CS3642-W02 Programming Assignment #4 (Spring 2024)

Total Points: 100 + 60 (bonus)

Due: April 7, 2024 (11:59 PM, EST)

1. [65 points] Build a decision tree classifier using the ID3 algorithm with information gain. The dataset consists of 17 training samples, employed to train a decision tree classifier for predicting the ripeness of uncut watermelons (a binary classification task). The ripeness of a watermelon is determined by its color, root, sound, texture, umbilicus, and surface. The possible values for each attribute are shown in the table.

Table. Watermelon dataset.

ID	Color	Root	Sound	Texture	Umbilicus	Surface	Ripe
1	Green	Curly	Muffled	Clear	Hollow	Hard	True
2	Dark	Curly	Dull	Clear	Hollow	Hard	True
3	Dark	Curly	Muffled	Clear	Hollow	Hard	True
4	Green	Curly	Dull	Clear	Hollow	Hard	True
5	Light	Curly	Muffled	Clear	Hollow	Hard	True
6	Green	Slightly Curly	Muffled	Clear	Slightly Hollow	Soft	True
7	Dark	Slightly Curly	Muffled	Slightly Blurry	Slightly Hollow	Soft	True
8	Dark	Slightly Curly	Muffled	Clear	Slightly Hollow	Hard	True
9	Dark	Slightly Curly	Dull	Slightly Blurry	Slightly Hollow	Hard	False
10	Green	Straight	Crisp	Clear	Flat	Soft	False
11	Light	Straight	Crisp	Blurry	Flat	Hard	False
12	Light	Curly	Muffled	Blurry	Flat	Soft	False
13	Green	Slightly Curly	Muffled	Slightly Blurry	Hollow	Hard	False
14	Light	Slightly Curly	Dull	Slightly Blurry	Hollow	Hard	False
15	Dark	Slightly Curly	Muffled	Clear	Slightly Hollow	Soft	False
16	Light	Curly	Muffled	Blurry	Flat	Hard	False
17	Green	Curly	Dull	Slightly Blurry	Slightly Hollow	Hard	False

1) [5 points] What is the entropy of the root node? ($Entropy(D)$)

2) [5 points] Suppose that we have selected color, which has three possible values {Green, Dark, Light}. If dataset (D) is split by color, then there are three subsets: D_1 (Color=Green), D_2 (Color=Dark), and D_3(Color=Light), what is the entropy of the three child nodes?

3) [5 points] What is the information gain: $Gain(D, Color) = Entropy(D) - \sum_{s=1}^{3} \frac{|D^s|}{|D|} Ent(D^s)$?

4) [40 points] Iterate other attributes **and draw the final decision tree**. You need to show the *entropy* calculation of each child node and the *Gain* calculation of each splitting node.

5) [10 points] Given a new watermelon with attributes $Color = Green, Root = Slightly\ Curly, Sound = Dull, Texture = Clear, Umbilicus = Hollow, Surface = Hard$, is it a ripe watermelon? Please provide an explanation based on the decision tree built in 4).

2. [35 points] Using the dataset above, build a Naïve Bayes classifier, to predict the label of a new watermelon with attributes $Color = Green, Root = Slightly\ Curly, Sound = Dull, Texture = Clear, Umbilicus = Hollow, Surface = Hard$. Detailed explanation is required.

3. [30 bonus points]. Implement a function to automatically build the decision tree for watermelon ripeness classification. Show the running results for classifying the new watermelon with the attributes $Color = Green, Root = Slightly\ Curly, Sound = Dull, Texture = Clear, Umbilicus = Hollow, Surface = Hard$. Note that you must read the dataset provided in the CSV file and build the decision tree according to the CSV file. Show the screenshot of the prediction for the new watermelon.

4. [30 bonus points]. Implement a function to automatically predict the watermelon ripeness using Naïve Bayes

classifier. Show the running results for classifying the new watermelon with the attributes $Color = Green, Root = Slightly\ Curly, Sound = Dull, Texture = Clear, Umbilicus = Hollow, Surface = Hard$
Note that you must read the dataset provided in the CSV file and build the Naïve Bayes classifier according to the CSV file. Show the screenshot of the prediction for the new watermelon.

You may write your code in a contemporary language of your choice; typical languages would include C/C++, Python, Java, Ada, Pascal, Smalltalk, Lisp, and Prolog.

For Question 3 and 4, it is NOT allowed to use the machine learning library, such as Scikit-Learn to build the classifier.

Submission requirement:

1. Submit *a* ***PDF file*** of your well-commented source program, your design, and your printed outputs (**screen shots**). **Please include your codes in your PDF file.** It is plagiarism to take any codes from the website or others. Try to understand the algorithm and implement the algorithm by your own.
2. For the bonus questions, please submit ***your project in a zipped file*** with an organized structure.
3. Please upload items 1) and 2) above separately to D2L.

Adding the following 5 sections at the beginning of your PDF, including your code and outputs.

I. Your information

// Course: _______________

// Student name: _______________

// Student ID: _______________

// Assignment #: _______________

// Due Date: _______________

// Signature: _______________(Your signature assures that everything is your own work. Required.)

// Score: _______________(Note: Score will be posted on D2L)

II. [65 points]. Decision tree for watermelon ripeness classification.

1) [5 points] What is the entropy of the root node? ($Entropy(D)$)

2) [5 points] Suppose that we have selected color, which has three possible values {Green, Dark, Light}. If dataset (D) is split by color, then there are three subsets: D_1 (Color=Green), D_2 (Color=Dark), and D_3(Color=Light), what is the entropy of the three child nodes?

3) [5 points] What is the information gain: $Gain(D, Color) = Entropy(D) - \sum_{s=1}^{3} \frac{|D^s|}{|D|} Ent(D^s)$?

4) [40 points] Iterate other attributes **and draw the final decision tree**. You need to show the *entropy* calculation of each child node and the *Gain* calculation of each splitting node.

5) [10 points] Given a new watermelon with attributes $Color = Green, Root = Slightly\ Curly, Sound = Dull, Texture = Clear, Umbilicus = Hollow, Surface = Hard$, is it a ripe watermelon? Please provide an explanation based on the decision tree built in 4).

III. [35 points] Naïve Bayes classifier, to predict the label of a new watermelon with attributes $Color = Green, Root = Slightly\ Curly, Sound = Dull, Texture = Clear, Umbilicus = Hollow, Surface = Hard$

IV. [30 bonus points] Implement a function to automatically build the decision tree for watermelon ripeness classification.

1) [25 points] Implementation
2) [5 points] Screenshots

V. [30 bonus points] Implement a Naïve Bayes classifier to predict the watermelon ripeness.

1) [25 points] Implementation
2) [5 points] Screenshots