CSE352 AI

HOMEWORK 2 - 10pts

Homework 2 has 3 Problems in 2 PARTS and an extra credit problem.

SOLVE ONLY TWO problems of your choice

I will post Solutions so you could compare your solutions with mine.
PROBLEMS similar to Hmks will appear on your TEST

You must **TYPE the statement of each problem** you are solving; otherwise problem will not be considered for correction.
You can DRAW your TREES by hand.

Here are some DEFINITIONS from the Lecture Notes that YOU NEED for your Homework

Definition 1

Given a classification dataset DB with a set
A = {$a_1, a_2, \ldots, a_n$} of attributes and a **class attribute C** with values {$c_1, c_2, \ldots, c_k$} (k classes),
any expression
$a_1 = v_1 \wedge \ldots \wedge a_k = v_k$, where $a_i \in$ A and v_i are values of attributes is called a **DESCRIPTION**.

In particular, **C = c_k** is called a **CLASS DESCRIPTION**.

Definition 2

A **CHARACTERISTIC FORMULA** is any expression
$C = c_k \Rightarrow a_1 = v_1 \wedge \ldots \wedge a_k = v_k$,
We write it shortly as

CLASS ⇒ DESCRIPTION

Definition 3

A **DISCRIMINANT formula** is any expression
$a_1 = v_1 \wedge \ldots \wedge a_k = v_k \Rightarrow C = c_k$
written shortly as
DESCIPTION ⇒ CLASS

Definition 4

A characteristic formula **CLASS ⇒ DESCRIPTION** is called a **CHARACTERISITIC RULE** of the classification dataset DB iff it is **TRUE** in DB, i.e. when the following holds

{o: DESCRIPTION} ∩ {o: CLASS} not= ∅

where **{o: DESCRIPTION}** is the set of all records of DB corresponding to the description **DESCRIPTION** and
{o: CLASS} is the set of all records of DB corresponding to the description CLASS

Definition 5

A discriminant formula DESCRIPTION ⇒ CLASS is called a **DISCRIMINANT RULE** of DB
iff it is **TRUE in DB,** i.e. the following two conditions hold

1. **{o: DESCRIPTION} not= ∅**

2. **{o: DESCRIPTION} ⊆ {o: CLASS}**

PART ONE:

Classification: Characteristic and Discriminant Rules

Given a dataset:

Record	a_1	a_2	a_3	a_4	C
o_1	1	1	1	0	1
o_2	2	1	2	0	2
o_3	0	0	0	0	0
o_4	0	0	2	1	0
o_5	2	1	1	0	1

C – class attribute

Problem 1

1. Find sets {o :DESCRIPTION} for the following descriptions

Follow the Example below when writing your solutions

Example: for description a1 = 2 a Λ a_2 = 1 you have evaluate the set:

{o : $a_1 = 2 \Lambda a_2 = 1$}= { o_2, o_5 }

1) $a_3 = 1 \Lambda a_4 = 0$

2) $a_2 = 0 \Lambda a_3 = 2$

3) C=1

4) C=0

2. For the following formulas use proper definitions stated above to **determine,** it means use proper definitions to **prove** whether **they are or they are not DISCRIMINANT / CHARACTERISTIC RULES** of our dataset.

Example:

$a_1 = 2\ a \Lambda\ a_2 = 1 \Rightarrow C = 1$
is a DISCRIMANT Formula that is **NOT** DISCRIMINANT RULE because
$\{o : a_1 = 2 \Lambda a_2 = 1\} = \{ \mathbf{o_2}, \mathbf{o_5} \}$, $\{o: C=1\} = \{ \mathbf{o_1}, \mathbf{o_5} \}$
and $\{ \mathbf{o_2}, \mathbf{o_5} \}$ is **NOT a subset** of $\{ \mathbf{o_1}, \mathbf{o_5} \}$

5) $a_1 = 1 \Lambda a_2 = 1 \Rightarrow C = 1$

6) $C = 1 \Rightarrow a_1 = 0 \Lambda a_2 = 1 \Lambda a_3 = 1$

7) $C = 2 \Rightarrow a_1 = 1$

8) $C = 0 \Rightarrow a_1 = 1 \Lambda a_4 = 0$

9) $a_1 = 2 \Lambda a_2 = 1 \Lambda a_3 = 1 \Rightarrow C = 0$

10) $a_1 = 0 \Lambda a_3 = 2 \Rightarrow C = 1$

PART TWO:

Decision Tree Learning 1

Here is the TRAINING DATA SET FOR THE HOMEWORK:
Class Attribute: Buys Computer

Age	Income	Student	Credit Rating	Buys Computer
<=30	high	No	Fair	No
<=30	high	No	Excellent	No
31…40	high	No	Fair	Yes
>40	medium	No	Fair	Yes
>40	Low	Yes	Fair	Yes
>40	low	Yes	Excellent	No
31…40	low	Yes	Excellent	Yes
<=30	medium	No	Fair	No
<=30	low	Yes	Fair	Yes
>40	medium	Yes	Fair	Yes
<=30	medium	Yes	Excellent	Yes
31…40	medium	No	Excellent	Yes
31…40	high	Yes	Fair	Yes
>40	medium	No	Excellent	No

When building your DECISION TREES follow Examples in the Lectures; you must include ALL steps of their constructions not only the final results.

Problem 2

Use the Training Data to create **two decision trees** described as follows

Tree1 : Build the decision tree using **general majority voting heuristic**, defined as follows:
You CAN use MAJORITY Vote for the majority class at anytable at any level of the tree – when you choose so.
Use CREDIT RATING as the **root attribute**, and nodes attributes of your own choice;

YOU MUST use at least 3 attributes as nodes.

2. Write down all the **rules determined by your tree** in the **description** and in the **predicate forms**

3. **EVALUATE predictive accuracy** for the set of your rules with respect to the **TEST Dataset** below

Tree 2: Use **Basic ID3 algorithm**
Use INCOME as **root attribute**, and nodes attributes of your choice;

2. Write down all the **rules determined by your tree** in the **description** and **predicate forms**
3. **Evaluate correctness of your rules, i.e. the predictive accuracy with respect to the TRAINING data**
4. **Evaluate predictive accuracy** for the set of your rules with respect to the **TEST Dataset** below.
Must show work.

TEST DATA SET

Obj	Age	Income	Student	Credit_Rating	Class
1	<=30	High	Yes	Fair	Yes
2	31…40	Low	No	Fair	Yes

3	31…40	High	Yes	Excellent	No
4	>40	Low	Yes	Fair	Yes
5	>40	Low	Yes	Excellent	No
6	<=30	Low	No	Fair	No

Problem 3

Create test data set of at least 6 records for your sets of rules rules corresponding to **Tree 1** or **Tree 2** that **guarantees 100% predictive accuracy.**
Prove that your example is correct.

Extra Credit - 5pts

EVALUATE Information Gain for 2 attributes on one NODE of your choice of your tree **Tree 1 or 2**

You must show work, not a final number; in fact you can write proper formulas for its computation without evaluating (calculator) the numbers.
I want to SEE if you understand the formulas