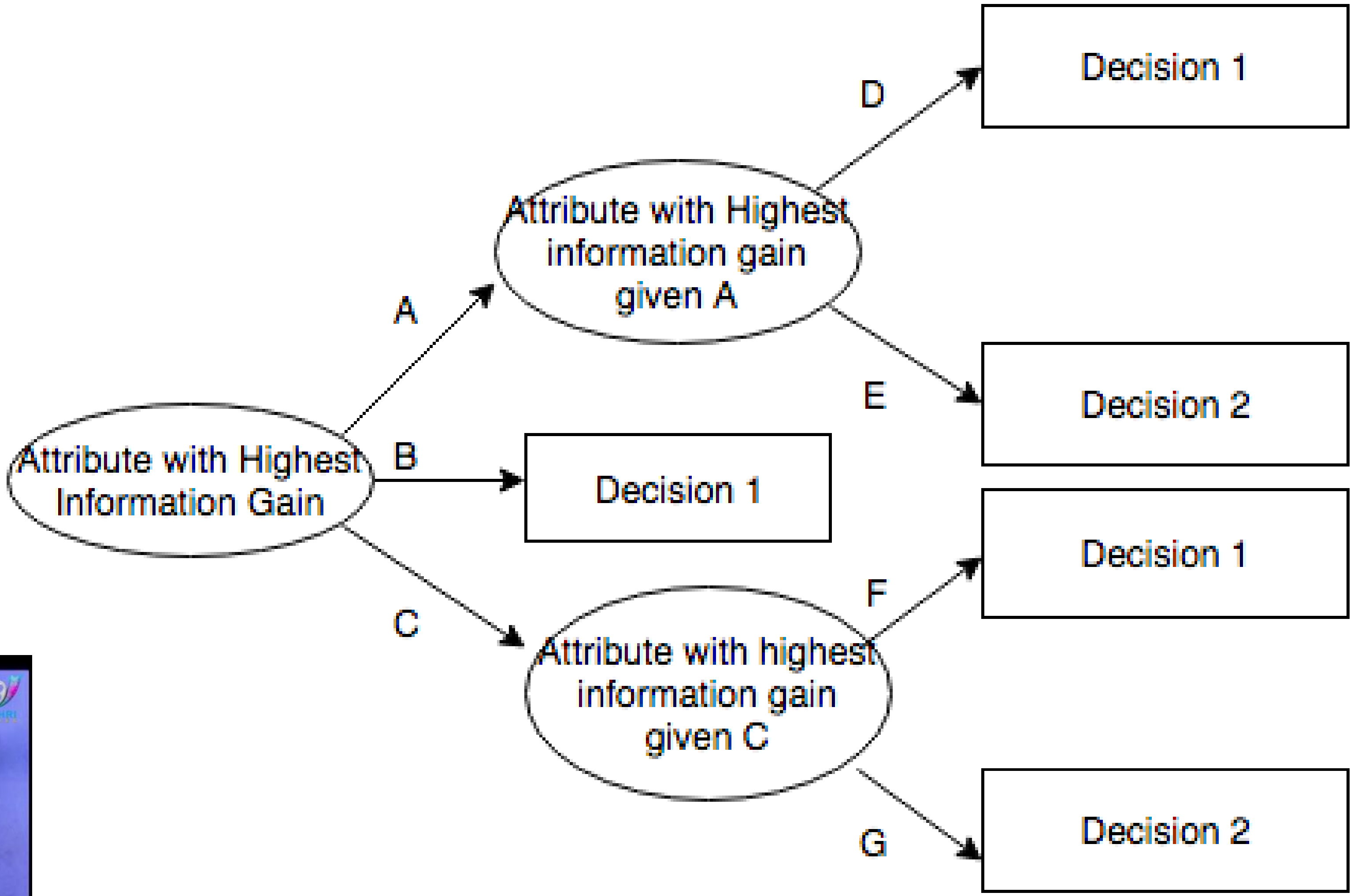
Decision Tree
C4.5



இதுவரை – ID3


Attribute with Highest Information Gain
A
B
C
Attribute with Highest information gain given A
Decision 1
Attribute with highest information gain given C
D
E
F
G
Decision 1
Decision 2
Decision 1
Decision 2


Savvy

Tennis Classification

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	85	85	Weak	No
2	Sunny	80	90	Strong	No
3	Overcast	83	78	Weak	Yes
4	Rain	70	96	Weak	Yes
5	Rain	68	80	Weak	Yes
6	Rain	65	70	Strong	No
7	Overcast	64	65	Strong	Yes
8	Sunny	72	95	Weak	No
9	Sunny	69	70	Weak	Yes
10	Rain	75	80	Weak	Yes
11	Sunny	75	70	Strong	Yes
12	Overcast	72	90	Strong	Yes
13	Overcast	81	75	Weak	Yes
14	Rain	71	80	Strong	No

Decision Tree C4.5

We will do what we have done in ID3 example. Firstly, we need to calculate global entropy. There are 14 examples; 9 instances refer to yes decision, and 5 instances refer to no decisior

$Entropy(Decision) = \sum - p(I) . \log_2 p(I) = - p(Yes) . \log_2 p(Yes) - p(No) . \log_2 p(No) = - (9/14) . \log_2(9/14) - (5/14) . \log_2(5/14) = 0.940$

In ID3 algorithm, we've calculated gains for each attribute. Here, we need to calculate gain ratios instead of gains.

$GainRatio(A) = Gain(A) / SplitInfo(A)$

$SplitInfo(A) = -\sum |Dj|/|D| \times \log_2|Dj|/|D|$

Decision Tree C4.5

Wind Attribute

Wind is a nominal attribute. Its possible values are weak and strong.

Gain(Decision, Wind) = Entropy(Decision) – ∑ (p(Decision|Wind) . Entropy(Decision|Wind))

Gain(Decision, Wind) = Entropy(Decision) – [p(Decision|Wind=Weak) . Entropy(Decision|Wind=Weak)] + [p(Decision|Wind=Strong) . Entropy(Decision|Wind=Strong)]

There are 8 weak wind instances. 2 of them are concluded as no, 6 of them are concluded as yes.

Entropy(Decision|Wind=Weak) = – p(No) . $\log_2$p(No) – p(Yes) . $\log_2$p(Yes) = – (2/8) . $\log_2$(2/8) – (6/8) . $\log_2$(6/8) = 0.811

Decision Tree C4.5

$Entropy(Decision|Wind=Strong) = -(3/6) . \log_2(3/6) - (3/6) . \log_2(3/6) = 1$

$Gain(Decision, Wind) = 0.940 - (8/14).(0.811) - (6/14).(1) = 0.940 - 0.463 - 0.428 = 0.049$

There are 8 decisions for weak wind, and 6 decisions for strong wind.

$SplitInfo(Decision, Wind) = -(8/14).\log_2(8/14) - (6/14).\log_2(6/14) = 0.461 + 0.524 = 0.985$

$GainRatio(Decision, Wind) = Gain(Decision, Wind) / SplitInfo(Decision, Wind) = 0.049 / 0.985 = 0.049$

Decision Tree C4.5

Outlook Attribute

Outlook is a nominal attribute, too. Its possible values are sunny, overcast and rain.

Gain(Decision, Outlook) = Entropy(Decision) – ∑ (p(Decision|Outlook) . Entropy(Decision|Outlook)) =

Gain(Decision, Outlook) = Entropy(Decision) – p(Decision|Outlook=Sunny) . Entropy(Decision|Outlook=Sunny) – p(Decision|Outlook=Overcast) . Entropy(Decision|Outlook=Overcast) – p(Decision|Outlook=Rain) . Entropy(Decision|Outlook=Rain)

There are 5 sunny instances. 3 of them are concluded as no, 2 of them are concluded as yes.

Entropy(Decision|Outlook=Sunny) = – p(No) . $\log_2$p(No) – p(Yes) . $\log_2$p(Yes) = -(3/5).$\log_2$(3/5) – (2/5).$\log_2$(2/5) = 0.441 + 0.528 = 0.970

Entropy(Decision|Outlook=Overcast) = – p(No) . $\log_2$p(No) – p(Yes) . $\log_2$p(Yes) = -(0/4).$\log_2$(0/4) – (4/4).$\log_2$(4/4) = 0

Decision Tree C4.5

Entropy(Decision|Outlook=Rain) = – p(No) . $\log_2$p(No) – p(Yes) . $\log_2$p(Yes) = –(2/5).$\log_2$(2/5) – (3/5).$\log_2$(3/5) = 0.528 + 0.441 = 0.970

Gain(Decision, Outlook) = 0.940 – (5/14).(0.970) – (4/14).(0) – (5/14).(0.970) – (5/14).(0.970) = 0.246

There are 5 instances for sunny, 4 instances for overcast and 5 instances for rain

SplitInfo(Decision, Outlook) = –(5/14).$\log_2$(5/14) –(4/14).$\log_2$(4/14) –(5/14).$\log_2$(5/14) = 1.577

GainRatio(Decision, Outlook) = Gain(Decision, Outlook)/SplitInfo(Decision, Outlook) = 0.246/1.577 = 0.155

Decision Tree C4.5

Humidity Attribute

As an exception, humidity is a continuous attribute. We need to convert continuous values to nominal ones. C4.5 proposes to perform binary split based on a threshold value. Threshold should be a value which offers maximum gain for that attribute. Let's focus on humidity attribute. Firstly, we need to sort humidity values smallest to largest.

Day	Humidity	Decision
7	65	Yes
6	70	No
9	70	Yes
11	70	Yes
13	75	Yes
3	78	Yes
5	80	Yes
10	80	Yes
14	80	No
1	85	No
2	90	No
12	90	Yes
8	95	No
4	96	Yes

Decision Tree C4.5

Check 65 as a threshold for humidity

Entropy(Decision|Humidity<=65) = – p(No) . $\log_2$p(No) – p(Yes) . $\log_2$p(Yes) = -(0/1).$\log_2$(0/1) – (1/1).$\log_2$(1/1) = 0

Entropy(Decision|Humidity>65) = -(5/13).$\log_2$(5/13) – (8/13).$\log_2$(8/13) =0.530 + 0.431 = 0.961

Gain(Decision, Humidity<> 65) = 0.940 – (1/14).0 – (13/14).(0.961) = 0.048

** The statement above refers to that what would branch of decision tree be for less than or equal to 65, and greater than 65. It **does not** refer to that humidity is not equal to 65!*

SplitInfo(Decision, Humidity<> 65) = -(1/14).$\log_2$(1/14) -(13/14).$\log_2$(13/14) = 0.371

Decision Tree C4.5

Check 70 as a threshold for humidity

$Entropy(Decision|Humidity \le 70) = -(1/4).\log_2(1/4) - (3/4).\log_2(3/4) = 0.811$

$Entropy(Decision|Humidity > 70) = -(4/10).\log_2(4/10) - (6/10).\log_2(6/10) = 0.970$

$Gain(Decision, Humidity <> 70) = 0.940 - (4/14).(0.811) - (10/14).(0.970) = 0.940 - 0.231 = 0.014$

$SplitInfo(Decision, Humidity <> 70) = -(4/14).\log_2(4/14) - (10/14).\log_2(10/14) = 0.863$

$GainRatio(Decision, Humidity <> 70) = 0.016$

Decision Tree C4.5

Check 75 as a threshold for humidity

Entropy(Decision|Humidity<=75) = – (1/5).$\log_2$(1/5) – (4/5).$\log_2$(4/5) = 0.721

Entropy(Decision|Humidity>75) = – (4/9).$\log_2$(4/9) – (5/9).$\log_2$(5/9) = 0.991

Gain(Decision, Humidity<> 75) = 0.940 – (5/14).(0.721) – (9/14).(0.991) = 0.940 – 0.2575 – 0.637 = 0.045

SplitInfo(Decision, Humidity<> 75) = -(5/14).$\log_2$(4/14) -(9/14).$\log_2$(10/14) = 0.940

GainRatio(Decision, Humidity<> 75) = 0.047

Decision Tree C4.5

Gain(Decision, Humidity <> 78) =0.090, GainRatio(Decision, Humidity <> 78) =0.090

Gain(Decision, Humidity <> 80) = 0.101, GainRatio(Decision, Humidity <> 80) = 0.107

Gain(Decision, Humidity <> 85) = 0.024, GainRatio(Decision, Humidity <> 85) = 0.027

Gain(Decision, Humidity <> 90) = 0.010, GainRatio(Decision, Humidity <> 90) = 0.016

Gain(Decision, Humidity <> 95) = 0.048, GainRatio(Decision, Humidity <> 95) = 0.128

Here, I ignore the value 96 as threshold because humidity cannot be greater than this value.

Decision Tree C4.5

Temperature feature is continuous as well. When I apply binary split to temperature for all possible split points, the following decision rule maximizes for both gain and gain ratio.

Gain(Decision, Temperature <> 83) = 0.113, GainRatio(Decision, Temperature<> 83) = 0.305

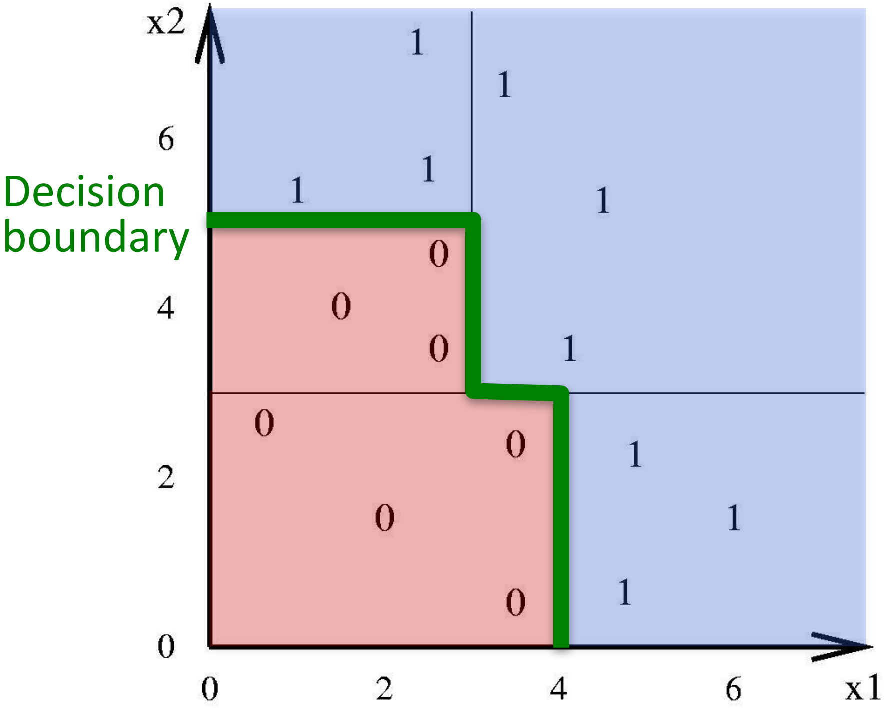
Decision Tree C4.5

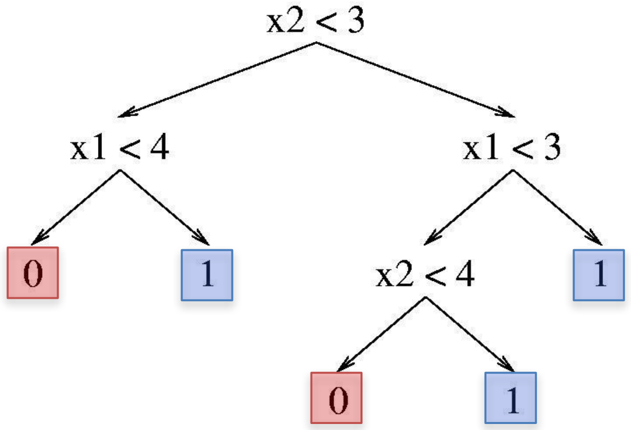
Let's summarize calculated gain and gain ratios. Outlook attribute comes with both maximized gain and gain ratio. This means that we need to put outlook decision in root of decision tree.

Attribute	Gain	GainRatio
Wind	0.049	0.049
Outlook	0.246	0.155
Humidity <> 80	0.101	0.107
Temperature <> 83	0.113	0.305

Decision Tree – Decision Boundary

- Decision trees divide the feature space into axis-parallel (hyper-)rectangles
- Each rectangular region is labeled with one label
  - or a probability distribution over labels





Decision Tree C4.5

https://sefiks.com/2018/05/13/a-step-by-step-c4-5-decision-tree-example/