Homework 04

Submission Notices:

- *Conduct your homework by filling answers into the placeholders given in this file (in Microsoft Word format). Questions are shown in black color, instructions/hints are shown in italic and blue color, and your content should use any color that is different from those.*
- *After completing your homework, prepare the file for submission by exporting the Word file (filled with answers) to a PDF file, whose filename follows the following format,*
  <StudentID-1>_<StudentID-2>_HW02.pdf   (Student IDs are sorted in ascending order)
  E.g., ***2312001_2312002_HW04.pdf***
  and then submit the file to Moodle directly WITHOUT any kinds of compression (.zip, .rar, .tar, etc.).
- *Note that you will get zero credit for any careless mistake, including, but not limited to, the following things.*
  1. *Wrong file/filename format, e.g., not a pdf file, use "-" instead of "_" for separators, etc.*
  2. *Disorder format of problems and answers*
  3. *Conducted not in English*
  4. *Cheating, i.e., copy other students' works or let the other student(s) copy your work.*

Problem 1. (2.5pts) Answer the following simple questions.

Please fill your answer in the table below.

Score	Questions	Answers
0.5pt	What is the purpose of an activation function in a neural network, and name two commonly used activation functions?	The activation function introduces non-linearity into the network, allowing it to learn complex patterns, Two commonly used activation function is step-activation function and sigmoid function
0.5pt	Explain overfitting in machine learning and describe one technique to reduce it.	Overfitting happens when a model learns the training data too well, including noises, and perform poorly on unseen data. One technique to reduce is dropout, which randomly deactives neurons during training to improve generalization
0.5pt	What is the difference between classification and regression tasks? Give an example of each.	Classification predicts a discrete categories (ex: predicating if an email is spam or not). Regression predicts continuous values (ex: predicating house prices based on feature).
0.5pt	Define the term "learning rate" in gradient descent and explain its impact on training.	The learning rate controls the step size at each iteration when updating weights. If it is too high, the model may diverge; if too low, training become very slow.

0.5pt	What is a confusion matrix, and how is it used to evaluate classification models?	A conclusion matrix is a table showing the counts of true positives, true negatives, false positives, false negatives. It helps evaluate classification performance by providing metric like accuracy, precision, recall and F1-score.

Problem 2. (1pt) Answer the following simple questions.

Please fill your answer in the table below.

Score	Questions	True / False	Explanation
0.25pt	In reinforcement learning, the policy defines the agent's behavior at each state.	True	A policy maps state to actions, defining how the agent behaves in each state.
0.25pt	Model-free RL methods require an explicit model of the environment's dynamics.	False	Model-free methods do not require knowledge of environment dynamics; they learn directly from experience.
0.25pt	Inverse reinforcement learning is mainly useful when designing a reward function manually is difficult.	True	IRL infers reward function from expert demonstrations, useful when manual reward design is complex.
0.25pt	In reinforcement learning, Q-learning is an example of a value-based method.	True	Q-learning learns the action-value function (Q-function) and selects actions based on maximizing it, which is a value-based approach.

Problem 3. (2pt) Decision tree for a binary classification task. We collected data on whether students decide to study in the library, based on three factors:

- ExamSoon? (Y/N) - whether they have an exam within 3 days.
- GroupStudy? (Y/N) - whether they have a study group session planned.
- Weather (R = Rainy, S = Sunny).

Your task: Determine the best decision tree to predict "Study in Library?" using the dataset below.

ExamSoon?	GroupStudy?	Weather	Study in Library?
Y	N	R	Y
Y	Y	R	Y
Y	Y	S	Y
Y	N	S	N
N	Y	S	Y

N	Y	R	N
N	N	S	N
N	N	R	N
Y	Y	S	Y
N	Y	S	Y

a) (1.5pts) Determine the best ID3 decision tree.
   - Calculate the entropy of the entire dataset regarding the target variable "Study in Library?".
   - Show calculations for Information Gain at each split, then
   - **Draw the final ID3 decision tree**

b) (0.25pt) According to your decision tree, what is the decision of a student studying in the library on a **Sunny day** when they **do not have an exam soon** and **do not have a group study session**?

c) (0.25pt) According to your decision tree, what is the decision of a student studying in the library when they **have an exam soon** and **the weather is rainy**?

Work Section

a) $H(Dataset) = -\frac{6}{10}.log_2\frac{6}{10} - \frac{4}{10}.log_2\frac{4}{10} = 0.971.$

$$AE(ExamSoon?) = \frac{5}{10}(-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}) + \frac{5}{10}(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}) = 0.846$$

.

$$AE(GroupStudy) = \frac{4}{10}(-\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}) + \frac{6}{10}(-\frac{5}{6}\log_2\frac{5}{6} - \frac{1}{6}\log_2\frac{1}{6}) = 0.714.$$

$$AE(Weather) = \frac{4}{10}(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}) + \frac{6}{10}(-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}) = 0.951.$$

$$IG(ExamSoon?) = 0.971 - 0.846 = 0.125$$

$$IG(GroupStudy?) = 0.971 - 0.714 = 0.257$$

$$IG(Weather) = 0.971 - 0.951 = 0.2$$

Choose "GroupStudy?" as a first node for decision tree.

As "GroupStudy?" we have 4N (1Y and 3N) (1) and 6Y (5Y and 1N) (2)

(1):

$$H(Dataset) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = 0.811$$

$$AE(ExamSoon?) = \frac{2}{4}(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}) + \frac{2}{4}(-\frac{2}{2}\log_2\frac{2}{2}) = 0.5$$

$$AE(Weather) = \frac{2}{4}\left(-\frac{1}{2}\log_2\frac{1}{2} + -\frac{1}{2}\log_2\frac{1}{2}\right) + \frac{2}{4}(-\frac{2}{2}\log_2\frac{2}{2}) = 0.5$$

$$IG(ExamSoon?) = 0.811 - 0.5 = 0.311.$$

$$IG(Weather) = 0.811 - 0.5 = 0.311.$$

Two IG value equals. Choose ExamSoon as new node for First Branch of tree.

- ExamSoon = N: all N.
- ExamSoon = Y: Y = 1 and N = 1. Spilit with remaining attributes is Weather (R = Y, S = N).

(2):

$$H(Dataset) = -\frac{5}{6}.\log_2\frac{5}{6} - \frac{1}{6}.\log_2\frac{1}{6} = 0.650.$$

$$AE(ExamSoon?) = \frac{3}{6}\left(-\frac{3}{3}\log_2\frac{3}{3}\right) + \frac{3}{6}\left(-\frac{2}{6}\log_2\frac{2}{6} - \frac{1}{6}\log_2\frac{1}{6}\right) = 0.480$$

$$AE(Weather) = \frac{2}{6}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{4}{6}\left(-\frac{4}{4}\log_2\frac{4}{4}\right) = 0.333.$$

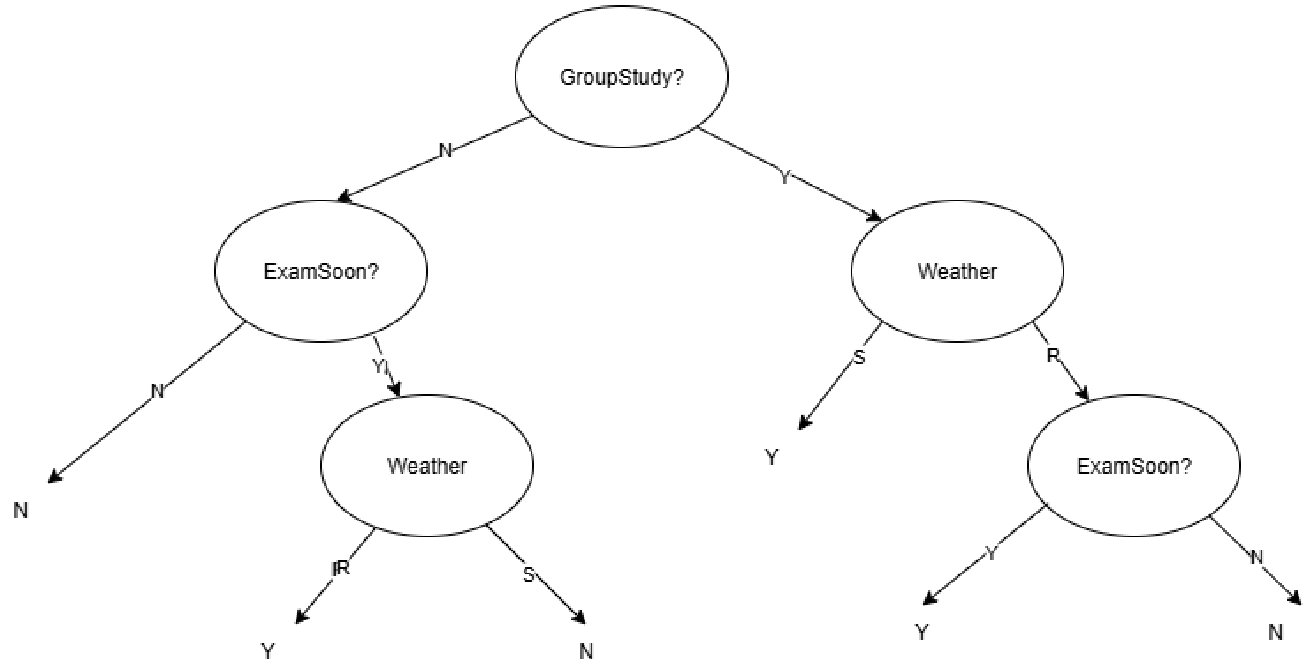
$$IG(ExamSoon?) = 0.650 - 0.480 = 0.17.$$

$$IG(Weather) = 0.650 - 0.333 = 0.317.$$

IG(Weather) > IG(ExamSoon?), Weather is the new node of the second branch.

- Weather = S → all Y.
- Weather = R → 2 samples ($Y = 1, N = 1$)
    - With ExamSoon? as the last:
        - $ExamSoon\ = Y\ \rightarrow Y$
        - $ExamSoon = N \rightarrow N$

Complete Decision Tree.



b) Students don't study as a group, exam isn't coming soon → Not study in the library.

c) If GroupStudy = Y; Weather = R → ExamSoon = Y → Y.

If GroupStudy = N; ExamSoon = Y → Weather = R → Y

No matter what GroupStudy is, students will study at library (Y).

Problem 4. (2.5pts) Consider the following dataset, in which **Class** is the target attribute.

Color	Size	Shape	Class
Red	Small	Round	Apple
Red	Large	Round	Blueberry
Red	Medium	Round	Apple
Blue	Small	Round	Blueberry
Green	Medium	Round	Cucumber
Red	Small	Oval	Apple
Blue	Medium	Oval	Blueberry
Green	Large	Oval	Cucumber
Red	Medium	Elongated	Apple
Blue	Large	Elongated	Cucumber

a) (1pt) Show your calculations to choose an attribute for the **root node** of the ID3 decision tree

b) (1pt) Show your calculations for selecting attributes at each branch of the ID3 decision tree until the tree is fully grown.

c) (0.5pt) Draw the resulting ID3 decision tree.

Work Section

a) Calculate for the root attribute.

$$H(Dataset) = -\frac{4}{10}\log_2\frac{4}{10} - \frac{3}{10}\log_2\frac{3}{10} - \frac{3}{10}\log_2\frac{3}{10} = 1.571$$

$$AE(Color) = \frac{5}{10}\left(-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}\right) + \frac{3}{10}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) + \frac{2}{10}\left(-\frac{2}{2}\log_2\frac{2}{2}\right) = 0.636.$$

$$AE(Size) = \frac{3}{10}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) + \frac{4}{10}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4}\right) + \frac{3}{10}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) = 1.1510.$$

$$AE(Shape) = \frac{5}{10}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{2}{5}\log_2\frac{2}{5} - \frac{1}{5}\log_2\frac{1}{5}\right) + \frac{3}{10}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{1}{3}\log_2\frac{1}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) + \frac{2}{10}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) = 1.436$$

$$IG(Color) = 1.571 - 0.636 = 0.935$$
$$IG(Size) = 1.571 - 1.1510 = 0.42$$
$$IG(Shape) = 1.571 - 1.436 = 0.135$$

Choose Color as a root of Decision tree.

b) Color = Green. 2 samples are cucumber.

Color = Blue (3 samples: Blueberry = 2, Cucumber = 1).

$$H(Dataset) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.918$$

$$AE(Size) = \frac{1}{3}\left(-\frac{1}{1}\log_2\frac{1}{1}\right) + \frac{1}{3}\left(-\frac{1}{1}\log_2\frac{1}{1}\right) + \frac{1}{3}\left(-\frac{1}{1}\log_2\frac{1}{1}\right) = 0$$

$$AE(Shape) = \frac{1}{3}\left(-\frac{1}{1}\log_2\frac{1}{1}\right) + \frac{1}{3}\left(-\frac{1}{1}\log_2\frac{1}{1}\right) + \frac{1}{3}\left(-\frac{1}{1}\log_2\frac{1}{1}\right) = 0$$

$$IG(Size) = 0.918 - 0 = 0.918$$
$$IG(Shape) = 0.918 - 0 = 0.918$$

Tie-breaker, choose Size:

- Blue & Size = Small → Blueberry.
- Blue & Size = Medium → Blueberry.
- Blue & Size = Large → Cucumber.

Color = Red (5 samples: apple: 4, Blueberry: 1)

$$H(Dataset) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0.722$$

$$AE(Size) = \frac{2}{5}\left(-\frac{2}{2}\log_2\frac{2}{2}\right) + \frac{2}{5}\left(-\frac{2}{2}\log_2\frac{2}{2}\right) + \frac{1}{5}\left(-\frac{1}{1}\log_2\frac{1}{1}\right) = 0$$

$$AE(Shape) = \frac{3}{5}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) + \frac{1}{5}\left(-\frac{1}{1}\log_2\frac{1}{1}\right) + \frac{1}{5}\left(-\frac{1}{1}\log_2\frac{1}{1}\right) = 0.551$$

$$IG(Size) = 0.722 - 0 = 0.722$$
$$IG(Shape) = 0.722 - 0.551 = 0.171$$

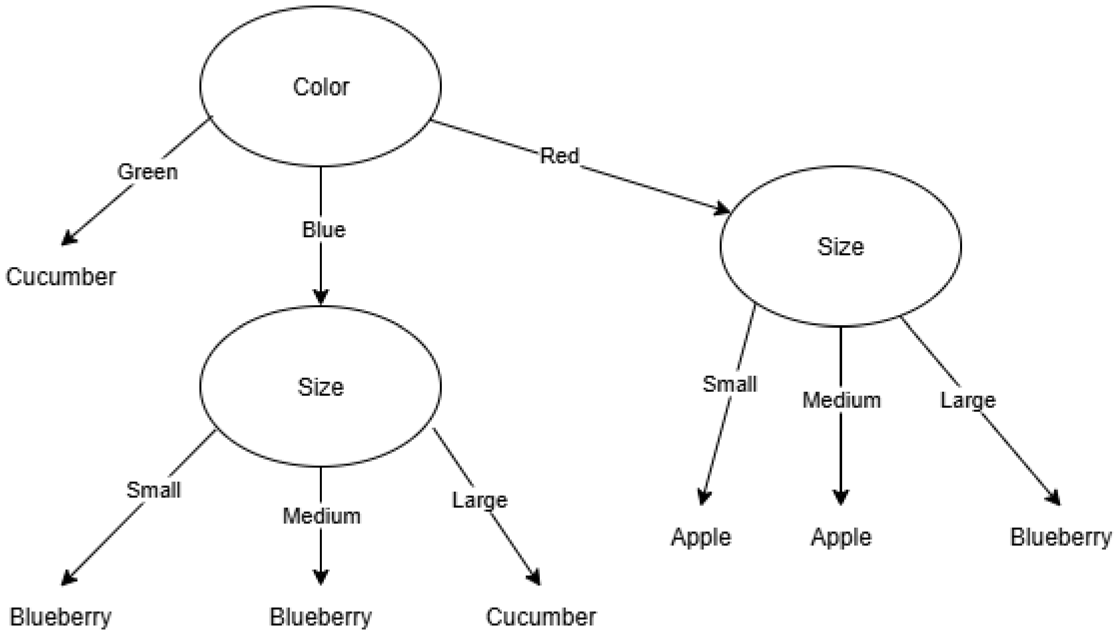
IG(Size) > IG(Shape), choose Size for next node.

Result:

- Red & Size = Small → Apple
- Red & Size = Medium → Apple

- Red & Size = Larger → Blueberry.

c) Complete decision tree.



Problem 5. (2pts) Gradient Descent

a) (1pt) Computational graph:

Consider a simple function:

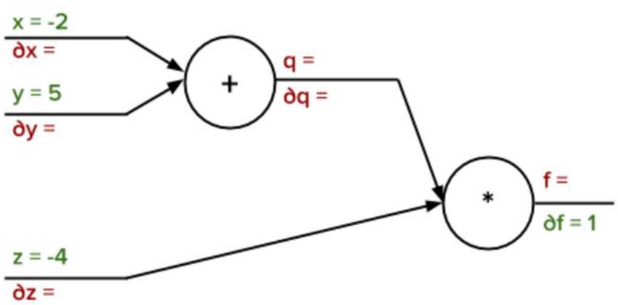
$$f(x, y, z) = (x + y) \cdot z$$

We can rewrite this as two separate equations:

$$q = x + y$$

$$f(x, y, z) = q \cdot z$$

Using this form, we can represent the process as a computation graph.



Now, suppose we evaluate the function at: $x = -2, y = 5, z = -4$. Additionally, let the upstream gradient - that is, the gradient of the loss L with respect to $f\,(\frac{\delta L}{\delta f})$ – be equal to 1. These values are already shown on the computation graph.

Your task: Solve for the required gradients both **symbolically** (without substituting numbers) and **numerically** (using the given values for $x = -2, y = 5, z = -4$ and $\frac{\delta L}{\delta f} = 1$).

1. ∂f / ∂q = $z = -4$

2. ∂q / ∂x = 1

3. $\partial q / \partial y = 1$

4. $\partial f / \partial z = x + y = -2 + 5 = 3$

5. $\partial f / \partial x = 1. z = 1 . -4 = (-4)$

6. $\partial f / \partial y = 1. z = 1. (-1) = -4$

b) (1pt) Computational graph:

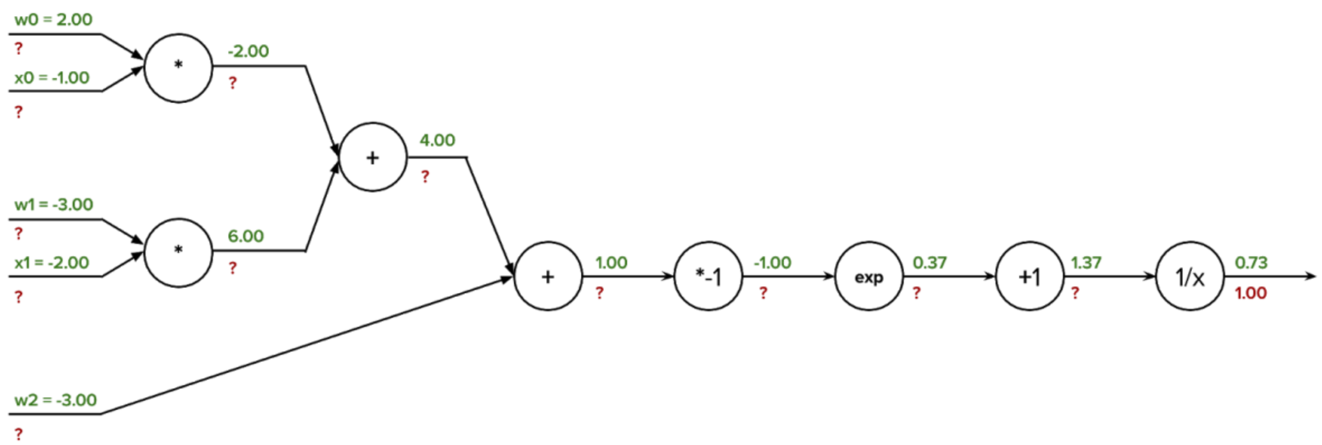
We will now carry out backpropagation for a single neuron in a neural network that uses a sigmoid activation function. First, define the pre-activation value:

$$z = w_0 x_0 + w_1 x_1 + w_2$$

The activation output is given by:

$$\alpha = \sigma(z) = \frac{1}{1 + e^{-z}}$$

The corresponding computation graph is shown below.



In this graph, the **forward pass activations** (values from the forward computation) are displayed above the edges, and the **upstream gradient** - the gradient of the loss L with respect to $\alpha \left(\frac{\delta L}{\delta \alpha}\right)$ is also provided.

Your task: Use this information to calculate all the remaining gradients (marked with question marks) in the computation graph. (Hint: You can use calculator here).

We have: $w_0 = 2, x_0 = -1, w_1 = -3, x_1 = -2, w_2 = -3.$

$$z = w_0 x_0 + w_1 x_1 + w_2 = -2 + 6 - 3 = 1$$

$$\alpha = \sigma(z) = \frac{1}{1 + e^{-x}} = \frac{1}{1 + e^{-1}} \approx 0.731$$

$$\left(\frac{\delta L}{\delta \alpha}\right) = 1.00$$

$$\frac{d\alpha}{dz} = \alpha(1 - \alpha)$$

1. $\frac{\partial \alpha}{\partial x_0} = \frac{d\alpha}{dz} . \frac{\partial z}{\partial x_0} = 0.731. (1 - 0.731). x_0 = 0.731. (1 - 0.731). (-1) = -0.197$

2. $\frac{\partial \alpha}{\partial w_0} = \frac{d\alpha}{dz} . \frac{\partial z}{\partial w_0} = 0.731 . (1 - 0.731). w_0 = 0.731(1 - 0.731). 2 = 0.393$

3. $\frac{\partial \alpha}{\partial x_1} = \frac{d\alpha}{dz} . \frac{\partial z}{\partial x_1} = 0.731 . (1 - 0.731). x_1 = 0.731(1 - 0.731). (-2) = -0.393$

4. $\frac{\partial \alpha}{\partial w_1} = \frac{d\alpha}{dz} . \frac{\partial z}{\partial w_1} = 0.731 . (1 - 0.731). w_1 = 0.731(1 - 0.731). (-3) = -0.5899$

5. $\frac{\partial \alpha}{\partial w_2} = \frac{d\alpha}{dz} . \frac{\partial z}{\partial w_2} = 0.731 . (1 - 0.731). w_2 = 0.731(1 - 0.731). (-3) = -0.5899$