Cse537
Artifficial Intelligence
Short Review 1
for Midterm 2

Professor Anita Wasilewska
Computer Science Department
Stony Brook University

Data Mining Process

- Questions:
- Describe and discuss all stages of the Data Mining Process
- Describe the role of Preprocessing stage and its main methods
- Discuss the Data Mining Proper stage
- Describe what is Descriptive/ non Descriptive Data Mining
- Which Models you would use for the Descriptive Data Mining and which for the non Descriptive Data Mining
- How and what decides which type of Data Mining is the best to use (implement)
- Give examples of types of applications and the best Models (algorithms) for them

Classification

- Describe what is CLASSIFICATION; type of data, goals and applications
- Describe all stages of the classification process
- Describe and discuss basic classification Models and their differences
- Discuss the Decision Tree Induction and its strengths and weaknesses
- Discuss the Neural Network Model and its strengths and weaknesses
- Define a CLASSIFIER
- Describe a process of building a CLASSIFIER

Classification Data and Rules

Given a classification dataset **DB** with a set

A = {*a1, a2,…, an*} of **attributes** and a **class** attribute **C** with values

{*c1, c2,…, ck*} - **k classes**

Definition 1

Any expression ***a1 = v1 & … & ak = vk*** where ***ai*** ∈ **A** and ***vi*** are corresponding values of attributes from **A**

is called a **DESCRIPTION**

Any expression **C = *ci*** is for ***ci*** ∈ **{*c1, c2,…, ck*}**

Is called a **CLASS DESCRIPTION**

Classification Data and Rules

Definition 2

A **CHARACTERISTIC FORMULA** is any expression

$$C = ck \Rightarrow a1 = v1 \ \& \ldots \& \ ak = vk$$

We write is as

CLASS ⇒ DESCRIPTION

Definition 3

A **DETERMINANT FORMULA** is any expression

$$a1 = v1 \ \Lambda \ldots \Lambda \ ak = vk \Rightarrow C = ck$$

We write it as

DESCIPTION ⇒ CLASS

Classification Data and Rules

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**

{o: CLASS} is the set of all records of DB corresponding to the **CLASS**

Classification Data and Rules

Definition 5

A discriminant formula

DESCRIPTION ⇒ CLASS

is called a **DISCRIMINANT RULE** of **DB**

iff

it is **TRUE in DB,** i.e. the following conditions hold

1. **{o: DESCRIPTION} not= ∅**

2. **{o: DESCRIPTION} ⊆ {o: CLASS}**

PROBLEM 1

Prove

that for any **classification** data base **DB**

and any of its **DISCRIMINANT RULES** of the form

DESCRIPTION ⇒ CLASS

the formula

CLASS ⇒ DESCRIPTION

is a **CHARACTERISTIC RULE** of the **DB**

PROBLEM 1 Solution

By **definition 5**, for any database DB :

DESCRIPTION ⇒ CLASS

is a **DISCRIMINANT RULE** **iff**

1. **{o: DESCRIPTION} not= ∅**

2. **{o: DESCRIPTION} ⊆ {o: CLASS}**

Therefore,

{o: DESCRIPTION} ∩ {o: CLASS} not= ∅

and by **Definition 4**

CLASS ⇒ DESCRIPTION

Is the **CHARACTERISITIC RULE**

PROBLEM 2

Given a dataset:

Record	A1	A2	A3	A4	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

Find the set **{o :DESCRIPTION}**
for the following descriptions

1) a1 = 2 & a2 = 1
2) a3 = 1 & a4 = 0
3) a2 = 0 & a3 = 2
4) c=1
5) c=0

PROBLEM 2 SOLUTION

Find the set **{o :DESCRIPTION}**

for the following descriptions

1) a1 = 2 & a2 = 1 Answer : {o1 }
2) a3 = 1 & a4 = 0 Answer : {o1 , o5}
3) a2 = 0 & a3 = 2 Answer : {o4}
4) c=1 Answer : {o1,o5}
5) c=0 Answer : {o3 ,o5}

PROBLEM 3

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

6) $a1 = 1 \ \& \ a2 = 1 \Rightarrow C = 1$

7) $C = 1 \Rightarrow a1 = 0 \ \& \ a2 = 1 \ \& \ a3 = 1$

8) $C = 2 \Rightarrow a1 = 1$

9) $C = 0 \Rightarrow a1 = 1 \ \& \ a4 = 0$

10) $a1 = 2 \ \& \ a2 = 1 \ \& \ a3 = 1 \Rightarrow C = 0$

11) $a1 = 0 \ \& \ a3 = 2 \Rightarrow C = 1$

PROBLEM 3 SOLUTION

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

6) $a1 = 1 \ \& \ a2 = 1 \Rightarrow C = 1$

{o1} is a subset of {o1 , o5} so **this is a DISCRIMINANT** rule

7) $C = 1 \Rightarrow a1 = 0 \ \& \ a2 = 1 \ \& \ a3 = 1$

{o: $a1 = 0 \ \& \ a2 = 1 \ \& \ a3 = 1$ } is an empty set so this is **not** a CHARACTERISTIC rule

8) $C = 2 \Rightarrow a1 = 1$

As the intersection is empty so this is **not** a **CHARACTERISTIC** rule

9) $C = 0 \Rightarrow a1 = 1 \ \& \ a4 = 0$ ----- {o3 , o4} /\ {o5} is empty set so this is **not a CHARACTERISTIC** rule

10) $a1 = 2 \ \& \ a2 = 1 \ \& \ a3 = 1 \Rightarrow C = 0$ ----- {o5} is not a subset of {o3 , o4} , so this is **not a DISCRIMINANT** rule

11) $a1 = 0 \ \& \ a3 = 2 \Rightarrow C = 1$ ----- {o4} is not a subset of {o1 , o5} , so this is **not a DISCRIMINANT** rule

Classification

- Describe what is Classification; which is the goal, what data one needs etc….
- Describe all stages of the Classification Process
- Describe basic methods of training and testing
- Describe the process of building a CLASSIFIER
- What is a CLASSIFIER?

Problem: Classification by Association

1. Use TRAIN data to find the set of classification rules using the Apriori Algorithm

   2. **Test** the rules with the TEST Data

   Use 2 different testing Method of your choice and compare the results

TRAINING DATA

Record	A1	A2	C
1	1	1	1
2	0	0	0
3	0	1	0
4	0	0	0
5	1	1	1
6	1	1	0
7	0	0	0
8	1	0	1

Transactional Data and Support calculations

	I1 (A1 =0)	I2(A1 = 1)	I3(A2 = 0)	I4(A2= 1)	I5(C=0)	I6(C=1)
1		+		+		+
2	+		+		+	
3	+			+	+	
4	+		+		+	
5		+		+		+
6		+		+	+	
7	+		+		+	
8		+	+			+
Count	4	4	4	4	5	3

Let the minimum support count = 3

L1:

Item set	Support Count
I1	4
I2	4
I3	4
I4	4
I5	5
I6	3

Candidate two item sets :

Item Set	Support Count
1,2	0
1,3	3
1,4	1
1,5	4
1,6	0
2,3	1
2,4	3
2,5	1
2,6	0
3,4	3
3,5	1
3,6	2
4,5	2
4,6	0

Classification by Association

Frequent 2 item set :

Item Set	Support Count
1,3	3
1,5	4
2,4	3
2,6	3
3,5	3

Classification by Association

Candidate 3 item set :

Item Set	Support Count
1,3,5	3
2,4,6	1

Classification by Association

Frequent 3 item Set :

Item set	Support Count
1,3,5	3

L = {(1,5),(2,6),(3,5),(1,3,5)}

This is the set used to find the classification rules by association

Don't forget to FIX and calculate Confidence and Support!

Testing :

Record	A1	A2	Expected class	Actual class	Correctly classified
1	1	1	1	1	Yes
2	1	0	0	?	No
3	0	0	1	0	No
4	1	0	0	0	Yes

Predictive accuracy = 3/4 * 100 = 75 %

PROBLEM:: BUILDING a CLASSIFIER

For a given data set **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.

Problem: Neural Networks

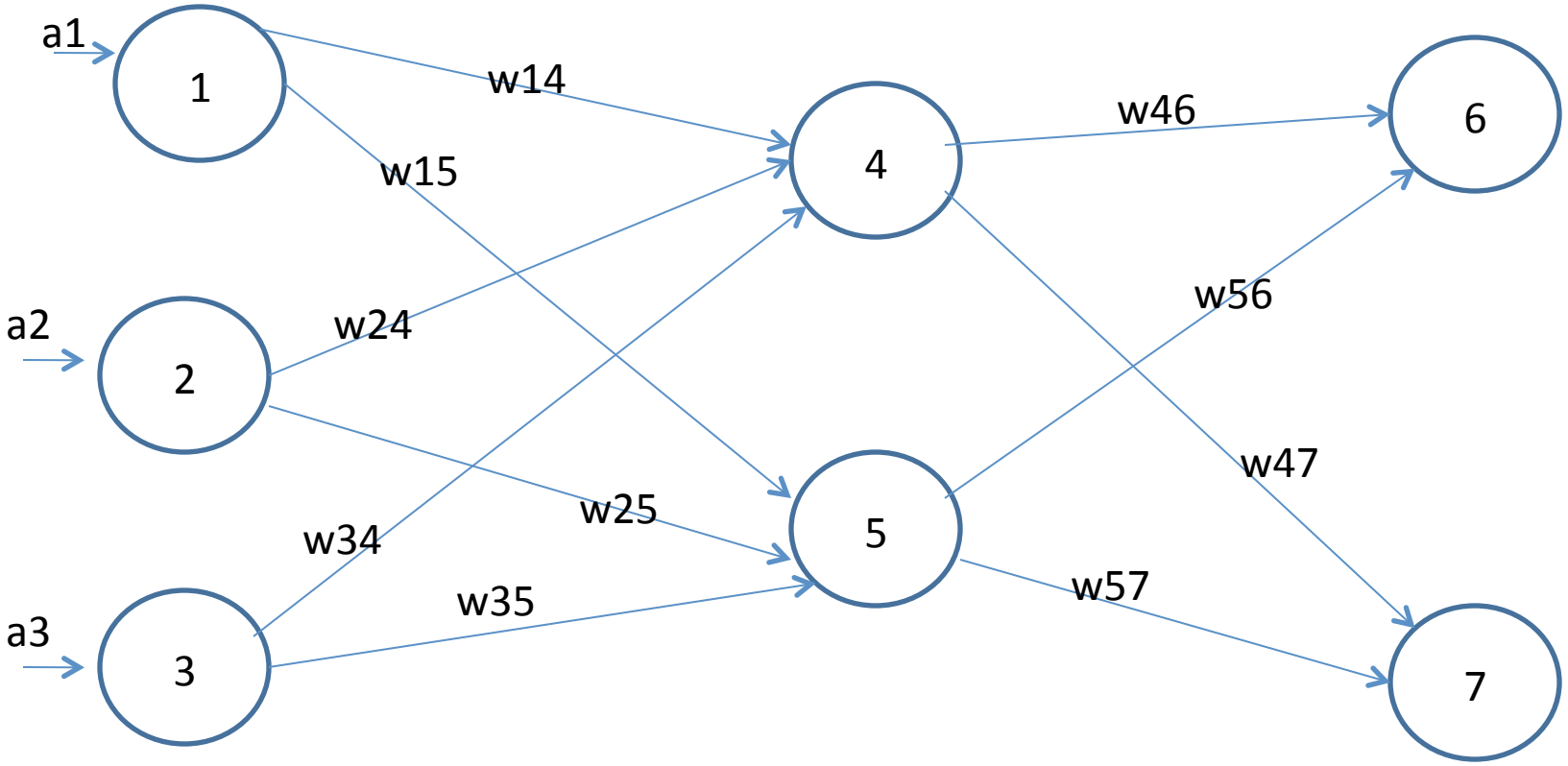
Given two records (Training Sample)

A1	A2	A3	Class
0.5	0	0.2	1
0	0.3	0.2	1
0.2	0.1	0	0

Construct a Neural Network with your own 2 different topologies and evaluate- **describe** a passage of ONE EPOCHS (use learning rate l = 0.7). Backpropagation formulas will be given

Topology :

Input = 3 , hidden = 2 and output = 2.



Problem: Neural Networks

For the **first iteration** we take the following values as input :

a1 = 0.5 , a2 = 0 , a3 = 0.2

w14 = 0.2 , w15 = -0.3 , w24 = 0.4 , w25 = 0.1

w34 = 0.2 , w35 = -0.3 , w46 = 0.4 , w56 = 0.1

w47 = 0.1 , w57 = 0.2

We take any random values for weights and BIASES