DDA3020 Homework 2

Instructions

- The **deadline** is 23:59, Nov 10, 2025.
- The weight of this assignment in the final grade is 20%.
- **Electronic submission**: Turn in solutions electronically via Blackboard. Be sure to submit your answers as one pdf file plus two python scripts for programming questions. Please name your solution files as"DDA3020HW2_studentID_name.pdf", "HW2_name_Q1.ipynb" and "HW2_name_Q2.ipynb" (".py" files are also acceptable).
- The complete and executable codes must be submitted. If you only fill in some of the results in your answer report for programming questions and do not submit the source code (.py or .ipynb files), you will receive 0 points for the question.
- Note that **late submissions** will result in discounted scores: 0-48 hours $\rightarrow$ 50%, more hours $\rightarrow$ 0%.
- Answer the questions in English. Otherwise, you'll lose half of the points.
- Collaboration policy: You need to solve all questions independently and collaboration between students is **NOT** allowed.

1 Written Problems (50 points)

1.1. (Tree-based Model, 25 points)

(5 points)Question 1: Node "Texture" splits into two leaves (Smooth and Rough), as shown in Figure 1. Analyze the performance on the training and validation sets, and decide whether the subtree should be pruned into a single leaf and justify your answer.

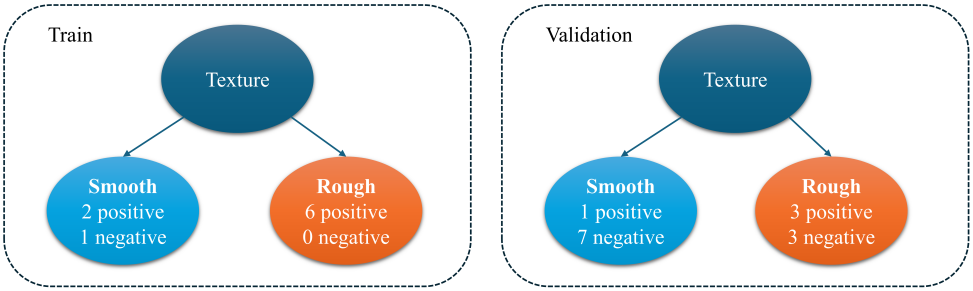

Figure 1: Training and validation distributions for node "Texture".

(20 points)Question 2: C4.5 is a decision tree algorithm that selects the feature with the maximum Gain Ratio as its splitting criterion. Given the training dataset in Table 1, we define

$$Gain(A) = H(D) - H(D|A) \textbf{ (or, } Gain(A) = Info(S) - Info(S|A) \textbf{ in lecture slides),}$$

$$GainRatio(A) = \frac{Gain(A)}{H(A)}.$$

where A_1, A_2, A_3, A_4 correspond to *Age*, *Employed*, *Own House*, and *Credit Status*, respectively. These features are used to predict whether a person can successfully obtain a loan application (*Class*: Yes/No). Answer the following sub-questions:

1. (5 points) Compute the *GainRatio* for features A_1, A_2, A_3, and A_4.
2. (5 points) According to the C4.5 algorithm, select the feature with the highest *GainRatio* as the root node. (For instance, if A_1 is chosen as the root, its child nodes correspond to the attribute values: *Youth*, *Middle-aged*, and *Elderly*).
3. (5 points) Iteratively construct the decision tree using *GainRatio* until 100% classification accuracy is achieved.
4. (5 points) Based on your results, draw the final decision tree and state whether all features are required to achieve perfect classification.

Table 1: Sample Data for Loan Applications

ID	Age	Employed	Own House	Credit Status	Class
1	Youth	No	No	Average	No
2	Youth	No	No	Good	No
3	Youth	Yes	No	Good	Yes
4	Youth	Yes	Yes	Average	Yes
5	Youth	No	No	Average	No
6	Middle-aged	No	No	Average	No
7	Middle-aged	No	No	Good	No
8	Middle-aged	Yes	No	Good	Yes
9	Middle-aged	No	Yes	Very Good	Yes
10	Middle-aged	No	Yes	Very Good	Yes
11	Elderly	No	Yes	Very Good	Yes
12	Elderly	No	No	Good	Yes
13	Elderly	Yes	No	Good	Yes
14	Elderly	Yes	No	Very Good	Yes
15	Elderly	No	No	Average	No

1.2. (Neural Network, 25 points) Consider the convolutional network defined by the layers below. The input shape is $32 \times 32 \times 3$ (regard it as a color image) and the output is 10 neurons (imagine in a 0-9 digit classification scenario). A CNN-based classifier can be defined as:

$$\text{Conv3}(16) + \text{Maxpool}_2 + \text{Conv5}(24) + \text{Maxpool}_2 + \text{FC10}$$

where

- Conv3(16): 16 filters with each size $3 \times 3 \times D$, where D is the depth of the activation volume at the previous layer, stride $= 1$, padding $= 1$;
- Conv5(24): 24 filters with each size $5 \times 5 \times D$, where D is the depth of the activation volume at the previous layer, stride $= 1$, padding $= 2$;
- Maxpool_2: 2×2 filter, stride $= 2$, padding $= 0$;
- FC10: A fully-connected layer with 10 output neurons.

Task: Compute the shape of activation map and the total number of parameters of each layer. Fill the results in the table.

NO.	Layer	Activation Shape	# Parameters	Mark
1	Input Layer	(32,32,3)	0	-
2	Conv3(16)			[3+2] pts
3	Maxpool2			[3+2] pts
4	Conv5(24)			[3+2] pts
5	Maxpool2			[3+2] pts
6	FC10			[3+2] pts

2 Programming (50 points)

Note that for programming you are supposed to write and run the code, and submit two notebook files with **the running result of each code block**.

2.1. (Tree-based Model, 25 points) Please download HW2_P1_DT.ipynb and adult.data and adult.test first and refer to the notebook file.

2.2. (Neural Network, 25 points) Please download HW2_P2_NN.ipynb and P2_data first and refer to the notebook file.