Kingdom of Saudi Arabia
Ministry of Education
College of Computer
Department of Computer Science

المملكة العربية السعودية
وزارة التعليم
كلية الحاسب
قسم علوم الحاسب

Course: CS 435 "Machine Learning" **Assignment # 1**
1st Semester 1445

Student Name:	Student ID:

Qs. 1: Can you think of a time when an ML-based system (employing ML algorithms) faced problem(s) because it might have caused some issues in the society? How would you handle such situations in a responsible way? **[2 Marks]**

Answer:

machine learning chatbot released by a company started to make offensive comments after learning from user interactions. To handle this, I would immediately pause its interactions, review and filter its training data, implement stricter learning guidelines, and regularly monitor its responses for any inappropriate behavior.

Qs. 2: Have you come across an example where technology employing ML algorithms influenced people's lives or opinions in a significant way? How do you think professionals should make sure the use of ML does not cause problems in society? **[1.5 Marks]**

Answer:

One example of ML in medicine is algorithms used for diagnosing diseases from medical images, which can greatly aid doctors but sometimes lead to misdiagnoses. Professionals should ensure these ML systems are rigorously tested, continuously updated, and used in conjunction with expert medical judgment to ensure patient safety and accuracy.

Qs. 3: Clustering **[1 Mark]**

Cluster the data: 1, 3, 15, 30, 45 using Single and Complete linkage clustering

Answer:

Distance(1, 3) = |1 - 3| = 2,Distance(1, 15) = |1 - 15| = 14,Distance(1, 30) = |1 - 30| = 29

Distance(1, 45) = |1 - 45| = 44,Distance(3, 15) = |3 - 15| = 12,Distance(3, 30) = |3 - 30| = 27

Distance(3, 45) = |3 - 45| = 42,Distance(15, 30) = |15 - 30| = 15,Distance(15, 45) = |15 - 45| = 30,Distance(30, 45) = |30 - 45| = 15

Single Linkage Clustering:

Start: {1},{3},{15},{30},{45}

Merge {1}and {3}: {1,3},{15},{30},{45}

Merge {15}and{30}: {1,3},{15,30},{45}

Merge {1,3}with {15,30}: {1,3,15,30},{45}

Merge all into one cluster: {1,3,15,30,45}

Complete Linkage Clusterin

Start: {1},{3},{15},{30},{45}

Merge {1} and{3}: {1,3},{15},{30},{45}

Merge {30}and {45}: {1,3},{15},{30,45}

Merge {1,3}with {15}: {1,3,15},{30,45}

Merge all into one cluster:{1,3,15,30,45}

Qs. 4: Decision Trees [1.5 Marks]

Download the dataset hwdata.csv from the Blackboard. This dataset has 250 records used by a mortgage company that says whether someone will purchase a second mortgage or not, based on several factors. The data has been presorted for your help and also summarized in a table alongside the data.

There are three attributes to consider when predicting whether purchase is true or false:

- *Region*: Suburban or Urban
- *Marital Status*: Married or Single
- *Income*: High, Medium, or Low

Answer the following questions:

a. What is the best single-attribute decision rule and what is its associated error?

b. What is the information measure of the entire data set?

c. What is the information gain for splitting on Region? Marital Status? Income? Based on this, pick the initial attribute to split on.

d. Complete one more level of the decision tree and stop (i.e., each leaf tests no more than 2 attributes).

A:

The best single-attribute decision rule for predicting mortgage purchase is based on income_Low, with an associated error rate of approximately 26%.

B:

$$H(T) = -(0.4 \times \log_2(0.4) + 0.6 \times \log_2(0.6))$$

entropy approximately =0.9933

C:

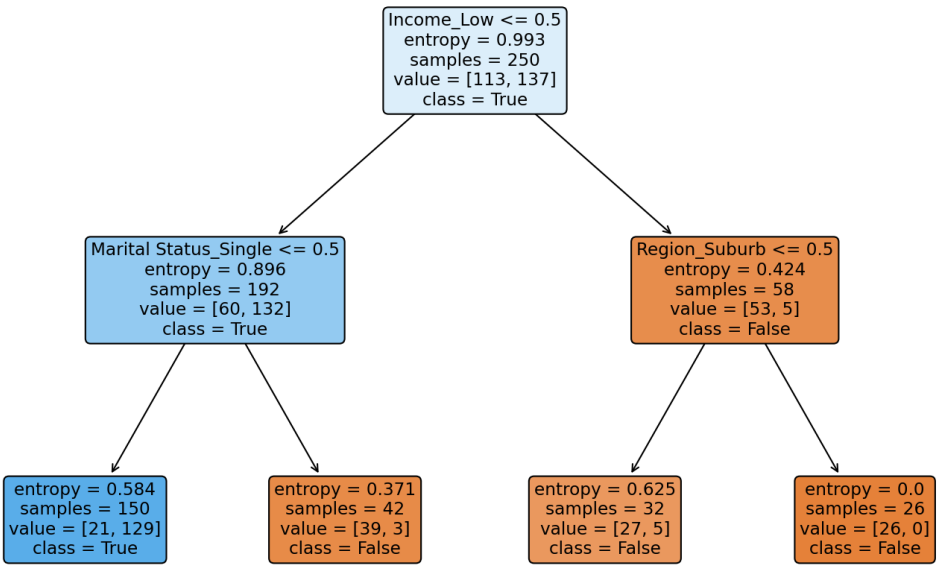
Region: Information gain = 0.000089

Marital Status: Information gain =0.1828

Income: Information gain =0.2098

The 'Income' attribute is the best initial attribute to split on due to the highest information gain.

D:



Qs. 5: Naïve Bayes Algorithm **[1 Mark]**

The following dataset contains loan information and can be used to try to predict whether a borrower will ***default***. Use the *Naïve Bayes* algorithm to determine whether:

loan *X = (Home Owner = No, Marital Status=Married, Income=High)*

loan *X = (Home Owner = Yes, Marital Status=Married, Income=Low)*

loan *X = (Home Owner = Yes, Marital Status=Divorced, Income=Low)*

should be classified as a Defaulted Borrower or not.

Home Owner	Marital Status	Annual Income	Defaulted Borrower
Yes	Single	High	No
No	Married	High	No
No	Single	Low	No
Yes	Married	High	No
No	Divorced	Low	Yes
No	Married	Low	No
Yes	Divorced	High	No
No	Single	Low	Yes
No	Married	Low	No
No	Single	Low	Yes

Answer:

P(DB =Yes) = 3/10

P(Defaulted Borrower=No) = 7/10

Home oner	y	n
y	0/3	3/3
n	3/7	4/7

Martial status	y	n
single	2/4	2/4
Married	0/4	4/4
divocred	1/2	1/2

Annual incom	y	n
y	0/4	4/4
n	3/6	3/6

loan X = (Home Owner = No, Marital Status=Married, Income=High)

(defultae= 3/10 * 3/7 * 0/4 * 0/4 = 0

(not defulte)= 7/10 * 4/7 * 4/4 * 4/4 = 0.4

P(DB=NO | X) is higher, we predict that loan X is Not Defaulted Borrower.

loan X = (Home Owner = Yes, Marital Status=Married, Income=Low)

default=3/10×0/3×0/4×3/6=0

Notdefulte=7/10×3/7×4/4×3/6=0.0735

P(DB=NO | X) is higher, we predict that loan X is Not Defaulted Borrower.

loan X = (Home Owner = Yes, Marital Status=Divorced, Income=Low)

default=3/10×0/3×1/2×3/6=0

notdefault=7/10×3/7×1/2×3/6=0.0184

P(DB=NO | X) is higher, we predict that loan X is Not Defaulted Borrower.