Assignment One
ECE 4200, Fall 2020

- 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)
- The questions marked with one asterisk can be slightly involved.
- The questions marked with two asterisks are OPTIONAL.
- **The due date is 9/20/2020, 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.

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 $0 \le \lambda \le 1$,

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

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,

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

Prove this using the concavity of logarithms.

Gender	**Car Type**	**Shirt Size**	**Shirt Size Fine**	**Label**
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

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!

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.

Problem 4. Decision Trees (30 points). See attached jupyter notebook for details.