CS 4824 / ECE 4424, Homework 1 (Written Portion), Due: Feb. 14, 2021

Question 1 [10 points]

Consider the dataset shown in Table 1 for a binary classification problem.

Customer ID	Housing Type	Gender	Marital Status	Class
1	Apartment ✓	Male	Married	C0
2	House 0	Male	Single	C1
3	House 0	Female	Married	C1
4	Apartment ✓	Female	Single	C0
5	Apartment ✓	Male	Married	C0
6	Hostel —	Male	Single	C1
7	House 0	Female	Married	C1
8	Apartment ✓	Female	Single	C0
9	Apartment ✓	Male	Married	C0
10	House 0	Male	Single	C1
11	Hostel —	Female	Married	C1
12	Hostel —	Female	Single	C0
13	House 0	Male	Married	C0
14	Hostel —	Male	Single	C1
15	Hostel —	Female	Married	C1
16	Apartment ✓	Female	Single	C0

Table 1

a. **[1 points]** Compute the entropy for the overall data.

$$Entropy(t) = -\sum_j p(j|t)\log_2 p(j|t) = -[(\tfrac{8}{16})\log_2(\tfrac{8}{16}) + (\tfrac{8}{16})\log_2(\tfrac{8}{16})] = 1$$

b. **[2 points]** Compute the entropy for each of the four attributes (consider a multi-way split using each unique value of an attribute).

$$H(Y) = -\sum_{i=1}^{k} P(Y=y_i)\log_2 P(Y=y_i)$$

Customer ID: $-[(\frac{0}{1})\log_2(\frac{0}{1}) + (\frac{1}{1})\log_2(\frac{1}{1})] = 0$

Housing: Apartment: $-[(\frac{6}{6})\log_2(\frac{6}{6}) + (\frac{0}{6})\log_2(\frac{0}{6})] = 0$

House: $-[(\frac{1}{5})\log_2(\frac{1}{5}) + (\frac{4}{5})\log_2(\frac{4}{5})] = 0.722$

Hostel: $-[(\frac{1}{5})\log_2(\frac{1}{5}) + (\frac{4}{5})\log_2(\frac{4}{5})] = 0.722$

Average: $(\frac{6}{16})(0) + (\frac{5}{16})(0.722) + (\frac{5}{16})(0.722) = 0.451$

Gender: Female: $-[(\frac{4}{8})\log_2(\frac{4}{8}) + (\frac{4}{8})\log_2(\frac{4}{8})] = 1$

Male: $-[(\frac{4}{8})\log_2(\frac{4}{8}) + (\frac{4}{8})\log_2(\frac{4}{8})] = 1$

Average: $(\frac{8}{16})(1) + (\frac{8}{16})(1) = 1$

Marital: Married: $-[(\frac{4}{8})\log_2(\frac{4}{8}) + (\frac{4}{8})\log_2(\frac{4}{8})] = 1$

Single: $-[(\frac{4}{8})\log_2(\frac{4}{8}) + (\frac{4}{8})\log_2(\frac{4}{8})] = 1$

Average: $(\frac{8}{16})(1) + (\frac{8}{16})(1) = 1$

c.

$$IG(X) = H(Y) - H(Y|X)$$

Customer ID: 1

Housing: $1-[(\frac{6}{16})(0) + (\frac{5}{16})(0.722) + (\frac{5}{16})(0.722)] = 0.549$

Gender: $1-[(\frac{8}{16})(1) + (\frac{8}{16})(1)] = 0$

Marital: $1-[(\frac{8}{16})(1) + (\frac{8}{16})(1)] = 0$

Highest: Customer ID

Lowest: Marital Status & Housing Type

c. **[3 points]** Compute the Information Gain (IG) obtained by splitting the overall data using each of the four attributes. Which attribute provides the highest IG, and which attribute provides the lowest IG.

d. **[2.5 points]** Compute the Gain Ratio for splitting over each of the four attributes. Which attribute provides the highest Gain Ratio?

e. **[1.5 points]** For splitting at the root node, would you choose the attribute that provides the maximum IG, or the attribute that provides maximum Gain Ratio? Briefly explain your choice.

d.

$$\text{Gain Ratio} = \frac{IG(X)}{-\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}}$$ Spilt INFO

Customer ID: $\frac{1}{-(\frac{1}{16})\log_2(\frac{1}{16}) \times 16} = \frac{1}{4} = 0.25$

Housing: $\frac{0.549}{-[(\frac{6}{16})\log_2(\frac{6}{16}) + 2(\frac{5}{16})\log_2(\frac{5}{16})]} = \frac{0.549}{1.579} = 0.348$

Gender: 0

Marital: 0

Highest: Housing

e. I would choose Housing Type for splitting since it has the maximum Gain Ratio. The customer ID has a large gain but contains too less child in each value. Therefore, Housing will have less biased on data.

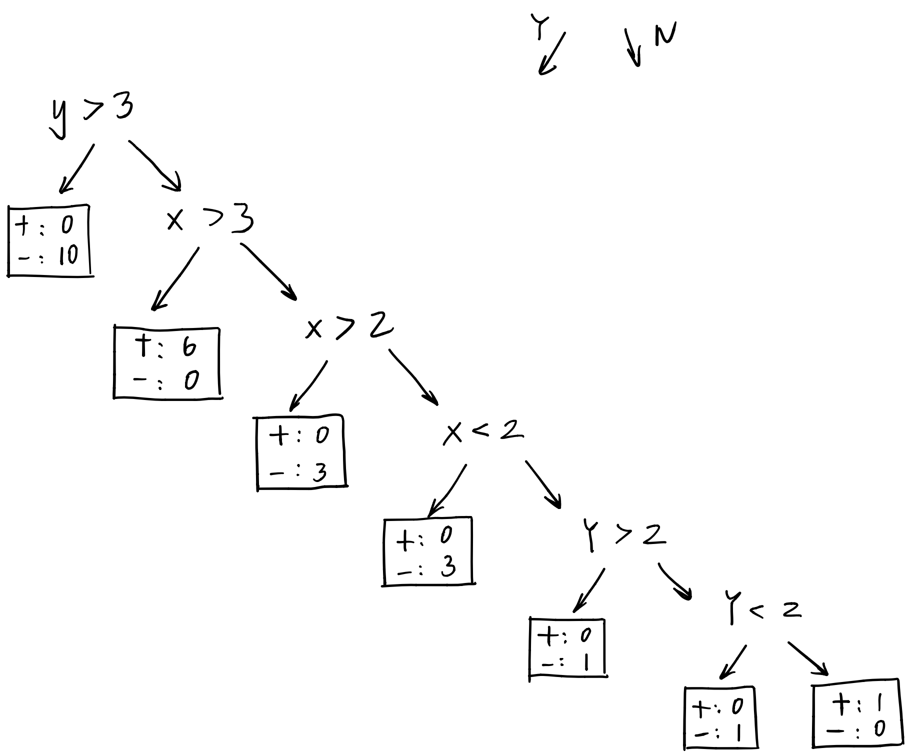
Question 2 [3 points]

Consider the training data given in Table 2 for classification, where the two classes of interest are '-' and '+'. We want to apply binary decision trees as our chosen algorithm for classifying this data.

	X=1	X=2	X=3	X=4	X=5
Y=5	-	-	-	-	-
Y=4	-	-	-	-	-
Y=3	-	-	-	+	+
Y=2	-	+	-	+	+
Y=1	-	-	-	+	+

Table 2

a. **[3 points]** Find a decision tree which uses minimum number of splits (decision boundaries at internal nodes) to perfectly classify each training data instance of Table 2. Hint: The minimum number of splits that you need to create a perfect classifier is 6. You are ***not*** required to compute the Information Gain at each split for constructing the decision tree, but to arrive at your solution by visually inspecting the data.



Question 3 [14 points]

Consider the dataset shown in Table 3.

Instance	A	B	C	Class
1	0	0	1	-
2	1	0	1	+
3	0	1	0	-
4	1	0	0	-
5	1	0	1	+
6	0	0	1	+
7	1	1	0	-
8	0	0	0	-
9	0	1	0	+
10	1	1	1	+

Table 3

a. **[3 points]** Estimate the conditional probabilities for $P(A = 1|+)$, $P(B = 1|+)$, $P(C = 1|+)$, $P(A = 1|-)$, $P(B = 1|-)$, and $P(C = 1|-)$.

b. **[2 points]** Use the conditional probabilities in part (a) to predict the class label for a test sample ($A = 1, B = 1, C = 1$) using the naïve Bayes approach.

c. **[2 points]** Compare $P(A = 1, B = 1|Class = +)$ against $P(A = 1|Class = +)$ and $P(B = 1|Class = +)$. Are the variables conditionally independent given the class?

d. **[3 points]** Let us consider the data instance (A=1, B=1, C=1). Compute the probability of this instance belonging to Class = + using

i. no attributes (i.e. calculate prior probability)

ii. attribute A [$P(Class = +|A=1)$]

iii. attributes A and B [$P(Class = +|A=1, B=1)$]

iv. attributes A, B and C [$P(Class = +|A=1, B=1, C=1)$]

Comment on the change in probability values as we proceed from (i) to (iv).

a. $P(A=1|+) = 3/5 = 0.6$ $\quad$ $P(A=1|-) = 2/5 = 0.4$

$P(B=1|+) = 2/5 = 0.4$ $\quad$ $P(B=1|-) = 2/5 = 0.4$

$P(C=1|+) = 4/5 = 0.8$ $\quad$ $P(C=1|-) = 1/5 = 0.2$

b. Let $P(A=1, B=1, C=1) = K$

$$
\begin{aligned}
&P(+ \mid A=1, B=1, C=1) \\
&= \frac{P(A=1, B=1, C=1 \mid +)\, P(+)}{P(A=1, B=1, C=1)} \\
&= \frac{P(A=1|+) P(B=1|+) P(C=1|+) P(+)}{P(A=1, B=1, C=1)} \\
&= \frac{0.6(0.4)(0.8)(0.5)}{K} \\
&= 0.096/K
\end{aligned}
$$

$$
\begin{aligned}
&P(- \mid A=1, B=1, C=1) \\
&= \frac{P(A=1, B=1, C=1 \mid -)\, P(-)}{P(A=1, B=1, C=1)} \\
&= \frac{P(A=1|-) P(B=1|-) P(C=1|-) P(-)}{P(A=1, B=1, C=1)} \\
&= \frac{(0.4)(0.4)(0.2)(0.5)}{K} \\
&= 0.016/K
\end{aligned}
$$

∴ class label should be "+"

c. $P(A=1 \mid Class=+) = 0.6$

$P(B=1 \mid Class=+) = 0.4$

$P(A=1, B=1 \mid Class=+) = 0.2$

$P(A=1 \mid Class=+) \times P(B=1 \mid Class=+) = (0.6)(0.4) = 0.24 \neq 0.2$

∴ So A and B are not conditionally independent.

d. i. $P(class=+) = \frac{5}{10} = 0.5$

ii. $P(class=+ \mid A=1) = \dfrac{P(A=1|+)P(+)}{P(A=1)} = \dfrac{0.6(0.5)}{(0.5)} = 0.6$

iii. $P(Class=+ \mid A=1, B=1) = \dfrac{P(A=1, B=1|+)P(+)}{P(A=1, B=1)} = \dfrac{P(A=1|+)P(B=1|+)P(+)}{(0.2)}$

$$= \frac{(0.6)(0.4)(0.5)}{(0.2)} = 0.6$$

iv.

$P(Class=+ \mid A=1, B=1, C=1) = \dfrac{P(A=1, B=1, C=1|+)P(+)}{P(A=1, B=1, C=1)} = \dfrac{P(A=1|+)P(B=1|+)P(C=1|+)P(+)}{(0.1)}$

$$= \frac{(0.6)(0.4)(0.8)(0.5)}{(0.1)} = 0.96$$

ii. attribute A [P(Class = +|A=1)]

iii. attributes A and B [P(Class = +|A=1, B=1)]

iv. attributes A, B and C [P(Class = +|A=1, B=1, C=1)]

Comment on the change in probability values as we proceed from (i) to (iv).

Now, consider Table 4

Instance	A	B	C	Class
1	0	0	1	-
2	0	0	1	+
3	0	0	0	-
4	1	0	0	-
5	0	0	1	+
6	0	0	1	+
7	1	0	0	-
8	0	0	0	-
9	0	1	0	+
10	0	1	1	+

Table 4.

e. **[3 points]** Estimate the conditional probabilities for P(A = 1|+), P(B = 1|+), P(C = 1|+), P(A = 1|−), P(B = 1|−), and P(C = 1|−) using Table 4.

f. **[1 point]** Based on Table 4, for a new data instance, **x** = (A = 1, B = 1, C = 1), compute the posterior probabilities, P(+|x) and P(-|x) using the Naïve Bayes approach.

e. $P(A=1|+) = 0/5 = 0$

$P(B=1|+) = 2/5 = 0.4$

$P(C=1|+) = 4/5 = 0.8$

$P(A=1|-) = 2/5 = 0.4$

$P(B=1|-) = 0/5 = 0$

$P(C=1|-) = 1/5 = 0.2$

f.

$$P(+ \mid A=1, B=1, C=1) = \frac{P(A=1, B=1, C=1 \mid +)\, P(+)}{P(A=1, B=1, C=1)} = \frac{P(A=1|+)\, P(B=1|+)\, P(C=1|+)\, P(+)}{P(A=1, B=1, C=1)} = \frac{(0)(0.4)(0.8)(0.5)}{(0)} = \frac{0}{0}$$

$$P(- \mid A=1, B=1, C=1) = \frac{P(A=1, B=1, C=1 \mid -)\, P(-)}{P(A=1, B=1, C=1)} = \frac{P(A=1|-)\, P(B=1|-)\, P(C=1|-)\, P(-)}{P(A=1, B=1, C=1)} = \frac{(0.4)(0)(0.2)(0.5)}{(0)} = \frac{0}{0}$$