EE16ML – Fall 2020 – Quiz

Team GAJEM, YOUR SID HERE

1. Suppose you have a biased 6-sided die, where there is a $\frac{1}{4}$ chance to roll a 1 and all other numbers have an equal chance to be rolled. What is the entropy and the Gini impurity of this die? No need to simplify the logs.

2. Given enough splits, can you classify any dataset using a decision tree with 100% accuracy?

   (a) Yes, because decision trees are deterministic

   (b) Yes, because given enough splits we can always uniquely identify a data point

   (c) No, because decision trees are probabilistic

   (d) No, because points with the identical features may belong to different classes

3. Calculate the information gain for each choice data split (leave your answer in terms of natural logs). Which choice of threshold produces a greater information gain?

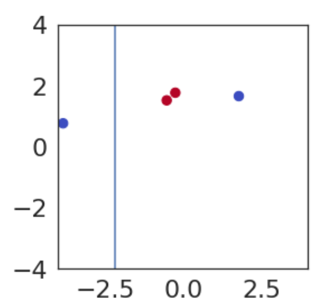

Figure 1

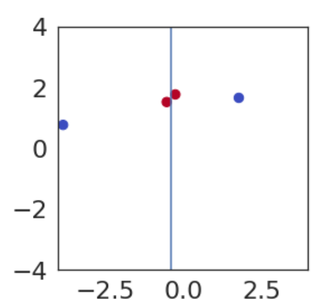

Figure 2

4. Let X be a random variable with discrete outcomes $\{x_1, x_2, ..., x_k\}$. We denote its probability mass function as $p(X)$. That is, for a specific outcome x_j, the probability that $X = x_j$ is $p(X = x_j)$. Recall entropy is defined as

$$H(X) = -\sum_{j=1}^{k} p(X = x_j) \ln p(X = x_j)$$

   (a) Show that $H(X) = \mathbb{E}[-\ln p(X)]$. Use the fact that $\mathbb{E}[g(X)] = \sum_{j=1}^{k} p(X = x_j)g(x_j)$ for discrete outcomes.

   (b) Given that $g(x) = \ln(x)$ is a concave function, and the Jensen inequality which states for a concave function $g(X)$,

$$\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$$

   find an upper bound for $H(X)$. Simplify as much as possible.

   (c) For what distribution of X is $H(X)$ equal to its upper bound?

5. Select which type of tree is MOST LIKELY to overfit:

   (a) Small tree

   (b) Large tree

   (c) Both are equally likely

6. Fill in the missing pseudo-code for the base cases in the pseudo code for the DECISION-TREE-LEARNING function (from Professor J. Listgarten's Nov. 16, 2020 lecture):

```
"""data_set is a nxk matrix for the n data samples at the current node, and outcomes
is a list of known outcomes for each data sample. Assume that unique(list) is a
function that returns the number of unique
objects in a list. Let majority_rule(list) be
a function that returns the object in a list
with the greatest occurance. """

function DECISION-TREE-LEARNING(data_set, outcomes)
        #create a new tree
        tree = new node()
        #base case 1
        if unique(outcomes) == 1
                tree.set_label(_______)
                return tree
        #base case 2
        else if unique(get_features_list(data_set)) == 1
                tree.set_label(_______)
                return tree
        else
                #select feature that maximizes information gaing
                best_feature = argmax(information_gain)
                for value v in best_feature:
                        indices = [index where feature_value(data, best_feature) == v]
                        subDataSet = data_set[indices]
                        subOutcomes = outcomes[indices]
                        subtree = DECISION-TREE-LEARNING(subDataPoints, subOutcomes)
                        tree.add_child(subtree)
                return tree
```

7. Which classification boundary(s) could NOT be from a decision tree? Ignore the dashed yellow line in option (b). Image credits to Professor J. Listgarten, from her Nov. 16, 2020 and Nov. 18, 2020 lectures.

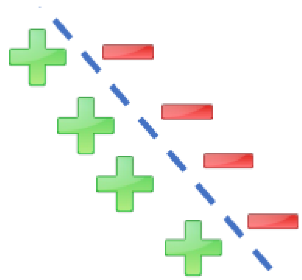

Figure 3

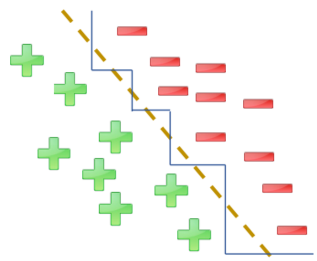

Figure 4

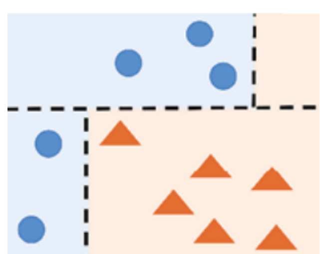

Figure 5

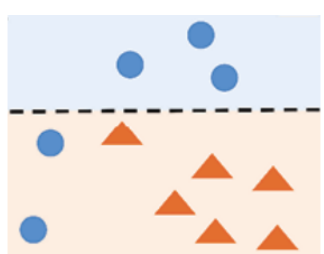

Figure 6

8. Choose all FALSE statements about information gain:

   (a) Knowing more information cannot decrease your current knowledge of a random variable.

   (b) Adversaries can cause negative information gain because they can use information against you.

   (c) The information gain between two random variables is zero if and only if the two variables are independent.

   (d) In the recursive DT algorithm, splitting on the feature with the largest information gain is equivalent to splitting on the feature with the lowest entropy.

9. Using the truth table below, construct a decision tree using the minimum number of a layers. Hint: given A is true does the label depend on C? Given A is false, does the label depend on B?

A	B	C	Label
T	T	T	T
F	T	T	T
T	T	F	T
F	T	F	F
T	F	T	F
F	F	T	T
T	F	F	F
F	F	F	F

Figure 7

10. The decision tree learning algorithm (described in question 6) is:

    (a) Optimal only

    (b) Complete only

    (c) Both optimal and complete

    (d) Neither optimal nor complete