CSE 151A Winter 2021 - Take Home Exam 3

1 hour **Name:** **PID:**

---

(1) This is an open book, open notes exam. You are free to consult any text book or notes. **You are not allowed to consult with any other person.**

(2) If you need any clarification, please post a private message to the instructors on Piazza.

(3) Remember that your work is graded on the *clarity* of your writing and explanation as well as the validity of what you write.

(4) This is a one-hour exam.

(1) Let X and Y be two random variables with the following joint distribution.

	$X = 1$	$X = 2$	$X = 3$
$Y = 0$	$1/2$	$1/10$	$1/10$
$Y = 1$	$1/10$	$1/5$	0

(a) (2 points) Calculate the marginal distributions of X and Y.

The marginal distribution of X is given by $\Pr(X = x) = \sum_{y=0}^{1} \Pr(X = x, Y = y)$:

$$\Pr(X = 1) = \frac{1}{2} + \frac{1}{10} = \frac{3}{5}$$
$$\Pr(X = 2) = \frac{1}{10} + \frac{1}{5} = \frac{3}{10}$$
$$\Pr(X = 3) = \frac{1}{10} + 0 = \frac{1}{10}$$

The marginal distribution of Y is given by $\Pr(Y = y) = \sum_{x=1}^{3} \Pr(X = x, Y = y)$:

$$\Pr(Y = 0) = \frac{1}{2} + \frac{1}{10} + \frac{1}{10} = \frac{7}{10}$$
$$\Pr(Y = 1) = \frac{1}{10} + \frac{1}{5} + 0 = \frac{3}{10}$$

(b) (4 points) Calculate the conditional distributions of $X|Y=0$ and $X|Y=1$.

Recall that $\Pr(X=x|Y=y) = \frac{\Pr(X=x,Y=y)}{\Pr(Y=y)}$.

So, the distribution of $X|Y=0$ is given by:

$$\Pr(X=1|Y=0) = \frac{\Pr(X=1,Y=0)}{\Pr(Y=0)} = \frac{1/2}{7/10} = \frac{5}{7}$$
$$\Pr(X=2|Y=0) = \frac{\Pr(X=2,Y=0)}{\Pr(Y=0)} = \frac{1/10}{7/10} = \frac{1}{7}$$
$$\Pr(X=3|Y=0) = \frac{\Pr(X=3,Y=0)}{\Pr(Y=0)} = \frac{1/10}{7/10} = \frac{1}{7}$$

The distribution of $X|Y=1$ is given by:

$$\Pr(X=1|Y=1) = \frac{\Pr(X=1,Y=1)}{\Pr(Y=1)} = \frac{1/10}{3/10} = \frac{1}{3}$$
$$\Pr(X=2|Y=1) = \frac{\Pr(X=2,Y=1)}{\Pr(Y=1)} = \frac{1/5}{3/10} = \frac{2}{3}$$
$$\Pr(X=3|Y=1) = \frac{\Pr(X=3,Y=1)}{\Pr(Y=1)} = \frac{0}{3/10} = 0$$

(c) (4 points) Now calculate the conditional entropy $H(X|Y)$. (It is okay to leave the final expression in terms of various logarithms).

Recall that the entropy of a random variable X is given by

$$H(X) = -\sum_x \Pr(X=x)\log\Pr(X=x)$$

Additionally, recall the formula for conditional entropy:

$$H(X|Y) = \sum_y \Pr(Y=y)H(X|Y=y)$$

Substituting in what we computed in the previous parts:

$$\begin{aligned} H(X|Y) &= \Pr(Y=0)H(X|Y=0) + \Pr(Y=1)H(X|Y=1) \\ &= -\frac{7}{10}\left(\frac{5}{7}\log\frac{5}{7} + \frac{1}{7}\log\frac{1}{7} + \frac{1}{7}\log\frac{1}{7}\right) - \frac{3}{10}\left(\frac{1}{3}\log\frac{1}{3} + \frac{2}{3}\log\frac{2}{3}\right) \\ &= -\frac{1}{2}\log\frac{5}{7} - \frac{1}{5}\log\frac{1}{7} - \frac{1}{10}\log\frac{1}{3} - \frac{1}{5}\log\frac{2}{3} \end{aligned}$$

(2) State whether the following statements are true or false. Justify your answer.

(a) Suppose S is a training dataset where each feature vector x_i is unique. Then the ID3 Decision tree constructed on S (without pruning) has zero training error.

This is true. The ID3 training algorithm terminates when each node satisfies one of two conditions: 1) it is pure with respect to training set S or 2) it contains a single feature vector. Since each feature vector x_i is unique, condition 2) implies condition 1) — if a leaf node has only one point, it must be pure.

Therefore, the un-pruned tree's accuracy on S will be 100%, since each feature $x_i \in S$ falls in a pure leaf node where all training set labels are identical.

(b) Pruning an ID3 Decision Tree will never increase its validation error.

This is true. Recall the pruning algorithm discussed in class for tree T containing nodes v_i:

- For each $v_i \in T$:
  - Create tree T' by replacing subtree rooted at v_i by a single node that predicts the majority label in v_i.
  - If validation error of tree T' is $\leq$ validation error of tree T, then set $T = T'$
- Repeat until there is no such node $v_i \in T$

Notice that we only update the tree, $T = T'$, if the validation error reduces or stays the same. As such, we are guaranteed that the validation error of the pruned tree at algorithm completion is $\leq$ that of the original tree.