ID3 Algorithm Manually Worked Through

Jared O'Toole
2019-11-18

Oracle Machine Learning
Section 4-3

Play Outdoor Sport

Num	Outlook	Temperature	Humidity	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

(1) Calculate Entropy on the results column "Play"

- there are 9 "yes"s and 5 "no"s (2 outcomes)
- $\text{Entropy}(S) = -p_{yes}\log_2(p_{yes}) + -p_{no}\log_2(p_{no})$

$$= \frac{-9}{14}\log_2\left(\frac{9}{14}\right) + \frac{-5}{14}\log_2\left(\frac{5}{14}\right)$$

$$\approx 0.94$$

(2) Find the next attribute with the most gain

- $\text{Gain}(S, A) = \text{Entropy}(s) - \sum_{v \in \text{values}(A)} \left[\frac{|S_v|}{|S|}\text{Entropy}(S_v)\right]$

(a) Test the "Outlook" Column

- there are 3 options: "Sunny", "Overcast", and "Rain" Total: 14

Outlook	# with "Yes" result		# with "No" result		Total #
Sunny	II	2	III	3	5
Overcast	IIII	4		0	4
Rain	III	3	II	2	5

- $\text{Entropy}(\text{Sunny}) = \frac{-2}{5}\log_2\left(\frac{2}{5}\right) + \frac{-3}{5}\log_2\left(\frac{3}{5}\right) \approx 0.971$
- $\text{Entropy}(\text{overcast}) = \frac{-4}{4}\log_2\left(\frac{4}{4}\right) + \frac{-0}{4}\log_2\left(\frac{0}{4}\right) = 0$

- Entropy(Rain) $= \frac{-3}{5}\log_2\left(\frac{3}{5}\right) + \frac{-2}{5}\log_2\left(\frac{2}{5}\right) \approx$ **0.971**

- Gain(S, Outlook) $= 0.94 - \left(Ve_{sunny} + Ve_{overcast} + Ve_{rain}\right)$

  with $Ve = \frac{|S_v|}{|S|} \cdot \text{Entropy}(S_v)$

  - $Ve_{sunny} = (5/14)(0.971) =$ **0.357**
  - $Ve_{overcast} = \cancel{(4/14)(0)} =$ **0.000**
  - $Ve_{rain} = (5/14)(0.971) =$ **0.357**

  $\ldots = 0.94 - 0.357 - 0.357 =$ **0.246**

(b) ▻ Test the "Temperature" Column

- There are 3 options: "cold", "mild", and "hot" Total: 14

Temperature	# with "Yes" result		# with "No" result		Total #
cold	III	3	I	1	4
mild	IIII	4	II	2	6
hot	II	2	II	2	4

- Entropy(cold) $= \frac{-3}{4}\log_2\left(\frac{3}{4}\right) + \frac{-1}{4}\log_2\left(\frac{1}{4}\right) = 0.8112781245\ldots$
- Entropy(mild) $= \frac{-4}{6}\log_2\left(\frac{4}{6}\right) + \frac{-2}{6}\log_2\left(\frac{2}{6}\right) = 0.9182958341\ldots$
- Entropy(hot) $= \frac{-2}{4}\log_2\left(\frac{2}{4}\right) + \frac{-2}{4}\log_2\left(\frac{2}{4}\right) = 1.0$
- $Ve_{cold} = (4/14)(0.8112781245\ldots) = 0.2317937499\ldots$
- $Ve_{mild} = (6/14)(0.9182958341\ldots) = 0.3935553575\ldots$
- $Ve_{hot} = (4/14)(1) = 0.2857142857$
- Gain(S, Temperature) $= 0.94 - \left(Ve_{cold} + Ve_{mild} + Ve_{hot}\right)$

  $= 0.0289366069\ldots$

(c) ▻ Test the "Humidity" Column

- There are 2 options: "high" and "normal" Total: 14

Humidity	# with "Yes" result		# with "No" result		Total #
High	III	3	IIII	4	7
Normal	IIII I	6	I	1	7

- Entropy(high) $= \frac{-3}{7}\log_2\left(\frac{3}{7}\right) + \frac{-4}{7}\log_2\left(\frac{4}{7}\right) = 0.985228136...$
- Entropy(normal) $= \frac{-6}{7}\log_2\left(\frac{6}{7}\right) + \frac{-1}{7}\log_2\left(\frac{1}{7}\right) = 0.591672778...$
- $VE_{high} = (7/14)(0.985228136...) = 0.462614068...$
- $VE_{normal} = (7/14)(0.591672778...) = 0.295836389...$
- Gain(S, Humidity) $= 0.94 - (VE_{high} + VE_{normal})$

$= 0.151549542...$

(d) ▻ Test the "Wind" Column

- There are 2 options: "Strong" and "Weak" Total: 14

Wind	# with "Yes" result		# with "No" result		Total #
Strong	IIII I	6	II	2	8
Weak	III	3	III	3	6

- Entropy(Strong) $= \frac{-6}{8}\log_2\left(\frac{6}{8}\right) + \frac{-2}{8}\log_2\left(\frac{2}{8}\right) = 0.811278124...$
- Entropy(weak) $= \frac{-3}{6}\log_2\left(\frac{3}{6}\right) + \frac{-3}{6}\log_2\left(\frac{3}{6}\right) = 1.0$
- $VE_{strong} = (8/14)(0.811278124...) = 0.463587499...$
- $VE_{weak} = (6/14)(1) = 0.428571428...$
- Gain(S, Wind) $= 0.94 - (VE_{strong} + VE_{weak})$

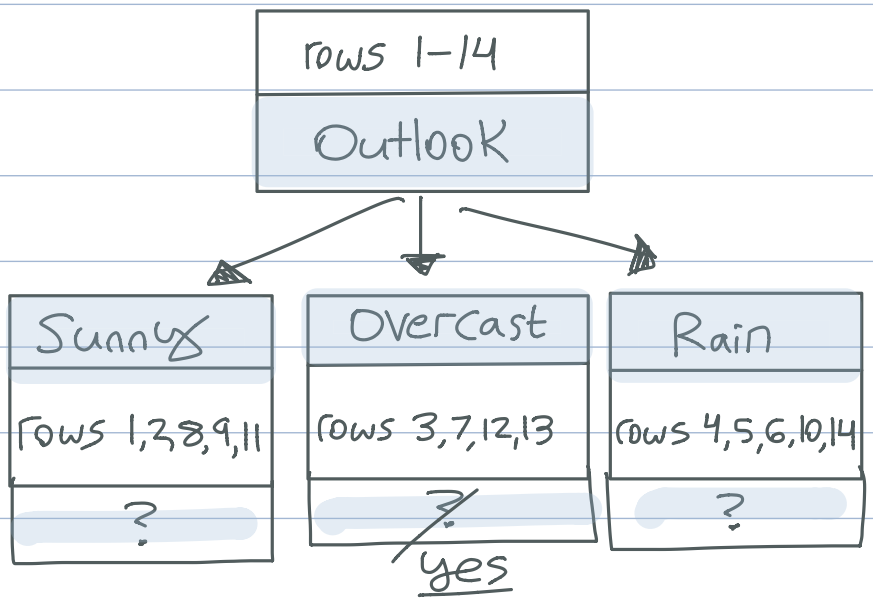
$= 0.047841071...$

(e) ▻ Choose the Column with the highest gain to be the root node

- Gain(S, Outlook) $= 0.246$ → Outlook
- Gain(S, Temperature) $= 0.028936606...$
- Gain(S, Humidity) $= 0.151549542...$

• Gain(S, Wind) = 0.0478410717... ➤ (g)(h)(i)

(f) ➤ Append the tree



• "Sunny" node has 2/5 "yes"s

$E(Sunny) = 0.9709505945...$

$\left(= \frac{-2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right)$

• "Overcast" node has 0/4 "yes"s

It can be categorized as "Yes"

• "Rain" node has 3/5 "Yes"s

$E(Rain) = 0.9709505945...$

$\left(= \frac{-3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right)$

(3) Find the next attribute with the most gain to split the "Sunny" node

➤ (a)(b)(c) Test the "Temperature", "Humidity", and "Wind" columns Total: 5 rows

Temp	#Yes	Totals
hot	0	2
mild	1	2
cold	1	1

$E(hot) = 0$

$\left(= \frac{-0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right)\right)$

$E(mild) = 1$

$\left(= \frac{-1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)$

$E(cold) = 0$

$\left(= \frac{-1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right)$

$VE_{hot} = \frac{2}{5} \cdot 0 = 0$

$VE_{mild} = \frac{2}{5} \cdot 1 = 0.40$

$VE_{cold} = \frac{1}{5} \cdot 0 = 0$

$Gain(^{Sunny}_{Temp.}) = 0.57095...$

Humid.	#Yes	Totals
High	0	3
Normal	2	2

$E(high) = 0$

$E(normal) = 0$

$VE_{high} = 0$

$VE_{normal} = 0$

$Gain(^{Sunny}_{Humid.}) = 0.9709...$

$(= 0.97 - (0+0))$

Wind	#Yes	Totals
weak	1	3
strong	1	2

$E(weak) = 0.9182958341...$

$\left(= \frac{-1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right)$

$E(strong) = 1.0$

$VE_{weak} = \frac{3}{5} \cdot E(weak) = 0.5509775...$

$VE_{strong} = \frac{2}{5} \cdot E(strong) = 0.40$

$Gain(^{Sunny}_{Wind}) = 0.019973094$

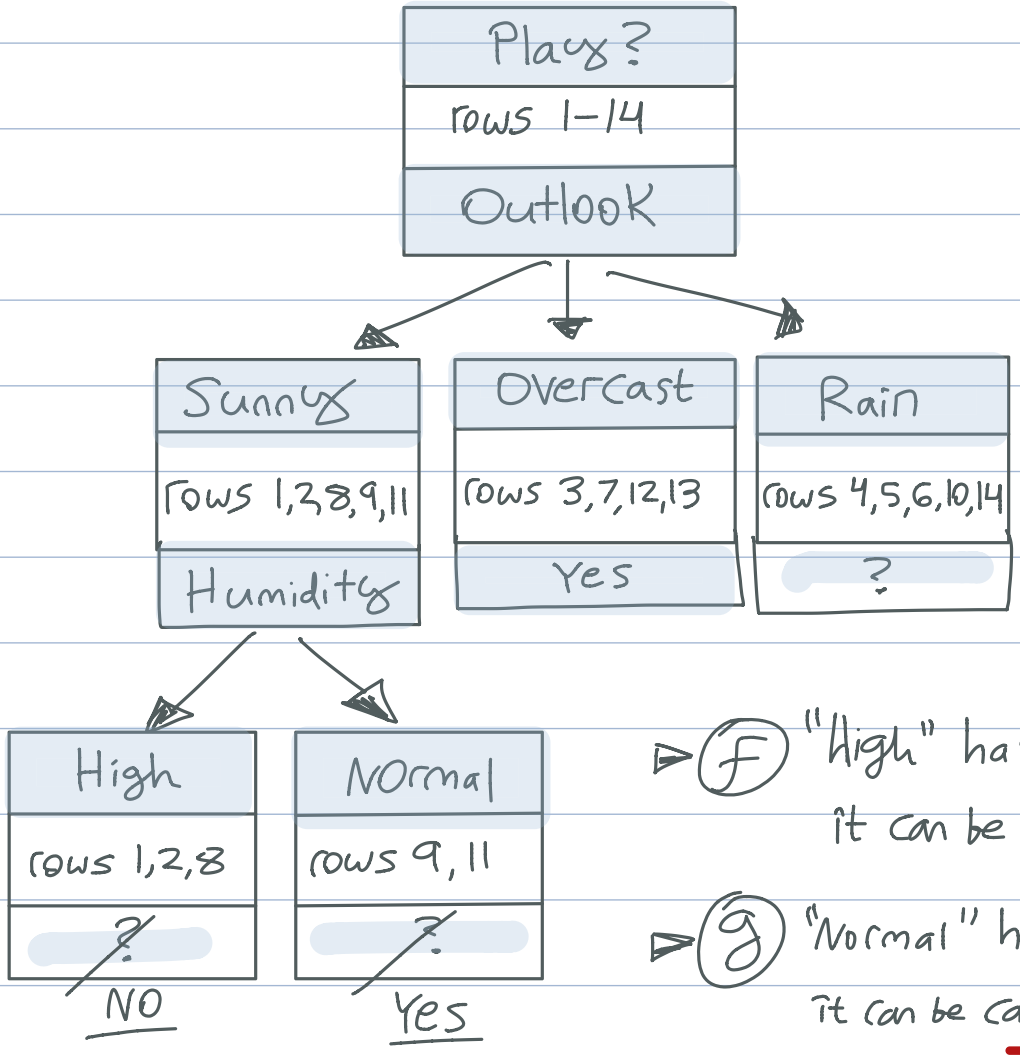
$(= 0.97 - (0.551 + 0.4))$

$(= 0.971 - (VE_{hot} + VE_{mild} + VE_{cold}))$

ⓓ Choose the Column with the highest gain to decide the child nodes of "Sunny"

- Gain(Sunny, Temp) = 0,57095...
- Gain(Sunny, Humid) = 0.9709... → Humidity
- Gain(sunny, wind) = 0.019973094

ⓔ Append the tree



ⓕ "High" has 0/3 "yes"s
it can be categorized as "No"

ⓖ "Normal" has 2/2 "yes"s
it can be categorized as "Yes"

④ Find the next attribute with the most gain to split the "Rain" node

ⓐⓑⓒ Test the "Temperature", "Humidity", and "wind" columns

Total: 5 rows

Temp	#Yes	Totals
~~hot~~	—	0
mild	2	3
Cold	1	2

Humid.	#Yes	Totals
High	1	2
Normal	2	3

Wind	#Yes	Totals
weak	3	3
Strong	0	2

$E(mild) = 0.9182958341...$

$E(cold) = 1$

$VE_{mild} = \frac{3}{5} E(mild) = 0.55097...$

$VE_{cold} = \frac{2}{5}(1) = 0.40$

$Gain(Rain, Temp) = 0.019973094...$

$E(high) = 1$

$E(normal) = 0.918...$

$VE_{high} = \frac{2}{5} \cdot 1 = 0.4$

$VE_{norm} = \frac{3}{5} \cdot E(norm) = 0.5509...$

$Gain(Rain, Humid) = 0.019973...$

$E(weak) = 0$

$E(strong) = 0$

$VE_{weak} = \frac{3}{5} \cdot 0 = 0$

$VE_{strong} = \frac{2}{5} \cdot 0 = 0$

$Gain(Rain, Wind) = 0.9709...$

ⓓ Choose the Column with the highest gain for "Rain"

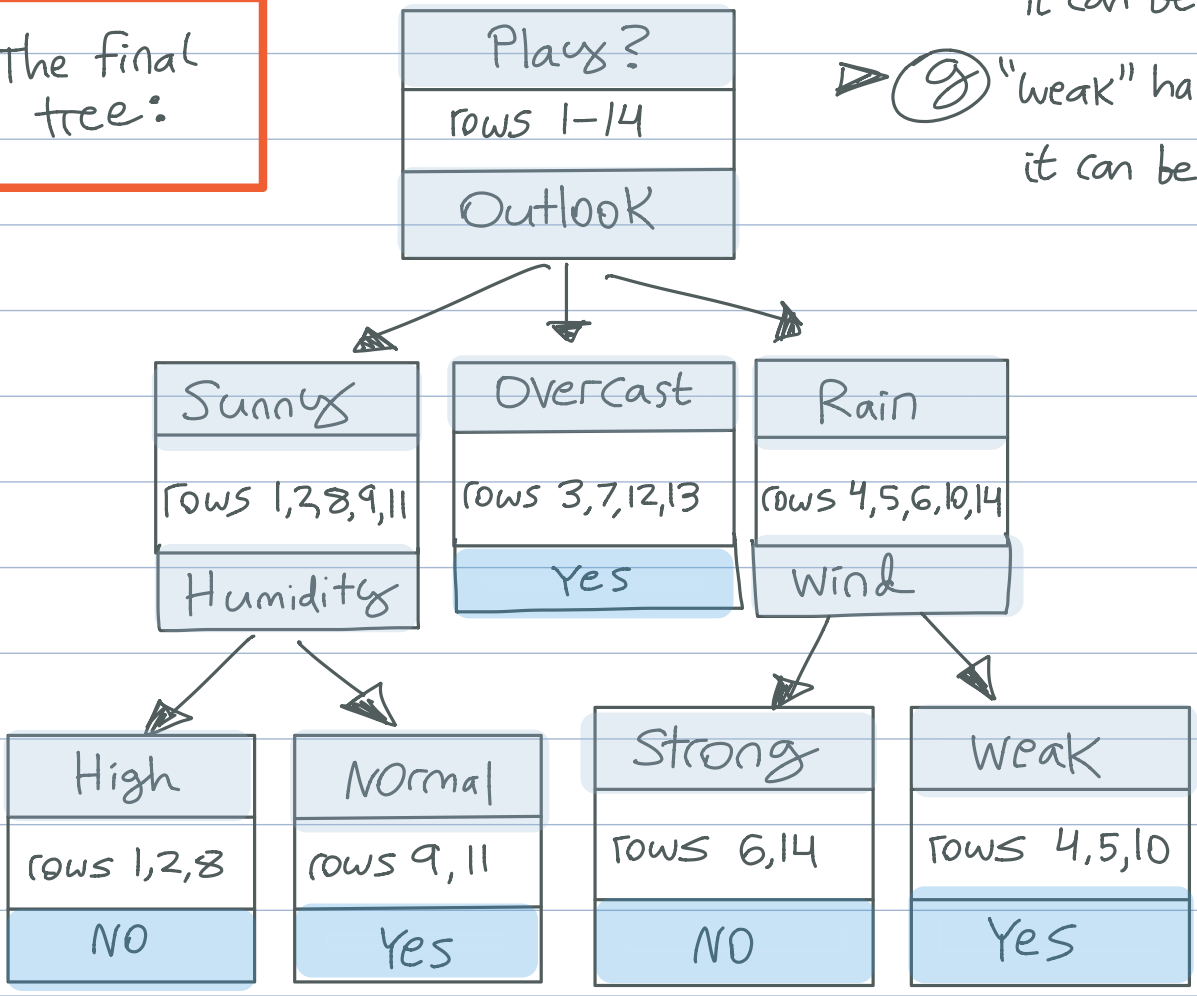
- Gain(wind, Temp) = Gain(wind, Humid) = 0.019973094
- Gain(wind, wind) = 0.9709... → Wind

ⓔ Append the tree

ⓕ "strong" has 0/2 "yes"s
it can be classified "No"

ⓖ "weak" has 3/3 "yes"s
it can be classified "Yes"

The final tree:



Pseudo-code for the ID3 algorithm:

```
# RECURSIVE FUNCTION
def branch(node, table)
    px = table.successes / table.total
    if px in (0, 1) # BASE CASE
        node.result = px
        return
    if not table.attrs # BASE CASE
        node.result = bool math.round(px)
        return
    # RECURSIVE CASE
    entropy_s = entropy(table)
    gains = dict()
    for attr in table.attrs
        gains[attr] = gain(entropy_s, attr)
    best, max = null, 0
    for attr, g in gains.items()
        if g > max
            best, max = attr, g
    sliced_table = table.without(best)
    for outcome in best
        sub_table = sliced_table.with_only(outcome)
        child = Node()
        node.add(child)
        branch(child, sub_table)
```