Classification
Short Review

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

\[ A = \{a_1, a_2, \ldots, a_n\} \] of attributes and a class attribute \( C \) 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 \( a_i \in A \) and \( v_i \) are corresponding values of attributes from \( A \) is called a DESCRIPTION

Any expression \( C = c_i \) 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

\[ C = c_k \Rightarrow a_1 = v_1 \ \& \ ... \ \& \ a_k = v_k \]

We write it as

CLASS \(\Rightarrow\) DESCRIPTION

Definition 3

A DETERMINANT FORMULA is any expression

\[ a_1 = v_1 \ \& \ ... \ \& \ a_k = v_k \Rightarrow C = c_k \]

We write it as

DESCRIPTION \(\Rightarrow\) 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

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

is a CHARACTERISTIC RULE of the DB.
PROBLEM 1 Solution

By definition 5, for any database DB:

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

is a \textbf{DISCRIMINANT RULE} \iff

1. \{o: DESCRIPTION\} \not= \emptyset

2. \{o: DESCRIPTION\} \subseteq \{o: CLASS\}

Therefore,

\[ \{o: \text{DESCRIPTION}\} \cap \{o: \text{CLASS}\} \not= \emptyset \]

and by Definition 4

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

Is the \textbf{CHARACTERISTIC RULE}
PROBLEM 2

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 : 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) \(c=0\)
PROBLEM 2 SOLUTION

Find the set \( \{ o : \text{DESCRIPTION} \} \)
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) \( a_1 = 1 \) & \( a_2 = 1 \) \( \Rightarrow \) \( C = 1 \)

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\} \( \setminus \) \{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

• What is a CLASSIFIER?

• Describe the process of building 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.

<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>
### Transactional Data and Support calculations

<table>
<thead>
<tr>
<th></th>
<th>I1 (A1 = 0)</th>
<th>I2 (A1 = 1)</th>
<th>I3 (A2 = 0)</th>
<th>I4 (A2 = 1)</th>
<th>I5 (C = 0)</th>
<th>I6 (C = 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>+</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Count</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>
Let the **minimum support count** = 3

**L1:**

<table>
<thead>
<tr>
<th>Item set</th>
<th>Support Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1</td>
<td>4</td>
</tr>
<tr>
<td>I2</td>
<td>4</td>
</tr>
<tr>
<td>I3</td>
<td>4</td>
</tr>
<tr>
<td>I4</td>
<td>4</td>
</tr>
<tr>
<td>I5</td>
<td>5</td>
</tr>
<tr>
<td>I6</td>
<td>3</td>
</tr>
</tbody>
</table>
Candidate two item sets:

<table>
<thead>
<tr>
<th>Item Set</th>
<th>Support Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2</td>
<td>0</td>
</tr>
<tr>
<td>1,3</td>
<td>3</td>
</tr>
<tr>
<td>1,4</td>
<td>1</td>
</tr>
<tr>
<td>1,5</td>
<td>4</td>
</tr>
<tr>
<td>1,6</td>
<td>0</td>
</tr>
<tr>
<td>2,3</td>
<td>1</td>
</tr>
<tr>
<td>2,4</td>
<td>3</td>
</tr>
<tr>
<td>2,5</td>
<td>1</td>
</tr>
<tr>
<td>2,6</td>
<td>0</td>
</tr>
<tr>
<td>3,4</td>
<td>3</td>
</tr>
<tr>
<td>3,5</td>
<td>1</td>
</tr>
<tr>
<td>3,6</td>
<td>2</td>
</tr>
<tr>
<td>4,5</td>
<td>2</td>
</tr>
<tr>
<td>4,6</td>
<td>0</td>
</tr>
</tbody>
</table>
Classification by Association

**Frequent 2 item set:**

<table>
<thead>
<tr>
<th>Item Set</th>
<th>Support Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,3</td>
<td>3</td>
</tr>
<tr>
<td>1,5</td>
<td>4</td>
</tr>
<tr>
<td>2,4</td>
<td>3</td>
</tr>
<tr>
<td>2,6</td>
<td>3</td>
</tr>
<tr>
<td>3,5</td>
<td>3</td>
</tr>
</tbody>
</table>
Classification by Association

Candidate 3 item set:

<table>
<thead>
<tr>
<th>Item Set</th>
<th>Support Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,3,5</td>
<td>3</td>
</tr>
<tr>
<td>2,4,6</td>
<td>1</td>
</tr>
</tbody>
</table>
Classification by Association

Frequent 3 item Set:

<table>
<thead>
<tr>
<th>Item set</th>
<th>Support Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,3,5</td>
<td>3</td>
</tr>
</tbody>
</table>

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

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

Don’t forget to FIX and calculate Confidence and Support!
Testing:

<table>
<thead>
<tr>
<th>Record</th>
<th>A1</th>
<th>A2</th>
<th>Expected class</th>
<th>Actual class</th>
<th>Correctly classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>?</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Predictive accuracy = \( \frac{3}{4} \times 100 = 75 \% \)
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)

<table>
<thead>
<tr>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0</td>
<td>0.2</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0.3</td>
<td>0.2</td>
<td>1</td>
</tr>
<tr>
<td>0.2</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Construct a Neural Network with your own topology and evaluate- describe a passage of ONE EPOCHS (use learning rate \( l = 0.7 \))
Topology:
Input = 3, hidden = 2 and output = 2.
Problem: Neural Networks

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

\[
\begin{align*}
  a_1 &= 0.5, \\
  a_2 &= 0, \\
  a_3 &= 0.2
\end{align*}
\]

\[
\begin{align*}
  w_{14} &= 0.2, \\
  w_{15} &= -0.3, \\
  w_{24} &= 0.4, \\
  w_{25} &= 0.1
\end{align*}
\]

\[
\begin{align*}
  w_{34} &= 0.2, \\
  w_{35} &= -0.3, \\
  w_{46} &= 0.4, \\
  w_{56} &= 0.1
\end{align*}
\]

\[
\begin{align*}
  w_{47} &= 0.1, \\
  w_{57} &= 0.2
\end{align*}
\]

We take any random values for **weights** and **biases**