Homework-1 Solutions

Question 1

You are given the following training data.

x	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
label	A	A	A	A	B	A	A	A	A	B	B	B	B	A	B	B	B	B

1. What would be the classification of a test sample with $x = 4.2$ according to 1-NN ?

Answer: B

2. What would be the classification of a test sample with $x = 4.2$ according to 2-NN ?

Answer: Either A or B.

3. What would be the classification of a test sample with $x = 4.2$ according to 3-NN ?

Answer: A

4 Use "leave-one-out" cross validation to estimate the error of 1-NN. If you need to choose between two or more examples of identical distance, make your choice so that the number of errors is maximized.

Answer: $\frac{8}{18}$.

5 Use "leave-one-out" cross validation to estimate the error of 2-NN. Whenever you need to make a choice between equal distance data or determining a majority, make your choice so that the number of errors is maximized.

Answer: $\frac{8}{18}$

6 Use "leave-one-out" cross validation to estimate the error of 3-NN. Whenever you need to make a choice between equal distance data or determining a majority, make your choice so that the number of errors is maximized.

Answer: $\frac{4}{18}$.

7 Use "leave-one-out" cross validation to estimate the error of 4-NN. Whenever you need to make a choice between equal distance data or determining a majority, make your choice so that the number of errors is maximized.

Answer: $\frac{4}{18}$.

8 Use "leave-one-out" cross validation to estimate the error of 17-NN. Whenever you need to make a choice between equal distance data or determining a majority, make your choice so that the number of errors is maximized.
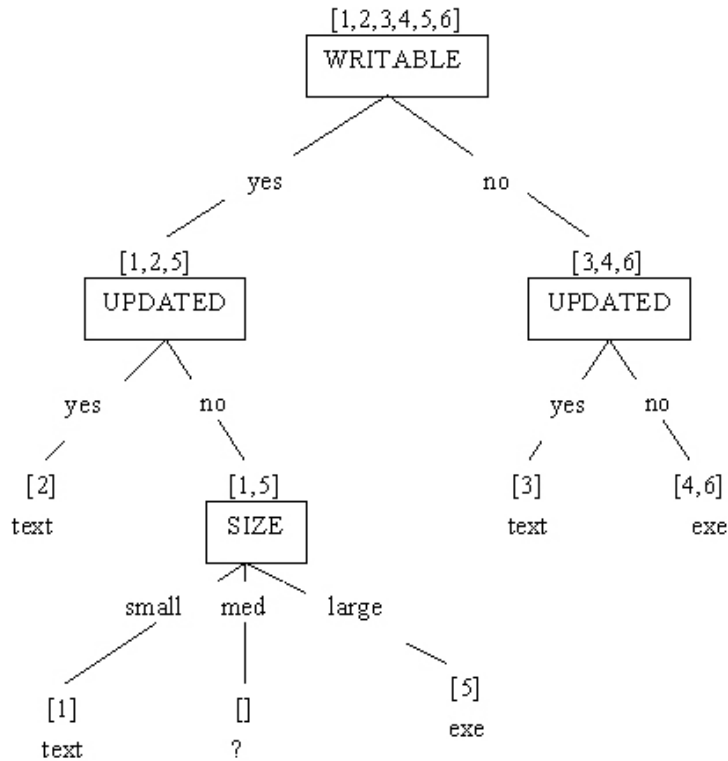
Answer: $\frac{18}{18}$.

Question 2

1. Compute the decision tree generated by first considering WRITABLE, then UPDATED, then SIZE.
2. What tree is generated by using ID3?
Starting with $s = \{1, 2, 3, 4, 5, 6\}$, we have:

$$
\begin{aligned}
Entropy(s) &= 1.0 \\
Entropy(s|WRITABLE) &= 0.9183 & Gain(s|WRITABLE) &= 0.0817 \\
Entropy(s|UPDATED) &= 0.54 & Gain(s|UPDATED) &= 0.46 \\
Entropy(s|SIZE) &= 0.7925 & Gain(s|SIZE) &= 0.2075
\end{aligned}
$$

[1,2,3,4,5,6]
WRITABLE

yes   no

[1,2,5]   [3,4,6]
UPDATED   UPDATED

yes   no   yes   no

[2]   [1,5]   [3]   [4,6]
text   SIZE   text   exe

small   med   large

[1]   []   [5]
text   ?   exe

Therefore, the best choice is UPDATED. This produces the following partition:

$$s_1 = \{2, 3\}, \quad s_2 = \{1, 4, 5, 6\}$$

we have

$$
\begin{aligned}
Entropy(s_1) &= 0 \\
Entropy(s_2) &= 0.811 \\
Entropy(s_2|SIZE) &= 0 & Gain(s_2|SIZE) &= 0.811 \\
Entropy(s_2|WRITABLE) &= 0.5 & Gain(s_2|WRITABLE) &= 0.311
\end{aligned}
$$

Therefore, the best choice is SIZE.
3. How would these trees classify the following two instances:
Tree 1: 1 — text 2 —exe
Tree 2: 1 — text 2 — text

[1,2,3,4,5,6]

UPDATED

yes   no

[2,3]

text

[1,4,5,6]

SIZE

small   med   large

[1]   [4]   [5,6]

text   exe   exe