SS-ZG548: ADVANCED DATA MINING

01 Logistics and Introduction

Dr. Kamlesh Tiwari
Assistant Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

July 24, 2021 ONLINE (WILP @ BITS-Pilani July-Dec 2021)

http://ktiwari.in/adm

Introduction

What is data?

Fact or values

What is Information?

Processed output of date

What is Knowledge?

Understanding of information.

X1	X2	X3	Y
2	3	4	1
8	7	5	10
9	6	1	14
7	4	9	2
1	5	5	1
5	3	6	2
7	4	4	7
6	3	5	4
3	2	3	

The Knowledge

What is **data-mining**?

Computation to facilitate Knowledge Discovery in Databases (**KDD**)

Goal of Data Mining (motivation of doing the same?)

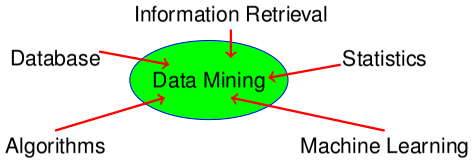
To provide efficient tools and techniques for KDD

Knowledge Discovering in Databases (KDD) involves

1. **Selection:** collection of data
2. **Preprocessing:** deal with incorrect/missing data
3. **Transformation:** common format and preprocessing
4. **Data Science:** algorithmic tools
5. **Interpretation/Evaluation:** presentation and visualization

Data Mining

Data mining is fairly involved discipline. It includes many fields such as database, information retrieval, statistics, and machine learning.



It differs from traditional query processing

- **Query:** not well formed. Miner may not know what he wants.
- **Data:** different version. Preprocessed and modified.
- **Output:** may not a subset. It could be an analysis.

Data Mining Parts

Data Mining has three parts

1. **Model:** is to be fit on data
2. **Search:** technique to evaluate data point
3. **Preference:** criteria to select one model over other

Example:

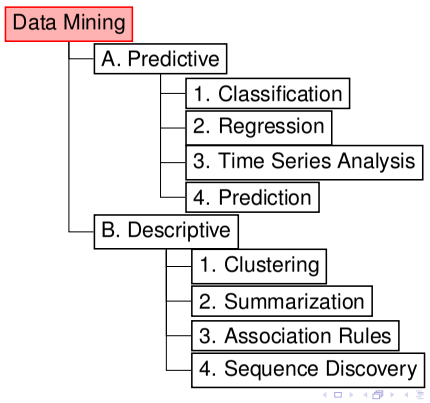
Assume a credit card company wants to decide whether a transaction should be

1. Authorized
2. Ask for more information
3. Decline

Search requires evaluation of past data. **Model** associates with the criteria to decide for one of the categories. **Preference** is given to criteria that suits the data best (want to reduce number of frauds or amount of fraud).

Data Mining: Tasks

Two broad categories of data mining models are *Predictive* and *Descriptive*. Some of the related tasks are



Classification

Classification maps data into *predefined* labels.

Example: Lots of mails are there in my mail box. Can you tell me which are SPAM?

- Task of supervised learning
- Often based on some patterns or characteristics
- We can use the frequency of words
- Assumption is that some words appears more or less frequently in SPAM

Regression

Regression is used to map data into *real valued* variable.

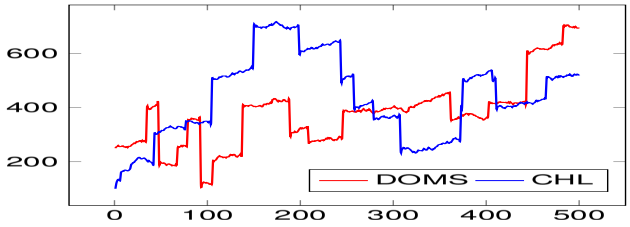
Example: What is the cost of my house?

- Task of supervised learning
- We have data about the cost of house based on features such as
  - location
  - Plot area
  - number of rooms
  - garden available or not
  - how old it is
- Current economical conditions can also matter
- Dimensionality is high

Time Series Analysis

In time series analysis the value of attribute is examined over time.

Example: Which stock is better?



- The values are obtained as evenly spaced time points (daily, weekly, hourly, etc.)
- Distance measures are used to find similarity
- Structural analysis is done

Prediction

Predicting future data states based on current or historical data.

Example: What comes next?

$$2, 4, 6, 8, 10, ...?...$$

$$2, 3, 5, 7, ...?..., 13$$

$$(10jul, rain), (11jul, rain), (12jul, no - rain), (13jul, ...?...)$$

- Predication can sometimes be seen as classification
- Application includes weather, flood, pattern recognition.

Clustering

Clustering is similar to classification except the groups are not pre-defined.

Example: How many kind of files are there in my directory?

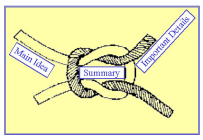
- Unsupervised learning setting
- We can use file name
- Words it has

Example: Who would take my offer?

- The database has information about age, gender, income, location, .. etc.

Summarization

Summarization maps data into subsets with associated simple descriptions. It is also called characterization or generalization.



Example: How to compare two universities?

- Average JEE rank
- Average number of publication
- Student/Faculty ratio
- Combination

Association Rules

Association rules tries to do linked analysis.

Example: Whether sames products are selling together?

- $I = \{i_1, i_2, i_3, ...i_m\}$, $T = \{t_1, t_2, t_3, ..., t_n\}$ and $t_i \subseteq I$
- Minimum support count should be maintained
- Can you see: Subset of frequent items is also frequent
- Apriori analysis

Let's do it:

$t_1 = (1,3,4)$, $t_2 = (2,3,5)$, $t_3 = (1,2,3,5)$, $t_4 = (2,5)$, $t_5 = (1,3,5)$ and minimum support count be 2

Association Rules

$t_1 = (1,3,4)$, $t_2 = (2,3,5)$, $t_3 = (1,2,3,5)$, $t_4 = (2,5)$, $t_5 = (1,3,5)$

Symb
{1}
{2}
{3}
{4}
{5}

Association Rules

$t_1 = (1,3,4)$, $t_2 = (2,3,5)$, $t_3 = (1,2,3,5)$, $t_4 = (2,5)$, $t_5 = (1,3,5)$

Symb	Sup
{1}	3 ✓
{2}	3 ✓
{3}	4 ✓
{4}	1
{5}	4 ✓

Association Rules

$t_1 = (1,3,4)$, $t_2 = (2,3,5)$, $t_3 = (1,2,3,5)$, $t_4 = (2,5)$, $t_5 = (1,3,5)$

Symb	Sup
{1}	3 ✓
{2}	3 ✓
{3}	4 ✓
{4}	1
{5}	4 ✓

Symb	Sup
{1,2}	1
{1,3}	3 ✓
{1,5}	2 ✓
{2,3}	2 ✓
{2,5}	3 ✓
{3,5}	3 ✓

Association Rules

$t_1 = (1,3,4)$, $t_2 = (2,3,5)$, $t_3 = (1,2,3,5)$, $t_4 = (2,5)$, $t_5 = (1,3,5)$

Symb	Sup
{1}	3 ✓
{2}	3 ✓
{3}	4 ✓
{4}	1
{5}	4 ✓

Symb	Sup
{1,2}	1
{1,3}	3 ✓
{1,5}	2 ✓
{2,3}	2 ✓
{2,5}	3 ✓
{3,5}	3 ✓

Symb	Sup
{1,2,3}	1
{1,2,5}	1
{1,3,5}	2 ✓
{2,3,5}	2 ✓

Symb	Sup
{1,2,3,5}	1

Association Rules

$t_1 = (1,3,4)$, $t_2 = (2,3,5)$, $t_3 = (1,2,3,5)$, $t_4 = (2,5)$, $t_5 = (1,3,5)$

Symb	Sup
{1}	3 ✓
{2}	3 ✓
{3}	4 ✓
{4}	1
{5}	4 ✓

Symb	Sup
{1,2}	1
{1,3}	3 ✓
{1,5}	2 ✓
{2,3}	2 ✓
{2,5}	3 ✓
{3,5}	3 ✓

Symb	Sup
{1,2,3}	1
{1,2,5}	1
{1,3,5}	2 ✓
{2,3,5}	2 ✓

Symb	Sup
{1,2,3,5}	1

Try yourself:

Let $I = \{A, B, C, D, E, F\}$ and $T = \{ t_1 = (A, B, C), t_2 = (A, F), t_3 = (A, B, C, E), t_4 = (A, B, D, F), t_5 = (C, F), t_6 = (A, B, C), t_7 = (A, B, C, E), t_8 = (C, D, E), t_9 = (B, D, E) \}$ and min support 3

Sequence Discovery

Sequence Discovery is used to discover sequential patterns in the data.

Example: what is my website access pattern?

- Pattern is based on a time sequence of a action
- It is pattern discovery problem

KDD Issues

- Human interaction
- Overfitting, Outliers
- Large dataset
- High dimension
- Multimedia data
- Missing data
- Irrelevant data
- Noisy data
- Changing data

And much more...

Syllabus

Introduction and basics | Stream data mining | Text mining
Distributed data mining | Sequence mining | Web Search
Mining complex structures (Trees, Graphs)
Case study: information retrieval, social network mining

Evaluation Scheme (July-Nov 2019)

- 3 Quiz/Assignment: 5% Each. Aug 16, Sept 16, Oct 16
- Mid-Semester Test: 35% (2H, Closed Book) 24 Sept 2021 (FN)
- Comprehensive Exam: 50% (3H, Open Book) 12 Nov 2021 (FN)

NOTE: All evaluation components are to be attempted individually. Plagiarism of any form is not accepted.

Thank You!

Thank you very much for your attention!

Queries ?