Assignment One Solutions
ECE 4950

- Problems 1, 2, and 3 will be self-graded by students. Teaching staff will grade problem 4.
- For each sub-part, assign full points if answer is correct (or you think is approximately correct), assign zero otherwise.
- Write the total for first 3 Problems, your name, and net-id on top of the solutions (that you scanned and uploaded). Hand in the graded solutions on or before **Wednesday, 15th March**. This can be done after the class or during office hours on any day before 15th March.

Problem 1. (5 points). In class, we computed information gain for Humidity, and Wind (for the tennis example). Complete the calculations to obtain the information gain values with respect to Outlook and Temperature. (The values to expect are on page 60 of Mitchell's book).

Solution (2.5 Points each, Total 5 Points):

$$\begin{aligned}
Gain(S, Outlook) &= Entropy(S) - \sum_{\nu \in \{Sunny, Overcast, Rain\}} Entropy(S_\nu) \\
&= 0.940 - \frac{5}{14} Entropy(S_{Sunny}) - \frac{4}{14} Entropy(S_{Overcast}) - \frac{5}{14} Entropy(S_{Rain}) \\
&= 0.940 - \frac{5}{14} Entropy(2+, 3-) - \frac{4}{14} Entropy(4+, 0-) - \frac{5}{14} Entropy(3+, 2-) \\
&= 0.940 - \frac{5}{14} \times 0.971 - \frac{4}{14} \times 0 - \frac{5}{14} \times 0.971 \\
&= 0.246
\end{aligned}$$

Similarly,

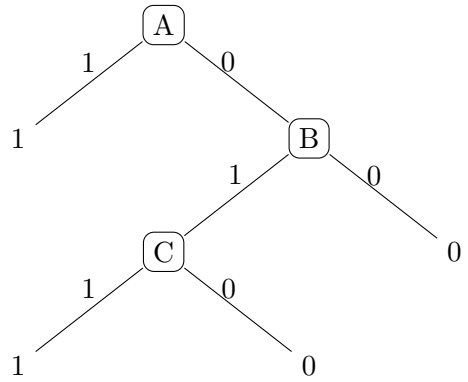
$$\begin{aligned}
Gain(S, Temperature) &= Entropy(S) - \sum_{\nu \in \{Hot, Mild, Cool\}} Entropy(S_\nu) \\
&= 0.940 - \frac{4}{14} Entropy(S_{Hot}) - \frac{6}{14} Entropy(S_{Mild}) - \frac{4}{14} Entropy(S_{Cool}) \\
&= 0.940 - \frac{4}{14} Entropy(2+, 2-) - \frac{6}{14} Entropy(4+, 2-) - \frac{4}{14} Entropy(3+, 1-) \\
&= 0.940 - \frac{4}{14} \times 1 - \frac{6}{14} \times 0.918 - \frac{4}{14} \times 0.811 \\
&= 0.029
\end{aligned}$$

Problem 2. (15 points). In general, decision trees represent a disjunction ("OR") of conjunctions ("AND") of constraints on the attribute values of instances. Each path from the tree root to a leaf corresponds to a conjunction of attribute tests, and the tree itself to a disjunction of these conjunctions. For example, the decision tree for the "Play Tennis" example in the class corresponds to the following Boolean function:

Play Tennis=(Outlook = Sunny $\bigwedge$ Humidity = Normal)
$\bigvee$ (Outlook = Overcast)
$\bigvee$ (Outlook = Rain $\bigwedge$ Wind = Weak)

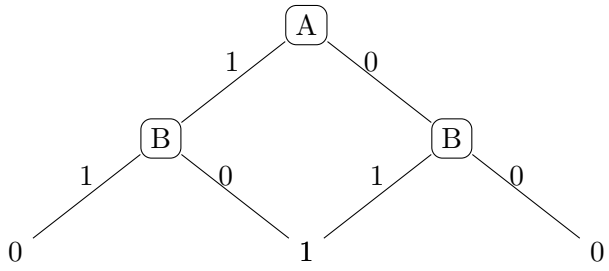
Suppose A, B, C, and D are four boolean variables. Draw decision trees that represent the following boolean functions:

1. A $\bigvee$ (B $\bigwedge$ C)
   Solution: (5 Points)
   By symmetry, you can exchange the B's and C's and get another correct tree.

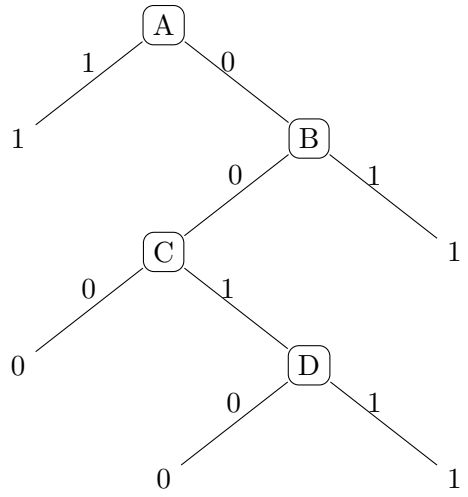


2. A `XOR` B

   Solution: (5 Points)
   One possible decision tree for this problem is given below. You can exchange the A's and B's by symmetry and get another correct tree.



3. (A $\bigvee$ B) $\bigvee$ (C $\bigwedge$ D)
   Solution: (5 Points)
   By symmetry, you can exchange the A's and B's, and C's and D's and get another correct tree.



Problem 3. (30 points). Consider the training set given in Table 1. There are nine examples, each with three features. Feature 1 and Feature 2 are binary, and Feature 3 is continuous.

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 1: Training Data

1. What is the entropy of the labels?
   Solution: (1 Point)
   $Entropy(Label) = Entropy(4+, 5-) = 0.9911$

2. For Feature 1 and Feature 2, compute the information gain with respect to the examples.
   Solution: (2.5 Points each, Total 5 Points)

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

Similarly,

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

3. 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?
Solution: (2 Points each, Total 12 Points)

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

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

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

4. Which feature will be chosen first?
Solution: (2 Points)
Feature 1.

5. Compute the gini impurity measure for Feature 1 and Feature 2.
Solution: (2.5 Points each, total 5 Points)

$$Gini(Label) = Gini(4+, 5-) = 1 - \left(\frac{4}{9}\right)^2 - \left(\frac{5}{9}\right)^2 = 0.494$$

$$\begin{aligned} GiniGain(Label, Feature\,1) &= Gini(Label) - \frac{4}{9}Gini(Label_{Feature\,1=T}) - \frac{5}{9}Gini(Label_{Feature\,1=F}) \\ &= 0.494 - \frac{4}{9}Gini(3+, 1-) - \frac{5}{9}Gini(1+, 4-) \\ &= 0.149 \end{aligned}$$

Similarly,

$$\begin{aligned} GiniGain(Label, Feature\,2) &= Gini(Label) - \frac{4}{9}Gini(Label_{Feature\,2=T}) - \frac{5}{9}Gini(Label_{Feature\,2=F}) \\ &= 0.494 - \frac{5}{9}Gini(2+, 3-) - \frac{4}{9}Gini(2+, 2-) \\ &= 0.005 \end{aligned}$$

6. Construct any decision tree that gives correct answers for all the training examples.
   Solution: (5 Points)
   There are many solutions. Any decision tree that gives correct answer is acceptable.