Jonathan Nusantara (jan 265) Assignment #1

Problem 1. (10 points). Design the decision tree for the tennis data using Gini impurity measure. Compute the Gini measure for all attributes at each node, and continue until all the examples are correctly labeled by the tree.

Playing Tennis Example

Day	Outlook	Temp	Humid	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	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	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

$$Gini(S) = 1 - \sum_{i=1}^{l} \left(\frac{n_i}{n}\right)^2 \qquad \{v_1, \dots, v_a\} \quad \text{val}$$

$$Gini(S;A) = \underbrace{\sum_{j=1}^{a} \frac{|S_{vj}|}{|S|}}_{\text{weighted ave}} Gini(S_{vj})$$

$S = \{1, \dots, 14\} \rightarrow \{5N, 9Y\}$, n_1 n_2, $l = 2$

$$Gini(S;\text{Outlook}) = \frac{5}{14} Gini(S_{\text{Sunny}}) + \frac{5}{14} Gini(S_{\text{Rain}}) + \frac{4}{14} Gini(S_{\text{Overcast}})$$

$$= \frac{5}{14}\left(1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right) + \frac{5}{14}\left(1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right) + \frac{4}{14}\left(1 - \left(\frac{0}{4}\right)^2 - \left(\frac{4}{4}\right)^2\right)$$

$$= \frac{6}{35} + \frac{6}{35} + \frac{4}{14}(0) \qquad (1)$$

$$= \frac{12}{35} \approx 0.343$$

$$Gini(S;\text{Temp}) = \frac{4}{14} Gini(S_{\text{Hot}}) + \frac{6}{14} Gini(S_{\text{Mild}}) + \frac{4}{14} Gini(S_{\text{Cool}})$$

$$= \frac{4}{14}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right) + \frac{6}{14}\left(1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2\right) + \frac{4}{14}\left(1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2\right)$$

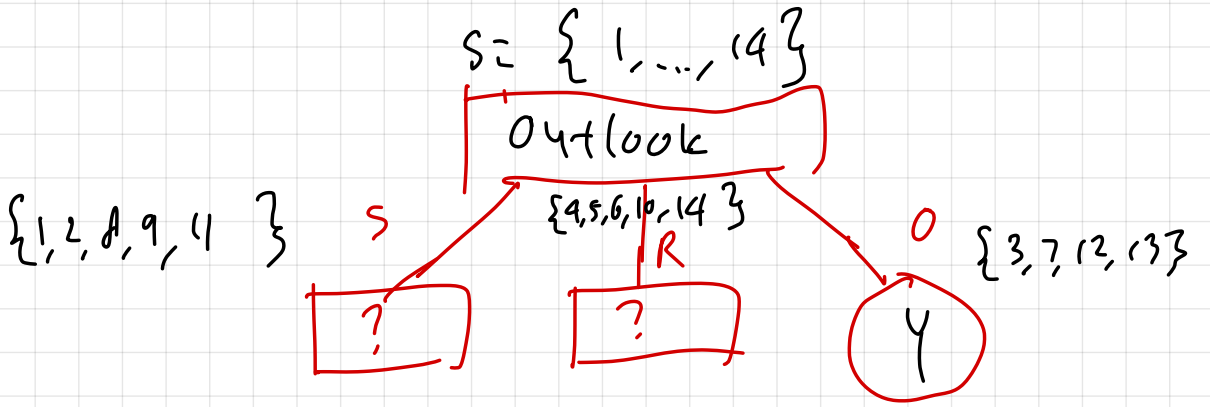
$$= \frac{1}{7} + \frac{4}{21} + \frac{3}{28}$$

$$= \frac{37}{84} \approx 0.440$$

$$Gini(S; Humid) = \frac{7}{14} Gini(S_{High}) + \frac{7}{14} Gini(S_{Normal})$$
$$= \frac{7}{14}\left(1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2\right) + \frac{7}{14}\left(1 - \left(\frac{1}{7}\right)^2 - \left(\frac{6}{7}\right)^2\right)$$
$$= 12/49 + \frac{6}{49}$$
$$= \underline{0.367}$$

$$Gini(S; wind) = \frac{8}{14} Gini(S_{weak}) + \frac{6}{14} Gini(S_{strong})$$
$$= \frac{8}{14}\left(1 - \left(\frac{2}{8}\right)^2 - \left(\frac{6}{8}\right)^2\right) + \frac{6}{14}\left(1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2\right)$$
$$= \frac{3}{14} + \frac{3}{14}$$
$$= \underline{0.429}$$ → minimize impurity

Since $Gini(S; Outlook)$ is lowest, use 'Outlook'



$$Gini(S_{sunny}, Temp) = \frac{2}{5} Gini(S_{Hot}) + \frac{2}{5} Gini(S_{mild}) + \frac{1}{5} Gini(S_{cool})$$
$$= \frac{2}{5}\left(1 - \left(\frac{2}{2}\right)^2 - \left(\frac{0}{2}\right)^2\right) + \frac{2}{5}\left(1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right) + \frac{1}{5}\left(1 - \left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2\right)$$
$$= 0 + \frac{1}{5} + 0$$
$$= \underline{0.2}$$

$$Gini(S_{sunny}, Humid) = \frac{3}{5} Gini(S_{High}) + \frac{2}{5} Gini(S_{Normal})$$
$$= \frac{2}{5}\left(1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2\right) + \frac{2}{5}\left(1 - \left(\frac{2}{2}\right)^2 - \left(\frac{0}{2}\right)^2\right)$$
$$= 0 + 0$$
$$= \underline{0}$$

$$Gini(S_{sunny}; Wind) = \frac{3}{5} Gini(S_{weak}) + \frac{2}{5} Gini(S_{strong})$$

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

$$= 4/15 + 1/5$$

$$= 7/15 \approx \underline{0.467}$$

Since $Gini(S_{sunny}; Humid) = 0 \Rightarrow$ smallest, use 'Humid'

$$Gini(S_{Rain}; Temp) = \frac{3}{5} Gini(S_{mild}) + \frac{2}{5} Gini(S_{cool})$$

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

$$= 4/15 + 1/5$$

$$= \underline{0.467}$$

$$Gini(S_{Rain}; Humid) = \frac{3}{5}\left(1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2\right) + \frac{2}{5}\left(1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right)$$

$$= 4/15 + 1/5$$

$$= \underline{0.467}$$

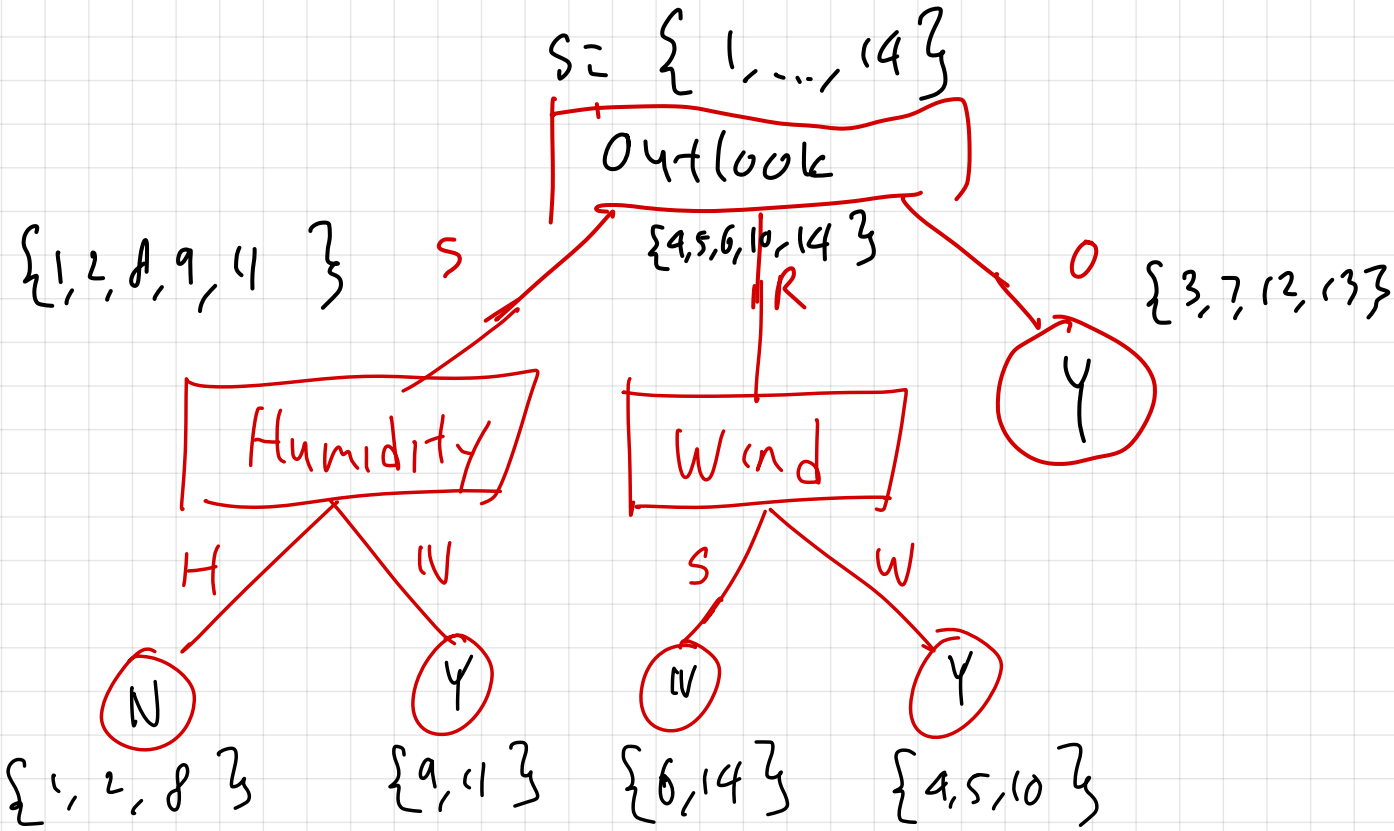
$$Gini(S_{Rain}; Wind) = \frac{3}{5} Gini(S_{weak}) + \frac{2}{5} Gini(S_{strong})$$

$$= \frac{3}{5}\left(1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2\right) + \frac{2}{5}\left(1 - \left(\frac{2}{2}\right)^2 - \left(\frac{0}{2}\right)^2\right)$$

$$= 0 + 0$$

$$= \underline{0}$$

$Gini(S_{Rain}; Wind)$ is lowest, choose 'Wind'


S = {1, ..., 14}
Outlook
{1,2,8,9,11}
S
{4,5,6,10,14}
R
O
{3,7,12,13}
Humidity
Wind
Y
H
N
Y
S
W
N
Y
{1,2,8}
{9,11}
{6,14}
{4,5,10}

↑

final tree

Problem 2 (20 points). Consider the training set given in Table 1. The attribute "Shirt Size Fine" is a **refinement** of the attribute "Shirt Size", wherein the value "Medium" has been further categorized into two values "Small-Medium" and "Large-Medium". The goal of this problem is to see the reasonably intuitive assertion that the information gain is higher for an attribute that is a refinement of another.

Note: when computing information gain, use base 2 logarithm.

1. What is the entropy of the labels?
2. Compute the information gain for "Shirt Size" and "Shirt Size Fine". Which is higher?
3. A function f is called concave if for x, y, and any $0 \leq \lambda \leq 1$,

$$f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda) f(y). \tag{1}$$

The `logarithm` function is concave. You can assume that as a fact for all other parts of this assignment. For this part, you have to show that Equation (1) holds for $\lambda = 1/2$, and f is the logarithm function.

4 The following inequality is called as the log-sum inequality. For positive x_1, x_2, y_1, y_2,

$$x_1 \log \frac{x_1}{y_1} + x_2 \log \frac{x_2}{y_2} \geq (x_1 + x_2) \log \frac{x_1 + x_2}{y_1 + y_2}. \tag{2}$$

Prove this using the concavity of logarithms.

	Gender	Car Type	Shirt Size	Shirt Size Fine	Label
1	M	Family	Small 1	Small	+ 1
2	M	Sports	Medium	Small-Medium	+ 2
3	M	Family	Large	Large	−
4	F	Sports	Small 2	Small	+ 3
5	M	Sports	Extra Large	Extra Large	+ 4
6	F	Luxury	Small 3	Small	−
7	M	Family	Medium	Large-Medium	−
8	M	Sports	Extra Large	Extra Large	+ 5
9	M	Sports	Large	Large	+ 6
10	F	Luxury	Medium	Large-Medium	−
11	F	Sports	Medium	Large-Medium	+ 7
12	F	Family	Small 4	Small	+ 8
13	F	Luxury	Large	Large	−
14	M	Luxury	Medium	Small Medium	−
15	F	Family	Medium	Small-Medium	+ 9

Table 1: Training Data

5* We will show that part 2 of this problem can be generalized as follows. Consider a training set of any size with the four features as in Table 1, again with the property that "Shirt Size Fine" is a **refinement** of the attribute "Shirt Size". Show that the information gain for "Shirt Size Fine" is always at least that for "Shirt Size". This is a heuristic justification for the fact that IG picks attributes that have more possibilities.
(**hint:** Suppose n_m are the number of medium's, and n_{ml} and n_{ms} are the number of small-medium, and large medium respectively. then $n_{ml} + n_{ms} = n_m$. You may also want to define terms such at n_m^+ which are the number of medium's with +ve labels). You may have to use Equation (2) carefully!

① $H(S) = \sum_{i=1}^{l} \frac{n_i}{n} \log_2\left(\frac{n}{n_i}\right)$

$= \frac{9}{15} \log_2\left(\frac{15}{9}\right) + \frac{6}{15} \log_2\left(\frac{15}{6}\right)$

$= 0.44218 + 0.52877$

$= \underline{0.97095} \quad < 1$

$n_1 \rightarrow +$
$n_2 \rightarrow -$

② $IG(S;A) = H(S) - \sum_{j=1}^{a} \frac{|S_{vj}|}{|S|} \times H(S_{vj})$

$IG(S;\text{Shirt Size}) = 0.97095 - \sum_{j=1}^{a} \frac{|S_{vj}|}{|S|} \times H(S_{vj})$

$H(S_{small}) = \frac{3}{4} \log_2\left(\frac{4}{3}\right) + \frac{1}{4} \log_2\left(\frac{4}{1}\right) = 0.81128$

$H(S_{medium}) = \frac{3}{6} \log_2\left(\frac{6}{3}\right) + \frac{3}{6} \log_2\left(\frac{6}{3}\right) = 1$

$H(S_{large}) = \frac{2}{3} \log_2\left(\frac{3}{2}\right) + \frac{1}{3} \log_2\left(\frac{3}{1}\right) = 0.91830$

$H(S_{xLarge}) = \frac{2}{2} \log_2\left(\frac{2}{2}\right) + \cancel{\frac{0}{2} \log_2\left(\frac{2}{0}\right)} = 0$

$IG(S;\text{Shirt Size}) = 0.97095 - \frac{4}{15}(0.81128) - \frac{6}{15}(1)$
$- \frac{3}{15}(0.91830) - \frac{2}{15}(0)$

$= \underline{0.170949}$

$$H(S_{small}) = \frac{3}{4}\log_2\left(\frac{4}{3}\right) + \frac{1}{4}\log_2\left(\frac{4}{1}\right) = 0.81128$$

$$H(S_{S-M}) = \frac{2}{3}\log_2\left(\frac{3}{2}\right) + \frac{1}{3}\log_2\left(\frac{3}{1}\right) = 0.91830$$

$$H(S_{L-M}) = \frac{2}{3}\log_2\left(\frac{3}{2}\right) + \frac{1}{3}\log_2\left(\frac{3}{1}\right) = 0.91830$$

$$H(S_{large}) = \frac{2}{3}\log_2\left(\frac{3}{2}\right) + \frac{1}{3}\log_2\left(\frac{3}{1}\right) = 0.91830$$

$$H(S_{XLarge}) = \frac{2}{2}\log_2\left(\frac{2}{2}\right) + \frac{0}{2}\log_2\left(\frac{2}{0}\right) = 0$$

$$IG(S; \underset{Final}{Shirt\ Size}) = 0.97095 - \frac{4}{15}(0.81128) - \frac{3}{15}(0.91830) - \frac{3}{15}(0.91830) - \frac{3}{15}(0.91830) - \frac{2}{15}(0)$$

$$= 0.203629$$

$$IG(S; \underset{Final}{Shirt\ Size}) > IG(S; Shirt\ Size)$$

$$0.203629 > 0.170949$$

(3) $f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda) f(y).$ (1) $\lambda = \frac{1}{2}$

$$f\left(\tfrac{1}{2}x + \tfrac{1}{2}y\right) \geq \tfrac{1}{2}f(x) + \tfrac{1}{2}f(y)$$

$$\log\left(\tfrac{x}{2} + \tfrac{y}{2}\right) \geq \tfrac{1}{2}\log(x) + \tfrac{1}{2}\log(y)$$

$$\log\left(\tfrac{x+y}{2}\right) \geq \tfrac{1}{2}\left(\log(x) + \log(y)\right)$$

$$\log\left(\tfrac{x+y}{2}\right) \geq \tfrac{1}{2}\left(\log(xy)\right)$$

$$\cancel{\log}\left(\tfrac{x+y}{2}\right) \geq \cancel{\log}\left(\sqrt{xy}\right) \qquad \times \cancel{e^{\log}} \ldots\ldots$$

$$\tfrac{x+y}{2} \geq \sqrt{xy} \qquad \leftarrow \ldots \; e$$

$$\frac{x^2+2xy+y^2}{4} \geq xy$$

$$x^2+y^2 \geq 4xy - 2xy$$

$$x^2+y^2 \geq 2xy$$

$$(x^2+y^2) - 2xy \geq 0$$

$$x^2 - 2xy + y^2 \geq 0$$

$$(x-y)^2 \geq 0$$

Since $(x-y)$ is squared, so no matter the value of x and y is, $(x-y)^2$ will be ≥ 0. So, the equation holds!

④

$$x_1 \log \frac{x_1}{y_1} + x_2 \log \frac{x_2}{y_2} \geq (x_1 + x_2) \log \frac{x_1 + x_2}{y_1 + y_2}. \quad (2)$$

$\div (x_1+x_2)$

$$\frac{x_1}{x_1+x_2} \log\left(\frac{x_1}{y_1}\right) + \frac{x_2}{x_1+x_2} \log\left(\frac{x_2}{y_2}\right) \geq \log\left(\frac{x_1+x_2}{y_1+y_2}\right)$$

$\log\left(\left(\frac{a}{b}\right)^Y\right) = \log\left(\frac{a}{b}\right) = -\log\left(\frac{b}{a}\right)$

$$-\frac{x_1}{x_1+x_2} \log\left(\frac{y_1}{x_1}\right) - \frac{x_2}{x_1+x_2} \log\left(\frac{y_2}{x_2}\right) \geq -\log\left(\frac{y_1+y_2}{x_1+x_2}\right)$$

$$\frac{x_1}{x_1+x_2} \log\left(\frac{y_1}{x_1}\right) + \frac{x_2}{x_1+x_2} \log\left(\frac{y_2}{x_2}\right) \leq \log\left(\frac{y_1+y_2}{x_1+x_2}\right) \quad \div -1$$

$$\log\left(\frac{y_1+y_2}{x_1+x_2}\right) \geq \frac{x_1}{x_1+x_2} \log\left(\frac{y_1}{x_1}\right) + \frac{x_2}{x_1+x_2} \log\left(\frac{y_2}{x_2}\right) \quad \text{flip}$$

From question ③, we get the equation below:

$$\log\left(\sum_{i=1}^{n} \alpha_i x_i\right) \geq \sum_{i=1}^{n} \alpha_i \log(x_i) \quad \ldots (3)$$

concavity of logarithm

Let $x_i = \frac{y_i}{x_i}$ and $\alpha_i = \frac{x_i}{\sum_{i=1}^{n} x_i}$, $n = 2$

$$\log\left(\sum_{i=1}^{n} \frac{x_i}{\sum_{i=1}^{n} x_i} \times \frac{y_i}{x_i}\right) \geq \sum_{i=1}^{n} \frac{x_i}{\sum_{i=1}^{n} x_i} \log\left(\frac{y_i}{x_i}\right)$$

$$\log\left(\left(\frac{x_1}{x_1+x_2} \times \frac{y_1}{x_1}\right) + \left(\frac{x_2}{x_1+x_2} \times \frac{y_2}{x_2}\right)\right) \geq \frac{x_1}{x_1+x_2} \log\left(\frac{y_1}{x_1}\right) + \frac{x_2}{x_1+x_2} \log\left(\frac{y_2}{x_2}\right)$$

$$\log\left(\frac{y_1}{x_1+x_2} + \frac{y_2}{x_1+x_2}\right) \geq \frac{x_1}{x_1+x_2} \log\left(\frac{y_1}{x_1}\right) + \frac{x_2}{x_1+x_2} \log\left(\frac{y_2}{x_2}\right)$$

$$\log\left(\frac{y_1+y_2}{x_1+x_2}\right) \geq \frac{x_1}{x_1+x_2} \log\left(\frac{y_1}{x_1}\right) + \frac{x_2}{x_1+x_2} \log\left(\frac{y_2}{x_2}\right)$$

Thus, we've just proven log-sum inequality using concavity of logarithm

(5) $$x_1 \log \frac{x_1}{y_1} + x_2 \log \frac{x_2}{y_2} \geq (x_1 + x_2) \log \frac{x_1 + x_2}{y_1 + y_2}. \quad (2)$$

$$\log\left(\frac{y_1+y_2}{x_1+x_2}\right) \geq \frac{x_1}{x_1+x_2}\log\left(\frac{y_1}{x_1}\right) + \frac{x_2}{x_1+x_2}\log\left(\frac{y_2}{x_2}\right)$$

$$IG(S; \underset{\text{Final}}{\text{Shirt Size}}) \geq IG(S; \text{Shirt Size})$$

$$H(S) = \sum_{i=1}^{l} \frac{n_i}{n} \log_2\left(\frac{n}{n_i}\right)$$

$$IG(S;A) = H(S) - \sum_{j=1}^{a} \frac{|S_{vj}|}{|S|} \times H(S_{vj})$$

$$IG(S;A) = \sum_{i=1}^{l} \frac{n_i}{n} \log_2\left(\frac{n}{n_i}\right) - \sum_{j=1}^{a} \frac{|S_{vj}|}{|S|} \times H(S_{vj})$$

$IG(S; \text{Shirt Size Final})$ $\qquad$ $IG(S; \text{Shirt Size})$

$$\underbrace{\sum_{i=1}^{l} \frac{n_i}{n} \log_2\left(\frac{n}{n_i}\right)}_{H(S)} - \sum_{j=1}^{a} \frac{|S_{vj}|}{|S|} \times H(S_{vj}) \geq \underbrace{\sum_{i=1}^{l} \frac{n_i}{n} \log_2\left(\frac{n}{n_i}\right)}_{H(S)} - \sum_{j=1}^{a} \frac{|S_{vj}|}{|S|} \times H(S_{vj})$$

$\Rightarrow$ we also cancel out weighted sum of $H(S_{small})$, $H(S_{Large})$, $H(S_{xLarge})$ from both sides, as they are equal

Let n_m, n_{ml}, n_{ms} be number of medium, medium-large, and medium-small respectively. Terms are also defined with (+) and (-), such that n_m^+ means n_m with label (+).

$$-\frac{n_{ml}}{n}\left(\frac{n_{ml}^+}{n_{ml}} \log \frac{n_{ml}}{n_{ml}^+} + \frac{n_{ml}^-}{n_{ml}} \log \frac{n_{ml}}{n_{ml}^-}\right) - \frac{n_{ms}}{n}\left(\frac{n_{ms}^+}{n_{ms}} \log \frac{n_{ms}}{n_{ms}^+} + \frac{n_{ms}^-}{n_{ms}} \log \frac{n_{ms}}{n_{ms}^-}\right) \geq$$

$$-\frac{n_m}{n}\left(\frac{n_m^+}{n_m} \log \frac{n_m}{n_m^+} + \frac{n_m^-}{n_m} \log \frac{n_m}{n_m^-}\right)$$

$$-\left(\frac{n^+_{mL}}{n}\log\frac{n_{mL}}{n^+_{mL}}+\frac{n^-_{mL}}{n}\log\frac{n_{mL}}{n^-_{mL}}\right)-\left(\frac{n^+_{ms}}{n}\log\frac{n_{ms}}{n^+_{ms}}+\frac{n^-_{ms}}{n}\log\frac{n_{ms}}{n^-_{ms}}\right)\geq-\left(\frac{n^+_m}{n}\log\frac{n_m}{n^+_m}+\frac{n^-_m}{n}\log\frac{n_m}{n^-_m}\right)$$

$\Rightarrow$ multiply by n on both side

$$-n^+_{mL}\log\frac{n_{mL}}{n^+_{mL}}-n^-_{mL}\log\frac{n_{mL}}{n^-_{mL}}-n^+_{ms}\log\frac{n_{ms}}{n^+_{ms}}-n_{ms}\log\frac{n_{ms}}{n^-_{ms}}\geq n^+_m\log\frac{n_m}{n^+_m}-n^-_m\log\frac{n_m}{n^-_m}$$

$$n^+_{mL}\log\frac{n^+_{mL}}{n_{mL}}+n^-_{mL}\log\frac{n^-_{mL}}{n_{mL}}+n^+_{ms}\log\frac{n^+_{ms}}{n_{ms}}+n_{ms}\log\frac{n^-_{ms}}{n_{ms}}\geq n^+_m\log\frac{n^+_m}{n_m}+n^-_m\log\frac{n^-_m}{n_m}$$

If we separate this equation into two based on + and − labels:

$$\underbrace{n^+_{mL}}_{x_1}\log\frac{\overbrace{n^+_{mL}}^{x_1}}{\underbrace{n_{mL}}_{y_1}}+\underbrace{n^+_{ms}}_{x_2}\log\frac{\overbrace{n^+_{ms}}^{x_2}}{\underbrace{n_{ms}}_{y_2}}\geq\underbrace{n^+_m}_{x_1+x_2}\log\frac{\overbrace{n^+_m}^{x_1+x_2}}{\underbrace{n_m}_{y_1+y_2}}$$ $\rightarrow$ similar to (2)

$n^+_{mL}+n^+_{ms}=n^+_m$.

$$\underbrace{n^-_{mL}}_{x_1}\log\frac{\overbrace{n^-_{mL}}^{x_1}}{\underbrace{n_{mL}}_{y_1}}+\underbrace{n^-_{ms}}_{x_2}\log\frac{\overbrace{n^-_{ms}}^{x_2}}{\underbrace{n_{ms}}_{y_2}}\geq\underbrace{n^-_m}_{x_1+x_2}\log\frac{\overbrace{n^-_m}^{x_1+x_2}}{\underbrace{n_m}_{y_1+y_2}}$$ $\rightarrow$ similar to (2)

$n^-_{mL}+n^-_{ms}=n^-_m$

Since we have proven the equation above, so

$$n^+_{mL}\log\frac{n^+_{mL}}{n_{mL}}+n^-_{mL}\log\frac{n^-_{mL}}{n_{mL}}+n^+_{ms}\log\frac{n^+_{ms}}{n_{ms}}+n_{ms}\log\frac{n^-_{ms}}{n_{ms}}\geq n^+_m\log\frac{n^+_m}{n_m}+n^-_m\log\frac{n^-_m}{n_m}$$

equation above is true, and:

$$IG(S;\underset{\text{Final}}{\text{Shirt Size}})\geq IG(S;\text{Shirt Size})$$

Problem 3. (10 points). Consider the training set given in Table 2. There are nine examples, each with three features. Feature 1 and Feature 2 are binary, and Feature 3 is continuous.

1. For Feature 1 and Feature 2, compute the information gain with respect to the examples.

2. To use a feature that takes continuous values, we fix a threshold and categorize the continuous feature depending on whether it is greater than the threshold or not. For example, in the given example, if the threshold is fixed at 4.5, we convert Feature 3 into a categorical feature, $\{S, G, G, S, G, S, G, G, G\}$ where S, G imply that the value is smaller than and greater than the threshold respectively.

   For Feature 3, compute the information gain with respect to the threshold values 2.5, 3.5, 4.5, 5.5, 6.5, and 7.5. Which threshold has the highest information gain?

3. Which feature will be chosen as the root node according to the information gain criterion, feature 1, 2, or 3? If feature 3 is chosen, please specify the threshold.

Feature 1	Feature 2	Feature 3	Label
T	T	1.0	+
T	T	6.0	+
T	F	5.0	−
F	F	4.0	+
F	T	7.0	−
F	T	3.0	−
F	F	8.0	−
T	F	7.0	+
F	T	5.0	−

Table 2: Training Data

4. Construct any decision tree that gives correct answers for all the training examples. You don't need to follow the information gain or Gini criterion. You can use different thresholds for feature 3 in different branches of the tree.

(1)

$$H(S) = \sum_{i=1}^{l} \frac{n_i}{n} \log_2\left(\frac{n}{n_i}\right)$$

$$IG(S;A) = H(S) - \sum_{j=1}^{a} \frac{|S_{vj}|}{|S|} \times H(S_{vj})$$

$$H(S) = \frac{4}{9}\log_2\left(\frac{9}{4}\right) + \frac{5}{9}\log_2\left(\frac{9}{5}\right) = 0.99108$$

Feature 1:

$$H(S_T) = \frac{3}{4}\log_2\left(\frac{4}{3}\right) + \frac{1}{4}\log_2\left(\frac{4}{1}\right) = 0.81128$$

$$H(S_F) = \frac{4}{5}\log_2\left(\frac{5}{4}\right) + \frac{1}{5}\log_2\left(\frac{5}{1}\right) = 0.72193$$

$$IG(S;1) = 0.99108 - \frac{4}{9}(0.81128) - \frac{5}{9}(0.72193) = 0.22944$$

Feature 2:

$$H(S_T) = \frac{2}{5}\log_2\left(\frac{5}{2}\right) + \frac{3}{5}\log_2\left(\frac{5}{3}\right) = 0.97095$$

$$H(S_F) = \frac{2}{4}\log_2\left(\frac{4}{2}\right) + \frac{2}{4}\log_2\left(\frac{4}{2}\right) = 1$$

$$IG(S;2) = 0.99108 - \frac{5}{9}(0.97095) - \frac{4}{9}(1) = 0.00722$$

$$\boxed{IG(S;1) > IG(S;2) \quad 0.22944 > 0.00722}$$

(2) $$H(S) = \frac{4}{9}\log_2\left(\frac{9}{4}\right) + \frac{5}{9}\log_2\left(\frac{9}{5}\right) = 0.99108$$

Feature 1	Feature 2	Feature 3	Label
T	T	1.0	+
T	T	6.0	+
T	F	5.0	−
F	F	4.0	+
F	T	7.0	−
F	T	3.0	−
F	F	8.0	−
T	F	7.0	+
F	T	5.0	−

2.5	3.5	4.5	5.5	6.5	7.5
S	S	S	S	S	S
G	G	G	G	S	S
G	G	G	S	S	S
G	G	S	S	S	S
G	G	G	G	G	S
G	S	S	S	S	S
G	G	G	G	G	G
G	G	G	G	G	S
G	G	G	S	S	S

$$H(S_S) = \frac{1}{1}\log_2\left(\frac{1}{1}\right) + \frac{0}{1}\log_2\left(\frac{1}{0}\right) = 0$$

2.5 $$H(S_G) = \frac{8}{8}\log_2\left(\frac{8}{3}\right) + \frac{5}{8}\log_2\left(\frac{8}{5}\right) = 0.95493$$

$$IG(S;2.5) = 0.99108 - 0 - \frac{8}{9} \times .95493 = 0.14270$$

$$H(S_S) = \frac{1}{2}\log_2\left(\frac{2}{1}\right) + \frac{1}{2}\log_2\left(\frac{2}{1}\right) = 1$$

3.5 $$H(S_G) = \frac{3}{7}\log_2\left(\frac{7}{3}\right) + \frac{4}{7}\log_2\left(\frac{7}{4}\right) = .98523$$

$$IG(S;3.5) = 0.99108 - \frac{2}{9} \times 1 - \frac{7}{9} \times .98523 = .00257$$

$$H(S_5) = \frac{2}{3}\log_2\left(\frac{3}{2}\right) + \frac{1}{3}\log_2\left(\frac{3}{1}\right) \approx .9183$$

4.5 $$H(S_6) = \frac{2}{6}\log_2\left(\frac{6}{2}\right) + \frac{4}{6}\log_2\left(\frac{6}{4}\right) \approx .91830$$

$$IG(S; 4.5) = 0.99108 - \frac{3}{9} \times .91830 - \frac{6}{9} \times .91830 \approx .07278$$

$$H(S_5) = \frac{2}{5}\log_2\left(\frac{5}{2}\right) + \frac{3}{5}\log_2\left(\frac{5}{3}\right) \approx .97095$$

5.5 $$H(S_6) = \frac{2}{4}\log_2\left(\frac{4}{2}\right) + \frac{2}{4}\log_2\left(\frac{4}{2}\right) \approx 1$$

$$IG(S; 5.5) = 0.99108 - \frac{5}{9} \times .97095 - \frac{4}{9} \times 1 \approx .00722$$

$$H(S_5) = \frac{3}{6}\log_2\left(\frac{6}{3}\right) + \frac{3}{6}\log_2\left(\frac{6}{3}\right) \approx 1$$

6.5 $$H(S_6) = \frac{2}{3}\log_2\left(\frac{3}{2}\right) + \frac{1}{3}\log_2\left(\frac{3}{1}\right) \approx .91830$$

$$IG(S; 6.5) = 0.99108 - \frac{6}{9} \times 1 - \frac{3}{9} \times .91830 \approx .01831$$

$$H(S_5) = \frac{4}{8}\log_2\left(\frac{8}{4}\right) + \frac{4}{8}\log_2\left(\frac{8}{4}\right) \approx 1$$

7.5 $$H(S_6) = \frac{1}{1}\log_2\left(\frac{1}{1}\right) + \frac{0}{1}\log_2\left(\frac{1}{0}\right) \approx 0$$

$$IG(S; 7.5) = 0.99108 - \frac{8}{9} \times 1 - 0 \approx .10219$$

Threshold 2.5 has highest IG

③ Feature 1 will be chosen, as it has highest IG compared to Feature 2 and Feature 3

④

Feature 1	Feature 2	Feature 3	Label
T	T	1.0	+
T	T	6.0	+
T	F	5.0	−
F	F	4.0	+
F	T	7.0	−
F	T	3.0	−
F	F	8.0	−
T	F	7.0	+
F	T	5.0	−