Cse352
Artificial Intelligence
Short Review for Midterm

Professor Anita Wasilewska
Computer Science Department
Stony Brook University
Midterm

• Midterm INCLUDES
• CLASSIFICATION
• CLASSIFICATION by Decision TREES
• and
• PRODUCTION SYSTEMS
• PROPOSITIONAL RESOLUTION
• material you needed to review for Q1
Learning Process

• Questions:
• Describe and discuss all stages of the Data Mining-Learning Process
• Describe the role of Preprocessing stage and its main methods
• Discuss the Data Mining –Learning Proper stage
• Describe what is Descriptive/ non Descriptive Learning Which Models you would use for the Descriptive Learning and which for the non Descriptive Learning
• How and what decides which type of Learning 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
• Define a CLASSIFIER
• Describe a process of building a CLASSIFIER
Given a classification dataset DB with a set

\[ \mathbf{A} = \{a_1, a_2, \ldots, a_n\} \] of attributes and a class attribute \( \mathbf{C} \) with values

\[ \{c_1, c_2, \ldots, c_k\} \] - \( k \) classes

**Definition 1**

Any expression \( a_1 = v_1 \& \ldots \& a_k = v_k \) where \( a_i \in \mathbf{A} \) and \( v_i \) are corresponding values of attributes from \( \mathbf{A} \) is called a **DESCRIPTION**

Any expression \( \mathbf{C} = c_i \) is for \( c_i \in \{c_1, c_2, \ldots, c_k\} \) is called a **CLASS DESCRIPTION**
Classification Data and Rules

Definition 2

A CHARACTERISTIC FORMULA is any expression

\[ C = c_k \implies a_1 = v_1 \land \ldots \land a_k = v_k \]

We write it as

\[ \text{CLASS} \implies \text{DESCRIPTION} \]

Definition 3

A DETERMINANT FORMULA is any expression

\[ a_1 = v_1 \land \ldots \land a_k = v_k \implies C = c_k \]

We write it as

\[ \text{DESCRIPTION} \implies \text{CLASS} \]
Classification Data and Rules

Definition 4

A characteristic formula

\[ \text{CLASS} \rightarrow \text{DESCRIPTION} \]

is called a CHARACTERISTIC 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

\[ \{ o : \text{CLASS} \} \]

is the set of all records of DB corresponding to the CLASS
Definition 5
A discriminant formula

\[ \text{DESCRIPTION} \implies \text{CLASS} \]

is called a **DISCRIMINANT RULE** of DB
iff it is **TRUE in DB**, i.e. the following conditions hold

1. \( \{o:\text{DESCRIPTION}\} \neq \emptyset \)

2. \( \{o:\text{DESCRIPTION}\} \subseteq \{o:\text{CLASS}\} \)
PROBLEM 1

Prove that for any **classification** data base **DB** and any of its **DISCRIMINANT RULES** of the form

\[
\text{DESCRIPTION} \Rightarrow \text{CLASS}
\]

the formula

\[
\subseteq
\]

\[
\text{CLASS} \Rightarrow \text{DESCRIPTION}
\]

is a **CHARACTERISTIC RULE** of the **DB**
PROBLEM 1 Solution

By definition 5, for any database DB:

\[
\text{DESCRIPTION} \Rightarrow \text{CLASS}
\]

is a **DISCRIMINANT RULE** iff

1. \(\{o: \text{DESCRIPTION}\} \neq \emptyset\)

2. \(\{o: \text{DESCRIPTION}\} \subseteq \{o: \text{CLASS}\}\)

Therefore,

\[
\{o: \text{DESCRIPTION}\} \cap \{o: \text{CLASS}\} \neq \emptyset
\]

and by **Definition 4**

\[
\text{CLASS} \Rightarrow \text{DESCRIPTION}
\]

Is the **CHARACTERISTIC RULE**
Given a dataset:

<table>
<thead>
<tr>
<th>Record</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>O2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>O3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>O4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>O5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Find the set \( \{ o : \text{DESCRIPTION} \} \) for the following descriptions:

1) \( a_1 = 2 \) & \( a_2 = 1 \)
2) \( a_3 = 1 \) & \( a_4 = 0 \)
3) \( a_2 = 0 \) & \( a_3 = 2 \)
4) \( c=1 \)
5) \( c=0 \)
Find the set \(\{o : \text{DESCRIPTION}\}\) for the following descriptions

1) \(a_1 = 2 \& a_2 = 1\)  \hspace{1cm} \text{Answer}: \\{o_1\}
2) \(a_3 = 1 \& a_4 = 0\)  \hspace{1cm} \text{Answer}: \\{o_1, o_5\}
3) \(a_2 = 0 \& a_3 = 2\)  \hspace{1cm} \text{Answer}: \\{o_4\}
4) \(c=1\)  \hspace{1cm} \text{Answer}: \\{o_1, o_5\}
5) \(c=0\)  \hspace{1cm} \text{Answer}: \\{o_3, o_5\}
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) \(a_1 = 1 \& a_2 = 1 \Rightarrow C = 1\)

7) \(C = 1 \Rightarrow a_1 = 0 \& a_2 = 1 \& a_3 = 1\)

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

9) \(C = 0 \Rightarrow a_1 = 1 \& a_4 = 0\)

10) \(a_1 = 2 \& a_2 = 1 \& a_3 = 1 \Rightarrow C = 0\)

11) \(a_1 = 0 \& a_3 = 2 \Rightarrow C = 1\)
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) \( a_1 = 1 \) & \( a_2 = 1 \) \( \Rightarrow \) \( C = 1 \)
   \{o1\} is a subset of \{o1, o5\} so this is a DISCRIMINANT rule

7) \( C = 1 \) \( \Rightarrow \) \( a_1 = 0 \) & \( a_2 = 1 \) & \( a_3 = 1 \)
   \{o: a_1 = 0 & a_2 = 1 & a_3 = 1\} is an empty set so this is not a CHARACTERISTIC rule

8) \( C = 2 \) \( \Rightarrow \) \( a_1 = 1 \)
   As the intersection is empty so this is not a CHARACTERISTIC rule

9) \( C = 0 \) \( \Rightarrow \) \( a_1 = 1 \) & \( a_4 = 0 \) ----- \{o3, o4\} \( \cap \) \{o5\} is empty set so this is not a CHARACTERISTIC rule

10) \( a_1 = 2 \) & \( a_2 = 1 \) & \( a_3 = 1 \) \( \Rightarrow \) \( C = 0 \) ----- \{o5\} is not a subset of \{o3, o4\}, so this is not a DISCRIMINANT rule

11) \( a_1 = 0 \) & \( a_3 = 2 \) \( \Rightarrow \) \( C = 1 \) ----- \{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 DTREE

1. Use the data below to build a CLASSIFIER by basic DTREE algorithm.
2. Use 2 different testing methods of your choice and compare the results.

<table>
<thead>
<tr>
<th>Record</th>
<th>A1</th>
<th>A2</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Data Mining Process

DATA MINING (proper)

INTERPRETATION AND EVALUATION

Transformed data

Processed Data

Target data

Data Preparation

CLEANING

SELECTION

Data
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)
Learning Proper

• **Learning 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** - 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 – can 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: $c_1$, $c_2$, ...$c_n$ of the class attribute $C$ 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:**
- **Dtreen:** some Attribute Selection Measures
- **Information Gain, Gini Index**

- We use them for selecting the *attribute* that “best” discriminates the given tuples according to *class*
Rule Based Systems REVIEW

Exercise

• Exercise 1
• Here are three simple expert rules
• R1: If your savings are small, then don’t invest in stocks
• R2: If you have no children and large income, then invest in stocks
• R3: If you have children and small income, then invest in savings
Exercise 1

• Conceptualize rules R1, R2, R3 in Predicate Form using predicates
  
  attribute(x, value of attribute)
  attribute(object, value of attribute)

  WRITE a format of a database TABLE needed for your conceptualization

REMARK: In order to express the rules Predicate Form, we must first define appropriate ATTRIBUTES and their values
Exercise 1

- We have the following ATTRIBUTES:

  - Savings
    Values: small, large
  - Income
    Values: small, large
  - InvestStocks
    Values: yes, no
  - InvestSavings
    Values: yes, no
  - Children
    Values: yes, no
Exercise 2

• Exercise 2
• The initial database has the following FACTS
  • F1: Savings(John, small)
  • F2: Children(John, no)
  • F3: Income(John, large)
• Are these FACTS true in Exercise 1 Data Table for some record o = John?
• Design a Data Table 2 in which the above FACTS are true
• Can you deduce InvestStocks(John, yes) on the base of the Data Table 2
Exercise 1: Predicate Form Conceptualization
Data Table Example

<table>
<thead>
<tr>
<th>Records</th>
<th>Savings</th>
<th>Income</th>
<th>InvesrStocks</th>
<th>InvestSavings</th>
<th>Children</th>
</tr>
</thead>
<tbody>
<tr>
<td>O₁</td>
<td>small</td>
<td>small</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>O₂</td>
<td>large</td>
<td>small</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>O₃</td>
<td>small</td>
<td>large</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>
Rules in Predicate Form

• RULES:

• **R1**: Savings(x, small) $\rightarrow$ InvestStock (x, no)

• **R2**: Children(x, no) $\land$ Income(x, large) $\rightarrow$ InvestStocks(x, yes)

• **R3**: Children(x, yes) $\lor$ Income(x, small) $\rightarrow$ InvestSavings(x, yes)
Rules in Predicate Form

- **RULES:**
  - **R1:** Savings(x, small) $\rightarrow$ InvestStock (x, no)
  - **R2:** Children(x, no) $\land$ Income(x, large) $\rightarrow$ InvestStocks(x, yes)
  - **R3:** Children(x, yes) $\lor$ Income(x, small) $\rightarrow$ InvestSavings(x, yes)
Exercise 1: Rules in Predicate Form

• RULES:

• R1: Savings(x, small) $\rightarrow$ InvestStock (x, no)

• R2: Children(x, no) $\land$ Income(x, large) $\rightarrow$ InvestStocks(x, yes)

• R3: Children(x, yes) $\lor$ Income(x, small) $\rightarrow$ InvestSavings(x, yes)
PART3: Exercise 2

• Exercise 2
• The initial database has the following FACTS
  • F1: Savings(John, small)
  • F2: Children(John, no)
  • F3: Income(John, large)
• 1. Are these FACTS true in Exercise 1 Data Table for a record o = John?
• 2. Design a Data Table 2 in which the above FACTS are true
• 3. Can you deduce InvestStocks(John, yes) on the base of the Data Table 2
Part 3: Exercise 3

• Given rules from Exercise 1:
  • R1: If your savings are small, then don’t invest in stocks
  • R2: If you have no children and large income, then invest in stocks
  • R3: If you have children and small income, then invest in savings
Exercise 3

• Conceptualize rules R1, R2, R3

In Propositional Logic in two ways:

1. Rules admit only atomic formulas; i.e. rules are built from propositional variables only – call the set of rules \( PR1 \)

2. Rules admit atomic formulas and negation of atomic formulas – call obtained set of rules \( PR2 \)
Part 3: Exercise 3

• Write initial databases B1 and B2 of facts corresponding to the facts F1, F2, F3 from Exercise 2 for
  
• (1) propositional conceptualization 1.
• (2) propositional conceptualization 2.
• (3) use corresponding rules from sets PR1, PR2 to deduce all facts from B1 and B2, respectively

Use Conflict Resolution from Busse Handout