CS 6350 Fall 2024 Machine Learning

Homework 1

Gnanambhal Shivarraman
UID: U1465199

September 20, 2024

Decision Tree

(1)

(a)

y, $p(\text{pos}) = \frac{2}{7}$, and $p(\text{neg}) = \frac{5}{7}$, then:

$$H(y) = -\frac{2}{7}\log_2\left(\frac{2}{7}\right) - \frac{5}{7}\log_2\left(\frac{5}{7}\right) \approx 0.863$$

x_1:

$x_1 = 0$, $p(\text{pos}) = \frac{1}{5}$, $p(\text{neg}) = \frac{4}{5}$, $H(x_1 = 0) = -\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\log_2\left(\frac{4}{5}\right) = 0.722$
$x_1 = 1$, $p(\text{pos}) = \frac{1}{2}$, $p(\text{neg}) = \frac{1}{2}$, $H(x_1 = 1) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$

Expected entropy: $\frac{5}{7} \times 0.72 + \frac{2}{7} \times 1 = 0.8$

Information gain: $0.86 - 0.8 = 0.06$

x_2:

$x_2 = 0$, $p(\text{pos}) = \frac{2}{3}$, $p(\text{neg}) = \frac{1}{3}$, $H(x_2 = 0) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) \approx 0.918$

$x_2 = 1$, $p(\text{pos}) = 0$, $p(\text{neg}) = 1$, $H(x_2 = 1) = 0$

Expected entropy: $\frac{3}{7} \times 0.918 + \frac{4}{7} \times 0 = 0.393$

Information gain: $0.86 - 0.393 = 0.467$

x_3:

$x_3 = 0$, $p(\text{pos}) = \frac{1}{4}$, $p(\text{neg}) = \frac{3}{4}$, $H(x_3 = 0) = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right) \approx 0.811$

$x_3 = 1$, $p(\text{pos}) = \frac{1}{3}$, $p(\text{neg}) = \frac{2}{3}$, $H(x_3 = 1) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) \approx 0.918$

Expected entropy: $\frac{4}{7} \times 0.811 + \frac{3}{7} \times 0.918 \approx 0.857$

Information gain: $0.86 - 0.857 = 0.003$

x_4:

$x_4 = 0$, $p(\text{pos}) = 0$, $p(\text{neg}) = 1$, $H(x_4 = 0) = 0$

$x_4 = 1$, $p(\text{pos}) = \frac{2}{3}$, $p(\text{neg}) = \frac{1}{3}$, $H(x_4 = 1) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) \approx 0.918$

Expected entropy: $\frac{3}{7} \times 0 + \frac{4}{7} \times 0.918 \approx 0.393$

Information gain: $0.86 - 0.393 = 0.467$

Use x_2 first. $x_2 = 1$, $y : p(\text{pos}) = \frac{2}{3}, p(\text{neg}) = \frac{1}{3}$:

$H(y) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} \approx 0.918.$

$x_1 = 0$: $H(x_1 = 0|y) = 1$; $x_1 = 1$: $H(x_1 = 1|y) = 0$.

Expected entropy: $\frac{2}{3} \times 1 + \frac{1}{3} \times 0 = 0.667.$

Information Gain: $0.918 - 0.667 = 0.251.$

x_3: $H(x_3 = 0|y) = 0$; $H(x_3 = 1|y) = 1$.

Expected entropy: $\frac{2}{3} \times 1 + \frac{1}{3} \times 0 = 0.667.$

Information Gain: $0.918 - 0.667 = 0.251.$

x_4: $H(x_4 = 0|y) = 0, H(x_4 = 1|y) = 0.$

Expected entropy: $\frac{2}{3} \times 0 + \frac{1}{3} \times 0 = 0.$

Information Gain: $0.918 - 0 = 0.918.$

I choose x_4: if $x_4 = 1$, $y = 1$; if $x_4 = 0$, $y = 0$.

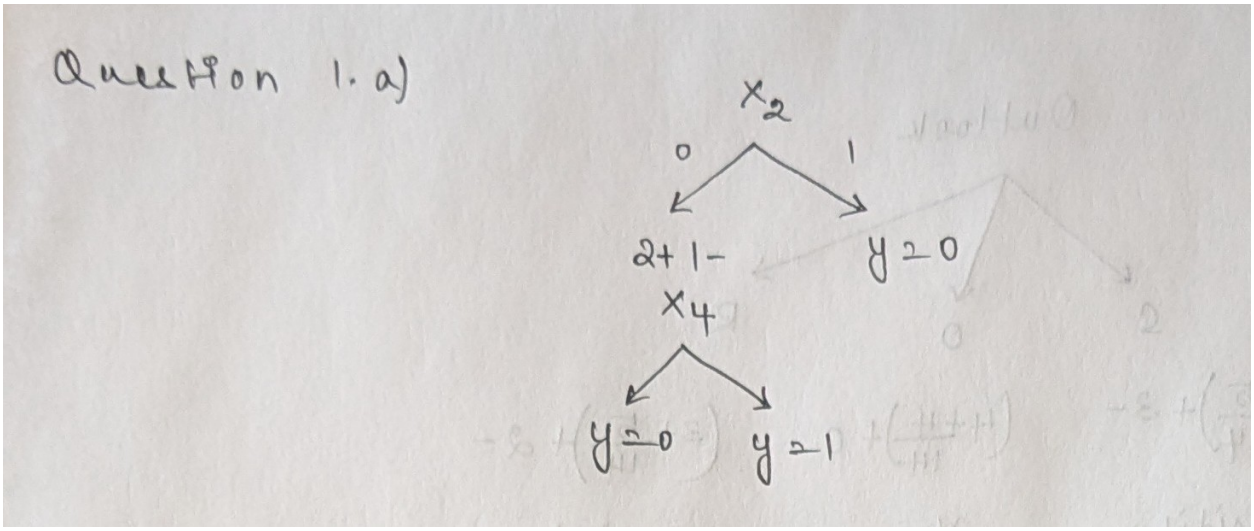


Table 1: Boolean Function Table

x1	x2	x3	x4	y
F	F	F	F	F
F	F	F	T	T
F	F	T	F	F
F	T	T	T	T
F	T	F	F	F
F	T	F	T	F
F	T	T	F	F
F	T	T	T	T
T	F	F	F	F
T	F	F	T	T
T	F	T	F	F
T	F	T	T	T
T	T	F	F	F
T	T	F	T	F
T	T	T	F	F
T	T	T	T	T

(b)

(2)

(a)

Answer: play, $p(\text{pos}) = \frac{9}{14}, p(\text{neg}) = \frac{5}{14}$, then ME(y) = $\frac{5}{14}$. outlook S, $p(\text{pos}) = \frac{2}{5}, p(\text{neg}) = \frac{3}{5}$, then ME(S) = $\frac{2}{5}$. outlook R, $p(\text{pos}) = \frac{3}{5}, p(\text{neg}) = \frac{2}{5}$, then ME(R) = $\frac{2}{5}$. outlook O, $p(\text{pos}) = 1, p(\text{neg}) = 0$, then ME(O) = 0.
Expected ME: $\frac{5}{14} \cdot \frac{2}{5} + \frac{5}{14} \cdot \frac{2}{5} = 0.29$.
Information gain of outlook: $0.36 - 0.29 = 0.07$.

Humidity H, $p(\text{pos}) = \frac{3}{7}, p(\text{neg}) = \frac{4}{7}$, then ME(H) = $\frac{3}{7}$. Humidity N, $p(\text{pos}) = \frac{6}{7}, p(\text{neg}) = \frac{1}{7}$, then ME(N) = $\frac{1}{7}$.
Expected ME: $\frac{7}{14} \cdot \frac{3}{7} + \frac{7}{14} \cdot \frac{1}{7} = 0.29$.
Information gain of humidity: $0.36 - 0.29 = 0.07$.

Wind W, $p(\text{pos}) = \frac{6}{8}, p(\text{neg}) = \frac{2}{8}$, then ME(W) = $\frac{2}{8}$. Wind S, $p(\text{pos}) = \frac{3}{6}, p(\text{neg}) = \frac{3}{6}$, then ME(S) = $\frac{3}{6}$.
Expected ME: $\frac{8}{14} \cdot \frac{2}{8} + \frac{6}{14} \cdot \frac{3}{6} = 0.357$.
Information gain of wind: $0.36 - 0.357 = 0.003$.

Temperature H, $p(\text{pos}) = \frac{2}{4}, p(\text{neg}) = \frac{2}{4}$, then ME(H) = $\frac{2}{4}$. Temperature M, $p(\text{pos}) = \frac{4}{6}, p(\text{neg}) = \frac{2}{6}$, then ME(M) = $\frac{2}{6}$. Temperature C, $p(\text{pos}) = \frac{3}{4}, p(\text{neg}) = \frac{1}{4}$, then ME(C) = $\frac{1}{4}$.
Expected ME: $\frac{4}{14} \cdot \frac{2}{4} + \frac{6}{14} \cdot \frac{2}{6} + \frac{4}{14} \cdot \frac{1}{4} = 0.357$.
Information gain of temperature: $0.36 - 0.357 = 0.003$. Thus, we choose outlook or humidity to split first.

S subset: $p(\text{pos}) = \frac{2}{5}, p(\text{neg}) = \frac{3}{5}$, then ME(y) = $\frac{2}{5}$. Humidity H, $p(\text{pos}) = 0, p(\text{neg}) = 1$, then ME(H) = 0. Humidity N, $p(\text{pos}) = 1, p(\text{neg}) = 0$, then ME(N) = 0.
Expected ME: $\frac{3}{5} \cdot 0 + \frac{2}{5} \cdot 0 = 0$.
Information gain of humidity: $0.4 - 0 = 0.4$.

Wind W, $p(\text{pos}) = \frac{1}{3}, p(\text{neg}) = \frac{2}{3}$, then ME(W) = $\frac{1}{3}$. Wind S, $p(\text{pos}) = \frac{1}{2}, p(\text{neg}) = \frac{1}{2}$, then ME(S) = $\frac{1}{2}$.
Expected ME: $\frac{3}{5} \cdot \frac{1}{3} + \frac{2}{5} \cdot \frac{1}{2} = 0.4$.
Information gain of wind: $0.4 - 0.4 = 0$.

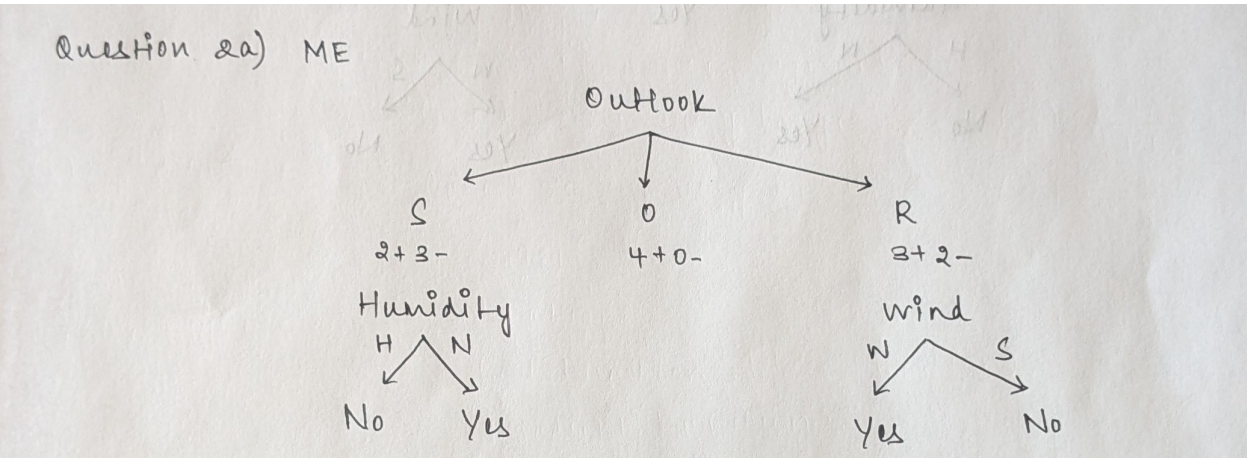
Temperature H, $p(\text{pos}) = 0, p(\text{neg}) = 1$, then ME(H) = 0. Temperature M, $p(\text{pos}) = \frac{1}{2}, p(\text{neg}) = \frac{1}{2}$, then ME(M) = $\frac{1}{2}$. Temperature C, $p(\text{pos}) = 1, p(\text{neg}) = 0$, then ME(C) = 0.
Expected ME: $\frac{2}{5} \cdot \frac{1}{2} = 0.2$.

Information gain of temperature: $0.4 - 0.2 = 0.2$.
Thus, we choose humidity to split S subset, Normal yes and High no.

Wind W, $p(\text{pos}) = 1, p(\text{neg}) = 0$, then ME(W) = 0. Wind S, $p(\text{pos}) = 0, p(\text{neg}) = 1$, then ME(S) = 0.
Expected ME: $\frac{3}{5} \cdot 0 + \frac{2}{5} \cdot 0 = 0$.
Information gain of wind: $0.4 - 0 = 0.4$.

Temperature M, $p(\text{pos}) = \frac{2}{3}, p(\text{neg}) = \frac{1}{3}$, then ME(M) = $\frac{1}{3}$. Temperature C, $p(\text{pos}) = \frac{1}{2}, p(\text{neg}) = \frac{1}{2}$, then ME(C) = $\frac{1}{2}$.
Expected ME: $\frac{2}{5} \cdot \frac{1}{2} + \frac{3}{5} \cdot \frac{1}{3} = 0.4$.
Information gain of temperature: $0.4 - 0.4 = 0$.

Therefore, we select wind to divide the R subset, with Weak indicating yes and Strong indicating no. All play labels have been assigned.



(b)

play, $p(pos) = \frac{9}{14}$, $p(neg) = \frac{5}{14}$. Then $GI(y) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.46$. outlook S, $p(pos) = \frac{2}{5}$, $p(neg) = \frac{3}{5}$, then $GI_S(y) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48$. outlook R, $p(pos) = \frac{3}{5}$, $p(neg) = \frac{2}{5}$, then $GI_R(y) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48$. outlook O, $p(pos) = 1$, $p(neg) = 0$, then $GI_O(y) = 1 - 1 = 0$.
Expected GI: $\frac{5}{14} \cdot 0.48 + \frac{5}{14} \cdot 0.48 = 0.205$.
Information gain of outlook: $0.46 - 0.205 = 0.255$.

Humidity H, $p(pos) = \frac{3}{7}$, $p(neg) = \frac{4}{7}$, then $GI_H(y) = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0.49$. Humidity N, $p(pos) = \frac{6}{7}$, $p(neg) = \frac{1}{7}$, then $GI_N(y) = 1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2 = 0.245$.
Expected GI: $\frac{7}{14} \cdot 0.49 + \frac{7}{14} \cdot 0.245 = 0.37$.
Information gain of humidity: $0.46 - 0.37 = 0.09$.

Wind W, $p(pos) = 68$, $p(neg) = 28$, then $GI_W(y) = 1 - (68)^2 - (28)^2 = 0.375$. Wind S, $p(pos) = 63$, $p(neg) = 36$, then $GI_S(y) = 1 - (36)^2 - (36)^2 = 0.5$.
Expected GI: $8 \cdot 0.375 + 6 \cdot 0.5 = 0.43$.
Information gain of wind: $0.46 - 0.43 = 0.03$.

Temperature H, $p(pos) = \frac{2}{4}$, $p(neg) = \frac{2}{4}$, then $GI_H(y) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0.5$. Temperature M, $p(pos) = \frac{4}{6}$, $p(neg) = \frac{2}{6}$, then $GI_M(y) = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = 0.44$. Temperature C, $p(pos) = 4$, $p(neg) = 1$, then $GI_C(y) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0.375$.
Expected GI: $4 \cdot 0.5 + 6 \cdot 0.44 + 4 \cdot 0.375 = 0.44$.

Information gain of temperature: $0.46 - 0.44 = 0.02$. Therefore, we choose Outlook.

Wind W, $p(pos) = \frac{1}{3}$, $p(neg) = \frac{2}{3}$, then $GI_W(y) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = 0.44$. Wind S, $p(pos) = \frac{1}{2}$, $p(neg) = \frac{1}{2}$, then $GI_S(y) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$.
Expected GI: $53 \cdot 0.44 + 25 \cdot 0.5 = 0.46$.
Information gain of wind: $0.48 - 0.46 = 0.02$.

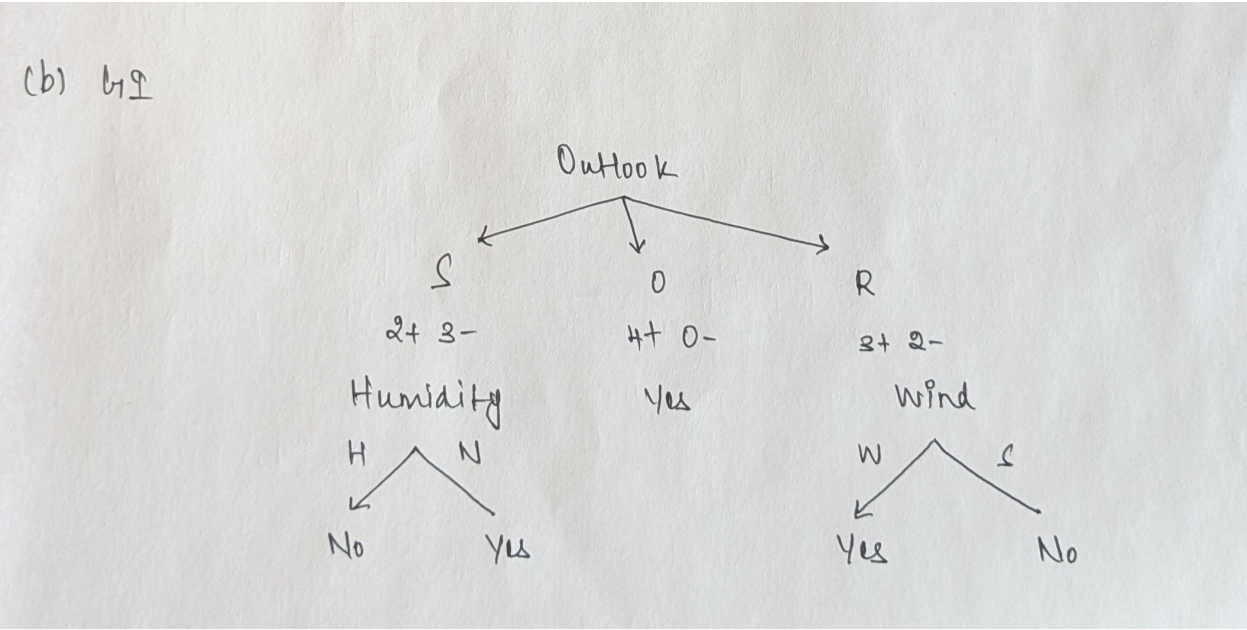
Temperature H, $p(pos) = 0$, $p(neg) = 1$, then $GI_H(y) = 0$. Temperature M, $p(pos) = \frac{1}{2}$, $p(neg) = \frac{1}{2}$, then $GI_M(y) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$. Temperature C, $p(pos) = 1$, $p(neg) = 0$, then $GI_C(y) = 0$.
Expected GI: $\frac{2}{5} \cdot 0.5 = 0.2$.
Information gain of humidity: $0.48 - 0.2 = 0.28$. Therefore, we choose Humidity to split the S subset.

the R subset: $p(pos) = 3$, $p(neg) = 2$, then $GI(y) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$. Humidity H, $p(pos) = 1$, $p(neg) = 2$, then $GI_H(y) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$. Humidity N, $p(pos) = \frac{2}{3}$, $p(neg) = \frac{1}{3}$, then $GI_N(y) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0.44$.
Expected GI: $35 \cdot 0.5 + 25 \cdot 0.44 = 0.476$.
Information gain of humidity: $0.48 - 0.476 = 0.004$.

Wind W, $p(pos) = 1$, $p(neg) = 0$, then $GI_W(y) = 0$. Wind S, $p(pos) = 0$, $p(neg) = 1$, then $GI_S(y) = 0$.
Expected GI: $35 \cdot 0 + 25 \cdot 0 = 0$.
Information gain of wind: $0.48 - 0 = 0.48$.

Temperature M, $p(pos) = \frac{2}{3}$, $p(neg) = \frac{1}{3}$, then $GI_M(y) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0.44$. Temperature C, $p(pos) = \frac{1}{2}$, $p(neg) = \frac{1}{2}$, then $GI_C(y) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$.
Expected GI: $53 \cdot 0.44 + 25 \cdot 0.5 = 0.46$.
Information gain of temperature: $0.48 - 0.46 = 0.02$.

Therefore, we select Wind to divide the R subset. All play labels have been assigned. The



(c)

The trees are identical. Although the Information gains vary, they reflect differences in the purity of the samples. In our dataset, despite these variations, the same attributes are consistently chosen to split the

data and grow the decision tree. As a result, the final structure of the tree remains the same.

(3)

(a)

Since "Sunny" appears 5 times, it can be chosen as the most frequent value. Hence, we now have an additional data point: Outlook: Sunny, Temperature: Mild, Humidity: Normal, Wind: Weak, Play: Yes.

New Information Gain Calculations:
play, $p(\text{pos}) = 10/15$ and $p(\text{neg}) = 5/15$: $H(y) = -\frac{10}{15}\log\frac{10}{15} - \frac{5}{15}\log\frac{5}{15} = 0.92$.
outlook is S, $p(\text{pos}) = 3/6$ and $p(\text{neg}) = 3/6$: $H(y) = -\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6}$.
outlook is R, $p(\text{pos}) = 3/5$ and $p(\text{neg}) = 2/5$: $H(y) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.97$.
outlook is O, $p(\text{pos}) = 1$ and $p(\text{neg}) = 0$: $H(y) = 0$.
Expected entropy outlook: $6/15 \cdot 1 + 5/15 \cdot 0.97 = 0.723$.
Information gain outlook: $0.92 - 0.723 = 0.197$.

Humidity:
humidity is H, $p(\text{pos}) = 3/7$ and $p(\text{neg}) = 4/7$: $H(y) = -\frac{3}{7}\log\frac{3}{7} - \frac{4}{7}\log\frac{4}{7} = 0.985$.
humidity is N, $p(\text{pos}) = 7/8$ and $p(\text{neg}) = 1/8$: $H(y) = -\frac{7}{8}\log\frac{7}{8} - \frac{1}{8}\log\frac{1}{8} = 0.54$.
Expected entropy humidity: $7/15 \cdot 0.985 + 8/15 \cdot 0.54 = 0.746$.
Information gain humidity: $0.92 - 0.746 = 0.174$.

Wind:
wind is W, $p(\text{pos}) = 7/9$ and $p(\text{neg}) = 2/9$: $H(y) = -\frac{7}{9}\log\frac{7}{9} - \frac{2}{9}\log\frac{2}{9} = 0.764$.
wind is S, $p(\text{pos}) = 3/6$ and $p(\text{neg}) = 3/6$: $H(y) = -\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6} = 1$.
Expected entropy wind: $9/15 \cdot 0.764 + 6/15 \cdot 1 = 0.86$.
Information gain wind: $0.92 - 0.86 = 0.06$.

Temperature:
temperature is H, $p(\text{pos}) = 2/4$ and $p(\text{neg}) = 2/4$: $H(y) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = 1$.
temperature is M, $p(\text{pos}) = 5/7$ and $p(\text{neg}) = 2/7$: $H(y) = -\frac{5}{7}\log\frac{5}{7} - \frac{2}{7}\log\frac{2}{7} = 0.86$.
temperature is C, $p(\text{pos}) = 3/4$ and $p(\text{neg}) = 1/4$: $H(y) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.811$.
Expected entropy temperature: $4/15 \cdot 1 + 7/15 \cdot 0.86 + 4/15 \cdot 0.811 = 0.884$.
Information gain temperature: $0.92 - 0.884 = 0.036$.

Conclusion: Based on the calculations, Outlook is chosen as the best feature.

(b)

Since Overcast appears 4 times the "Yes" label, it is the most common value. Therefore, the new data point is added: {Outlook: Overcast, Temperature: Mild, Humidity: Normal, Wind: Weak, Play: Yes}.

New Information Gain Calculations:
Play, $p(\text{pos}) = \frac{10}{15}$, $p(\text{neg}) = \frac{5}{15}$: $H(y) = -\frac{10}{15}\log_2\frac{10}{15} - \frac{5}{15}\log_2\frac{5}{15} = 0.92$.
Outlook = Sunny, $p(\text{pos}) = \frac{2}{5}$, $p(\text{neg}) = \frac{3}{5}$: $H(y) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$.
Outlook = Rainy, $p(\text{pos}) = \frac{3}{5}$, $p(\text{neg}) = \frac{2}{5}$: $H(y) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.97$.
Outlook = Overcast, $p(\text{pos}) = 1$, $p(\text{neg}) = 0$: $H(y) = 0$.
Expected entropy Outlook: $\frac{5}{15} \cdot 0.971 + \frac{5}{15} \cdot 0.97 = 0.647$.

Information gain Outlook: $0.92 - 0.647 = 0.27$.

Humidity: Humidity = High, $p(\text{pos}) = \frac{3}{7}$, $p(\text{neg}) = \frac{4}{7}$: $H(y) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.985$.
Humidity = Normal, $p(\text{pos}) = \frac{7}{8}$, $p(\text{neg}) = \frac{1}{8}$: $H(y) = -\frac{7}{8}\log_2\frac{7}{8} - \frac{1}{8}\log_2\frac{1}{8} = 0.54$.
Expected entropy Humidity: $\frac{7}{15} \cdot 0.985 + \frac{8}{15} \cdot 0.54 = 0.746$.

Information gain Humidity: $0.92 - 0.746 = 0.174$.

Wind:
Wind = Weak, $p(\text{pos}) = \frac{7}{9}$, $p(\text{neg}) = \frac{2}{9}$: $H(y) = -\frac{7}{9}\log_2\frac{7}{9} - \frac{2}{9}\log_2\frac{2}{9} = 0.764$.
Wind = Strong, $p(\text{pos}) = \frac{3}{6}$, $p(\text{neg}) = \frac{3}{6}$: $H(y) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$.
Expected entropy Wind: $\frac{9}{15} \cdot 0.764 + \frac{6}{15} \cdot 1 = 0.86$.

Information gain Wind: $0.92 - 0.86 = 0.06$.

Temperature: Temperature = Hot, $p(\text{pos}) = \frac{2}{4}$, $p(\text{neg}) = \frac{2}{4}$: $H(y) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$.
Temperature = Mild, $p(\text{pos}) = \frac{5}{7}$, $p(\text{neg}) = \frac{2}{7}$: $H(y) = -\frac{5}{7}\log_2\frac{5}{7} - \frac{2}{7}\log_2\frac{2}{7} = 0.86$.
Temperature = Cool, $p(\text{pos}) = \frac{3}{4}$, $p(\text{neg}) = \frac{1}{4}$: $H(y) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.811$.
Expected entropy Temperature: $\frac{4}{15} \cdot 1 + \frac{7}{15} \cdot 0.86 + \frac{4}{15} \cdot 0.811 = 0.884$.

Information gain Temperature: $0.92 - 0.884 = 0.036$.

Therefore, Outlook can still be chosen as the first feature.

(c)

Since Overcast appears 4 times in the "Yes" label, it is selected as the most common value. Therefore, we add one more data point: {Outlook: Overcast, Temperature: Mild, Humidity: Normal, Wind: Weak, Play: Yes}.

New Information Gain Calculations:
Play, $p(\text{pos}) = \frac{10}{15}$, $p(\text{neg}) = \frac{5}{15}$: $H(y) = -\frac{10}{15}\log_2\frac{10}{15} - \frac{5}{15}\log_2\frac{5}{15} = 0.92$.
Outlook = Sunny, $p(\text{pos}) = \frac{5}{14}$, $p(\text{neg}) = \frac{9}{14}$: $H(y) = -\frac{5}{14}\log_2\frac{5}{14} - \frac{9}{14}\log_2\frac{9}{14} = 0.99$.
Outlook = Rainy, $p(\text{pos}) = \frac{5}{14}$, $p(\text{neg}) = \frac{9}{14}$: $H(y) = -\frac{5}{14}\log_2\frac{5}{14} - \frac{9}{14}\log_2\frac{9}{14} = 0.953$.
Outlook = Overcast, $p(\text{pos}) = 1$, $p(\text{neg}) = 0$: $H(y) = 0$.
Expected entropy Outlook: $\frac{5}{15} \cdot 0.99 + \frac{5}{15} \cdot 0.953 = 0.693$.

Information gain Outlook: $0.92 - 0.693 = 0.227$.

Humidity: Humidity = High, $p(\text{pos}) = \frac{3}{7}$, $p(\text{neg}) = \frac{4}{7}$: $H(y) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.985$.
Humidity = Normal, $p(\text{pos}) = \frac{7}{8}$, $p(\text{neg}) = \frac{1}{8}$: $H(y) = -\frac{7}{8}\log_2\frac{7}{8} - \frac{1}{8}\log_2\frac{1}{8} = 0.54$.
Expected entropy Humidity: $\frac{7}{15} \cdot 0.985 + \frac{8}{15} \cdot 0.54 = 0.746$.

Information gain Humidity: $0.92 - 0.746 = 0.174$.

Wind: Wind = Weak, $p(\text{pos}) = \frac{7}{9}$, $p(\text{neg}) = \frac{2}{9}$: $H(y) = -\frac{7}{9}\log_2\frac{7}{9} - \frac{2}{9}\log_2\frac{2}{9} = 0.764$.
Wind = Strong, $p(\text{pos}) = \frac{3}{6}$, $p(\text{neg}) = \frac{3}{6}$: $H(y) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$.
Expected entropy Wind: $\frac{9}{15} \cdot 0.764 + \frac{6}{15} \cdot 1 = 0.86$.

Information gain Wind: $0.92 - 0.86 = 0.06$.

Temperature: Temperature = Hot, $p(\text{pos}) = \frac{2}{4}$, $p(\text{neg}) = \frac{2}{4}$: $H(y) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$.
Temperature = Mild, $p(\text{pos}) = \frac{5}{7}$, $p(\text{neg}) = \frac{2}{7}$: $H(y) = -\frac{5}{7}\log_2\frac{5}{7} - \frac{2}{7}\log_2\frac{2}{7} = 0.86$.
Temperature = Cool, $p(\text{pos}) = \frac{3}{4}$, $p(\text{neg}) = \frac{1}{4}$: $H(y) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.811$.
Expected entropy Temperature: $\frac{4}{15} \cdot 1 + \frac{7}{15} \cdot 0.86 + \frac{4}{15} \cdot 0.811 = 0.884$.

Information gain Temperature: $0.92 - 0.884 = 0.036$.

Therefore, I can still choose Outlook as the first feature, which remains the best.

(d)

Continuing from the previous discussion, we select Outlook as the first attribute. If the Outlook is Overcast, the label is "Yes".

the Sunny subset: $p(\text{pos}) = \frac{14}{17}$ and $p(\text{neg}) = \frac{3}{17}$, then $H(y) = -\frac{14}{17}\log_2\frac{14}{17} - \frac{3}{17}\log_2\frac{3}{17} = 0.99$.
Humidity = High: $p(\text{pos}) = 0$, $p(\text{neg}) = 1$, thus $H_H(y) = 0$.
Humidity = Normal: $p(\text{pos}) = 1$, $p(\text{neg}) = 0$, Therefore $H_N(y) = 0$.
Expected entropy is: $\frac{3}{5}\cdot 0 + \frac{2}{5}\cdot 0 = 0$.

Information gain Outlook: $0.99 - 0 = 0.99$.

Next, Temperature = Hot: $p(\text{pos}) = 0$, $p(\text{neg}) = 1$, thus $H_H(y) = 0$.
Temperature = Mild, $p(\text{pos}) = \frac{5}{14}$ and $p(\text{neg}) = \frac{1}{14}$, leading to $H_M(y) = -\frac{5}{14}\log_2\frac{5}{14} - \frac{1}{14}\log_2\frac{1}{14} = 0.983$.
Temperature = Cool: $p(\text{pos}) = 1$, $p(\text{neg}) = 0$, thus $H_C(y) = 0$.
Expected gain is: $\frac{2}{5}\cdot 0.983 + \frac{1}{5}\cdot 0 = 0.433$.

Information gain Outlook: $0.99 - 0.433 = 0.557$.
Thus, we choose Humidity to split the Sunny subset, where Normal leads to "Yes" and High to "No".

the Rainy subset: $p(\text{pos}) = \frac{3}{14}$ and $p(\text{neg}) = \frac{2}{14}$, so $H(y) = -\frac{3}{14}\log_2\frac{3}{14} - \frac{2}{14}\log_2\frac{2}{14} = 0.953$.
Humidity = High: $p(\text{pos}) = \frac{1}{2}$ and $p(\text{neg}) = \frac{1}{2}$ gives $H_H(y) = 1$.
Humidity = Normal: $p(\text{pos}) = \frac{3}{14}$ and $p(\text{neg}) = 1$, leading to $H_N(y) = -\frac{3}{14}\log_2\frac{3}{14} - 0 = 0.879$.
Expected entropy is: $2\cdot 1 + \frac{14}{14}\cdot 0.879 = 0.924$.

Information gain Rainy: $0.953 - 0.924 = 0.029$.

Wind = Weak: $p(\text{pos}) = 1$, $p(\text{neg}) = 0$, so $H_W(y) = 0$.
Wind = Strong: $p(\text{pos}) = 0$, $p(\text{neg}) = 1$, thus $H_S(y) = 0$.
Expected entropy is: 0.

Information gain Wind: $0.953 - 0 = 0.953$.
The tree structure is finalized based on these calculations.

(4)

We can prove this by examining the second derivative of the entropy function defined as f:

$$f(x) = -x\log_2(x) - (1-x)\log_2(1-x)$$

Calculating the second derivative gives:

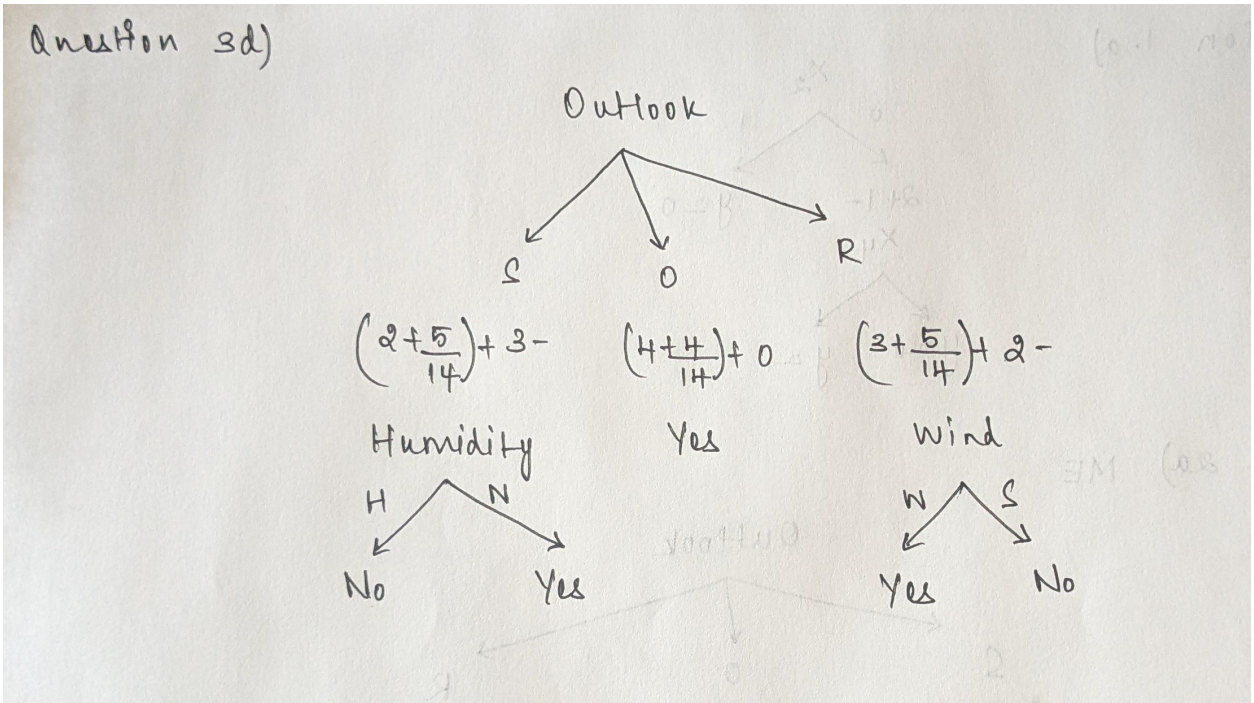
$$\nabla^2 f(x) = \nabla(\nabla f(x)) = \nabla\left(-\log_2(x) - \log_2(1-x)\right) = -\frac{1}{\ln(2)}\cdot\frac{1}{x(1-x)}$$

This expression is negative $0 < x < 1$, indicating that f is concave.
Now, the Information gain can be expressed as:

$$\text{GAIN} = f(S) - \sum_{i=1}^{k}\frac{|S_i|}{|S|}f(S_i)$$

where $|S| = N$, with p representing positives and n representing negatives. By applying Jensen's Inequality, we have:

$$\sum_{i=1}^{k}\frac{|S_i|}{|S|}f(S_i) \le f\left(\frac{|P|}{N}\right)N$$



Since f is concave, we conclude that Information gain is always non-negative, confirming that splitting the data improves or maintains purity.

(5)

When working with regression datasets, the measure of purity is associated with the mean or variance of the data. Therefore, the gain can be defined as:

$$\text{GAIN} = \text{Variance}(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} \text{Variance}(S_i)$$

This mula calculates the reduction in variance as a result of splitting the dataset, guiding us in selecting the optimal attribute the decision tree.

Decision Tree Practice

1

I have created a code repository on GitHub. The repository link is as follows:

`https://github.com/GnanambhalSvRam/CS6350`

In the repository, I have committed a README.md file and created a folder named "DecisionTree."

Table 2: Train and Test Errors by Heuristic and Maximum Depth

Heuristic Type	Depth	Train Error	Test Error
Information Gain	1	0.302	0.2967
	2	0.222	0.2225
	3	0.181	0.1964
	4	0.082	0.1470
	5	0.027	0.1044
	6	0.000	0.1250
Gini Index	1	0.302	0.2967
	2	0.222	0.2225
	3	0.176	0.1841
	4	0.089	0.1332
	5	0.027	0.1044
	6	0.000	0.1250
Majority Error	1	0.302	0.2967
	2	0.301	0.3159
	3	0.221	0.2651
	4	0.107	0.1854
	5	0.038	0.1360
	6	0.000	0.1538

2

(a) `https://github.com/GnanambhalSvRam/CS6350/tree/main/DecisionTree`

(b) Table shown in this page (Table 2: Train and Test Errors by Heuristic and Maximum Depth

(c)Answer:

Using a car dataset, we observe that Information Gain methods based on entropy and the Gini index outperm the ME gain in terms of accuracy. As the depth increases, training error steadily decreases, but this leads to overfitting. Notably, the permance at depth 6 seems to degrade compared to depth 5 when evaluated on the test data.

3

(a) Please refer Table 3: Results with 'unknown' as a specific value in Page 11

(b) Please refer Table 4: Results with 'unknown' as a missing value in Page 12

(c) Answer:

As the tree depth increases, the training error tends to decrease; however, after reaching a specific depth, the test error begins to rise, indicating that overfitting has occurred. When implementing our method for handling unknown attribute values, there is a slight improvement in performance, although the enhancement is not significant. In banking datasets with a higher number of features, using Majority Error with Information Gain yields better results on the test datasets.

Table 3: Results with 'unknown' as a specific value

Heuristic	Depth	Train Error	Test Error
information_gain	1	0.1192	0.1248
	2	0.1060	0.1114
	3	0.1006	0.1080
	4	0.0800	0.1266
	5	0.0624	0.1426
	6	0.0480	0.1510
	7	0.0372	0.1612
	8	0.0292	0.1666
	9	0.0222	0.1718
	10	0.0182	0.1770
	11	0.0150	0.1790
	12	0.0136	0.1824
	13	0.0132	0.1838
	14	0.0132	0.1852
	15	0.0132	0.1852
	16	0.0132	0.1852
gini_index	1	0.1088	0.1166
	2	0.1042	0.1088
	3	0.0936	0.1156
	4	0.0754	0.1318
	5	0.0606	0.1500
	6	0.0484	0.1636
	7	0.0366	0.1738
	8	0.0268	0.1776
	9	0.0214	0.1828
	10	0.0172	0.1864
	11	0.0140	0.1904
	12	0.0136	0.1914
	13	0.0132	0.1922
	14	0.0132	0.1942
	15	0.0132	0.1942
	16	0.0132	0.1942
majority_error	1	0.1088	0.1166
	2	0.1042	0.1088
	3	0.0966	0.1146
	4	0.0840	0.1264
	5	0.0686	0.1374
	6	0.0626	0.1472
	7	0.0572	0.1576
	8	0.0524	0.1664
	9	0.0490	0.1744
	10	0.0440	0.1854
	11	0.0384	0.2008
	12	0.0314	0.2108
	13	0.0256	0.2278
	14	0.0200	0.2432
	15	0.0168	0.2446
	16	0.0132	0.2484

Table 4: Training and Testing Errors for Different Heuristics

Heuristic	Depth	Train Error	Test Error
information_gain	1	0.1192	0.1248
	2	0.1060	0.1114
	3	0.1022	0.1108
	4	0.0876	0.1286
	5	0.0698	0.1406
	6	0.0560	0.1460
	7	0.0444	0.1554
	8	0.0366	0.1580
	9	0.0304	0.1636
	10	0.0244	0.1688
	11	0.0206	0.1702
	12	0.0182	0.1740
	13	0.0180	0.1752
	14	0.0180	0.1784
	15	0.0180	0.1792
	16	0.0180	0.1792
gini_index	1	0.1088	0.1166
	2	0.1052	0.1104
	3	0.1010	0.1096
	4	0.0882	0.1276
	5	0.0720	0.1440
	6	0.0564	0.1512
	7	0.0444	0.1628
	8	0.0352	0.1652
	9	0.0296	0.1704
	10	0.0236	0.1748
	11	0.0200	0.1766
	12	0.0184	0.1802
	13	0.0180	0.1816
	14	0.0180	0.1848
	15	0.0180	0.1858
	16	0.0180	0.1858
majority_error	1	0.1088	0.1166
	2	0.1050	0.1102
	3	0.0976	0.1162
	4	0.0866	0.1230
	5	0.0782	0.1292
	6	0.0730	0.1346
	7	0.0658	0.1472
	8	0.0582	0.1608
	9	0.0522	0.1758
	10	0.0470	0.1864
	11	0.0416	0.1984
	12	0.0366	0.2128
	13	0.0304	0.2244
	14	0.0244	0.2354
	15	0.0216	0.2362
	16	0.0180	0.2406