Data Mining - Lecture 1
Introduction to Data Mining

Dr. Dewan Md. Farid

Associate Professor, Department of Computer Science & Engineering
United International University

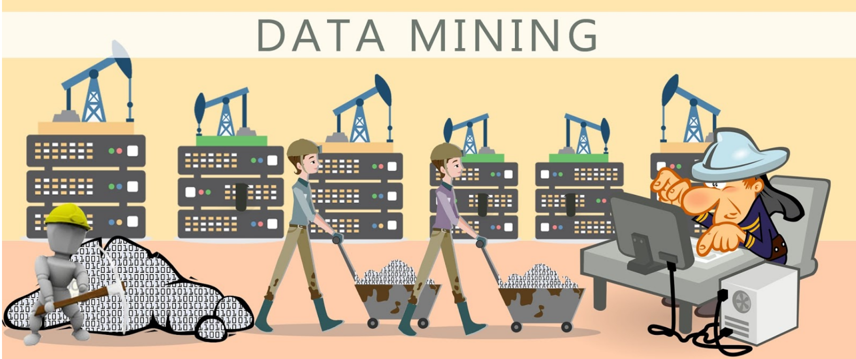
March 03, 2018

Introduction

Tasks & Issues

Simple Example: The Weather Problem

Weka

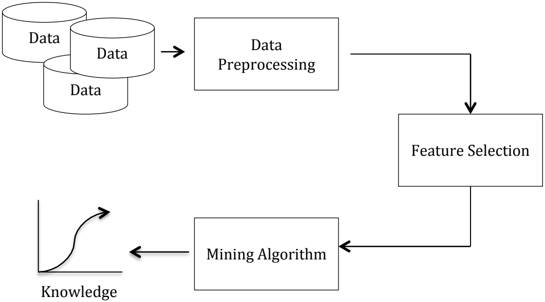

DATA MINING

What is Data Mining?

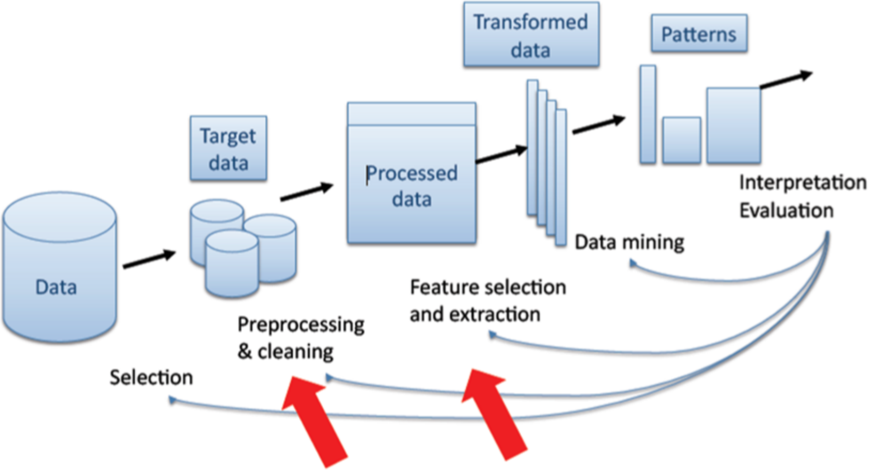
Data mining is also known as *Knowledge Discovery from Data*, or *KDD* for short, which turns a large collection of data into knowledge. Data mining is a multidisciplinary field including machine learning, artificial intelligence, pattern recognition, knowledge-based systems, high-performance computing, database technology, and data visualisation.

- Data mining is the process of analysing data from different perspectives and summarising it into useful information.
- Data mining is the process of finding hidden information and patterns in a huge database.
- Data mining is the extraction of implicit, previously unknown, and potentially useful information from data.

Data Mining Process



Knowledge Discovery in Databases (KDD) process



Text Books

1. **Data Mining Concepts and Technique**, Jiawei Han, Micheline Kamber, and Jian Pei (Third Edition)

2. **Data Mining Practical Machine Learning Tools and Techniques**, Ian H. Witten, Eibe Frank, and Mark A. Hall (Third Edition)

Reference Books

1. **Data Mining Introductory and Advanced Topics**, Margaret H. Dunham
2. **Introduction to Data Mining**, Pang-Ning Tan, Michael Steinbach, and Vipin Kumar
3. **Principles of Data Mining**, David Hand, Heikki Mannila, and Padhraic Smyth
4. **Data Mining Knowledge Discovery and Applications**, Edited by A. Karahoca
5. **Mining Complex Data**, D. A. Zighed, S. Tsumoto, Z. W. Ras, and H. Hacid

Data, Information, and Knowledge

Data
: Data are any recorded facts, numbers, or text that can be processed by a computer - scientific data, medical data, demographic data, financial data, and marking data.

Information
: The patterns, associations, or relationships among all this data can provide information.

Knowledge
: Information can be converted into knowledge about historical patterns and future trends.

What Kinds of Data can be Mind?

The most basic forms of data for mining are come from:

1. Database Data
2. Data Warehouses
3. Transactional Data

Machine Learning

Machine learning provides the technical basis of data mining. It is a branch of artificial intelligence, which concerns the construction and study of systems that can learn from data.

For example, a machine learning system could be trained on email messages to learn to distinguish between spam and non-spam messages. After learning, it can then be used to classify new email messages into spam and non-spam folders.

Types of Learning

Supervised learning is basically a synonym of classification. The supervision in the learning comes from the labeled instances in the training data.

Unsupervised learning is essentially a synonym of clustering. The learning process is unsupervised since the input instances are not class labeled.

Semi-supervised learning is a class of machine learning technology that make use of both labeled and unlabelled instances when learning a model.

Active learning is a machine learning approach that lets users play an active role in the learning process. An active learning approach can ask a user (e.g., a domain expert) to label an instance, which may be from a set of unlabelled instances.

Machine Learning & Data Mining

These two terms are commonly confused, as they often employ the same methods and overlap significantly.

They can be roughly defined as follows:

- Machine learning focuses on prediction, based on known properties learned from the training data.
- Data mining focuses on the discovery of unknown properties in the data.

Data Mining Task

Data mining is defined as the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful in that they lead some advantage, usually an economic one.

Data mining have two major functions:

1. Classification
2. Clustering

Classification

- Classification maps data into predefined groups or classes.
- It is often referred to as supervised learning because the classes are determined before examining the data.
- Classification creates a function from training data. The training data consist of pairs of input objects, and desired output. The output of the function can be a continuous value, or can predict a class label of the input object.
- The task of the classification is to predict the value of the function for any valid input object after having seen only a small number of training examples.

Clustering

- Clustering is similar to classification except that the groups are not predefined, but rather defined by the data alone.
- Clustering is alternatively referred to as unsupervised learning or segmentation.
- It can be thought of as partitioning or segmenting the data into groups that might or might not be disjointed.
- The clustering is usually accomplished by determining the similarity among the data on predefined attributes. The most similar data are grouped into clusters.

The Weather Problem

To illustrate an example, we consider a small weather data.

- It has four attributes/ features that represent the weather condition of a particular day:
  1. Outlook
  2. Temperature
  3. Humidity
  4. Wind
- Each attribute has several unique attribute values.
- The **Play** column represents the class category of each instance. It indicates whether a particular weather condition is suitable or not for playing tennis.

Table: Weather Data

Outlook	**Temperature**	**Humidity**	**Wind**	**Play**
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Extracting Rules from Weather Data

A set of rules learned from the Table 1.

1. If Outlook = Sunny and Humidity = High then Play = No
2. If Outlook = Sunny and Humidity = Normal then Play = Yes
3. If Outlook = Overcast then Play = Yes
4. If Outlook = Rain and Wind = Strong then Play = No
5. If Outlook = Rain and Wind = Weak then Play = Yes

Discrete Vs. Continuous Attributes

A **discrete attribute** has a finite or countably infinite set of values, which may or may not be represented as integers.
On the other side, a *continuous attribute* has a numeric or continuous attribute values.

Weather Data with Numeric Values

In the slightly more complex data, two of the attributes - temperature and humidity have numeric values.

Table: Weather Data with some Numeric Attributes

Outlook	Temperature	Humidity	Wind	Play
Sunny	85	85	Weak	No
Sunny	80	90	Strong	No
Overcast	83	86	Weak	Yes
Rain	70	96	Weak	Yes
Rain	68	80	Weak	Yes
Rain	65	70	Strong	No
Overcast	64	65	Strong	Yes
Sunny	72	95	Weak	No
Sunny	69	70	Weak	Yes
Rain	75	80	Weak	Yes
Sunny	75	70	Strong	Yes
Overcast	72	90	Strong	Yes
Overcast	81	75	Weak	Yes
Rain	71	91	Strong	No

Rules with Numeric Attribute Values

A set of rules learned from the Table 2.

1. If Outlook = Sunny and Humidity $>$ 75 then Play = No
2. If Outlook = Sunny and Humidity $\leq$ 75 then Play = Yes
3. If Outlook = Overcast then Play = Yes
4. If Outlook = Rain and Wind = Strong then Play = No
5. If Outlook = Rain and Wind = Weak then Play = Yes

Association Analysis

Suppose that, you want to know which items are frequently purchased together (i.e., within the same transaction).
An example of such a rule is,
$buys(X, ?computer?) \rightarrow buys(X, ?software?)[support = 1\%, confidence = 50\%]$

- Where X is a variable representing a customer.
- A **confidence**, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that he/she will buy software as well.
- A 1% **support** means that 1% of all the transactions under analysis show that computer and software are purchased together.

Input: Concepts, Instances, and Attribute

- The input takes the form of *concepts*, *instances*, and *attributes*.
- We call the thing that is to be learned a *concept description*.
- Each instance is characterised by the values of attributes that measure different aspects of the instance.
- There are many different types of attributes, although typical data mining schemes deal only with numeric and nominal, or categorical attributes.

Concepts, Instances, and Attribute

Concept is the thing to be learned.

Concept description is the output produced by a learning scheme or classifier.

Instances are the things that are to be classified or associated or clustered. Each dataset is represented as a matrix of instances versus attributes, which in database terms is a single relation, or a *flat file*.

Attribute is a data field, representing a characteristic or feature of a data object. The nouns *attribute, dimension, feature*, and *variable* are often us interchangeably in the literature. The value of an attribute for a particular instance is a measurement of the quantity to which the attribute refers.

Data Preprocessing

1. Data Cleaning - Missing Values, Noisy Data
2. Data Integration - Redundancy and Correlation Analysis, Instance Duplication
3. Data Reduction - Dimensionality Reduction
4. Attribute subset Selection

Weka: Data Mining Software in Java

Weka (**Waikato Environment for Knowledge Analysis**) is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualisation.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Attribute-Relation File Format (ARFF)

An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software.

ARFF files have two distinct sections. The first section is the **Header** information, which is followed the Data information. The Header of the ARFF file contains the name of the relation, a list of the attributes (the columns in the data), and their types. Lines that begin with a Textbf% are comments. The **@RELATION**, **@ATTRIBUTE** and **@DATA** declarations are case insensitive.

Example of ARFF file

```
@relation weather
@attribute outlook Sunny, Overcast, Rainy
@attribute temperature Hot, Mild, Cool
@attribute humidity High, Normal
@attribute windy Strong, Weak
@attribute play Yes, No
@data
Sunny,Hot,High,Weak,No
Sunny,Hot,High,Strong,No
Overcast,Hot,High,Weak,Yes
Rainy,Mild,High,Weak,Yes
Rainy,Cool,Normal,Weak,Yes
Rainy,Cool,Normal,Strong,No
Overcast,Cool,Normal,Strong,yes
Sunny,Mild,High,Weak,No
Sunny,Cool,Normal,Weak,Yes
Rainy,Mild,Normal,Weak,Yes
Sunny,Mild,Normal,Strong,Yes
Overcast,Mild,High,Strong,Yes
Overcast,Hot,Normal,Weak,Yes
Rainy,Mild,High,Strong,No
```

*** THANK YOU ***