Data Structures and Algorithms
with applications in Machine Learning
- MCQ 3 -

NAME: ____________________ **GROUP:** ____________________

Each Question: 1 Mark **Duration:** 20 Minutes

Completely fill the circles as shown: ○○●○

Answer sheet

Q1. ○ a. ○ b. ○ c. ○ d.

Q2. ○ a. ○ b. ○ c. ○ d.

Q3. ○ a. ○ b. ○ c. ○ d.

Q4. ○ a. ○ b. ○ c. ○ d.

Q5. ○ a. ○ b. ○ c. ○ d.

Q6. ○ a. ○ b. ○ c. ○ d.

Q7. ○ a. ○ b. ○ c. ○ d.

Q8. ○ a. ○ b. ○ c. ○ d.

Q9. ○ a. ○ b. ○ c. ○ d.

Q10. ○ a. ○ b. ○ c. ○ d.

The Quiz

Credit risk prediction aims to build a model capable of determining the quality of a loan. Loans are classified into two categories:

- **Good (G)**: Indicates a low risk of default, encoded as 1.
- **Bad (B)**: Indicates a high risk of default, encoded as 0.

The prediction is based on the following features:

- **Age**: A numerical value ranging from 18 to 100.
- **Geography**: A categorical variable with possible values `Germany`, `France`, `UK`, `China`, and `Japan`.
- **Sex**: A categorical variable, either `male` or `female`.
- **Purpose**: A categorical variable with values `P1`, `P2`, `P3`, and `P4`.
- **Credit**: A numerical value representing the loan amount in dollars.
- **Duration**: A numerical value indicating the loan term in months.
- **Education**: A categorical variable, classified as either `unskilled` or `skilled`.

Preprocessing

Q. 1 To normalize numerical features such as `Age`, `Credit`, and `Duration`, the following transformation is applied:

$$x_{\text{normalized}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

What is the name of this transformation?

○ a. Logarithmic Transformation

○ b. Min-Max Scaling (scaling to [0,1])

○ c. Standardization (scaling to zero mean and unit variance)

○ d. No transformation is required

Q. 2 Consider the categorical variable `Geography`, which has the following categories and their associated indices:

- `Germany`: 0
- `France`: 1
- `UK`: 2
- `China`: 3
- `Japan`: 4

After applying one-hot encoding, one column is dropped to avoid multicollinearity. If a sample has the value `Germany`, what would the one-hot encoded transformation look like?

○ a. `[1, 0, 0, 0, 0]`

- ○ b. `[1, 0, 0, 0]`
- ○ c. `[Germany]`
- ○ d. `[0, 0, 1, 0]`

Q. 3 After normalizing numerical features and one-hot encoding categorical features, what will be the shape of the feature matrix `X` and target vector `y`, assuming 1000 samples? The table below describes the number of categories for each categorical variable:

Categorical Variable	Number of Categories
`Geography`	5
`Sex`	2
`Purpose`	4
`Education`	2

Table 1: Number of Categories for Categorical Variables

Assume that one column is dropped for each one-hot encoded categorical variable to avoid multicollinearity.

- ○ a. `X.shape = (1000, 12)`, `y.shape = (1000,)`
- ○ b. `X.shape = (1000, 15)`, `y.shape = (1000,)`
- ○ c. `X.shape = (1000, 13)`, `y.shape = (1000,)`
- ○ d. `X.shape = (1000, 9)`, `y.shape = (1000,)`

In the next section, we will explore how to train a Decision Tree model to classify loans as either low-risk (`Good`) or high-risk (`Bad`).

Training a Decision Tree Model

Q. 4 Let Y be a random variable that can take values from the finite set $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$. The probability distribution of Y is denoted as $p(y) = \mathbb{P}(Y = y)$, where $y \in \mathcal{Y}$. The entropy of Y is defined as:

$$H(Y) = -\sum_{y \in \mathcal{Y}} p(y) \log_2(p(y)).$$

Which of the following statements best explains the meaning of entropy?

- ○ a. Entropy measures the likelihood of the most probable value of Y.
- ○ b. Entropy represents the total uncertainty across all possible outcomes of Y, without averaging.
- ○ c. Entropy is the measure of how many outcomes Y can possibly take.
- ○ d. Entropy quantifies the average uncertainty reduction (in bits) when the value of Y is revealed.

Q. 5 Which of the following equations correctly defines the information gain IG for a split based on X?

- ○ a. $IG(D_p, X) = I(D_p) - \frac{N_{\text{left}}}{N_p} I(D_{\text{left}}) - \frac{N_{\text{right}}}{N_p} I(D_{\text{right}})$
- ○ b. $IG(D_p, X) = I(D_p) + \frac{N_{\text{left}}}{N_p} I(D_{\text{left}}) + \frac{N_{\text{right}}}{N_p} I(D_{\text{right}})$
- ○ c. $IG(D_p, X) = \frac{N_{\text{left}}}{N_p} I(D_{\text{left}}) + \frac{N_{\text{right}}}{N_p} I(D_{\text{right}})$
- ○ d. $IG(D_p, X) = I(D_{\text{left}}) - I(D_p) - I(D_{\text{right}})$

Where:

- D_p: The dataset at the parent node.
- D_{left}, D_{right}: The datasets at the left and right child nodes.
- N_p: The total number of samples at the parent node.
- N_{left}, N_{right}: The total number of samples in the left and right child nodes.
- $I(D)$: The impurity measure of a dataset D.

Q. 6 Consider the following dataset of 10 samples with the binary feature "Sex" and the target variable Y:

Sex (X)	**Target** (Y)
Male	0
Female	1
Female	1
Male	0
Female	1
Male	0
Male	0
Female	1
Male	0
Female	0

Table 2: Dataset of 10 samples with binary feature "Sex" and target variable Y.

Using this dataset:

- Y is the target variable, where 1 indicates "Good" (low risk) and 0 indicates "Bad" (high risk).
- X is the feature "Sex," where "Male" and "Female" are the two possible values.

What is the information gain $IG(D, X)$ for a split based on X ?

- ○ a. $IG(D, X) = 0.5$
- ○ b. $IG(D, X) = 0.0$
- ○ c. $IG(D, X) = 0.8$
- ○ d. $IG(D, X) = 1.0$

Q. 7 The following pseudo-code implements a method to find the best feature and threshold for splitting a dataset in a Decision Tree.

Algorithm 1 Finding the Best Split

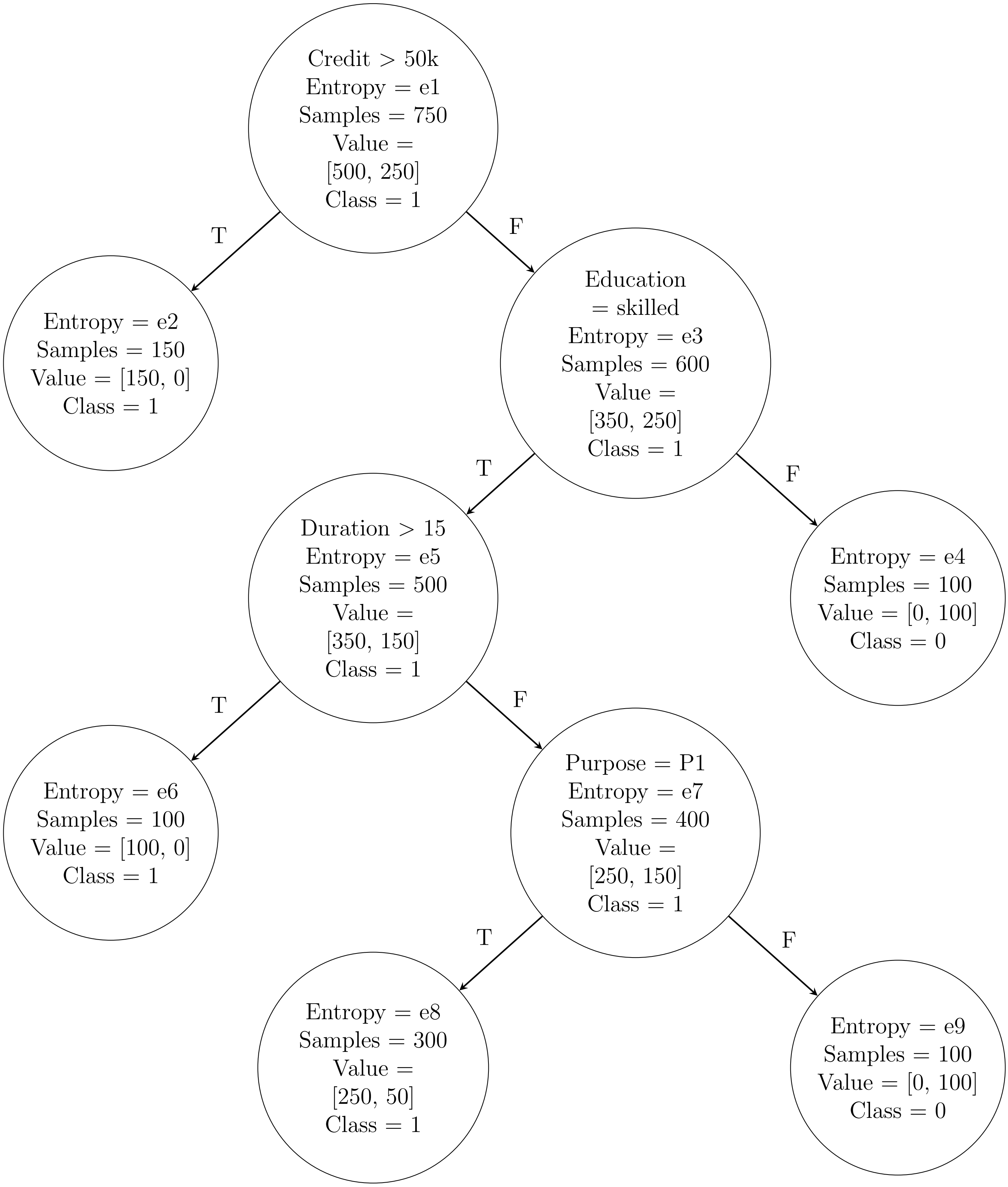
```
Require: Feature matrix X, target labels y
Ensure: Best feature, best threshold, and corresponding information gain
Initialize best_gain ← 0, best_feature ← None, best_threshold ← None
for each feature c in X do
    thresholds ← unique_values(X[:, c])
    for each threshold in thresholds do
        left ← X[:, c] < threshold
        right ← X[:, c] ≥ threshold
        if len(y[left]) == 0 or len(y[right]) == 0 then
            Continue
        end if
        gain ← compute_information_gain(y, y[left], y[right])
        if gain > best_gain then
            . . .                                         ▷ Fill in the blank
        end if
    end for
end for
return best_feature, best_threshold, best_gain
```

What should replace the blank to correctly update the variables?

- ○ a. best_gain $\leftarrow$ gain, best_feature $\leftarrow c$, best_threshold $\leftarrow$ threshold
- ○ b. Continue
- ○ c. best_feature $\leftarrow c$
- ○ d. gain $\leftarrow 0$

Trained Decision Tree and Prediction Process

The following graph illustrates the trained decision tree obtained after fitting the model to the dataset. Each node represents a decision point based on a specific feature, displaying the entropy, sample size, and class distribution.


Credit > 50k
Entropy = e1
Samples = 750
Value = [500, 250]
Class = 1
T
F
Entropy = e2
Samples = 150
Value = [150, 0]
Class = 1
Education = skilled
Entropy = e3
Samples = 600
Value = [350, 250]
Class = 1
T
F
Duration > 15
Entropy = e5
Samples = 500
Value = [350, 150]
Class = 1
Entropy = e4
Samples = 100
Value = [0, 100]
Class = 0
T
F
Entropy = e6
Samples = 100
Value = [100, 0]
Class = 1
Purpose = P1
Entropy = e7
Samples = 400
Value = [250, 150]
Class = 1
T
F
Entropy = e8
Samples = 300
Value = [250, 50]
Class = 1
Entropy = e9
Samples = 100
Value = [0, 100]
Class = 0

Q. 8 In the decision tree graph, which of the entropies e_i, where $i \in \{1, 2, \ldots, 9\}$, will be equal to zero?

- ○ a. e_1, e_3, e_5, e_7
- ○ b. e_2, e_3, e_5, e_7
- ○ c. e_2, e_4, e_6, e_9
- ○ d. e_8, e_9

Q. 9 Using the trained decision tree and the information provided in the nodes, determine the confusion matrix. Which of the following correctly represents the values for TP, TN, FP, and FN ?

	Predicted Positive	Predicted Negative
Actual Positive	TP = __	FN = __
Actual Negative	FP = __	TN = __

- ○ a. TP = 300, TN = 200, FP = 100, FN = 150
- ○ b. TP = 350, TN = 150, FP = 150, FN = 100
- ○ c. TP = 500, TN = 200, FP = 50, FN = 0
- ○ d. TP = 500, TN = 200, FP = 0, FN = 50

Q. 10 We have a new sample with the following features:

- **Age:** 24
- **Geography:** UK
- **Sex:** Female
- **Credit:** 35k$
- **Education:** skilled
- **Duration:** 11

What should be the value of **"Purpose"** to ensure that the Decision Tree algorithm predicts a **'Good'** target for this sample?

- ○ a. $P1$
- ○ b. $P2$
- ○ c. $P3$
- ○ d. $P4$