Cse534
Data Mining

Test Review 1

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

Data Mining Process

- **Questions:**
- 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

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

We write is as

CLASS ⇒ DESCRIPTION

Definition 3

A **DETERMINANT FORMULA** is any expression

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

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:: 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.

Algorithm for Decision Tree Induction

book slide

- Basic algorithm (a greedy algorithm)
  - Tree is constructed in a top-down recursive divide-and-conquer manner
  - At start, all the training examples are at the root
  - Attributes are categorical (if continuous-valued, they are discretized in advance)
  - Examples are partitioned recursively based on selected attributes
  - Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)
- Conditions for stopping partitioning
  - All samples for a given node belong to the same class
  - There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf
  - There are no samples left

Classification—A Two-Step Process

Book slide

- **Model construction:** describing a set of predetermined classes
  - Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
  - The set of tuples used for model construction is training set
  - The model is represented as classification rules, decision trees, or mathematical formulae
- **Model usage:** for classifying future or unknown objects
  - Estimate accuracy of the model
    - The known label of test sample is compared with the classified result from the model
    - Accuracy rate is the percentage of test set samples that are correctly classified by the model
    - Test set is independent of training set (otherwise overfitting)
  - If the accuracy is acceptable, use the model to classify new data
- Note: If *the test set* is used to select models, it is called validation (test) set

Comparing Attribute Selection Measures book slide

- The three measures, in general, return good results but
  - **Information gain**:
    - biased towards multivalued attributes
  - **Gain ratio**:
    - tends to prefer unbalanced splits in which one partition is much smaller than the others
  - **Gini index**:
    - biased to multivalued attributes
    - has difficulty when # of classes is large
    - tends to favor tests that result in equal-sized partitions and purity in both partitions

Problem: Neural Networks

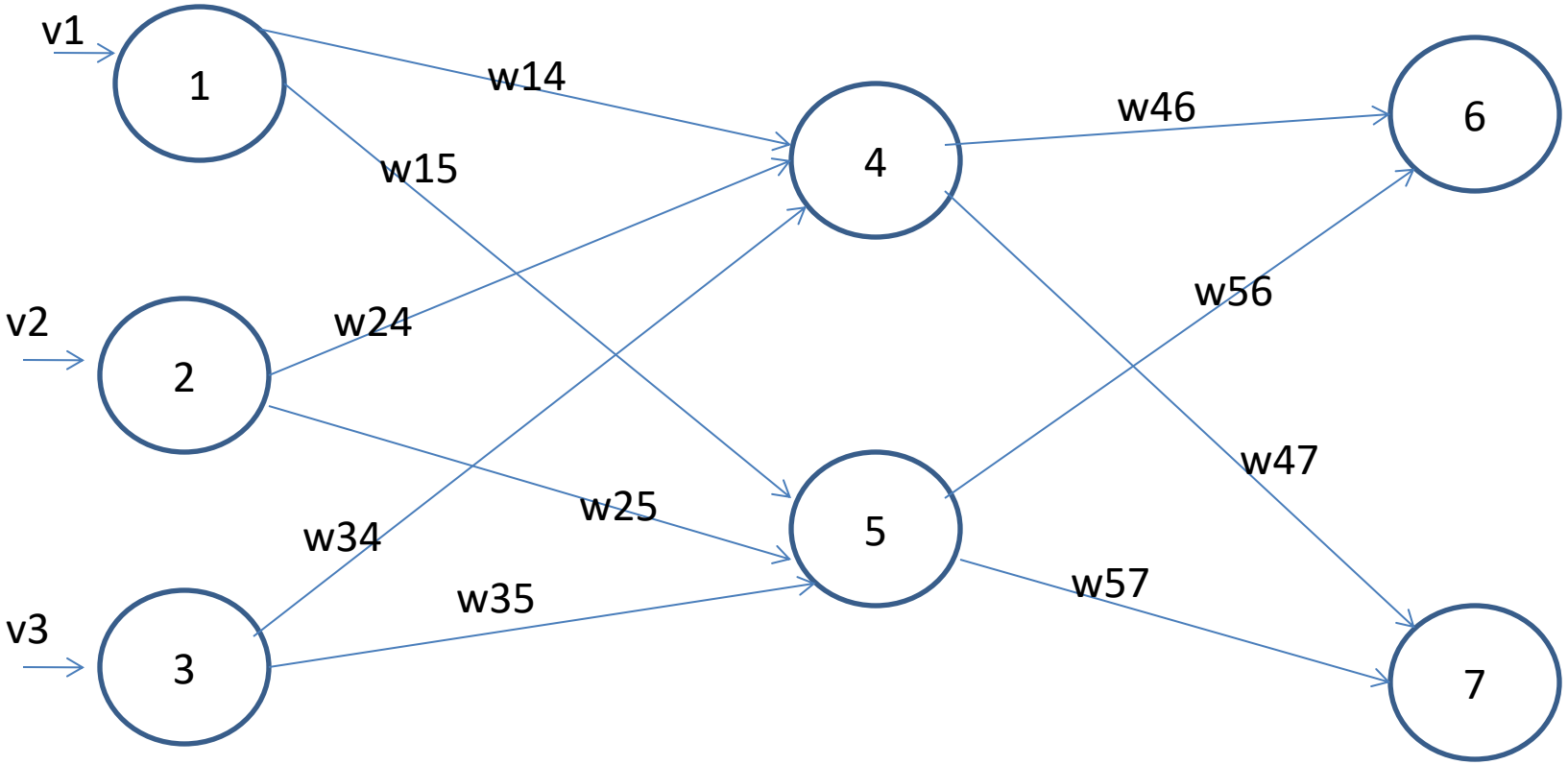
Given two records (Training Sample)

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 EPOCH (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

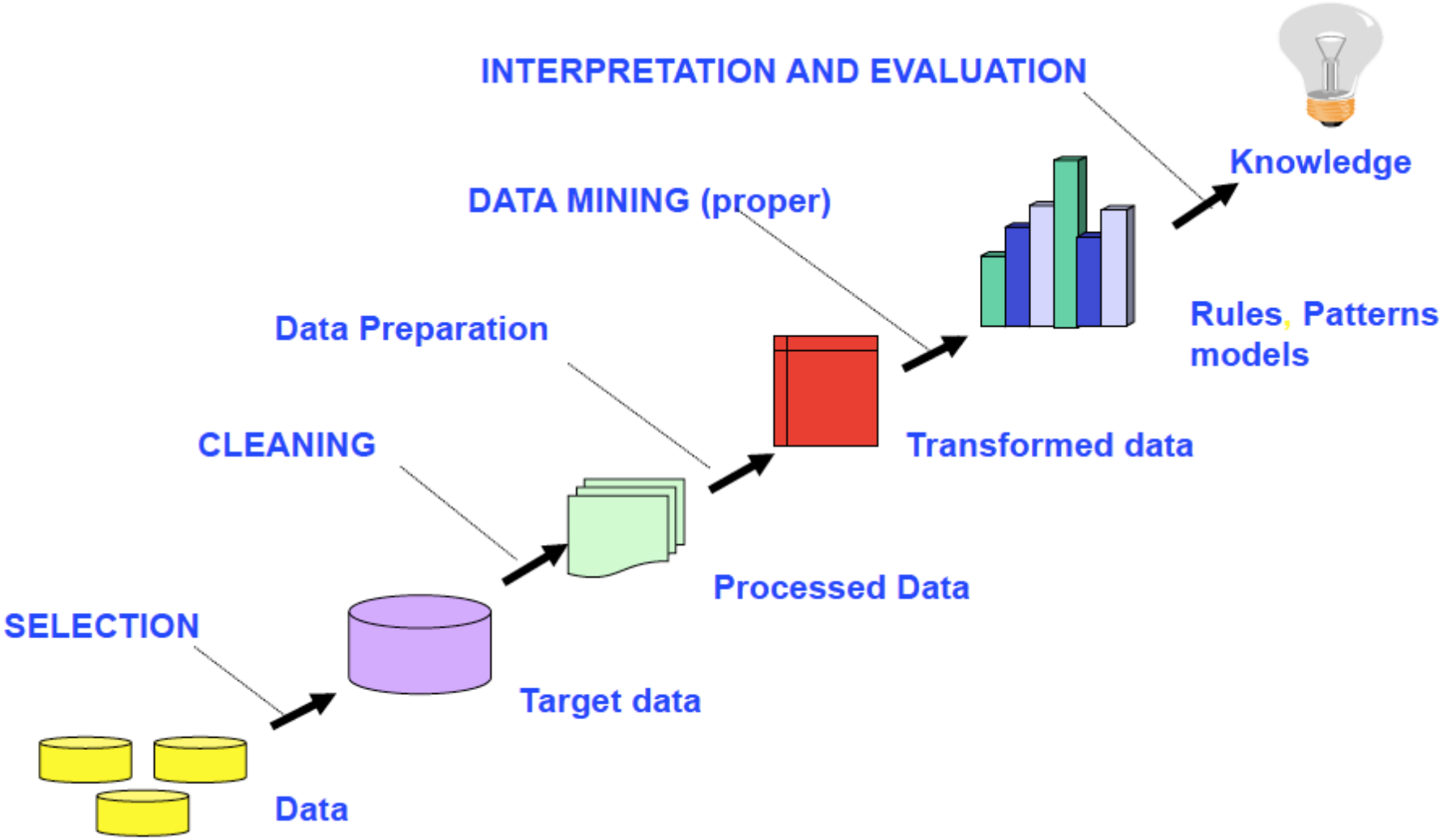
w14 = 0.2 , w15 = -0.3 , w24 = 0.4 , w25 = 0.1

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

Data Mining Process



Data Mining Process

- **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

- **Data transformation**
- Normalization and aggregation

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.

Data Mining Process

- **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 binning, attributes values clustering
-

Data Mining Process

- **Data mining 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 Models

- **Statistical** and **Descriptive**
- **Statistical models** use data to predict some unknown or missing numerical values
- **Descriptive models** aim to find patterns in the data that provide some information about what the data contains
- In case of Classification they often present the knowledge as a set of **rules** of the form IF.... THEN…

Descriptive/Non-Descriptive Models

- **Descriptive:**
- Decision Trees, Rough Sets, Classification by Association, Genetic Algorithms

- **Statistical:**
- Neural Networks, Bayesian Networks, Cluster, Outlier analysis, Trend and evolution analysis
- **Optimization method:**
- Genetic Algorithms

Which type of Data Mining to use

- Different Data Mining methods are required for different kind of data
  and different kinds of goals

- **Application** and **algorithms** for the
  Business Advantages
- Data Mining uses gathered data to predict tendencies and waves,
- classify new data,
- Find previously unknown patterns for the use for business advantages,
- Discover unknown relationships

Market Analysis and Management

- **Target marketing**
- – DM finds clusters of “model” customers who share the same characteristics: interest,
- income level, spending habits, etc. Determine customer purchasing patterns over time

- **Customer profiling**
- – data mining can tell you what types of customers buy what products (clustering or
- classification)

- and OTHERS

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

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

- In case od descriptive DM we also want the found rules to involve as few attributes as it is possible (minimal length of the rules)

Classification Process

- **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
- **Stage 4:** build the **classifier**

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 of building a classifier
- if it has been **trained** and **tested** and the **predictive accuracy** is on an acceptable level.
- CLASSIFIER is a **final product** of the process.
-
- PREDICTIVE ACCURACY of a classifier is a percentage of well classified data in the test data set.

Training and Testing

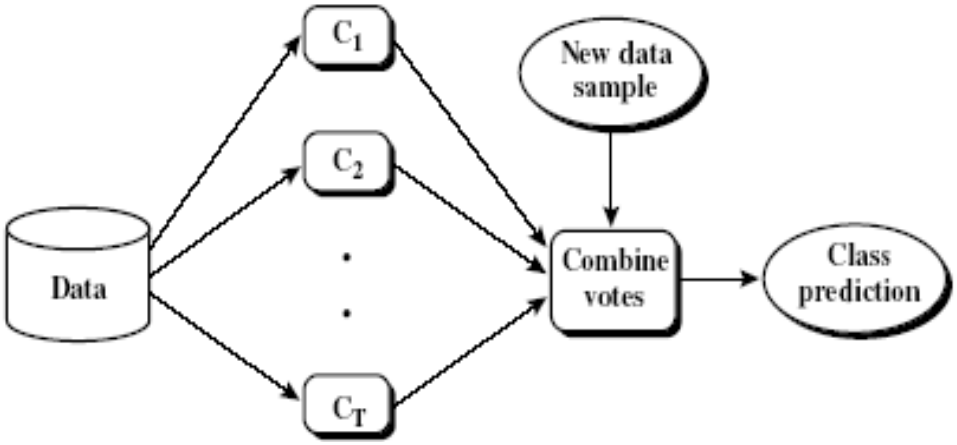
- The **main methods** of predictive accuracy evaluations are:
- Re-substitution (N ; N) – for parameters tuning
- Holdout (2N/3 ; N/3)
- k-fold cross-validation (N- N/k ; N/k)
- Leave-one-out (N-1 ; 1)

Metrics for Evaluating Classifier Performance

- The predictive accuracy is one of basic performance measures of a classifier (model) learned in Stages 1-3 when applied to predict the class label of unknown records

- You must be able to list and shortly DESCRIBE other metrics
- Lecture 4-testing

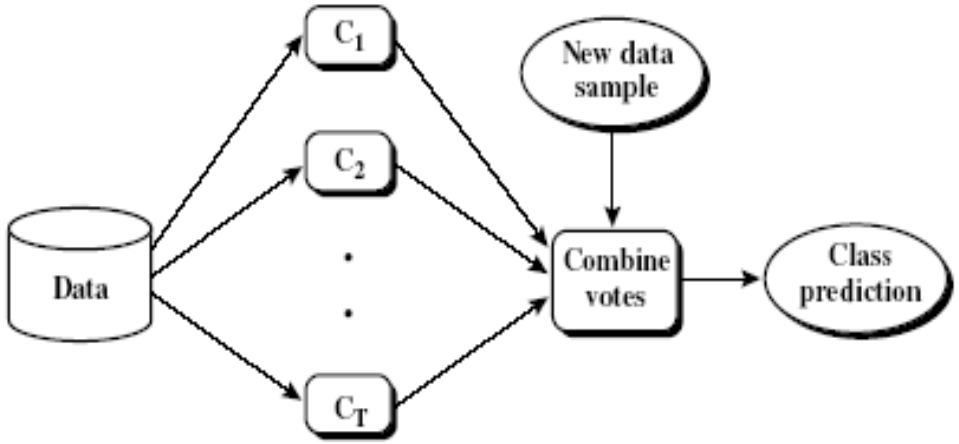
Ensemble Methods



- Ensemble methods
  - Use a combination of models to increase accuracy
  - Combine a series of k learned models

M_1, M_2, …, M_k, with the aim of creating an improved model M* as a CLASSIFIER

Building the CLASSIFIER



- Popular ensemble methods of **building the CLASSIFIER**
  - Bagging: averaging the prediction over a collection of classifiers
  - Boosting: weighted vote with a collection of classifiers
  - Erandom Forest: *decision tree* classifier

Random Forest

- Random Forest:

  each classifier in the ensemble is a

  decision tree classifier

  It is **generated** using a random selection

  of attributes at each **node** of the tree to determine the **split**

  In **final classifier**, each tree **votes** and

  the most popular class is **returned**

Neural Network

- Neural Network is a set of connected INPUT/OUTPUT UNITS, where each connection has a WEIGHT associated with it

- Neural Network learning is also called CONNECTIONIST **learning** due to the connections between units

- **Neural Network is always fully connected**
- It is a case of SUPERVISED, INDUCTIVE or CLASSIFICATION learning

Neural Network Learning

- **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

- **Neural Network** has a high tolerance to noisy and incomplete data.

Classification by Backpropagation

- **Backpropagation:** a **neural network** learning algorithm
- Started by psychologists and neurobiologists to develop and test computational analogues of neurons
- **A neural network:** a set of connected input/output units where each connection has a **weight** associated with it
- During the learning phase, the **network learns by adjusting the weights** so as to be able to predict the correct class label of the input tuples
- Also referred to as **connectionist learning** due to the connections between units

How A Multi-Layer Neural Network Works?

- 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**

How A Multi-Layer Neural Network Works?

- The network is **feed-forward** - it means that **none** of the **weights** **cycles back** to an **input unit** or to an **output unit** of a **previous layer**

- From a statistical point of view, networks perform **nonlinear regression**

-

- Given enough hidden units and enough training samples, they can closely **approximate** any function

MLFF Network Topology

- **Network topology:**
- We define the **network topology** by setting the following

1. number of units in the input layer

2. number of hidden layers

3. number of units in each hidden layer

4. number of units in the output layer

Classification by Backpropagation

- **Backpropagation** is a neural network **learning algorithm**

- **It learns** by iteratively processing a set of training data **comparing** the network's prediction for **each record** with the actual **known target value**

- **The target** value may be the **known** class label of the **training tuple** or a **continuous value** for prediction

Classification by Backpropagation

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**

- These **weights modifications** are propagated in **"backwards"** direction, that is,
  from the **output layer**, through **each hidden layer**
  down to the **first hidden layer**

- Hence the name **backpropagation**

Steps in Backpropagation Algorithm

- **STEP ONE:**

  initialize the **weights and biases**

- **The weights** in the network are initialized to small random numbers ranging for example from -1.0 to 1.0, or -0.5 to 0.5

- **Each unit** has **a bias** associated with it
- **The biases** are similarly initialized to small random numbers

- **STEP TWO:** **feed** the **training sample**

Steps in Backpropagation Algorithm

- **STEP THREE**:
- **propagate** the inputs forward by applying **activation function**
- We compute the net input and output of each unit in the hidden and output layers
- **STEP FOUR: backpropagate** the **error**
- **STEP FIVE:**
- **update** weights and biases to reflect the propagated errors
- **STEP SIX:**
- **repeat** and apply **terminating conditions**