DDA3020 Homework 2

Due date: April 04, 2025

Instructions

- The **deadline** is 23:59, April 04, 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_Q3.ipynb" and "HW2_name_Q4.ipynb". Please do NOT include the data in your submission.
- 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, you will receive 0 points for the question.
- Note that **late submissions** will result in discounted scores: 0-48 hours → 50%, more hours → 0%. And coding files **without running output** will lead to additional point deduction.
- 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. (Decision Trees, 30 points)

(2 points)Question 1: In a decision tree, what is the purpose of the leaf nodes?

A) To represent the class label or value to be predicted

B) To store the conditions for splitting the data

C) To indicate the importance of a feature

D) To represent the depth of the tree

(2 points)Question 2: Which of the following algorithms can be used for both classification and regression tasks?

A) Decision trees

B) Logistic Regression

C) Support Vector Machines

D) All of the above

(2 points)Question 3: How can decision trees be made more robust to noise in the data?

A) By increasing the maximum depth of the tree

B) By using a smaller minimum samples per leaf

C) By using ensemble techniques like bagging or boosting

D) By removing features with low importance

(2 points)Question 4: How do decision trees handle continuous variables?

A) By discretizing the continuous variables into intervals

B) By using one-hot encoding

C) By normalizing the continuous variables

D) By ignoring the continuous variables

(2 points)Question 5: What is entropy in the context of decision trees?

A) A measure of disorder or impurity in a node

B) A measure of the complexity of a decision tree

C) The difference between the predicted and actual values in a node

D) The rate at which information is gained in a decision tree

Day	Weather	Temperature	Humidity	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	Weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	Rainy	Mild	High	Strong	No

Take a look at the following data. Here, the Y label indicates whether a child goes out to play or not.

1. (12 points, 3 points for each attribute) Calculate the information gain for each attribute and select the optimal splitting attribute (to construct the first layer of the tree).
2. (8 points) Draw the final tree structure without writing out the calculation process. (Hint: The error rate of the final training set is 0, that is, all samples are correctly classified.)

1.2.(CNN, 20 points) Consider the convolutional network defined by the layers below. The input shape is $28 \times 28 \times 1$ and the output is 5 neurons. Consider layers:

$$\text{Conv4(15)} + \text{Maxpool}_2 + \text{Conv2(25)} + \text{Maxpool}_2 + \text{FC5}$$

where

- Conv4(15): 15 filters with each size $4 \times 4 \times D$, where D is the depth of the activation volume at the previous layer, stride = 1, padding = 1;
- Conv2(25): 25 filters with each size $2 \times 2 \times D$, where D is the depth of the activation volume at the previous layer, stride = 1, padding = 0;
- Maxpool$_2$: 2×2 filter, stride = 2, padding = 0;
- FC5: A fully-connected layer with 5 output neurons.

1. (10 pts) Compute the shape of activation map of each layer.
2. (10 pts) Compute the total number of parameters of each layer.

Table 1: Understanding and calculating the number of parameters in CNN

NO.	**Layer**	**Activation Shape**	**# Parameters**	Mark
1	Input Layer	(28,28,1)	0	-
2	Conv4(15)			4 pts
3	Maxpool2			4 pts
4	Conv2(25)			4 pts
5	Maxpool2			4 pts
6	FC5			4 pts

2 Programming (50 points)

2.1. Important Notes

1. Detailed instructions and questions are in notebook (.ipynb) files. Remember to fill all blanks labeled with [TASK]. They include not only code tasks but also analysis tasks.

2. Don't need to include your answer for these two programming questions in your PDF report, but submit two notebook (.ipynb) files.

3. Please do NOT include the data files in your submission.

4. Please ensure that all running outputs are displayed in submitted code files. Otherwise, there will be point deduction for the coding part.

2.2. (Tree-based Models, 30 points) In this question, we will implement several tree-based models using *sklearn* package. Follow instructions and codes provided in **Q3.ipynb** to solve this question.

2.3. (Convolutional Neural Network, 20 points) In this question, we will explore the CNN models with the help of *TensorFlow* package, as outlined in the **Q4.ipynb** file. By following the instructions in the notebook, we will implement a simple and an advanced CNN models to solve the real problem.