

**CSE 590/634 DATA MINING
MIDTERM FALL 2009
(100pts + 15extra)**

NAME: _____

ID: _____

PART ONE: Classification Data and Rules (30pts)

Definition 1

Given a classification dataset DB with a set $A = \{a_1, a_2, \dots, a_n\}$ of attributes and a class attribute C with values $\{c_1, c_2, \dots, c_k\}$ (k classes),
any expression;

$a_1 = v_1 \wedge \dots \wedge a_k = v_k$, where $a_i \in A \cup C$ 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
 $C = c_k \Rightarrow a_1 = v_1 \wedge \dots \wedge a_k = v_k$, or shortly we write is as
CLASS \Rightarrow DESCRIPTION

Definition 3

A DETERMINANT formula is any expression

$a_1 = v_1 \wedge \dots \wedge a_k = v_k \Rightarrow C = c_k$, or shortly
DESCRIPTION \Rightarrow CLASS

Definition 4

A characteristic formula CLASS \Rightarrow DESCRIPTION is called a CHARACTERISITIC RULE of the classification dataset DB iff it is **TRUE** in DB, i.e. when the following holds
 $\{o: \text{DESCRIPTION}\} \cap \{o: \text{CLASS}\} \neq \emptyset$,
where $\{o: \text{DESCRIPTION}\}$ is the set of all records of DB corresponding to the description DESCRIPTION, $\{o: \text{CLASS}\}$ is the set of all records of DB corresponding to the description CLASS.

Definition 5

A discriminant formula $\text{DESCRIPTION} \Rightarrow \text{CLASS}$ is called a **DISCRIMINANT RULE** of DB iff it is **TRUE in DB**, i.e. the following holds

1. $\{o: \text{DESCRIPTION}\} \neq \emptyset$
2. $\{o: \text{DESCRIPTION}\} \subseteq \{o: \text{CLASS}\}$

Given a dataset:

Record	a_1	a_2	a_3	a_4	C
o₁	1	1	1	0	1
o₂	2	1	2	0	2
o₃	0	0	0	0	0
o₄	0	0	2	1	0
o₅	2	1	1	0	1

C – class attribute

Q1. (5pts) Find $\{o: \text{DESCRIPTION}\}$ for the following descriptions

1) $a_1 = 2 \wedge a_2 = 1$

2) $a_3 = 1 \wedge a_4 = 0$

3) $a_2 = 0 \wedge a_3 = 2$

4) $c=1$

5) $c=0$

Q2. (5pts) For the following formulae use proper definitions to determine (it means prove) whether they are / are not **DISCRIMINANT / CHARACTERISTIC RULES** of our dataset.

6) $a_1 = 1 \wedge a_2 = 1 \Rightarrow C = 1$

7) $C = 1 \Rightarrow a_1 = 0 \wedge a_2 = 1 \wedge a_3 = 1$

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

9) $C = 0 \Rightarrow a_1 = 1 \wedge a_4 = 0$

10) $a_1 = 2 \wedge a_2 = 1 \wedge a_3 = 1 \Rightarrow C = 0$

11) $a_1 = 0 \wedge a_3 = 2 \Rightarrow C = 1$

Q3. (10pts)

1. Prove that for any classification data base DB, for any of its DISCRIMINANT rules
DESCRIPTION \Rightarrow CLASS, the formula CLASS \Rightarrow DESCRIPTION is a characteristic
RULE of the DB.

2. Prove that the inverse statement to Q3 is not true, i.e show that there is a classification data base DB with a characteristic rule $\text{CLASS} \Rightarrow \text{DESCRIPTION}$, such that the formula $\text{DESCRIPTION} \Rightarrow \text{CLASS}$ is not a discriminant rule of that DB.

Q3. (10pts)

Prove that our definition of discriminant rule is correct; i.e. that the discriminant rules as we defined do *discriminate records belonging to a given class from the rest of classes*.

PART TWO: Classification by Decision Tree Induction (20pts) and Classification by Association (20pts)

Given a TRAINING and TEST classification data as follows:

TRAIN

Record	a_1	a_2	C
o₁	1	1	1
o₂	0	0	0
o₃	0	1	0
o₄	0	0	0
o₅	1	1	1
o₆	1	1	0
o₇	0	0	0
o₈	1	0	1

TEST

Record	a_1	a_2	C
o₁	1	1	1
o₂	1	0	0
o₃	0	0	1
o₄	0	0	0

Q1. (5pts)

1. Construct a Decision Tree with \mathbf{a}_1 as the ROOT.

Q2. (5pts)

Write down all the discriminant rules determined by your tree using a PREDICATE form.

Q3. (5pts)

Use the TEST data to evaluate predictive accuracy of your set of rules generated by the given TRAIN data.

Q3. (20pts)

Use TRAIN data to find the set of classification rules by Apriori Algorithm. Test the rules with the TEST Data. Compare the results with Decision Tree results.

Q3 Solution space

PART THREE: BUILDING a CLASSIFIER (30pts)

For the data set given below build a classifier following all steps needed in the constructions: preprocessing, training, and testing.

Describe and motivate your choice of algorithms and methods used at each step.

CLASSIFICATION DATA:

Age	Income	Student	Credit Rating	Buys Computer
21	60,000	yes	3	No
30	70,000	No	5	No
38	65,000	No	2	Yes
45	45,000	yes	3	Yes
46	25,000	no	2	Yes
47	30,000	Yes	6	No
39	28,000	Yes	5	No
29	48,000	Yes	3	No
50	75,000	Yes	2	No
48	41,000	Yes	3	No
30	37,000	Yes	6	Yes
51	46,000	No	4	Yes
32	80,000	Yes	2	No
45	50,000	No	4	No

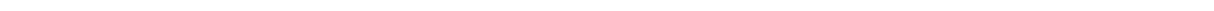
1. Data after Preprocessing (5pts)

Age	Income	Student	Credit Rating	Buys Computer
		yes		No
		No		No
		No		Yes
		yes		Yes
		no		Yes
		Yes		No
		Yes		No
		Yes		No
		Yes		No
		Yes		No
		Yes		Yes
		No		Yes
		Yes		No
		No		No

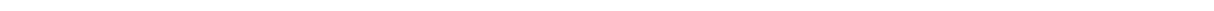
2. Training: data and method (15pts)



3. Testing: Data and method (5pts)



4. MY CLASSIFIER (5pts)



PART FOUR: Neural Networks (15extra points)

Given two records (Training Sample)

a_1	a_2	a_3	Class
0.5	0	0.2	1
0	0.3	0.2	1
0.2	0.1	0	0

1. Construct a Neural Network with one hidden layer (your own topology) and evaluate a passage of ONE EPOCHS.

Use Learning Rate $\ell = 0.7$

2. Write you're the terminating conditions for your network
3. Write a condition for success; i.e. how you decide that the record is well classified.

Extra page