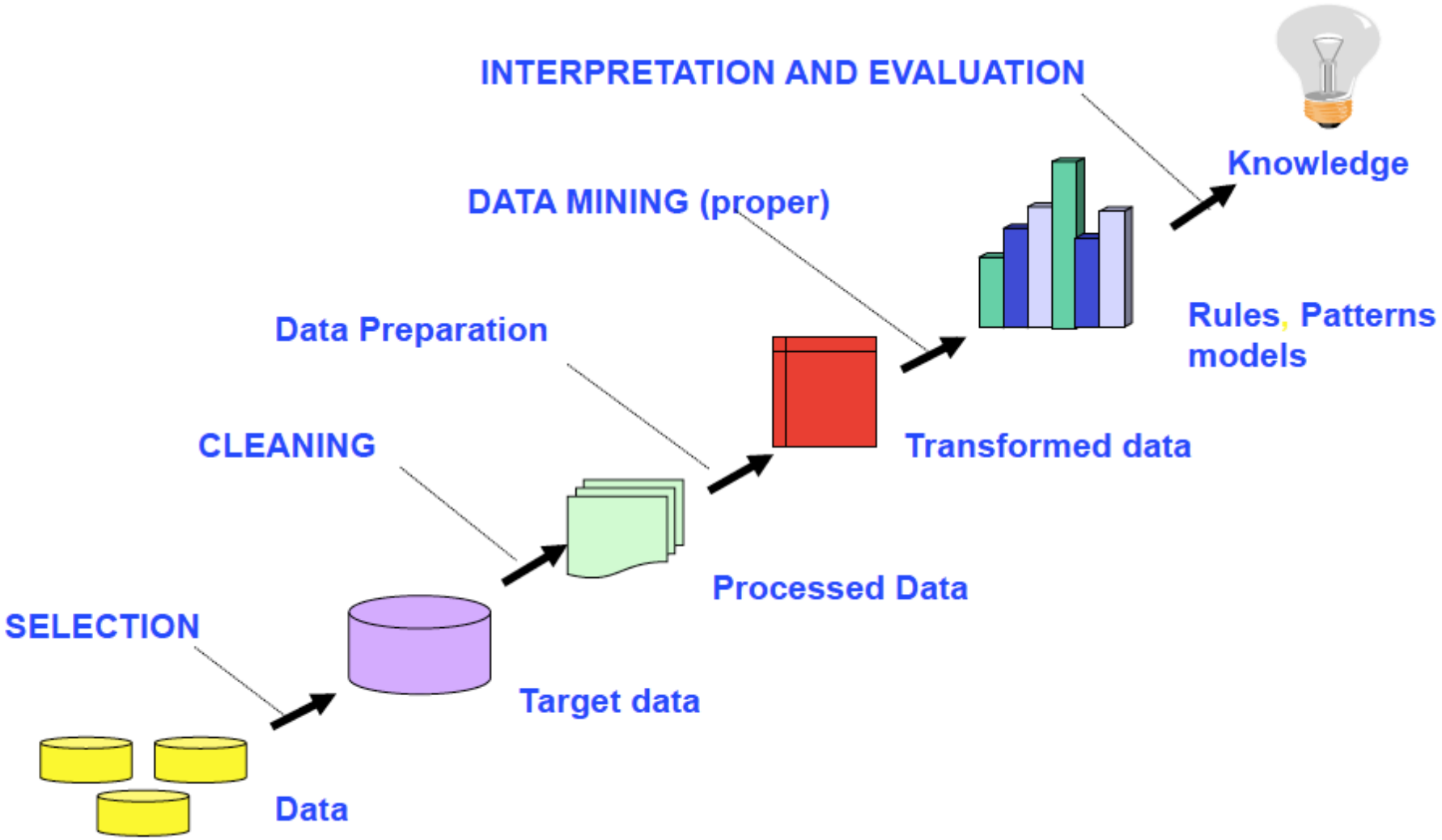
Cse634

DATA MINING

TEST REVIEW

Professor Anita Wasilewska
Computer Science Department
Stony Brook University

Data Mining Process



Preprocessing stage

- **Preprocessing:**
- includes all the operations that have to be performed before a data mining algorithm is applied

- Data in the real world is dirty: incomplete, noisy and inconsistent.
- Quality decisions must be based on quality Data.

Preprocessing stage

- Data cleaning
- – Fill in missing values, smooth noisy data(binning, clustering, regression), identify or remove outliers, and resolve inconsistencies
- Data integration
- – Integration of multiple databases, data cubes, or files

Preprocessing stage

- **Data transformation**

- Normalization and aggregation
- Data reduction and attribute selection
- Obtains reduced presentation in volume but produces the same or similar analytical results (stratified sampling, PCA, cluster)

- Data discretization
- Part of data reduction but reduces the number of values of the attributes by dividing the range of attributes into intervals (segmentation by natural partition, hierarchy generation)
-

DM Proper

- **DM proper** is a step in the **DM process** in which algorithms are applied to obtain patterns in data.

- It can be re-iterated- and usually is

Descriptive / non descriptive data mining and models

- Statistical - descriptive
- Statistical data mining uses historical data to predict some unknown or missing numerical values
- Descriptive data mining aims to find patterns in the data that provide some information about what the data contains
- often presents the knowledge as a set of rules of the form IF.... THEN…

Models

- **Discriptive:** Decision Trees, Rough Sets, Classification by Association
- **Statistical:** Neural Networks, Bayesian Networks, Cluster, Outlier analysis, Trend and evolution analysis
- **Optimization method:** Genetic Algorithms – but it can also be descriptive

Classification

- **Classification:**
- Finding models (rules) that describe (characterize) or/ and distinguish (discriminate) classes or concepts for future prediction
- **Classification Data Format**:
- a data table with key attribute removed.
- Special attribute, called a class attribute must be distinguished.
- The values: c1, c2, ...cn of the class atrribute C are called **class labels**
- The class label attributes are discrete valued and unordered.

Classification

- **Goal:**
- FIND a minimal set of characteristic and/or discriminant rules, or **other descriptions** of the class C, or all, or some other classes

- We also want the found rules to involve as few attributes as it is possible

Classification

- Stage 1: build the basic patterns structure- **training**
- Stage 2: optimize parameter settings; can use (N:N) re-substitution- **parameter tuning**
- Re-substitution error rate = training data error rate
- Stage 3: use **test data** to compute- predictive accuracy/error rate - **testing**

Decision Tree

- DECISION TREE
- A flow-chart-like tree structure；
- Internal node denotes an attribute;
- Branch represents the values of the node attribute;
- Leaf nodes represent class labels

DT Basic Algorithm

- The **basic DT algorithm** for decision tree construction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner

- Tree STARTS as a **single node** representing all training dataset (data table with records called samples)
- **IF** the **samples** (records in the data table) are all in the **same class**, **THEN** the node becomes a **leaf** and is labeled with that class
- The algorithm uses the same process recursively to form a **decision tree** at each partition

DT Basic Algorithm

- The recursive partitioning **STOPS** only when any one of the following conditions is TRUE
- **1.** All records (samples) for the given node belong to the same class
- **2.** There are no remaining attributes on which the samples (records in the data table) may be further partitioned – a **LEAF** is created with **majority vote** for training sample
- **3.** There is no records (samples) left – a **LEAF** is created with **majority vote** for training sample

- **Majority voting** involves converting node N into a leaf and labeling it with the most common class in **D** which is a set of training tuples and their associated class labels

Attribute Selection Measures

- **Some Heuristics:**
- Decision Tree: some Attribute Selection Measures are
- **Information Gain, Gini Index**

- **We use them for selecting the attribute that "best" discriminates the given tuples according to class**

Neural Networks

- Neural Network is a set of connected INPUT/OUTPUT UNITS, where each connection has a WEIGHT associated with it
- Neural Network **learns** by adjusting the weights so as to be able to correctly classify the training data and hence, after testing phase, to **classify unknown data**
- Neural Network needs long time for training Determining network topology is difficult
- Choosing single **learning rate** impossible (train with subset)
- Neural Network has a **high tolerance** to noisy and incomplete data
- NN is generally better with larger number of hidden units

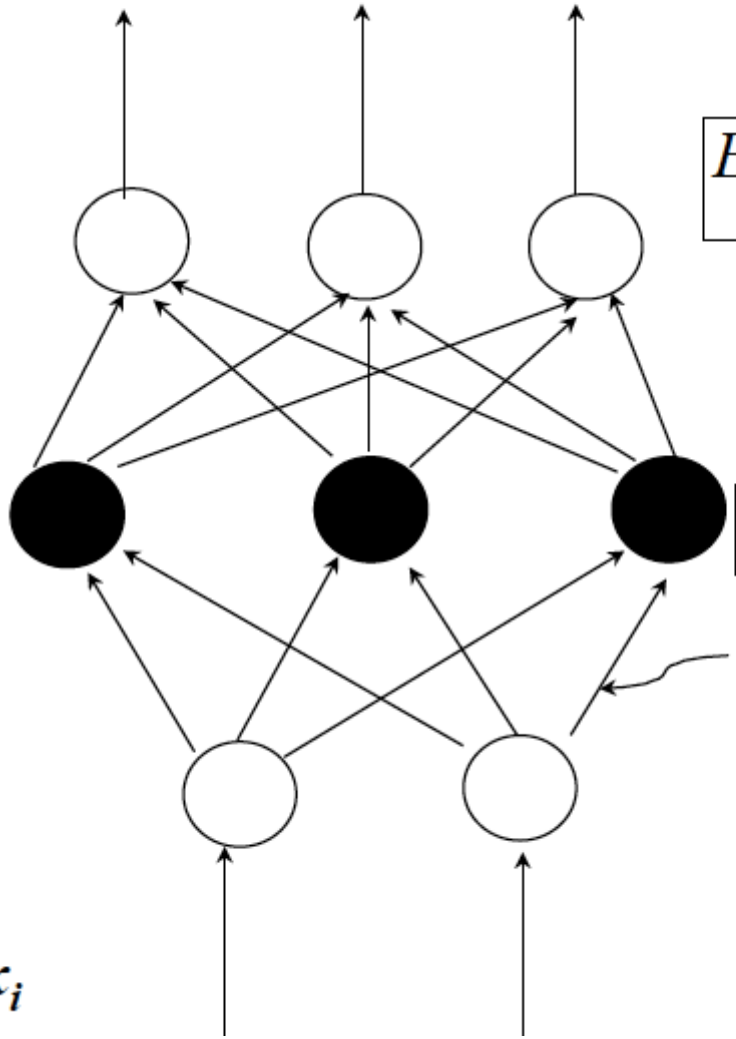
Neural Networks

- The inputs to the network correspond to the attributes and their values for each training tuple
- Inputs are fed simultaneously into the units making up the input layer
- Inputs are then weighted and fed simultaneously to a hidden layer
- The number of hidden layers is arbitrary, although often only one or two
- The weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network's prediction

Neural Networks

- For each training sample, the weights are first set random then they are modified as to minimize the mean squared error between the network's classification (prediction) and actual classification
- **Backpropagation Algorithm:**
- STEP ONE: initialize the weights and biases
- STEP TWO: feed the training sample
- STEP THREE: propagate the inputs forward
- STEP FOUR: backpropagate the error
- STEP FIVE: backpropagate the weights
- STEP SIX: repeat and apply terminating Conditions

Backpropagation Formulas



$$Err_j = O_j(1-O_j)\sum_k Err_k w_{jk}$$

$$\theta_j = \theta_j + (l)Err_j$$

$$w_{ij} = w_{ij} + (l)Err_j O_i$$

$$Err_j = O_j(1-O_j)(T_j - O_j)$$

$$O_j = \frac{1}{1+e^{-I_j}}$$

$$I_j = \sum_i w_{ij} O_i + \theta_j$$

Backpropagation

- **Terminating Conditions**:
- **Process Stops** when:
- All wij in the previous epoch are below some threshold
- The percentage of samples misclassified in the previous epoch is below some threshold
- a pre- specified number of epochs has expired

Building a classifier

- **Building a classifier** consists of two phases: training and testing.
- We use the training data set to **create patterns**: rules, trees, or to train a Neural or Bayesian network
- We evaluate created patterns with the use of **test data**
- We terminate the process
- if it has been **trained** and **tested** and the **predictive accuracy** is on an acceptable level.

- **Classifier** is a **final product** of this process ready to be used to **classify records** with unknown class attribute values

- PREDICTIVE ACCURACY of a classifier is a percentage of well classified data in the test data set.

Training and Testing

- The main methods of predictive accuracy evaluations are:
- Re-substitution (N ; N)
- Holdout (2N/3 ; N/3)
- k-fold cross-validation (N- N/k ; N/k)
- Leave-one-out (N-1 ; 1)

Association Analysis

- Finding frequent patterns called **associations**, among sets of items or objects in transaction databases, relational databases, and other information repositories
- **Confidence:**
- The rule X->Y holds in the database D with confidence c if the c% of the transactions in D that contain X also contain Y
- Support:
- The rule X->Y has support s in D if s% of the transaction in D contain XUY
- We (user) fix MIN support usually low and Confidence high

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

Association and Genetic Algorithms

- Describe the Apriori Algorithm and Association Analysis
- Discuss types of Association Analysis applications
- Describe classification by Association and compare it with the classification by or Neural Network
- Discuss types of Classification by Association applications

Association and Genetic Algorithms

- Describe principles of Genetic Algorithms
- Give examples of chromosomes encoding
- Describe GA operators and parameters
- Describe the role of fitness function
- Describe GA Reproduction Cycle
- Discuss types of GA applications
- Compare classification by GA with NN and DT classifications

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

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

TRAIN 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	Test Data Class	Rules assigned 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 = 2/4 * 100 = 50 %

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

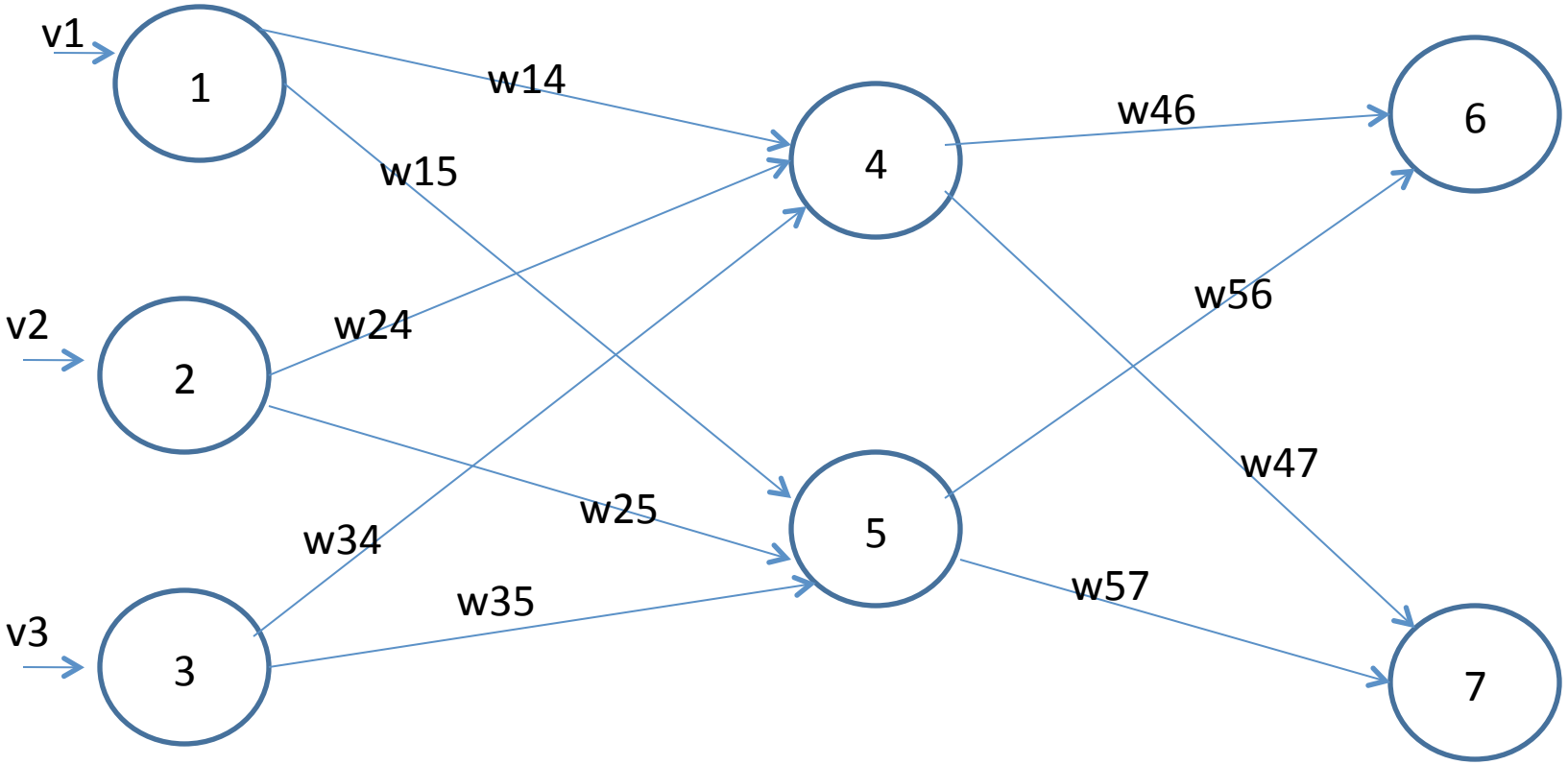
Given two records (Training Sample)

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). IF I ask you for that- the 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

GENERAL: We take any random values for weights and BIASES, and fix the learning rate