Introduction to Data Mining and Warehousing

1. Data Mining : What and Why

. Data Mining is the process of discovering patterns, insights, and Knowledge from large sets of data using computational techniques

. Why we need Data Mining

↳ Rapid Data Generation : Data is being produced at unprecendented rates.

. In 1 second there are :

7,998 Tweets
839 Instagram Uploads
66,335 Google Searchs
55,560 of Internet Traffic
2,681,874 Emails Sent

↳ Human Limitation : We need computational tools to help human to :

. Summarize large datasets
. Understand patterns
. Extract useful Knowledge

2. Data vs. Information

- **Data**: Raw, unprocessed facts
  - Numbers, records, logs
- **Information**: Patterns and insights derived from data

Why it Matters

- Data is abundant (business, science, medicine, etc.)
- Without processing, raw data is useless
- Extracting patterns transforms data into a valuable resource

3. Data Mining Examples

Raw Data

Data Mining

Patterns
Knowledge

Recommender Systems

Data on library books and users' past reading history

Data Mining

What book to recommend next to given user such that there is a high likehood that the user will like it?

Resource Allocation

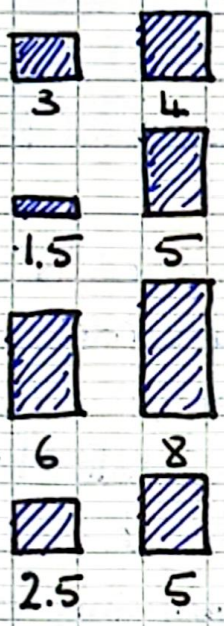
Data Mining

Given a newly acquired book, what is an accurate estimate of the nb of users who will read it in the next 12 months?

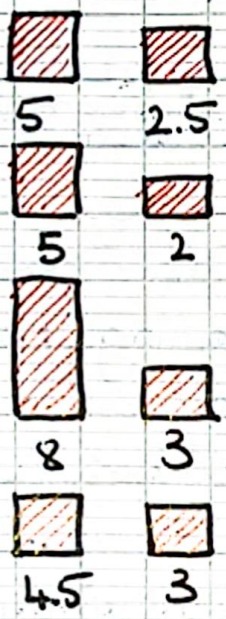
- Simple Classification Problem:
Decide if the object belongs to ClassA or Class B based on a simple rule.

↳ Example 1

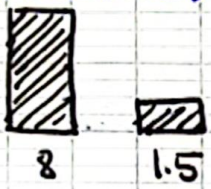
Class A



Class B



we need to classify the following object and specify to wich class it belong (A or B)



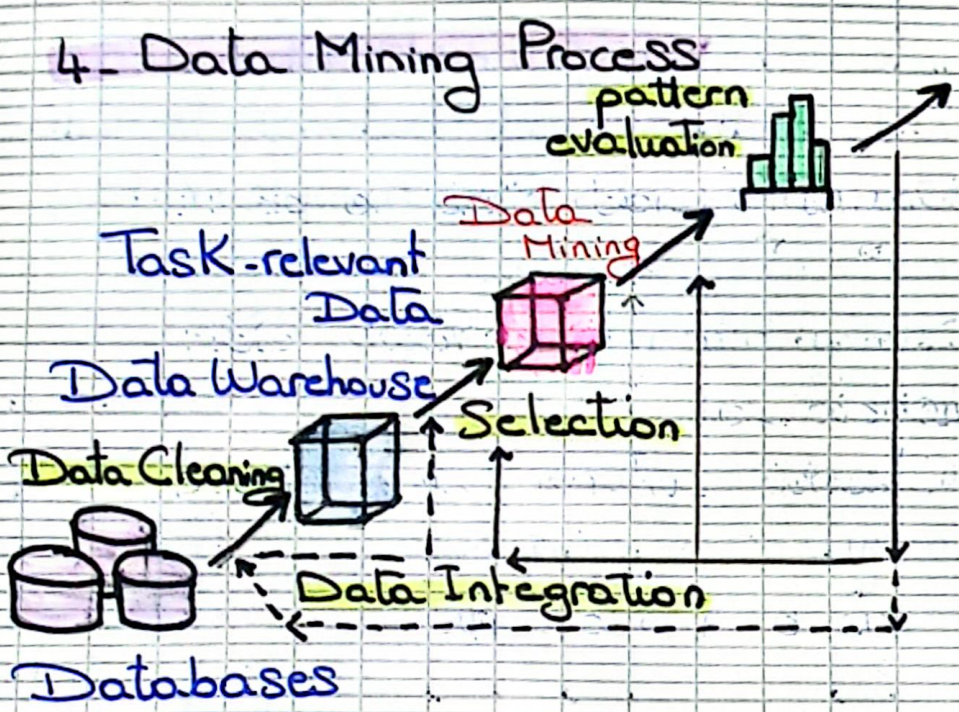
=> We can conclude a rule based on the data we have:

" If the left bar is smaller than the right bar, it's an A, otherwise it's a B"

As 8 (left) is bigger than 1.5 (right) then the object belong to the class B

This example demonstrates how data mining aims to extracts rules to classify data into categories, or to reveal patterns and insights from raw data.

4- Data Mining Process



1) Data Cleaning: Removing errors or inconsistencies from the data
2) Data Integration: Combining data from multiple sources into a unified dataset
3) Data Warehousing: Organizing and storing data in a warehouse (structured database)
4) Data Cube Construction: Creating multidimensional views of data for easier analysis
5) Data Selection: Extracting relevant data for mining tasks
6) Data Mining: Discovering patterns and relationships in the data
7) Result Presentation: Visualizing or reporting the findings for better understanding
8) Pattern and Knowledge Storage: Storing useful patterns in a knowledge base for future use.

5_ Data Mining Approaches and Techniques (What Kinds of patterns can be mined from data?)

1) Concept / Class Description : Characterization and Discrimination
2) Mining frequent patterns, associations, and Correlations
3) Classification and Regression
4) Cluster Analysis
5) Outlier Detection

5.1) Characterization / Discrimination (Data Warehousing Part)

→ Data Characterization

. Summarizes the general characteristics or features of a target class of data

. How it work?

Data is typically collected using queries (e.g. SQL) on a database

. Example:

To study software products with a 10% sales increase in the previous year, execute an SQL query to extract relevant data from the sales databases

. Technique Used:

↳ Statistical Summaries : Basic summaries like averages, total, and plots.

↳ OLAP Roll-Up Operations : Aggregating data into higher_level summaries (from data cubes in warehousing)

→ Data Discrimination:
- Compares the general features of a target class of data objects with one or more contrasting classes.
- How it works?
Compares patterns across classes (e.g. positive vs. negative sales trends)
- Example:
Compare software products with a 10% sales increase to those with a 30% sales decrease over the same period
- Techniques Used:
Methods are similar to data characterization (e.g. OLAP tools, statistical summaries)

5.2) Pattern Discovery

→ Frequent Patterns (Frequent Itemsets)
- Identify items or events that frequently occur together
- Example: Discover which items customers often purchase together in a store (e.g. bread and butter.)

→ Association, Correlation, vs. Causality
- Association Rule:
Example: Diaper → Coffee [0.5%, 75%]
  - Support: 0.5% (percentage of transactions containing both items)
  - Confidence: 75% (likehood of buying coffee if a diaper is purchased)

. Question : Are items with strong association also strongly correlated ?
  → Not always as correlation implies a stronger statistical relationship, while association may just reflect co-occurrence)

→ Challenge
  . Efficiency: How to mine frequent patterns and rules effectively from large datasets

→ Applications
  1) Classification : Categorize data based on discovered patterns
  2) Clustering : Group similar data items based on shared patterns
  3) Others : Fraud detection, recommendations

5.3) Classification

→ Data Example:
A large collection of books, each with:
  . Attributes : Title, author, information...
  . Category Labels: Art, History, Geography

→ Goal:
Create a Classification Model:
  . A set of patterns that can map books to their categories based on their features (e.g., text content, Keywords, etc)

→ Uses of the Model

1) Prediction: Given a new, uncategorized book, predict its category using the model
2) Description:
   . Provide insights into why certain books belong to specific categories
   . Example: "Books with keywords related to Renaissance art are often classified under Art."

5.4) Regression

→ Data:

A large collection of books with attributes: Title, Info, Full Text, Number of Users

→ Goal:

Build a regression model that maps books to their expected number of readers
↳ The model should learn patterns from the data, such as correlations between book features and the nb of users who accessed the book

→ Uses of the Model

1) Prediction: For a new book, predict the expected nb of readers in the next 12 months
2) Description: Extract insights from data, such as trends or factors influencing book popularity.

5.5) Clustering

→ Data:

A large collection of books with attributes Title, Info, Full Text

→ Task:

Clustering: Automatically group books into clusters based on similarity in their features

↳ The model will identify patterns in the data that allow books with similar characteristics to be grouped together

→ Uses of the Model:

1) Description: Gain insights into how books can be categorized by similarity, uncovering trends or hidden patterns in the data

2) Recommendations: Recommend books based on their preferences by grouping similar books together

5.6) Outlier Analysis

→ Data:

Customer Transactions, with attributes like:

. Transaction Amount, Date, Additional features (customer ... ategory)

→ Task:

Outlier Detection: Automatically identify data points (transactions) that significantly deviate from the rest of the data

↳ Outlier could represent unusual behaviors, such as unusually high or low transaction amounts, or transactions that occure on rare dates

6- Data Mining Tasks

6.1) Descriptive Data Mining

. Uncover patterns that describe behavior

↳ Concept / Class description

↳ Pattern discovery

↳ Cluster analysis

6.2) Predictive Data Mining

. Predict behavior of new entities

↳ Classification

↳ Regression

↳ Outlier detection

6.3) Hybrid Approach

Use both descriptive (understanding data) and predictive (forecasting future outcomes) techniques together

6.4) Descriptive and Predictive Patterns

Some patterns can serve both purposes (e.g. describing and predicting customer behavior

7. Data Mining Applications

→ Identifying important groups of microorganisms in the human body

→ Classifying galaxies in the universe

→ Email Spam Filtering

→ Document Sentiment Analysis

→ Images / video processing

→ Audio / Voice processing

→ Recommender Systems

→ Black and white image colorization

→ Image Classification / Object Recognition

→ Description generation using deep neural networks.