Practice Problem Set
Learning and Decision Tree
CSE422
Instructor: Ipshita Bonhi Upoma

Part 1: Short Calculation.

1. You are building a decision tree to predict whether a person will vote in an election based on their age group (18-30, 31-50, 51+), income level (Low, Medium, High), and interest in politics (Low, Medium, High).

   The dataset contains the following distribution:

   - 50 people in the 18-30 group: 30 vote, 20 do not vote.
   - 40 people in the 31-50 group: 35 vote, 5 do not vote.
   - 30 people in the 51+ group: 10 vote, 20 do not vote.

   Question 1.1 Compute the Entropy of the given dataset.
   Question 1.2 Compute the Conditional entropy of Age Group.
   Question 1.3 Compute the information gain if you were to split the data based on the age group.

   Suppose another attribute "Interest in Politics" with the following distribution:

   - 60 people have low interest: 15 vote, 45 do not vote.
   - 30 people have medium interest: 25 vote, 5 do not vote.
   - 30 people have high interest: 35 vote, 0 do not vote.

   Question 1.2 Calculate the information gain if you split the dataset based on "Interest in Politics."

   Question 1.3 Between “Interest in Politics” and “Age Group” which attribute should be in the root node of the decision tree.

2. A bank wants to predict if a customer will default on a loan based on their credit score (Low, Medium, High) and previous loan history (Good, Bad).

   The dataset has the following distribution:

   - 40 people with Low credit score: 10 default, 30 do not default.
   - 35 people with Medium credit score: 5 default, 30 do not default.
   - 25 people with High credit score: 1 default, 24 do not default.

   Question 2.1 What is the entropy of the dataset before the split, and what is the information gain when you split based on credit score?

Another attribute, "Loan History," has the following distribution:

- 50 people with Good loan history: 5 default, 45 do not default.
- 50 people with Bad loan history: 25 default, 25 do not default.

Question 2.2 Calculate the information gain for splitting the dataset based on loan history.

3. A factory is predicting whether a machine will fail based on temperature (Low, Medium, High) and maintenance frequency (Regular, Irregular).

   The distribution for the "Temperature" attribute is as follows:

   - 30 machines operate at low temperature: 5 fail, 25 do not fail.
   - 40 machines operate at medium temperature: 20 fail, 20 do not fail.
   - 30 machines operate at high temperature: 25 fail, 5 do not fail.

   Question: What is the information gain if you split the dataset based on temperature?

   Now, consider the "Maintenance Frequency" attribute:

   - 60 machines have regular maintenance: 10 fail, 50 do not fail.
   - 40 machines have irregular maintenance: 40 fail, 0 do not fail.

   Question: What is the information gain for splitting based on maintenance frequency?

Part 2: Constructing Decision Tree

Disclaimer: All of these dataset are fake and made for simulation purposes only.

For the following training datasets construct the full decision tree using the ID3 algorithm (shown in class). Show and explain all the steps of the ID3 algorithm using your work.

Dataset 1: This is a dataset of 14 records with features as age group, employment status, interest in community events and the class variable is Participation. Using this dataset as your training set and the ID3 algorithm, construct the decision tree to identify from these features if a person will participate in an event or not. Show all steps.

Attributes:
- **Age Group**: Young (Y), Middle-aged (M), Senior (S)
- **Employment Status**: Employed (E), Unemployed (U)
- **Interest in Community Events**: Low (L), Medium (M), High (H)

Target Variable

- **Participation**: Yes or No.

Age Group	Employment Status	Interest in Community Events	Participation
Young	Employed	High	Yes
Young	Employed	Medium	Yes
Young	Unemployed	Low	No
Young	Unemployed	High	Yes
Middle-aged	Employed	Medium	Yes
Middle-aged	Employed	Low	No
Middle-aged	Unemployed	Medium	No
Senior	Employed	High	Yes
Senior	Unemployed	Low	No
Senior	Unemployed	Medium	No
Senior	Employed	Low	No
Young	Employed	Low	No
Middle-aged	Unemployed	High	Yes
Senior	Unemployed	High	Yes

Dataset 2: This dataset focuses on analyzing factors that influence whether women feel safe in their workplace. The key attributes include Workplace Environment (supportive, neutral, hostile), Harassment Policies (strong, moderate, weak), Workplace Flexibility (high, medium, low), availability of Health Benefits (yes, no), and the presence of Workplace Safety Measures (yes, no). The target variable is Feeling Safe, which indicates whether a woman feels secure in her work environment (yes or no).

Attribute Summary:

- **Workplace Environment**: Supportive, Neutral, Hostile
- **Harassment Policies**: Strong, Moderate, Weak
- **Workplace Flexibility**: High, Medium, Low
- **Health Benefits**: Yes, No
- **Workplace Safety Measures**: Yes, No

Target Variable:

- **Feeling Safe**: Yes or No

Workplace Environment	Harassment Policies	Workplace Flexibility	Health Benefits	Safety Measures	Feeling Safe
Supportive	Strong	High	Yes	Yes	Yes
Hostile	Weak	Low	No	No	No
Neutral	Moderate	Medium	Yes	Yes	Yes
Neutral	Strong	Medium	Yes	Yes	No
Hostile	Weak	Low	No	No	No
Neutral	Moderate	High	Yes	No	No
Hostile	Strong	High	Yes	Yes	Yes
Neutral	Weak	Medium	No	No	No

Hostile	Moderate	Low	No	No	No
Supportive	Strong	High	Yes	Yes	Yes
Neutral	Moderate	Low	Yes	No	No
Hostile	Weak	Low	No	No	No
Supportive	Strong	High	Yes	Yes	Yes
Neutral	Moderate	Medium	No	Yes	Yes

Dataset 3: This dataset is focused on predicting whether a computer science student will have a successful career based on several behavior-related attributes such as Study hours per week, Project experience, Internship experience, Extracurricular Involvement, Networking efforts.

Study Hours per Week	**Project Experience**	**Internship Experience**	**Extracurricular Involvement**	**Networking Efforts**	**Successful Career**
High	High	Yes	High	High	Yes
Medium	Medium	Yes	Medium	Medium	Yes
Low	Low	No	Low	Low	No
High	Medium	Yes	High	Medium	Yes
Medium	Low	No	Medium	Low	No

Low	Low	No	Low	Low	No
High	High	Yes	Medium	High	Yes
Medium	Medium	Yes	High	Medium	Yes
Low	Medium	No	Medium	Low	No
High	High	Yes	High	High	Yes
Medium	Medium	Yes	Medium	High	Yes
Low	Low	No	Low	Low	No
High	High	Yes	High	High	Yes
Medium	Low	No	Medium	Low	No

Attribute Summary:

- Study Hours per Week: Low (0-10 hours), Medium (10-20 hours), High (20+ hours)
- Project Experience: Low, Medium, High (based on the number and complexity of projects completed)
- Internship Experience: Yes, No (whether the student has completed an internship)
- Extracurricular Involvement: Low, Medium, High (participation in clubs, hackathons, etc.)
- Networking Efforts: Low, Medium, High (attending events, meeting professionals, etc.)

Target Variable:

- Successful Career: Yes or No (based on factors such as securing a good job, advancing in the field, or achieving academic goals)

Dataset 4: This dataset focuses on predicting which type of crop—Cereal Crops, Vegetable Crops, or Fruit Crops—will grow best on a particular piece of farming land. The attributes include Soil Type (sandy, clay, loam), Water Availability (low, medium, high), Sunlight Exposure (low, medium, high), and Fertilizer Use (low, medium, high).

Soil Type	Water Availability	Sunlight Exposure	Fertilizer Use	Best Crop to Grow

Loam	High	High	Medium	Fruit Crops
Sandy	Low	High	Low	Cereal Crops
Clay	Medium	Medium	Medium	Vegetable Crops
Loam	Medium	High	High	Fruit Crops
Sandy	Low	Medium	Low	Cereal Crops
Clay	Medium	Low	Medium	Vegetable Crops
Loam	High	Medium	High	Fruit Crops
Sandy	Low	Medium	Low	Cereal Crops
Clay	High	Medium	Medium	Vegetable Crops
Loam	Medium	High	Medium	Fruit Crops
Sandy	Medium	High	Medium	Cereal Crops
Clay	Medium	Medium	High	Vegetable Crops
Loam	High	Low	Medium	Fruit Crops
Sandy	Low	High	Low	Cereal Crops
Clay	Medium	Low	Medium	Vegetable Crops

Attribute Summary:

- Soil Type: Sandy, Clay, Loam (the type of soil that affects crop growth)
- Water Availability: Low, Medium, High (the availability of water for crops)
- Sunlight Exposure: Low, Medium, High (the amount of sunlight the land receives)
- Fertilizer Use: Low, Medium, High (the amount of fertilizer applied to the land)

Target Variable:

- Best Crop to Grow:
  - Cereal Crops.

  - Vegetable Crops.
  - Fruit Crops.

Part 3: Calculate from given probabilities

1. The following information is collected from a dataset including three types of plants (Shrub, Flower and Tree) with three colors (Red, Yellow, Blue).

   Overall Type Probabilities: P(Shrub)=P(Flower)=P(Tree)= ⅓.

   Color Distributions for Each Type:

   - For Shrubs: P(Red | Shrub)=¼ , P(Yellow | Shrub)= ½ , P(Blue | Shrub)=¼.
   - For Flowers: P(Red | Flower)=½, P(Yellow | Flower)=0, P(Blue | Flower)=½.
   - For Trees: P(Red | Tree)=0, P(Yellow | Tree)=0, P(Blue | Tree)=1

   Consider plant type to be the class variable and colors to be an attribute.
   a. Calculate the Entropy of the dataset. [1]
   b. Calculate the Conditional Entropy, H(Type|Color) for attribute, Color with respect to predicting the type. [6]
   c. Calculate the Information Gain, I(Type, Color) for attribute, Color with respect to predicting the type.

Part 3: Conceptual Understanding

Find out the answer to the following questions.

1. What is the difference between supervised learning and unsupervised learning?
2. What is a training dataset, and why is it important in machine learning?
3. How does a machine learning model learn from data?
4. What is the purpose of splitting a dataset into training and test sets?
5. What is overfitting in machine learning, and how can it be prevented?
6. What is the difference between classification and regression?
7. What is the main objective of the ID3 algorithm in decision tree learning?
8. Write down the steps of constructing a decision tree using the ID3 algorithm
9. What role does entropy play in the ID3 algorithm?
10. How does ID3 determine the best attribute for splitting the dataset at each node?
11. What is information gain, and how is it used in the ID3 algorithm?
12. Why does the ID3 algorithm prefer attributes with the highest information gain?
13. How does the ID3 algorithm handle continuous attributes?
14. What are the potential drawbacks of using the ID3 algorithm in large datasets?
15. How does the ID3 algorithm handle missing or incomplete data?
16. Can the ID3 algorithm handle multi-class classification problems? How?

17. What is the stopping condition for the ID3 algorithm when building a decision tree?