Assignment One Solutions
ECE 4200/5420, Fall 2021

- Provide credit to **any sources** other than the course staff that helped you solve the problems. This includes the names of all students you talked to regarding the problems.
- You can look up definitions/basics online (eg wikipedia, stack-exchange, etc)
- Questions marked with an asterisk **is optional for students taking the class as ECE 4200**.
- **The due date is 9/10/2021, 23.59.59 Eastern time**.

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.

Solution:

Note : In the following solution, Gini $(S, \text{feature value})$ is to be interpreted as Gini $(S_{\text{feature value}})$ as defined in class. Let S be the set of all 14 samples.

$$\text{Gini}\,(S) = 1 - \left(\frac{5}{14}\right)^2 - \left(\frac{9}{14}\right)^2 = 0.4592$$

$$\begin{aligned}
\text{Gini}\,(S, \text{Outlook}) &= \frac{5}{14}\text{Gini}\,(S, \text{Sunny}) + \frac{4}{14}\text{Gini}\,(S, \text{Overcast}) + \frac{5}{14}\text{Gini}\,(S, \text{Rainy}) \\
&= \frac{5}{14}\text{Gini}\,(2Y, 3N) + \frac{4}{14}\text{Gini}\,(4Y, 0N) + \frac{5}{14}\text{Gini}\,(3Y, 2N) \\
&= 0.3429 \\
\text{Gini}\,(S, \text{Temperature}) &= \frac{4}{14}\text{Gini}\,(S, \text{Hot}) + \frac{6}{14}\text{Gini}\,(S, \text{Mild}) + \frac{4}{14}\text{Gini}\,(S, \text{Cool}) \\
&= \frac{4}{14}\text{Gini}\,(2Y, 2N) + \frac{6}{14}\text{Gini}\,(4Y, 2N) + \frac{4}{14}\text{Gini}\,(3Y, 1N) \\
&= 0.4405 \\
\text{Gini}\,(S, \text{Humidity}) &= \frac{7}{14}\text{Gini}\,(S, \text{High}) + \frac{7}{14}\text{Gini}\,(S, \text{Normal}) \\
&= \frac{7}{14}\text{Gini}\,(3Y, 4N) + \frac{7}{14}\text{Gini}\,(6Y, 1N) \\
&= 0.3674
\end{aligned}$$

$$\begin{aligned}\text{Gini}\left(S, \text{Wind}\right) &= \frac{7}{14}\text{Gini}\left(S, \text{Weak}\right) + \frac{7}{14}\text{Gini}\left(S, \text{Strong}\right)\\ &= \frac{7}{14}\text{Gini}\left(5Y, 2N\right) + \frac{7}{14}\text{Gini}\left(4Y, 3N\right)\\ &= 0.449\end{aligned}$$

Since $\text{Gini}\left(S, Outlook\right)$ is the lowest, we split the dataset using the Outlook attribute. All samples that have the Outlook attribute set to Overcast are correctly classified as "Yes".
Let S_{OS}, S_{OR} be the samples that have the Outlook attribute set to Sunny and Rainy respectively. We will now use Gini impurity to find the attribute over which we can split S_{OS} and S_{OR}. For S_{OS},

$$\text{Gini}\left(S_{OS}\right) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48$$

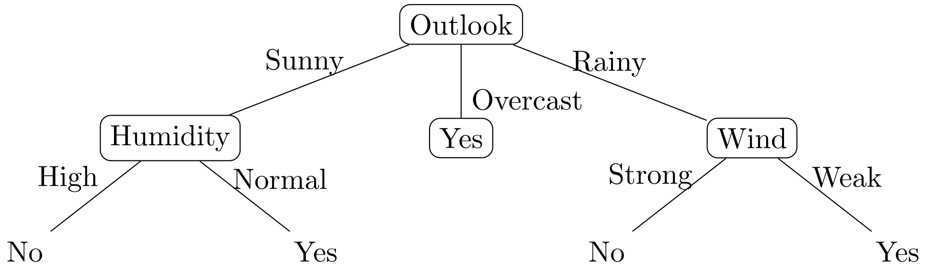
$$\begin{aligned}\text{Gini}\left(S_{OS}, \text{Temperature}\right) &= \frac{2}{5}\text{Gini}\left(S_{OS}, \text{Hot}\right) + \frac{2}{5}\text{Gini}\left(S_{OS}, \text{Mild}\right) + \frac{1}{5}\text{Gini}\left(S_{OS}, \text{Cool}\right)\\ &= \frac{2}{5}\text{Gini}\left(0Y, 2N\right) + \frac{2}{5}\text{Gini}\left(1Y, 1N\right) + \frac{1}{5}\text{Gini}\left(1Y, 0N\right)\\ &= 0.2\\ \text{Gini}\left(S_{OS}, \text{Humidity}\right) &= \frac{3}{5}\text{Gini}\left(S_{OS}, \text{High}\right) + \frac{2}{5}\text{Gini}\left(S_{OS}, \text{Normal}\right)\\ &= \frac{3}{5}\text{Gini}\left(0Y, 3N\right) + \frac{2}{5}\text{Gini}\left(2Y, 0N\right)\\ &= 0\\ \text{Gini}\left(S_{OS}, \text{Wind}\right) &= \frac{3}{5}\text{Gini}\left(S_{OS}, \text{Weak}\right) + \frac{2}{5}\text{Gini}\left(S_{OS}, \text{Strong}\right)\\ &= \frac{3}{5}\text{Gini}\left(1Y, 2N\right) + \frac{2}{5}\text{Gini}\left(1Y, 1N\right)\\ &= 0.4667\end{aligned}$$

Since $\text{GiniImp}\left(S_{OS}, \text{Humidity}\right)$ is the lowest, we split S_{OS} using the Humidity attribute. Since all samples in S_{OS} are correctly classified, we stop extending this part of the decision tree.
We now look at the remaining samples S_{OR},

$$\text{Gini}\left(S_{OR}\right) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48$$

$$\begin{aligned}
\text{Gini}\left(S_{OR}, \text{Temperature}\right) &= \frac{0}{5}\text{Gini}\left(S_{OS}, \text{Hot}\right) + \frac{3}{5}\text{Gini}\left(S_{OR}, \text{Mild}\right) + \frac{2}{5}\text{Gini}\left(S_{OR}, \text{Cool}\right) \\
&= \frac{3}{5}\text{Gini}\left(2Y, 1N\right) + \frac{2}{5}\text{Gini}\left(1Y, 1N\right) \\
&= 0.4667 \\
\text{Gini}\left(S_{OR}, \text{Humidity}\right) &= \frac{2}{5}\text{Gini}\left(S_{OR}, \text{High}\right) + \frac{3}{5}\text{Gini}\left(S_{OR}, \text{Normal}\right) \\
&= \frac{2}{5}\text{Gini}\left(1Y, 1N\right) + \frac{3}{5}\text{Gini}\left(2Y, 1N\right) \\
&= 0.4667 \\
\text{Gini}\left(S_{OR}, \text{Wind}\right) &= \frac{3}{5}\text{Gini}\left(S_{OR}, \text{Weak}\right) + \frac{2}{5}\text{Gini}\left(S_{OR}, \text{Strong}\right) \\
&= \frac{3}{5}\text{Gini}\left(3Y, 0N\right) + \frac{2}{5}\text{Gini}\left(0Y, 2N\right) \\
&= 0
\end{aligned}$$

Since GiniImp $\left(S_{OR}, \text{Wind}\right)$ is the lowest, we split S_{OR} using the Wind attribute. Since all samples in S_{OR} are correctly classified, we stop extending this part of the decision tree. Since all samples have been correctly classified, we end the algorithm. The final decision tree is the same as what we obtained when we used information gain as the impurity measure at nodes.



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.

Gender	Car Type	Shirt Size	Shirt Size Fine	Label
M	Family	Small	Small	+
M	Sports	Medium	Small-Medium	+
M	Family	Large	Large	−
F	Sports	Small	Small	+
M	Sports	Extra Large	Extra Large	+
F	Luxury	Small	Small	−
M	Family	Medium	Large-Medium	−
M	Sports	Extra Large	Extra Large	+
M	Sports	Large	Large	+
F	Luxury	Medium	Large-Medium	−
F	Sports	Medium	Large-Medium	+
F	Family	Small	Small	+
F	Luxury	Large	Large	−
M	Luxury	Medium	Small Medium	−
F	Family	Medium	Small-Medium	+

Table 1: Training Data

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.

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!

Solution:

1. There are 15 data points in the training set. 9 data points have label ”+” and 6 data points have label ”-”. Therefore,

$$\begin{aligned} H\left(\text{Labels}\right) &= \frac{9}{15} \log \frac{15}{9} + \frac{6}{15} \log \frac{15}{6} \\ &= 0.9709 \end{aligned}$$

2.

$$\begin{aligned}
IG\left(\text{Label,Shirt Size}\right) &= H\left(\text{Labels}\right) - \frac{4}{15}H\left(\text{Label}_{\text{S}}\right) - \frac{6}{15}H\left(\text{Label}_{\text{M}}\right) - \frac{3}{15}H\left(\text{Label}_{\text{L}}\right) - \frac{2}{15}H\left(\text{Label}_{\text{XL}}\right)\\
&= 0.9709 - \frac{4}{15}H\left(3+,1-\right) - \frac{6}{15}H\left(3+,3-\right) - \frac{3}{15}H\left(1+,2-\right) - \frac{2}{15}H\left(2+,0-\right)\\
&= 0.1709
\end{aligned}$$

$$\begin{aligned}
IG\left(\text{Label,Shirt Size Fine}\right) &= H\left(\text{Labels}\right) - \frac{4}{15}H\left(\text{Label}_{\text{S}}\right) - \frac{3}{15}H\left(\text{Label}_{\text{MS}}\right) - \frac{3}{15}H\left(\text{Label}_{\text{ML}}\right)\\
&\quad - \frac{3}{15}H\left(\text{Label}_{\text{L}}\right) - \frac{2}{15}H\left(\text{Label}_{\text{XL}}\right)\\
&= 0.9709 - \frac{4}{15}H\left(3+,1-\right) - \frac{3}{15}H\left(2+,1-\right) - \frac{3}{15}H\left(1+,2-\right)\\
&\quad - \frac{3}{15}H\left(1+,2-\right) - \frac{2}{15}H\left(2+,0-\right)\\
&= 0.2036
\end{aligned}$$

The information gain for Shirt Size Fine is higher than the information gain for Shirt Size. Intuitively, you would expect this since the refinement gives you more information about the attribute.

3. Taking exponent on both sides,

$$\begin{aligned}
&\log\left(\frac{x+y}{2}\right) \geq \frac{1}{2}\log(x) + \frac{1}{2}\log(y)\\
\iff &\frac{x+y}{2} \geq \sqrt{xy}\\
\iff &(\sqrt{x} - \sqrt{y})^2 \geq 0,
\end{aligned}$$

which always holds for $x, y > 0$.

4. Since $\log\left(x\right)$ is a concave function, for any $0 \leq \lambda \leq 1$, and for any $a, b > 0$,

$$\lambda \log\left(a\right) + \left(1-\lambda\right)\log b \leq \log\left(\lambda a + \left(1-\lambda\right)b\right)$$

Substitute $\lambda = \frac{x_1}{x_1+x_2}, a = \frac{y_1}{x_1}, b = \frac{y_2}{x_2}$ in this equation to get

$$\begin{aligned}
\frac{x_1}{x_1+x_2}\log\frac{y_1}{x_1} + \frac{x_2}{x_1+x_2}\log\frac{y_2}{x_2} &\leq \log\left(\frac{x_1}{x_1+x_2} * \frac{y_1}{x_1} + \frac{x_2}{x_1+x_2} * \frac{y_2}{x_2}\right)\\
&= \log\frac{y_1+y_2}{x_1+x_2}\\
&= -\log\frac{x_1+x_2}{y_1+y_2}\\
-\left(\frac{x_1}{x_1+x_2}\log\frac{y_1}{x_1} + \frac{x_2}{x_1+x_2}\log\frac{y_2}{x_2}\right) &\geq \log\frac{x_1+x_2}{y_1+y_2}\\
\frac{x_1}{x_1+x_2}\log\frac{x_1}{y_1} + \frac{x_2}{x_1+x_2}\log\frac{x_2}{y_2} &\geq \log\frac{x_1+x_2}{y_1+y_2}\\
x_1\log\frac{x_1}{y_1} + x_2\log\frac{x_2}{y_2} &\geq (x_1+x_2)\log\frac{x_1+x_2}{y_1+y_2}
\end{aligned}$$

5. Let's fix some notation first.

   - S_m : number of data points such that Shirt Size = Medium
   - S_m^+ : number of data points with +ve labels such that Shirt Size = Medium
   - S_m^- : number of data points with -ve labels such that Shirt Size = Medium
   - S_{ms} : number of data points such that Shirt Size = Small Medium
   - S_{ms}^+ : number of data points with +ve labels such that Shirt Size = Small Medium
   - S_{ms}^- : number of data points with -ve labels such that Shirt Size = Small Medium
   - S_{ml} : number of data points such that Shirt Size = Large Medium
   - S_{ml}^+ : number of data points with +ve labels such that Shirt Size = Large Medium
   - S_{ml}^- : number of data points with -ve labels such that Shirt Size = Large Medium
   - S : Set of data points
   - A : $\{\text{Small}, \text{Large}, \text{Extra Large}\}$
   - S_v : Set of data points such that Shirt Size = Shirt Size Fine = v, for all v $\in A$

   Clearly,

$$|S_m| = |S_m^+| + |S_m^-|, \qquad |S_{ms}| = |S_{ms}^+| + |S_{ms}^-|, \qquad |S_{ml}| = |S_{ml}^+| + |S_{ml}^-| \tag{3}$$

$$|S_m^+| = |S_{ms}^+| + |S_{ml}^+|, \qquad |S_m^-| = |S_{ms}^-| + |S_{ml}^-| \tag{4}$$

   In order to use results from part 3, we assume that each of the above quantities are positive.

$$IG\left(\text{Labels,Shirt Size}\right) = H\left(S\right) - \left(\sum_{v \in A} \frac{|S_v|}{|S|} H\left(S_v\right) + \frac{|S_m|}{|S|} H\left(S_m\right)\right)$$
$$IG\left(\text{Labels,Shirt Size Fine}\right) = H\left(S\right) - \left(\sum_{v \in A} \frac{|S_v|}{|S|} H\left(S_v\right) + \frac{|S_{ms}|}{|S|} H\left(S_{ms}\right) + \frac{|S_{ml}|}{|S|} H\left(S_{ml}\right)\right)$$

   Expanding $H\left(S_m\right), H\left(S_{ms}\right), H\left(S_{ml}\right)$

$$\frac{|S_m|}{|S|} H\left(S_m\right) = -\frac{|S_m|}{|S|} \left(\frac{|S_m^+|}{|S_m|} \log \frac{|S_m^+|}{|S_m|} + \frac{|S_m^-|}{|S_m|} \log \frac{|S_m^-|}{|S_m|}\right)$$
$$\frac{|S_{ms}|}{|S|} H\left(S_{ms}\right) + \frac{|S_{ml}|}{|S|} H\left(S_{ml}\right) = -\frac{|S_{ms}|}{|S|} \left(\frac{|S_{ms}^+|}{|S_{ms}|} \log \frac{|S_{ms}^+|}{|S_{ms}|} + \frac{|S_{ms}^-|}{|S_{ms}|} \log \frac{|S_{ms}^-|}{|S_{ms}|}\right)$$
$$- \frac{|S_{ml}|}{|S|} \left(\frac{|S_{ml}^+|}{|S_{ml}|} \log \frac{|S_{ml}^+|}{|S_{ml}|} + \frac{|S_{ml}^-|}{|S_{ml}|} \log \frac{|S_{ml}^-|}{|S_{ml}|}\right)$$

   Regrouping terms on the right hand side and simplifying, we get

$$\frac{|S_{ms}|}{|S|} H\left(S_{ms}\right) + \frac{|S_{ml}|}{|S|} H\left(S_{ml}\right) = -\frac{1}{|S|} \left(|S_{ms}^+| \log \frac{|S_{ms}^+|}{|S_{ms}|} + |S_{ml}^+| \log \frac{|S_{ml}^+|}{|S_{ml}|} + |S_{ms}^-| \log \frac{|S_{ms}^-|}{|S_{ms}|} + |S_{ml}^-| \log \frac{|S_{ml}^-|}{|S_{ml}|}\right)$$

Fearure 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

Applying the result in part 3 to the first two terms and the last two terms on the right hand side, and applying (3) and (4) we get

$$\begin{aligned}\frac{|S_{ms}|}{|S|}H\left(S_{ms}\right)+\frac{|S_{ml}|}{|S|}H\left(S_{ml}\right) &\leq -\frac{1}{|S|}\left(|S_m^+|\log\frac{|S_m^+|}{|S_m|}+|S_m^-|\log\frac{|S_m^-|}{|S_m|}\right)\\ &= -\frac{|S_m|}{|S|}\left(\frac{|S_m^+|}{|S_m|}\log\frac{|S_m^+|}{|S_m|}+\frac{|S_m^-|}{|S_m|}\log\frac{|S_m^-|}{|S_m|}\right)\\ &= \frac{|S_m|}{|S|}H\left(S_m\right)\end{aligned}$$

$$\text{Therefore, } IG\left(\text{Labels,Shirt Size}\right)\leq IG\left(\text{Labels,Shirt Size Fine}\right)$$

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. 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, if we use the threshold value with the highest information gain for the third feature?
4. Construct any decision tree that gives correct answers for all the training examples.

Solution:

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

$$\begin{aligned}IG(Label,Feature\,1) &= Entropy(Label)-\frac{4}{9}Entropy(Label_{Feature\,1=T})-\frac{5}{9}Entropy(Label_{Feature\,1=F})\\ &= 0.9911-\frac{4}{9}Entropy(3+,1-)-\frac{5}{9}Entropy(1+,4-)\\ &= 0.229\end{aligned}$$

Similarly,

$$\begin{aligned}
IG(Label, Feature\,2) &= Entropy(Label) - \frac{4}{9}Entropy(Label_{Feature\,2=T}) - \frac{5}{9}Entropy(Label_{Feature\,2=F}) \\
&= 0.9911 - \frac{5}{9}Entropy(2+, 3-) - \frac{4}{9}Entropy(2+, 2-) \\
&= 0.007
\end{aligned}$$

2. 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?

$$\begin{aligned}
&IG(Label, Feature3 \leq 2.5) \\
&\quad = Entropy(Label) - \frac{1}{9}Entropy(Label_{Feature3 \leq 2.5}) - \frac{8}{9}Entropy(Label_{Feature3 > 2.5}) \\
&\quad = 0.9911 - \frac{1}{9}Entropy(1+, 0-) - \frac{8}{9}Entropy(3+, 5-) \\
&\quad = 0.143
\end{aligned}$$

Similarly, the information gains for the other threshold values can be computed to get

$$\begin{aligned}
IG(Label, Feature3 \leq 3.5) = 0.003 \\
IG(Label, Feature3 \leq 4.5) = 0.073 \\
IG(Label, Feature3 \leq 5.5) = 0.007 \\
IG(Label, Feature3 \leq 6.5) = 0.018 \\
IG(Label, Feature3 \leq 7.5) = 0.102
\end{aligned}$$

3. Which feature will be chosen as the root node, if we use the threshold value with the highest information gain for the third feature?

   The answer would be $Feature1$.

4. Construct any decision tree that gives correct answers for all the training examples.

   There are many solutions. Any decision tree that gives correct answer is acceptable.