ELE 364: Assignment #1 Solutions

1. (10 pts)

   (a) p(C) = p(O) = p(T) = p(S) = 1/25; p(M) = p(H) = p(E) = p(L) = p(R) = p(G) = 2/25; p(A) = p(I) = p(N) = 3/25.
   Hence, $H = -\sum plog_2(p) = 3.5933$ bits.

   (b) New sets are $\{A, A, A, E, E, I, I, I, O\}$ and $\{C, G, G, H, H, L, L, M, M, N, N, N, R, R, S, T\}$. Hence, $H1 = 1.8911$ bits and $H2 = 3.0778$ bits. The overall entropy will be $H_{new} = H1 \times (9/25) + H2 \times (16/25) = 2.6506$ bits. Hence, $IG = 0.9427$ bits.

   (c) More entropy gives the player more options (also, rarer letters with higher points) and is hence preferable.

2. (10 pts) Bootstrap Sample A: Gini index of the dataset: 0.32

Split by feature	Level	Partition Gini index	Remainder	Information gain
EXERCISE	Daily	0		
	Weekly	0	0	0.320
	Rarely	0		
FAMILY	yes	0.444	0.267	0.053
	no	0		

Since EXERCISE has a higher information gain, it is used at the root. The tree is shown below.

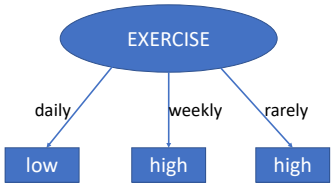

Figure 1: Tree A.

Bootstrap Sample B: Gini index of the dataset: 0.320

Split by feature	Level	Partition Gini index	Remainder	Information gain
SMOKER	False	0		0.320
	True	0	0	
OBESE	False	0.444	0.267	0.053
	True	0		

Since SMOKER has a higher information gain, it is used at the root. The tree is shown below.

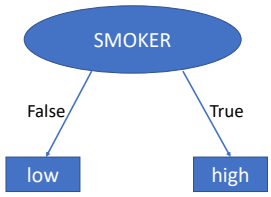

Figure 2: Tree B.

Bootstrap Sample C: Gini index of the dataset: 0.480

Split by feature	Level	Partition Gini index	Remainder	Information gain
FAMILY	Yes	0.5	0.4	0.080
	No	0	0	
OBESE	False	0.444	0.267	0.213
	True	0		

Since OBESE has a higher information gain, it is used at the root. This creates one pure partition with OBESE = true. However, for OBESE = false, the partition is not pure (the first three rows). Unfortunately, using FAMILY, which has yes in all three rows, does not lead to a pure partition either. So we just report the majority value (low) for OBESE = false. The tree is shown below.

For the query, Tree A, B, and C output high, low, and high, respectively. Thus, the majority is high and is the prediction.

3. (10 pts) Initially, the weight of every data instance is $\frac{1}{n}$. Since the total error is ϵ, there are ϵn misclassified instances. Their total weight is $\frac{1}{n} \times \frac{1}{2\epsilon} \times \epsilon \times n = \frac{1}{2}$. Since there are $n - \epsilon n = (1-\epsilon)n$ correctly classified instances, their total weight is $\frac{1}{n} \times \frac{1}{2(1-\epsilon)} \times (1-\epsilon) \times n = \frac{1}{2}$. Hence, the sum of all the weights is 1.0.

4. (10 pts) $\binom{7}{4} \times 0.9^4 \times 0.1^3 + \binom{7}{5} \times 0.9^5 \times 0.1^2 + \binom{7}{6} \times 0.9^6 \times 0.1 + \binom{7}{7} \times 0.9^7 = 0.9973$.

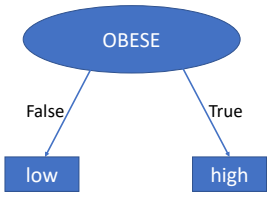

Figure 3: Tree C.