Decision Tree Quiz Answer Key

1.

The probability of 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
$5[-0.15\log(0.15)] - 0.25\log(0.25)$
$= 0.75\log(\frac{5}{0.75}) + 0.25\log(4)$.

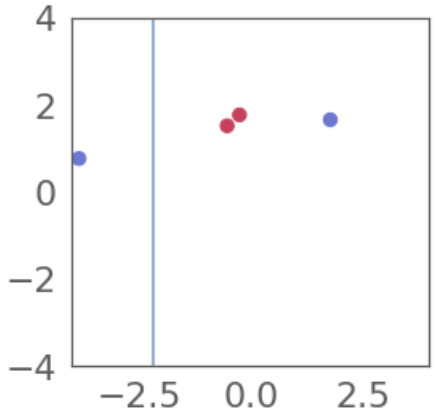
Plugging into the formula for Gini impurity, we have
$5[(0.15)(1-0.15)] + 0.25(1-0.25)$
$= 5(0.15)(0.85) + 0.25(0.75)$

2.

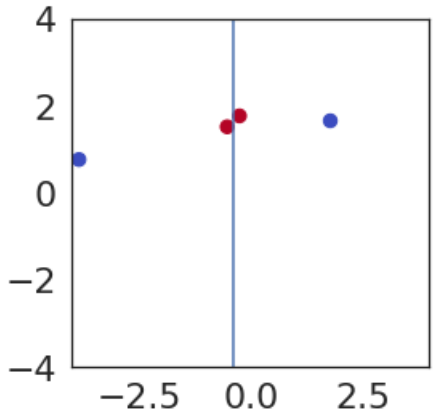
Answer: (d). Points with the same features but different class labels cannot be differentiated by our DT classifier.

3.

Prior to 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 $H(X) = \frac{1}{2}\ln(2) + \frac{1}{2}\ln(2) = \ln(2)$



$H(X|left) = 1\ln(1) = 0$
$H(X|right) = \frac{1}{3}\ln(3) + \frac{2}{3}\ln\frac{3}{2}$
$IG = H(X) - [\frac{1}{4}H(X|left) + \frac{3}{4}H(X|right)]$
$= \ln(2) - \frac{3}{4}(\frac{1}{3}\ln(3) + \frac{2}{3}\ln\frac{3}{2})$



$H(X|left) = \frac{1}{2}\ln(2) + \frac{1}{2}\ln(2) = \ln(2)$
$H(X|right) = \frac{1}{2}\ln(2) + \frac{1}{2}\ln(2) = \ln(2)$
$IG = H(X) - [\frac{1}{2}H(X|left) + \frac{1}{2}H(X|right)]$
$= \ln(2) - [\frac{1}{2}\ln(2 + \frac{1}{2}\ln(2)] = 0$

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

4.

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

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

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

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

3. The entropy of X is equal to $\ln k$ when X is discreetly uniform, that is $p(X) = \frac{1}{k}$ for all values of X. To see this,

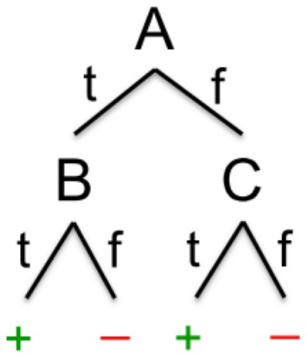
$$\begin{aligned} H(X) &= -\sum_{j=1}^{k} p(X = x_j) \ln p(X = x_j) \\ &= -\sum_{j=1}^{k} \frac{1}{k} \ln \frac{1}{k} \\ &= -\ln \frac{1}{k} \\ &= \ln k \end{aligned}$$

5.

Answer: (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.

6.

The following tree is the smallest



Source: 11/16 lec Prof. J. Listgarten, CS189
Other trees accepted for partial credit

7.

Answer: (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.

8.

Answer: (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”.

9.

1. `outcomes[0]` or similar pseudo code
2. `majority_rule(outcomes)` or similar pseudo code

10.

Answer: complete only. The greedy approach outlined is suboptimal, but still holds up pretty well empirically against popular DT algorithms C4.5 and CART (see https://www.aaai.org/Papers/KDD/1995/KDD95-054.pdf 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.