Mawaba Pascal Dao

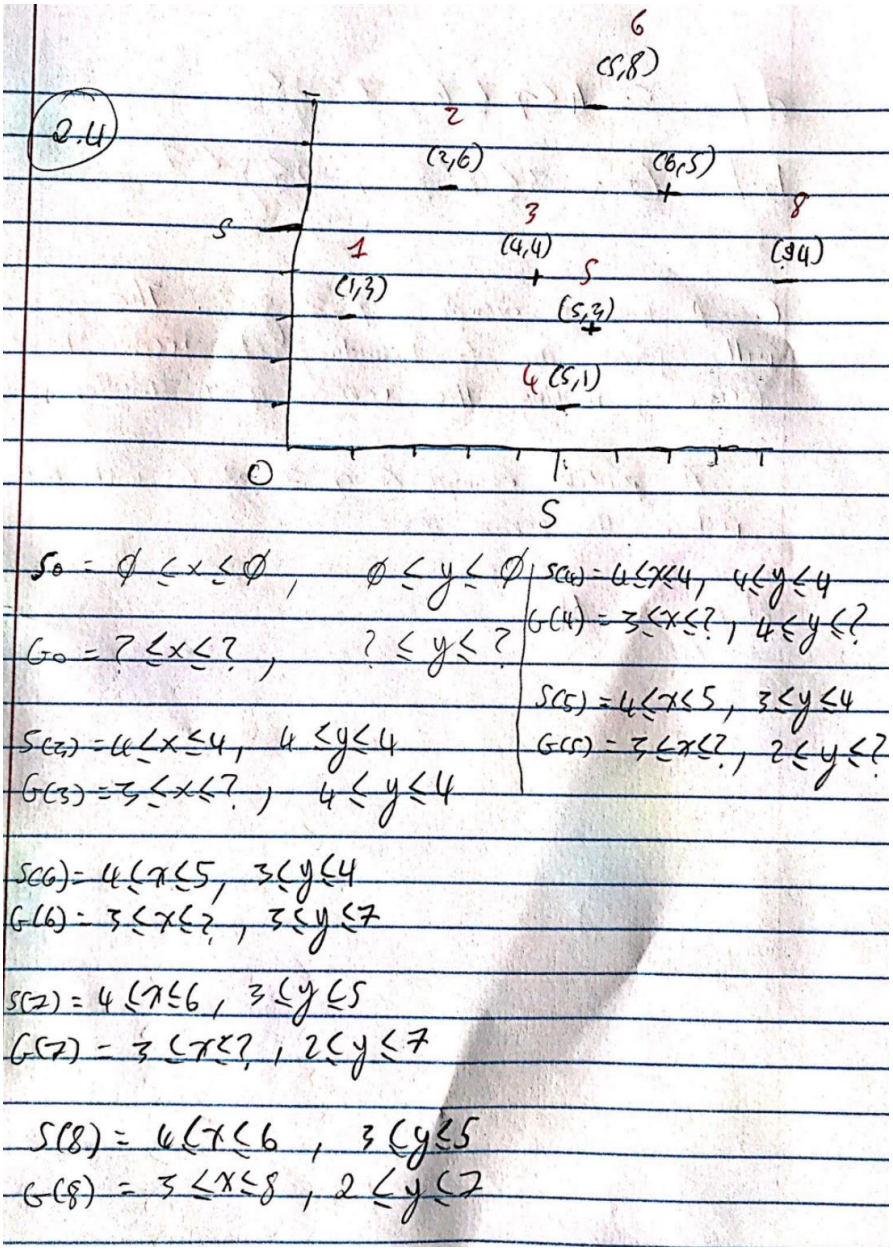
ML HW2

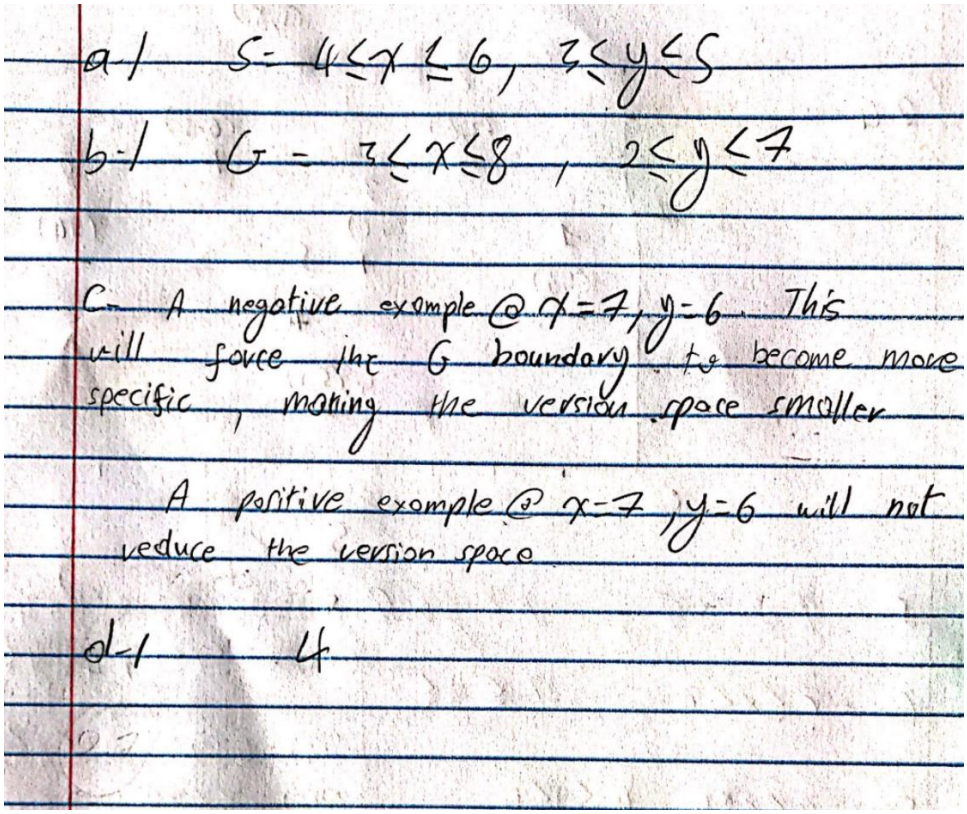
CSE5693

Dr. Chan

Feb 23, 2022

a.

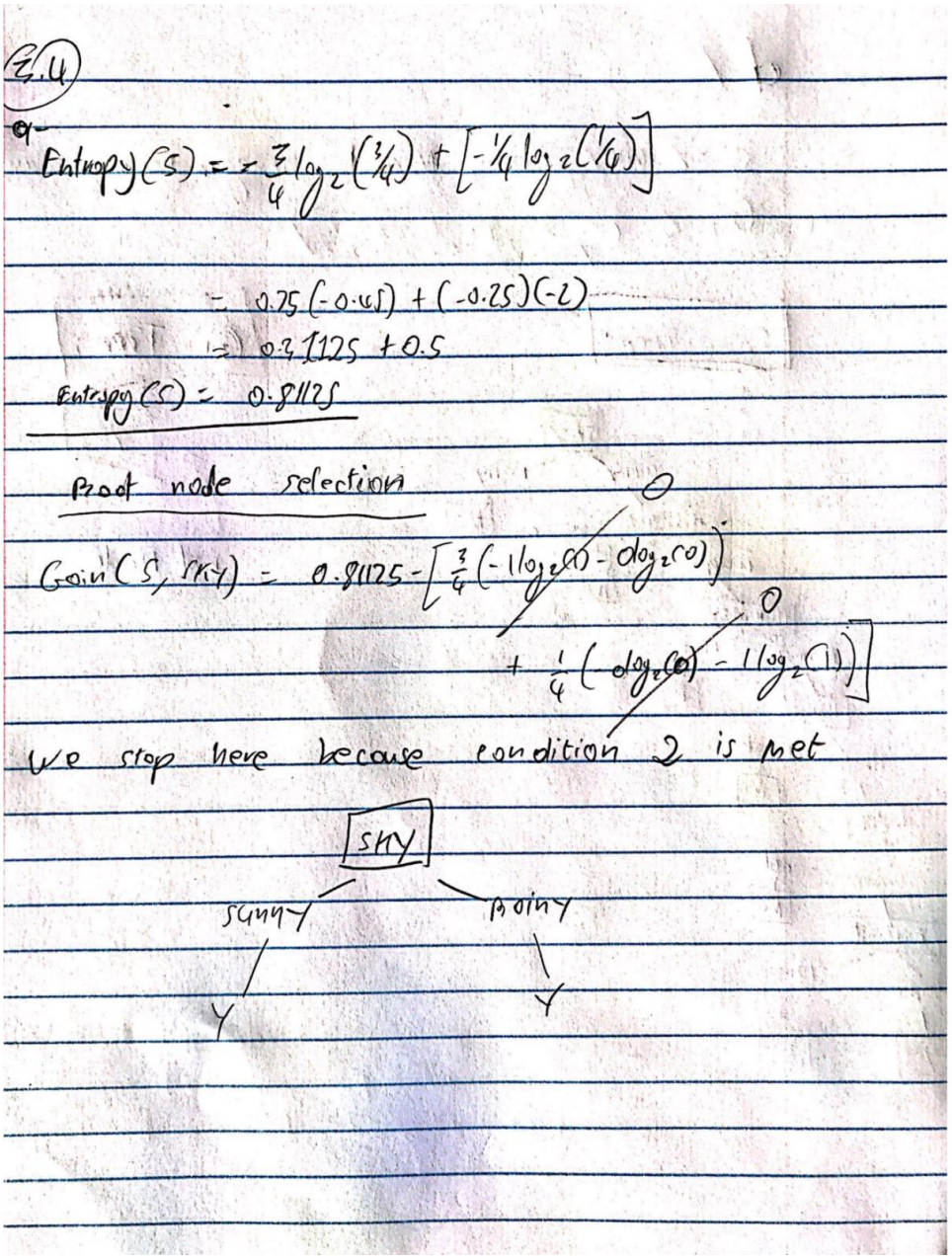




b.

There cannot be a maximally specific consistent hypothesis for classifying real numbers because the interval between any 2 real numbers is infinite. In the example of 4.5< x < 6.1, if x was 5 a more specific consistent hypothesis could be 4.0 < x < 6.0, and even more specific one could be 4.999 < x < 5.001. The hypothesis can be made infinitely more specific using real numbers. We can modify the hypothesis representation to converge to a maximally specific consistent hypothesis by fixing the decimal point precision of the natural numbers, or by using integers instead of natural numbers.

c.

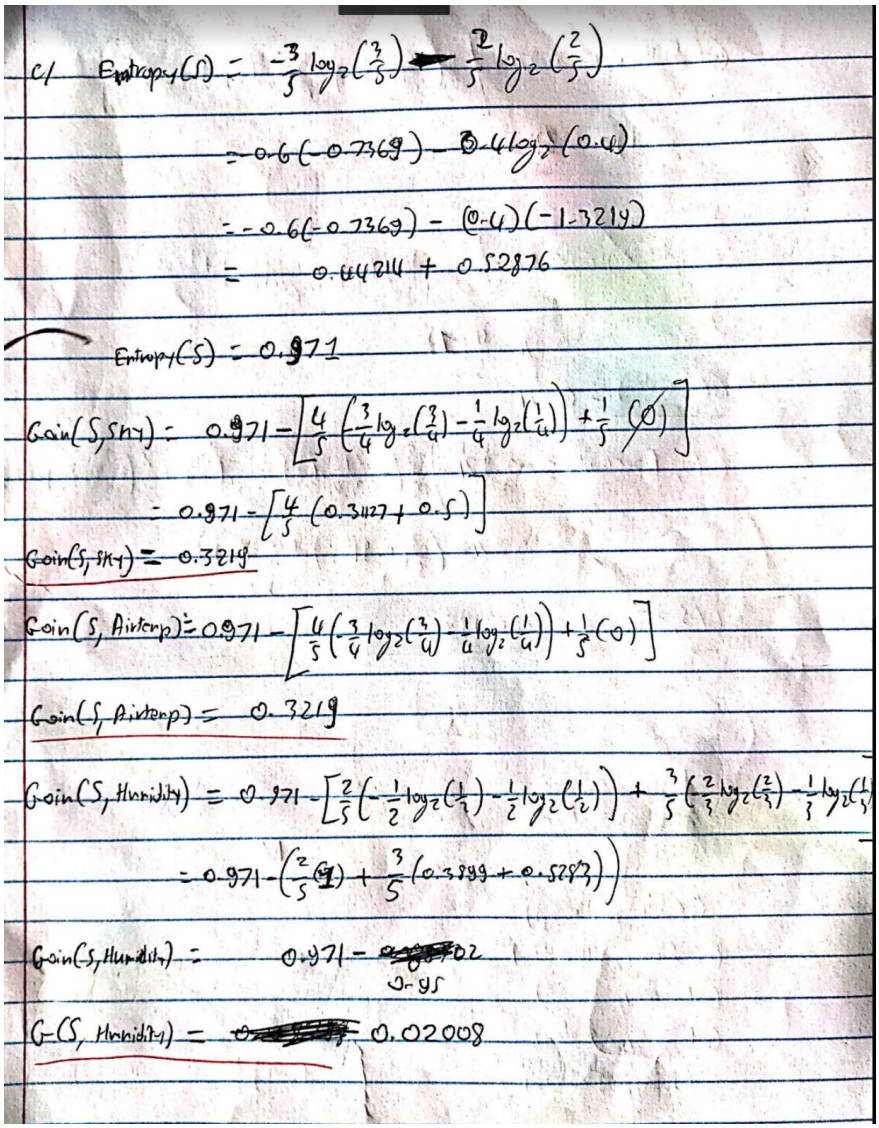

Q.4
Entropy(S) = $-\frac{3}{4}\log_2(\frac{3}{4}) + [-\frac{1}{4}\log_2(\frac{1}{4})]$
= 0.25(-0.65) + (-0.25)(-2)
= 0.31125 + 0.5
Entropy(S) = 0.81125
Root node selection
Gain(S, Play) = $0.81125 - [\frac{3}{4}(-1\log_2(1) - 0\log_2(0)) + \frac{1}{4}(-0\log_2(0) - 1\log_2(1))]$
0
0
We stop here because condition 2 is met
SKY
SUNNY
RAINY
Y
Y

3.4 -b

The decision tree learned in this case was one with the sky attribute at the root, and with sky value ending up as the yes leaf node

and the rainy value ends up with the N leaf node. This is equivalent to the following general boundary hypothesis from the version space in Figure 2.3: <Sunny,?,?,?,?,?>. This hypothesis will classify on Sunny samples as positive

and all non-sunny samples as negative, the decision tree and version space from figure 2.3 therefore both output the same classification on this training set.



$G(S, wind) = 0.971 - \left[\frac{4}{5}\left(-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) + \frac{1}{5}(0)\right]$

$G(S, wind) = 0.321$

$G(S, water) = 0.971 - \left[\frac{4}{5}\left(-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right) + \frac{1}{5}(0)\right]$

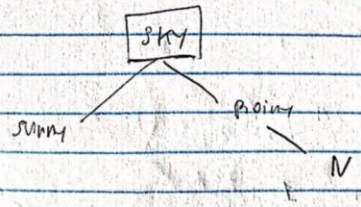
$0.971 - \frac{4}{5}(1)$

$G(S, water) = 0.171$

$G(S, Forecast) = 0.971 - \left[\frac{3}{5}\left(-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right) + \frac{2}{5}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right]$

$= 0.971 - \left(\frac{3}{5}(0.9185) + \frac{2}{5}(1)\right)$

$G(S, Forecast) = 0.0199$



$G(Sunny, Airtemp)$

$G(Sunny, Humidity)$

$G(Sunny, wind)$

$G(Sunny, water)$

$G(Sunny, Forecast)$

Entropy (Sunny) $= -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)$

$= 0.31127 + 0.5$

$E(\text{Sunny}) = 0.81127$

$\text{Gain}(\text{Sunny, Temp}) = 0.81127 - \left[\frac{1}{4}\left(-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right)\right]$

$\text{Gain}(\text{Sunny, Temp}) = 0.81127 - 0.81127 = 0$

$\text{Gain}(\text{Sunny, H}) = 0.81127 - \left[\frac{2}{4}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) + \frac{2}{4}\log_2(0)\right]$

$= 0.81127 - \frac{2}{4}$

$\text{Gain}(\text{Sunny, H}) = 0.31127$

$G(\text{Sunny, wind}) = 0.81127 - [0]$

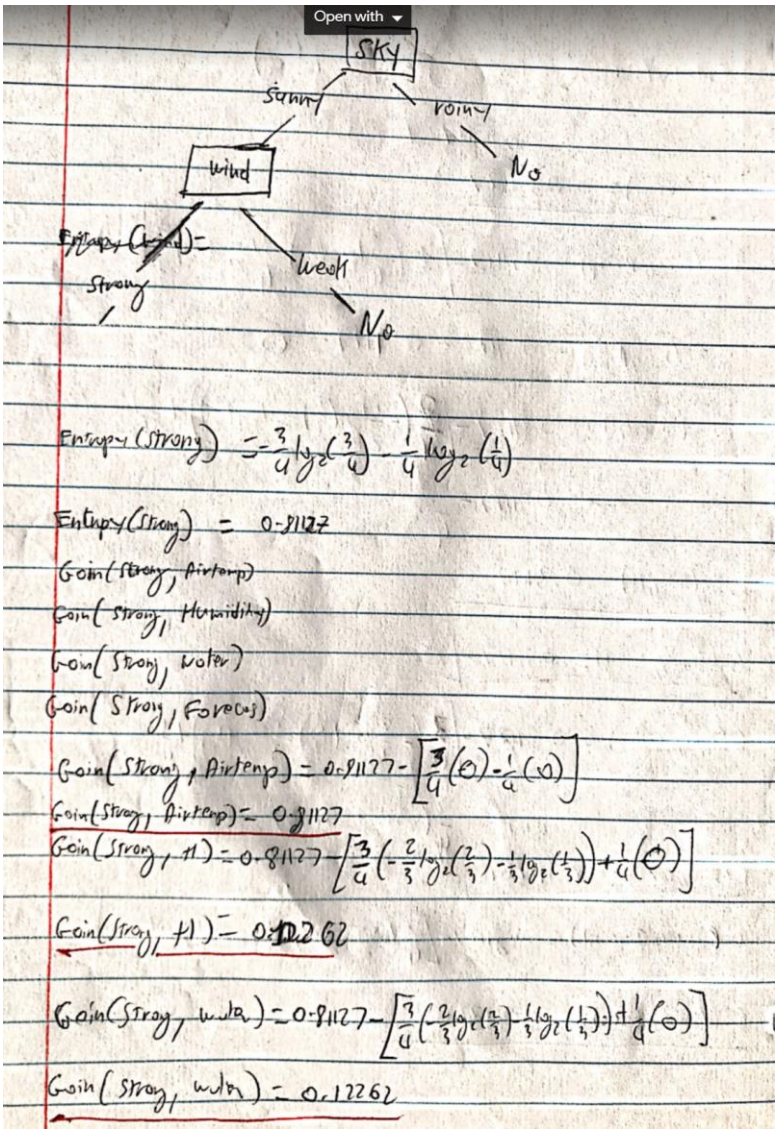
$G(\text{Sunny, water}) = 0.81127 - \left[\frac{3}{4}\left(-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right) + \frac{1}{4}\log_2(0)\right]$

$0.81127 - \frac{3}{4}(0.3999 + 0.5283)$

$G(\text{Sunny, water}) = 0.12262$

$G(\text{Sunny, Forecast}) = 0.81127 - \left[\frac{3}{4}\left(\frac{2}{3}\log_2\left(\frac{2}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right) + 0\right]$

$\text{Gain}(\text{Sunny, Forecast}) = 0.12262$



Entropy (Strong) $= -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)$

Entropy (Strong) = 0.81127

Gain (Strong, Airtemp)

Gain (Strong, Humidity)

Gain (Strong, water)

Gain (Strong, Forecas)

Gain (Strong, Airtemp) $= 0.81127 - \left[\frac{3}{4}(0) - \frac{1}{4}(0)\right]$

Gain (Strong, Airtemp) = 0.81127

Gain (Strong, H) $= 0.81127 - \left[\frac{3}{4}\left(-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right) + \frac{1}{4}(0)\right]$

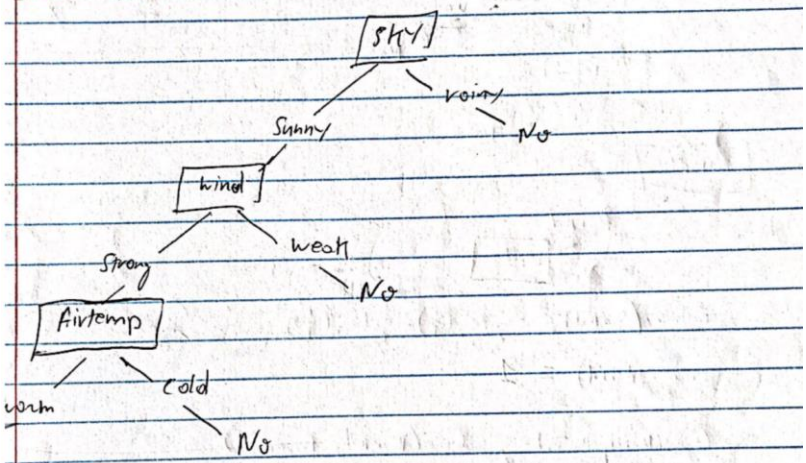
Gain (Strong, H) = 0.12262

Gain (Strong, water) $= 0.81127 - \left[\frac{3}{4}\left(-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right) + \frac{1}{4}(0)\right]$

Gain (Strong, water) = 0.12262

Gain(~~Strong~~, Forecast) = 0.81127 - $\left[\frac{2}{4}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right]$

Gain (Strong, Forecast) = 0.31127



Entropy (warm) = ~~0.971~~ 0.81127

Gain (warm, H)

Gain (warm, water)

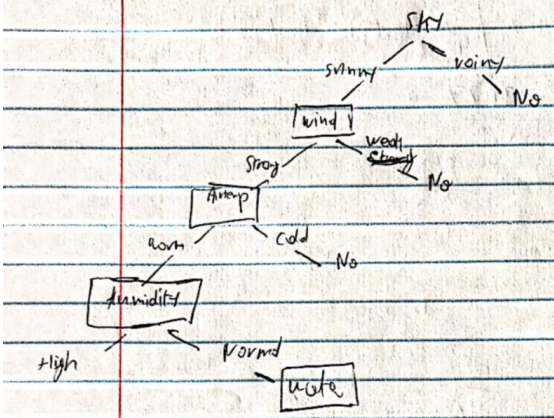
Gain (warm, Forecw)

~~Gain (warm, H) = 0.971~~

Gain (warm, H) = 0.31127

Gain (warm, wta) = 0.12262

Gain (warm, Forecst) = 0.12262



Entropy(Normal) = $-\frac{1}{2}\lg_2(\frac{1}{2}) - \frac{1}{2}\lg_2(\frac{1}{2})$

Entropy (Normal) = 1

Gain(Wand, water) ; Gain(Normal, Forecast)

Gain(Wand, water) = $1 - \left[\frac{4}{4}\left(-\frac{1}{2}\lg_2\frac{1}{2} - \frac{1}{2}\lg_2\frac{1}{2}\right)\right]$

Gain(Normal, water) = 0

Gain (Normal, forecast) = 0

Entropy (high) = $-\frac{2}{3}\lg_2(\frac{2}{3}) - \frac{1}{3}\lg_2(\frac{1}{3})$

Entropy (high) = 0.9185

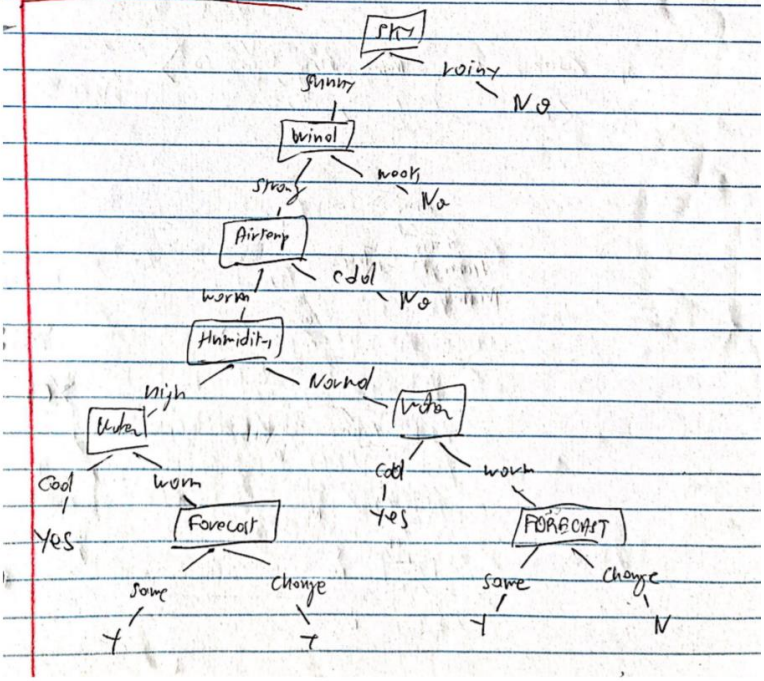
$Gain(high, water) = 0.9185 - \left[\frac{2}{3}\left(-\frac{1}{2}\lg_2\left(\frac{1}{2}\right) - \frac{1}{2}\left(\lg_2\left(\frac{1}{2}\right)\right)\right) + \frac{1}{3}(0)\right]$

$Gain(high, water) = 0.25193$

$Gain(high, Forecast) = 0.9185 - \left[\frac{1}{3}(0) + \frac{2}{3}\left(-\frac{1}{2}\lg_2\left(\frac{1}{2}\right) - \frac{1}{2}\lg_2\left(\frac{1}{2}\right)\right)\right]$

$= 0.9185 - \frac{2}{3}$

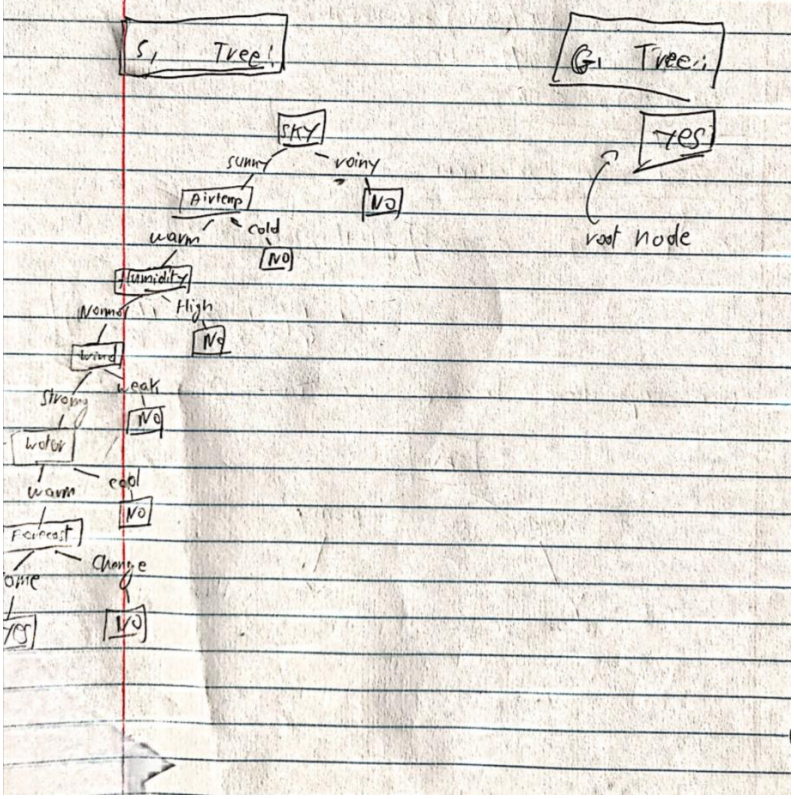
$Gain(high, Forecast) = 0.25193$

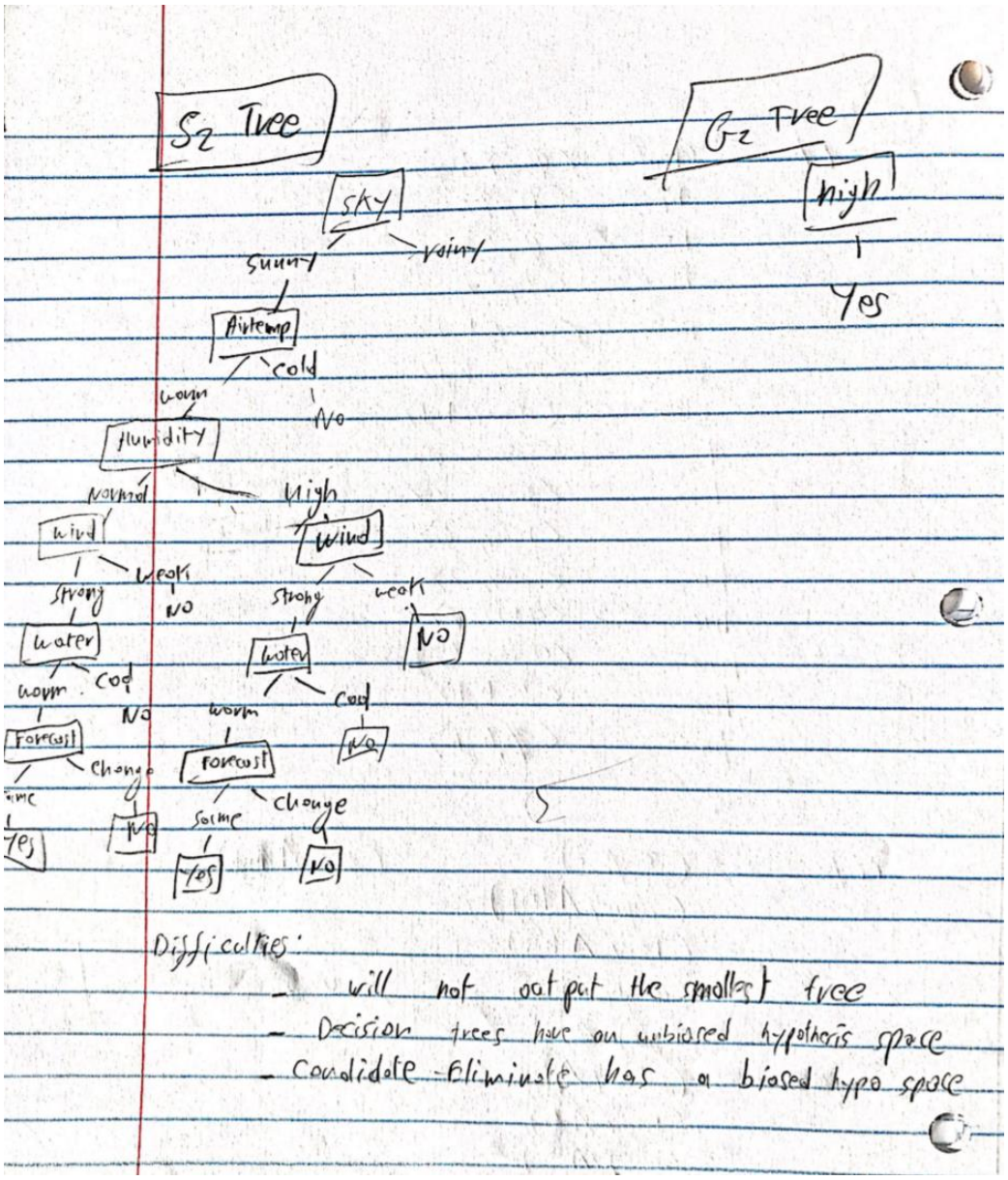


d-1

$S_1 = \langle$ Suun, worm, Normal, strong, warm, same$\rangle$

$G_1 = \langle ?, ?, ?, ?, ?, ? \rangle$



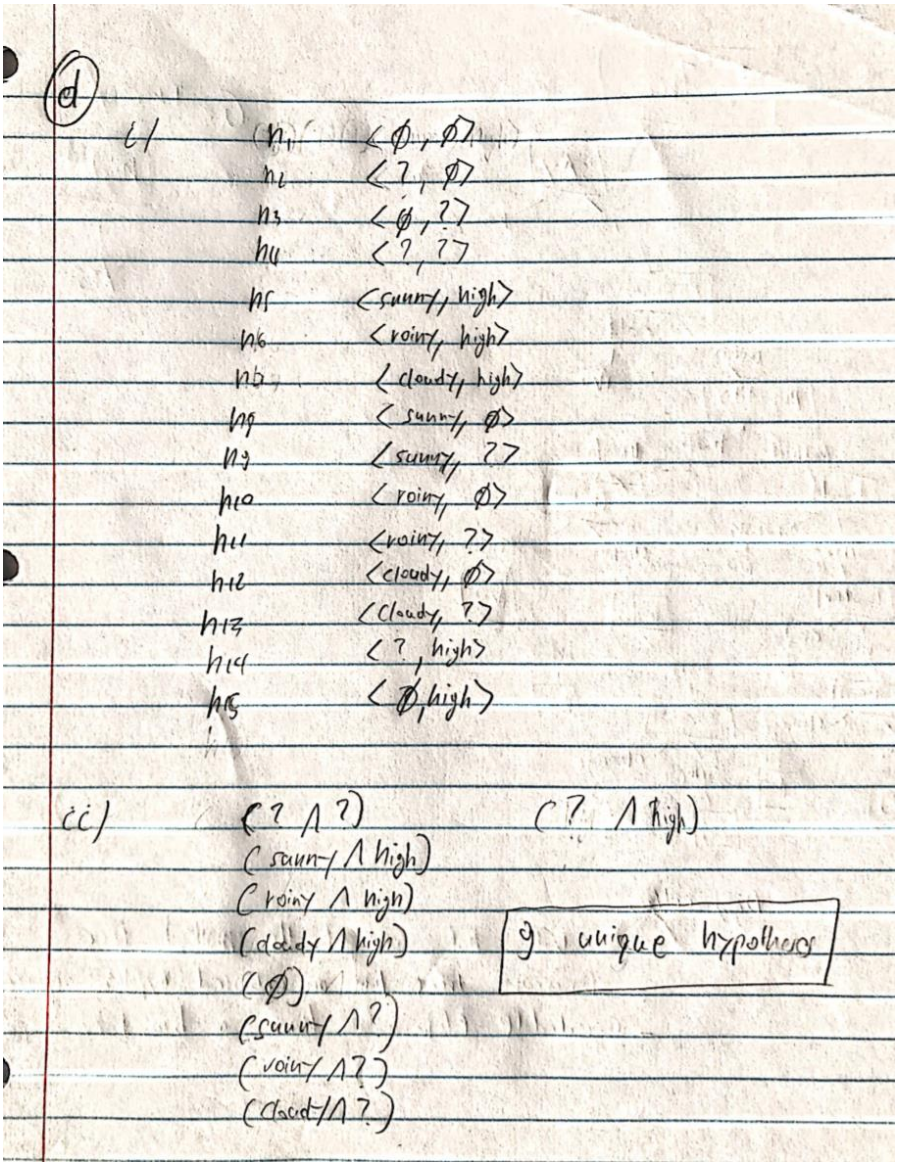


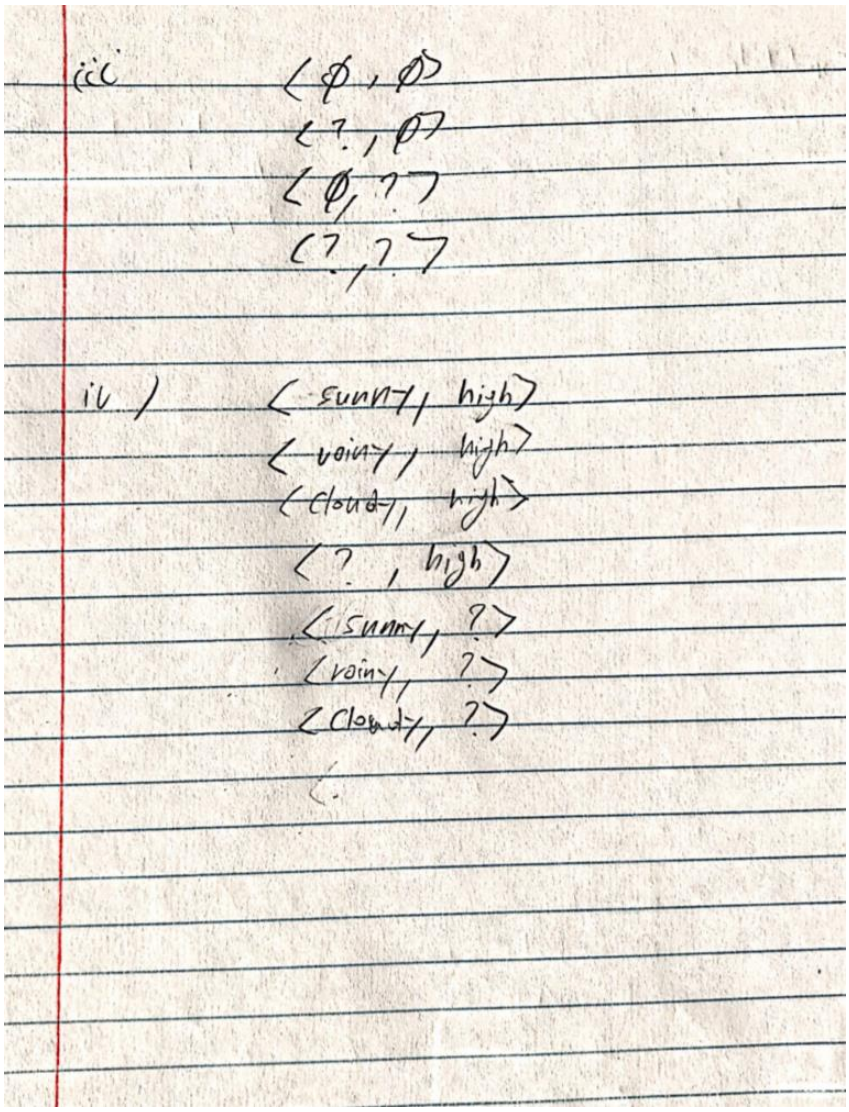
Difficulties:

- will not output the smallest tree
- Decision trees have an unbiased hypothesis space
- Candidate-Eliminate has a biased hypo space

Candidate Eliminate does not use the information gain heuristic to select the next attribute and find the smallest tree. Candidate-Eliminate will output unnecessarily large trees in its specific boundary, each w a large # of syntactical equivalents.

d.

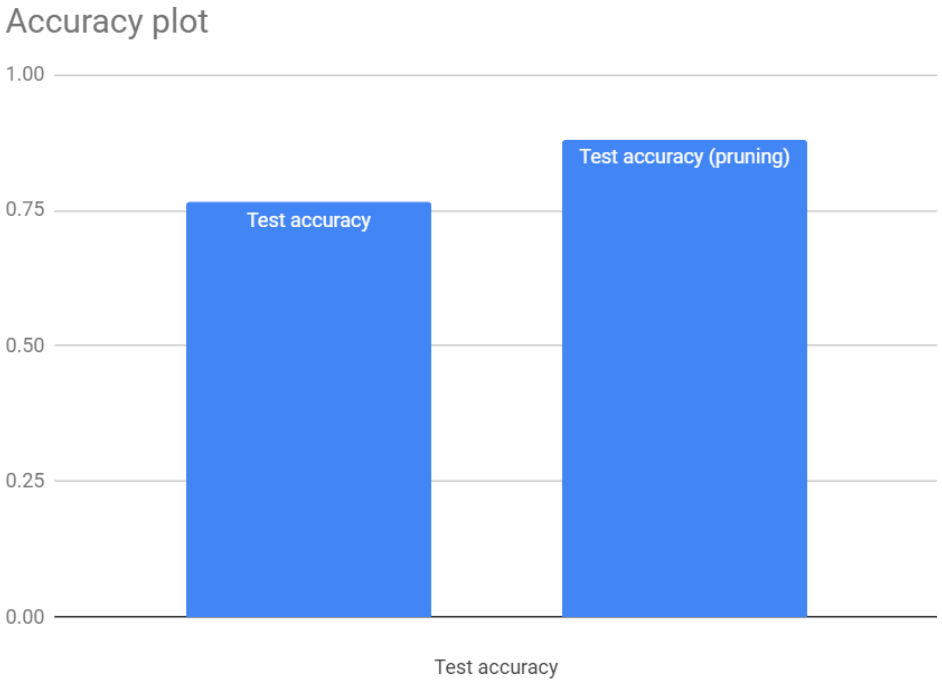

d
b/
h1 <∅, ∅>
h2 <?, ∅>
h3 <∅, ??
h4 <?, ??
h5 <sunny, high>
h6 <rainy, high>
h7 <cloudy, high>
h8 <sunny, ∅>
h9 <sunny, ??
h10 <rainy, ∅>
h11 <rainy, ??
h12 <cloudy, ∅>
h13 <cloudy, ??
h14 <?, high>
h15 <∅, high>
c/
(? ∧ ?)
(? ∧ high)
(sunny ∧ high)
(rainy ∧ high)
(cloudy ∧ high)
(∅)
(sunny ∧ ?)
(rainy ∧ ??)
(cloud ∧ ??)
9 unique hypotheses



e.

Pruning the tree yielded better results on the test set as expected. The accuracy went from 76.4% to 88% after pruning.

Increasing the percentage of the corrupted data, testIrisNoisy, yields the same test accuracy from 2% to 20% during experimentation. This is likely due to a bug in testIrisNoisy that I did not have enough time to fix.


Accuracy plot
1.00
0.75
0.50
0.25
0.00
Test accuracy
Test accuracy (pruning)
Test accuracy