

GroupD Practice

1. Find the information content of an outcome x of a random variable X such that $P(x) = \frac{1}{4}$.

$$I(x) = \boxed{}$$

ANSWER: $I(x) = -\log_2 \frac{1}{4} = \log_2 4 = 2$

2. For the outcome Z of two fair coin flips with outcomes $(0, 0)$, $(0, 1)$, $(1, 0)$ and $(1, 1)$, whose

probabilities are all $\frac{1}{4}$, the entropy $H(Z) = \boxed{}$

$4 \cdot \frac{1}{4} [-\log_2 \frac{1}{4}] = 2$.

3. Consider a classification decision tree node containing the set of examples $S = \{(\mathbf{x}, y)\}$ where $\mathbf{x} = (x_1, x_2, x_3)$:

S			
x_1	x_2	x_3	y
4	8	15	1
0	7	12	0
1	11	14	0
5	10	13	0
2	9	17	1
3	6	16	1

- (a) The entropy of this node in bits is $\boxed{}$.

ANSWER:

The node's y values are 1, 0, 0, 0, 1, 1, so $f_{ID3}(S) = \frac{1}{|S|} \sum_{(\mathbf{x}, y) \in S} y = \frac{1}{6}(1 + 0 + 0 + 0 + 1 + 1) = \frac{1}{2}$.

$H(S) = 1$

- (b) The (feature, threshold) pair (j, t) that yields the best split for this node is feature $j =$

$$\boxed{}$$

and threshold $t =$

$$\boxed{}$$

ANSWER:

Using feature $j = 3$ and threshold $t = 14.5$ (or any $t \in [14, 15)$) splits S into $S_- = \{(\mathbf{x}, y) \in S | x^{(j)} \leq t\} = \{0, 0, 0\}$ and its complement $S_+ = \{(\mathbf{x}, y) \in S | x^{(j)} > t\} = \{1, 1, 1\}$, each of which has entropy 0.

4. Consider a regression decision tree with `max_depth=1` (that is, the root node is split once into two leaves) made from the set of examples $S = \{(\mathbf{x}, y)\}$ where $\mathbf{x} = x_1$:

S	
x_1	y
0	10
1	11
2	12
3	13
4	14
5	15

What value does this tree predict for $x_1 = 4.5$?

ANSWER: 14

The best split uses feature $j = 1$ and threshold $t = 2.5$, yielding a left subtree containing the first 3 examples and a right subtree containing the last 3. A prediction with $x_1 = 4.5$ uses the right subtree, whose average y value is $\hat{y} = 14$.