Kyle Whynott
CPSC 483-01
#885909789

Spring 2024 CPSC 483.01 INTRODUCTION TO MACHINE LEARNING

HW1 (100 Points)

Submission deadline: April 14, 11:59PM, submit it on Canvas.

Any HW content shall NOT be made publicly accessible without the written consent of the instructor.

Part I: *k*-Nearest Neighbors (*k*-NN) (30 points)

You are given the following training data.

x_1	4.2	3.2	2.2	1.2	0.2	0.8	1.8	2.8	3.8	4.9	5.8	6.8	7.8	8.8	9.8	10.8	11.8	12.8
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: A (B)

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

Answer: A (B)

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

Answer: (A) B

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, prioritize your choice that maximizes the number of errors. Iterate!

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

Predicted Label	A	A	A	B	A	B	A	A	B	A	B	B	A	B	A	B	B	B
x_i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	✓	✓	✓	X	X	X	✓	✓	X	X	✓	✓	X	X	X	✓	✓	

$= \frac{8}{18}$

5. Use "leave-one-out" cross validation to estimate the error of 2-NN. Whenever you need to choose between two or more examples of identical distance or you need to determine a majority, prioritize your choice that maximizes the number of errors.

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

P.L.	A	A	A	B	A	B	A	A	B	A	B	B	A	B	A	B	B	B
*	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
				X	X	X			X	X			X	X	X			

6. Use "leave-one-out" cross validation to estimate the error of 3-NN. Whenever you need to choose between two or more examples of identical distance or you need to determine a majority, prioritize your choice that maximizes the number of errors.

Answer: $\frac{?}{18} = \frac{4}{18}$

P.L.	A	A	A	A	A	A	A	B	A	B	B	B	B	B	B	B
X	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
				X				X	X					X			

Part II: Naïve Bayesian Classifier (20 points)

Consider the following training data from an employee database. The target feature (attribute) is *salary*.

department	status	age	salary
sales	senior	31-40	Medium
sales	junior	21-30	Low
sales	junior	31-40	Low
systems	junior	21-30	Medium
systems	senior	31-40	High
systems	junior	21-30	Medium
systems	senior	41-50	High
marketing	senior	31-40	Medium
marketing	junior	31-40	Medium
secretary	senior	41-50	Medium
secretary	junior	21-30	Low

If given a test instance with values: systems, senior, and 21-30 for the descriptive attributes department, status, and age, respectively, what would be a Naïve Bayesian classification for the salary of the test instance? (*Show your calculation process*)

Answer: $P(x_1 = \text{systems}, x_2 = \text{senior}, x_3 = 21\text{-}30 \mid A)$

$=$ prob$(x_1 = a_1 \mid A)$ prob$(x_2 = a_2 \mid A)$ prob$(x_3 = a_3 \mid A)$ prob(A)

	prob$(x_1 = a_1 \mid A)$	prob$(x_2 = a_2 \mid A)$	prob$(x_3 = a_3 \mid A)$	prob(A)
high : A = high	2/4	2/5	1/4	2/11
med : A = med	2/4	3/5	1/4	6/11
low : A = low	0	0	2/4	3/11

high: $2/4 \cdot 2/5 \cdot 1/4 \cdot 2/11 = 0.009090909$

med: $2/4 \cdot 3/5 \cdot 1/4 \cdot 6/11 = 0.040909091$

low: $0 \cdot 0 \cdot 2/4 \cdot 3/11 = 0$

Of this test instance of attributes the value would be a salary of category medium for this Naive Bayes Classifier

Part III: Decision Tree (50 points).

Consider the following set of training data. The target feature (attribute) is *Class*, which can have values *text* (text file) or *.exe* (executable file) for different instances, is to be predicted based on other descriptive features (attributes) of the instance.

Instance	Writable	Updatable	Size	Class
1	yes	no	small	text
2	yes	yes	large	text
3	no	yes	medium	text
4	no	no	medium	.exe
5	yes	no	large	.exe
6	no	no	large	.exe

Note for Questions 1 and 2: Internal nodes of trees are annotated with descriptive feature names, leaf nodes are labeled with *Class* values. Sorted instances could be displayed within nodes of the trees.

Question 1 (10 points)

Construct the decision tree (called T1) generated by first considering the descriptive feature *Writable*, then *Updatable*, and then *Size*. See Back of this page.

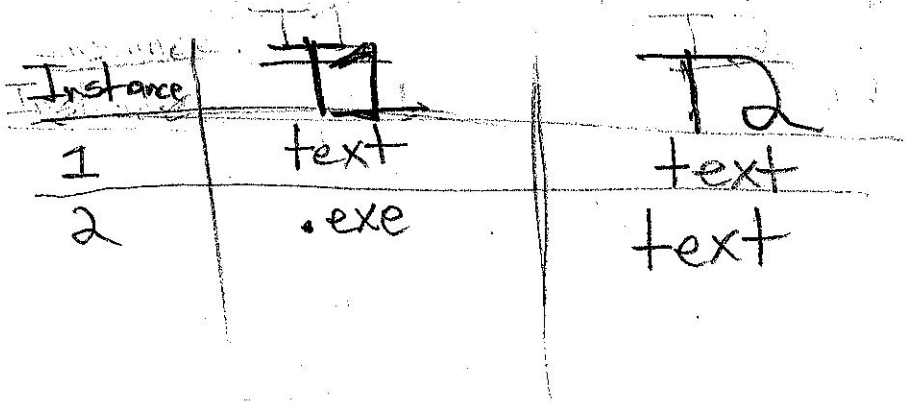
Question 2 (30 points)

Compute and construct the decision tree (called T2) using the ID3 algorithm. Show your work of calculations. See Last Page Front & Back Sides

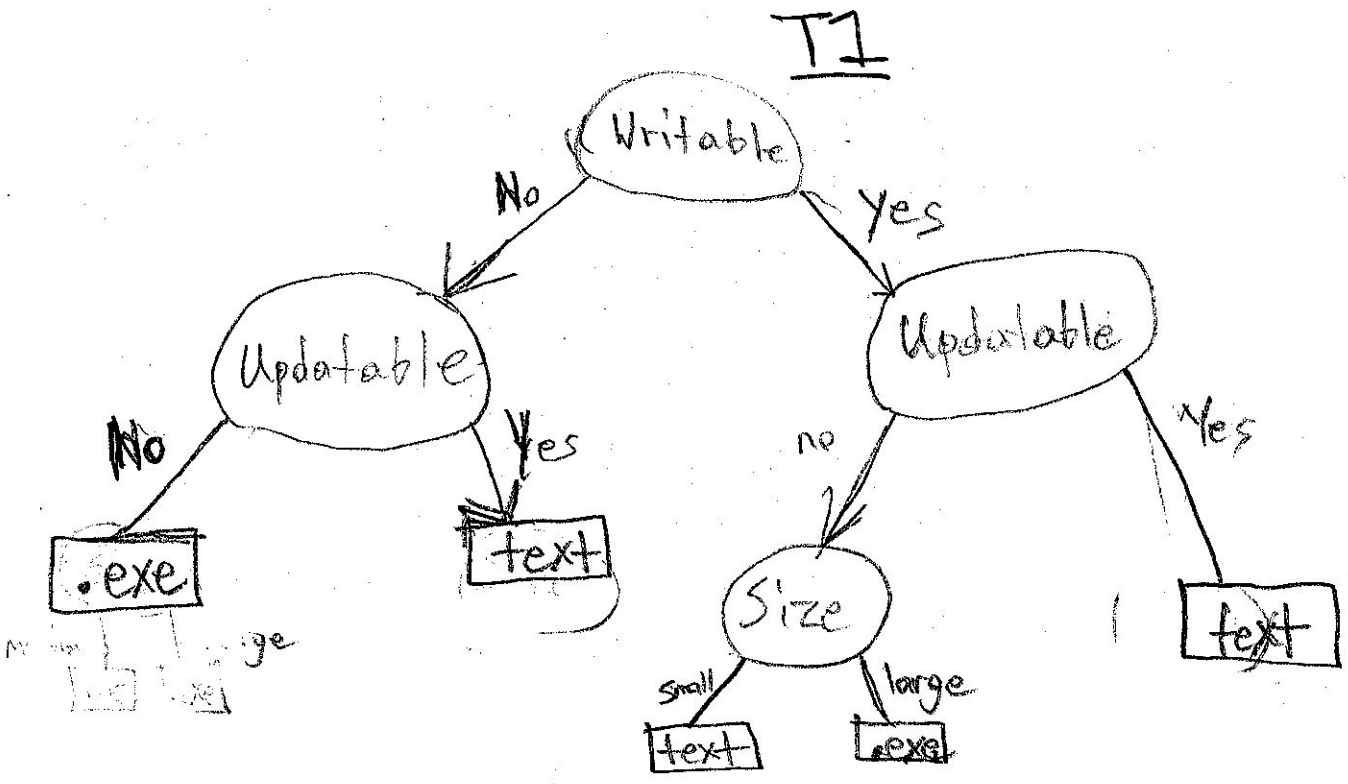
Question 3 (10 points)

How will T1 and T2 classify the following two new instances, respectively?

Instance	Writable	Updatable	Size	Class
1	yes	yes	large	?
2	no	no	small	?



Part III Question 1



Part III Question 2 ID3

Entropy of Dataset

Entropy(D)

$= ([3.exe, 3text]) = -\left[\frac{3}{6}\log_2\left(\frac{3}{6}\right) + \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right]$

Information Gain $= \boxed{1}$

Entropy(D) = 1

H(Writable) $= \left(\frac{3}{6}\left(-\left[\left[\frac{2}{6}\cdot\log_2\left(\frac{2}{6}\right)\right] + \left[\frac{1}{6}\log_2\left(\frac{1}{6}\right)\right]\right)\right) + \ldots$

Writable: Yes, Class: Text — Writable: Yes, Class: .exe

$\left(\frac{3}{6}\left(-\left[\frac{1}{6}\cdot\log_2\left(\frac{1}{6}\right)\right] + \left[\frac{2}{6}\log_2\left(\frac{2}{6}\right)\right]\right)\right) = 0.9591479$

Writable: No, Class: text — Writable: No, Class: .exe

Info. Gain (Writable) = 1 − 0.959147917 = 0.040852083

H(Updatable) $= \left(\frac{2}{6}\left(-\left(\left[\frac{2}{6}\cdot\log_2\left(\frac{2}{6}\right)\right] + [0]\right)\right)\right) + \ldots$

Writable: Yes, Class: Text — Writable: Yes, Class: .exe

$\left(\frac{4}{6}\left(-\left(\left[\frac{1}{6}\cdot\log_2\left(\frac{1}{6}\right)\right] + \left[\frac{3}{6}\cdot\log_2\left(\frac{3}{6}\right)\right]\right)\right)\right) = 0.75054$

Writable: No, Class: Text — Writable: No, Class: .exe

Info. Gain (Updatable) = 1 − 0.750543056 = 0.249456944

H(Size) $= \left(\frac{1}{6}\left(-\left(\left[\frac{1}{6}\cdot\log_2\left(\frac{1}{6}\right)\right] + [0]\right)\right)\right) + \left(\frac{2}{6}\left(-\left(\left[\frac{1}{6}\log_2\left(\frac{1}{2}\right)\right] + \left[\frac{1}{6}\log_2\left(\frac{1}{6}\right)\right]\right)\right)\right)$

Size: Small, Class: Text — Size: Small, Class: .exe — Size: Medium, Class: Text — Size: Medium, Class: .exe

$+\left(\frac{3}{6}\left(-\left(\left[\frac{1}{6}\log_2\left(\frac{1}{6}\right)\right] + \left[\frac{2}{6}\log_2\left(\frac{2}{6}\right)\right]\right)\right)\right) = 0.78984653$

Size: Large, Class: Text — Size: Large, Class: Exe

Info Gain(Size) = 1 − 0.78 = 0.2101

I03 T2

*Based off InfoGo!
Calculator