EE16ML – Fall 2020 – Quiz Solutions

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 11 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.

   Solution. The probability that the die is 1 is 0.25, so the probability that it is either 2,3,4,5 or 6 is $\frac{0.75}{5} = 0.15$. Plugging into the formula for entropy, we have:

   $$\begin{aligned} &5[-0.15\log(0.15)] - 0.25\log(0.25) \\ &\quad = 0.75\log\frac{1}{.15} + .25\log 4 \end{aligned}$$

   Plugging into the formula for Gini impurity, we have

   $$\begin{aligned} &5[(0.15)(1-0.15)] + 0.25(1-0.25) \\ &\quad = 5(0.15)(0.85) + 0.25(0.75) \end{aligned}$$

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

   Solution. (d). Points with the same features but different class labels cannot be differentiated by our DT classifier.

3. Suppose we have the decision tree from lecture, and Alice, Bob, and Carol have the features specified below. Classify whether or not Alice, Bob, and Carol should travel based on this decision tree.

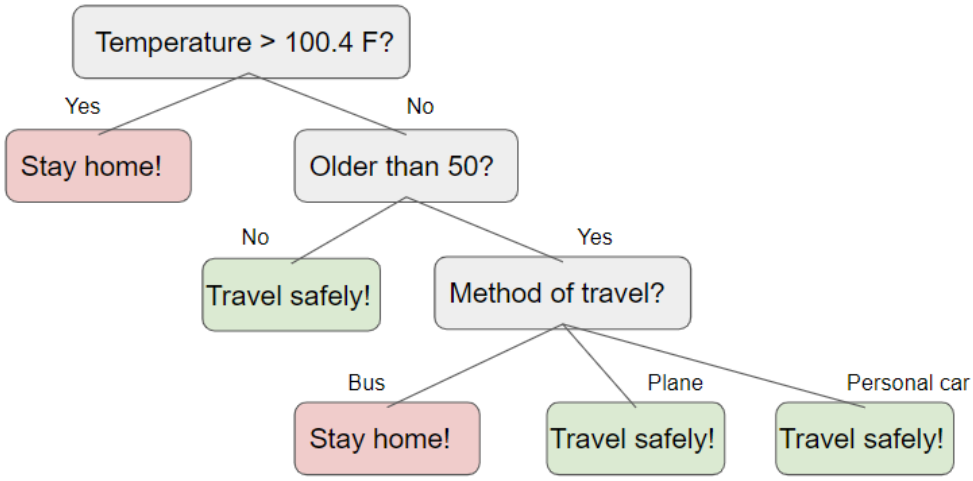


Figure 1

Name	Temperature (F)	Age (years)	Method of travel
Alice	101	27	Bus
Bob	99.7	51	Personal car
Carol	96	48	Plane

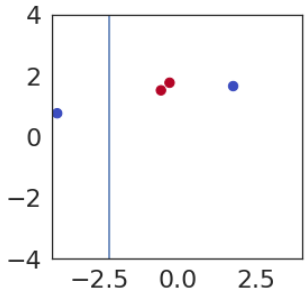
Now classify Alice, Bob and Carol if we modify our decision tree in the following way: now, the first internal node now says "Temperature $<$ 98 F or Temperature $>$ 100.4 F?" and the second internal node now says "Older than 45?", and Alice remeasures her temperature and now is at 100.1 F.

Solution. Before the modification, Alice would get classified to stay home, Bob would get classified to travel safely, and Carol would get classified to travel safely.

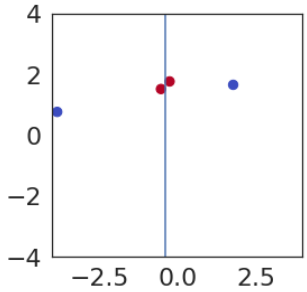
After the modification, Alice would now be classified to travel safely, Bob would still be classified to travel safely, and Carol would now be classified to stay home.

Fun fact: a temperature of below 95 F is considered hypothermia. The average normal temperature is considered to be 98.6 F; a temperature of 96 F could still be within normal fluctuation, but Carol might want to get her temperature re-checked! Low body temperature could be a sign of infection.

4. 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?



(a)



(b)

Solution. Prior to the split, the data set contains two red samples and two blue samples. Then the probability that a point is blue is $\frac{1}{2}$, likewise the probability for red is $\frac{1}{2}$. So $H(X) = \frac{1}{2}\ln(2) + \frac{1}{2}\ln(2) = \ln(2)H(X)$.

For the first figure, given the left half of the split, $H(X|left) = 1\ln 1 = 0$. Given the right half of the split, $H(X|right) = \frac{1}{3}\ln 3 + \frac{2}{3}\ln\frac{3}{2}$. Thus, our information gain $IG = H(X) - [\frac{1}{4}H(X|left) + \frac{3}{4}H(X|right)] = \ln 2 - \frac{1}{4}\ln 3 - \frac{1}{2}\ln\frac{3}{2}$.

For the second figure, given the left half of the split, $H(X|left) = \frac{1}{2}\ln 2 + \frac{1}{2}\ln 2 = \ln 2$. Given the right half of the split, $H(X|right) = \frac{1}{2}\ln 2 + \frac{1}{2} = \ln 2$. Then our information gain $IG = H(X) - \frac{1}{2}[H(X|left) + H(X|right)] = \ln 2 - \ln 2 = 0$.

One way to see that the first threshold has a higher information gain is to determine whether we know more or the same after the split. In the second plot, we can visually see that we have not gained any new information.

5. **Optional.** 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?

Solution.

(a) Consider $g(X) = -\ln p(X)$. Then:

$$\begin{aligned}
\mathbb{E}[g(X)] &= \sum_{j=1}^{k} p(X = x_j)g(x_j) \\
&= -\sum_{j=1}^{k} p(X = x_j) \ln p(X = x_j) \\
&= H(X)
\end{aligned}$$

(b) Let $g(x) = \ln x$, which we know is concave.

$$\begin{aligned}
\mathbb{E}[\ln x] &\leq \ln(\mathbb{E}[x]) \\
\mathbb{E}[\ln \frac{1}{p(X)}] &\leq \ln(\mathbb{E}[\frac{1}{p(X)}]) \\
\mathbb{E}[-\ln p(X)] &\leq \ln(\mathbb{E}[\frac{1}{p(X)}) \\
H(X) &\leq \ln sum_{j=1}^{k} p(X = x_j) \frac{1}{p(X = x_j} \\
H(X) &\leq \ln k
\end{aligned}$$

(c) $H(X) = \ln k$ when X is distributed as discrete uniform distribution, i.e. $p(X =$

$x_i) = \frac{1}{k}$ for all values of X, x_i. To see this,

$$\begin{aligned} H(X) &= -\sum_{j=1}^{k} p(X = x_j) \ln p(X = x_j) \\ &= -\sum_{j=1}^{k} \frac{1}{k} \ln \frac{1}{k} \\ &= -\ln \frac{1}{k} \\ &= \ln k \end{aligned}$$

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

   (a) Small tree

   (b) Large tree

   (c) Both are equally likely

   Solution. (b). As covered in lecture, larger trees are more prone to overfitting because they are able to model more complex functions of features than smaller trees can.

7. 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
```

Solution.

(a) `outcomes[0]` or similar pseudo code

(b) `majority-rule(outcomes)` or similar pseudo code

8. 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

   Solution. (b). The greedy approach outlined is suboptimal, but still holds up pretty well empirically against popular DT algorithms C4.5 and CART (see this paper for more details). As a test-taking tip, note that an algorithm that is optimal must also be complete, so (a) should be crossed off as an option.

9. 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.

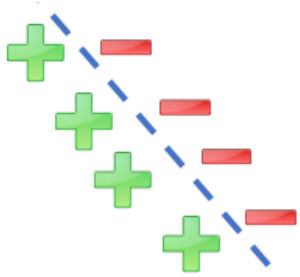


Figure 2

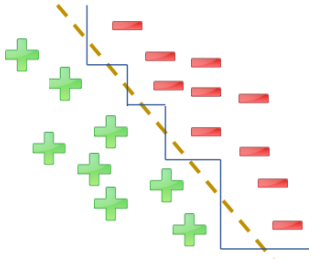


Figure 3

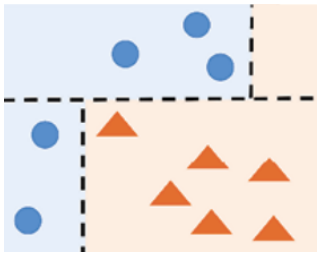


Figure 4

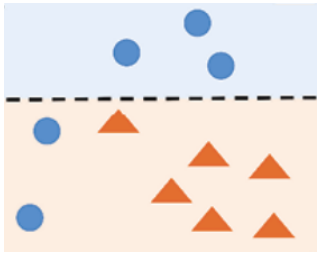


Figure 5

Solution. (a) and (c). For (a), recall that boundaries induced by DTs are axis-aligned, so this diagonal decision boundary could not have been found by a DT. For (c), notice

that there is a vertical split in the upper right corner. Recall that in our recursive DT algorithm, we stop splitting when either all points have identical labels or all points have identical features, i.e. our DT algorithm will never create a split that does not divide any points.

10. 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.

    Solution. (b). The definition of information gain simply measures the difference in entropy before and after knowing some information; it does not measure whether knowing that information is "good" or "bad".

11. 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 6

Solution. The following tree is the smallest. Other trees accepted for partial credit.

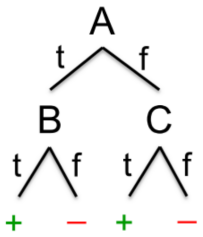


Figure 7