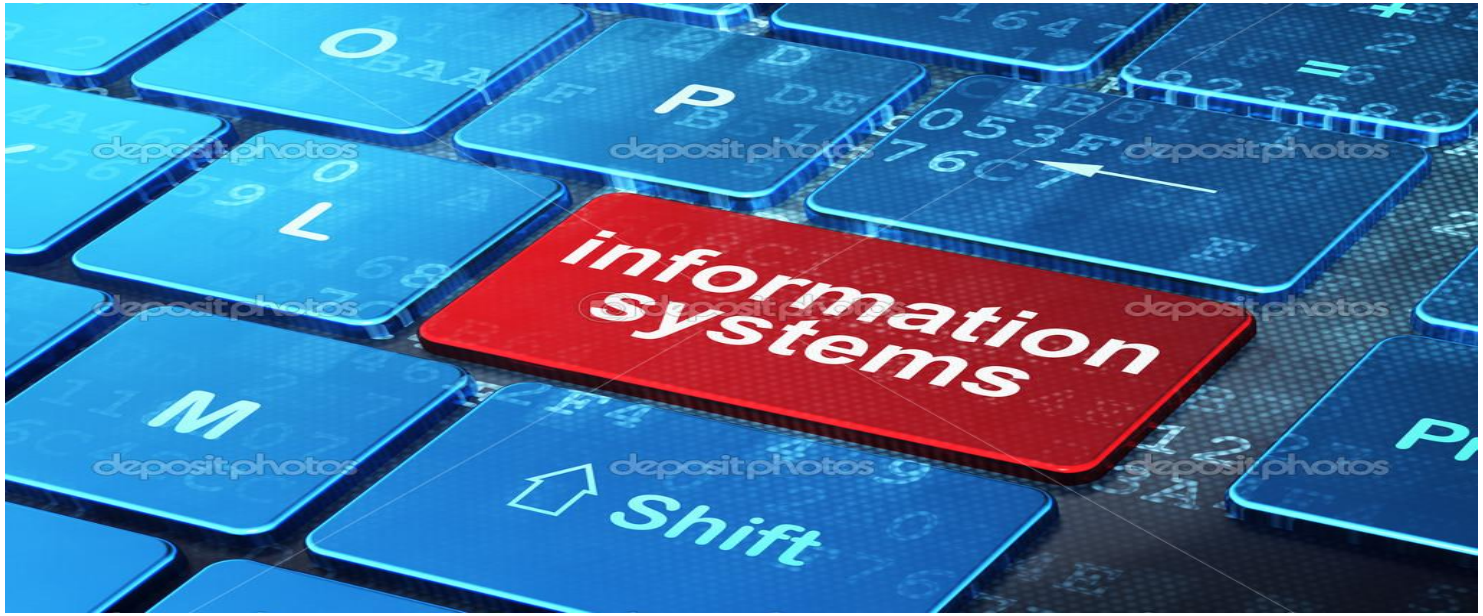
Advanced Topics in Information systems



SE204

Lecture 4

Dr. Nelly Amer

Data mining.

- What is data mining?
- Why data mining?
- Data Mining steps.
- Data analysis, Data Mining and Data science.
- Data Mining Models.
- Frequent pattern mining.

What is data mining?

Data mining is the process of discovering hidden patterns, correlations, trends, and knowledge from large amounts of data .

The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

Examples:

- In market basket analysis can discover that customers who buy bread are also likely to purchase milk.

- Watch a cooking video, and YouTube suggests more recipes.

- Facebook show ads based on what you like.

Data mining.

Not all information discovery tasks are considered to be data mining, Examples include queries, e.g., looking up individual records in a database or finding web pages that contain a particular set of keywords. This is because such tasks can be accomplished through simple interactions with a database management system or an information retrieval system. These systems rely on traditional computer science techniques, which include query processing algorithms, for retrieving information from data, **while data mining** focus on discovering **hidden, interesting, and useful patterns**, or **hidden correlations** in data, **not Simple queries, reports, or statistics that retrieve existing information**.

≠

Why data mining.

traditional data analysis techniques have often encountered practical difficulties in meeting the challenges posed by big data applications. The following are some of the specific challenges that motivated the development of data mining.

- Scalability
- High Dimensionality
- Heterogeneous
- Complex Data
- Data Ownership and Distribution
- Non-traditional Analysis

Why data mining.

Scalability

Because of advances in data generation and collection, data sets with sizes of terabytes, petabytes, or even Exabyte are becoming common. If data mining algorithms are to handle these massive data sets, they must be scalable.

Many data mining algorithms employ special search strategies to handle exponential search problems.

High Dimensionality

It is now common to encounter data sets with hundreds or thousands of attributes instead of the handful common a few decades ago, and Traditional data analysis techniques that were developed for low-dimensional data often do not work well for such high-dimensional data.

Examples:

1- In bioinformatics, progress in microarray technology has produced gene expression data involving thousands of features.

2- Data sets with spatial components also tend to have high dimensionality. For example, consider a data set that contains measurements of temperature at various locations. If the temperature measurements are taken repeatedly for an extended period, the number of dimensions (features) increases in proportion to the number of measurements taken

Why data mining.

Heterogeneous

Traditional data analysis methods often deal with data sets containing attributes of the same type, either continuous or categorical. As the role of data mining in business, science, medicine, and other fields has grown, so has the need for techniques that can handle heterogeneous attributes.

Complex Data

Recent years have also seen the emergence of more complex data objects. Examples of such non-traditional types of data include **web and social media** data containing text, hyperlinks, images, audio, videos, DNA data with complex structure, and climate data that consists of measurements (temperature, pressure, etc.) at various times and locations on the Earth’s surface.

Why data mining.

Data Ownership and Distribution

Sometimes, the data needed for an analysis is not stored in one location or owned by one organization. Instead, the data is geographically distributed among resources belonging to multiple entities. This requires the development of distributed data mining techniques. The key challenges faced by distributed data mining algorithms include the following:

(1) how to reduce the amount of communication needed to perform the distributed computation.

(2) how to effectively consolidate the data mining results obtained from multiple sources.

(3) how to address data security and privacy issues.

Why data mining.

Non-traditional Analysis

The traditional statistical approach is based on a hypothesize-and-test paradigm. In other words, a hypothesis is proposed, an experiment is designed to gather the data, and then the data is analyzed with respect to the hypothesis. Unfortunately, this process is extremely labor-intensive. Current data analysis tasks often require the generation and evaluation of thousands of hypotheses, and consequently, the development of some data mining techniques has been motivated by the desire to automate the process of hypothesis generation and evaluation.

Data Mining steps

There are seven steps as the following :

1- Data cleaning

Data cleaning: it is the process to handle noise, incomplete and inconsistent data.

Incomplete. When some of the attribute values are missed.

Noisy. When data contains errors or outliers, for example, some of the data points in a dataset may contain extreme values that can severely affect the dataset's range.

Inconsistent. refers to the lack of uniformity in data format or content, for example if records do not start with a capital letter, discrepancies are present.

2- **Data integration**

The multiple data sources are **combined from different databases**. It is the mixture of heterogeneous and homogeneous types of data which gets stored in data warehouse.

3-Data selection

Data relevant to the analysis task are retrieved from the database.

Data Mining steps

4-Data transformation

In this data transformed into forms appropriate for mining by performing summary or aggregation operations.

5-Data mining

An essential process where intelligent methods are applied in order to extract data patterns

6-Pattern evaluation

To identify the truly interesting patterns representing knowledge base on some interesting measures.

7-Knowledge

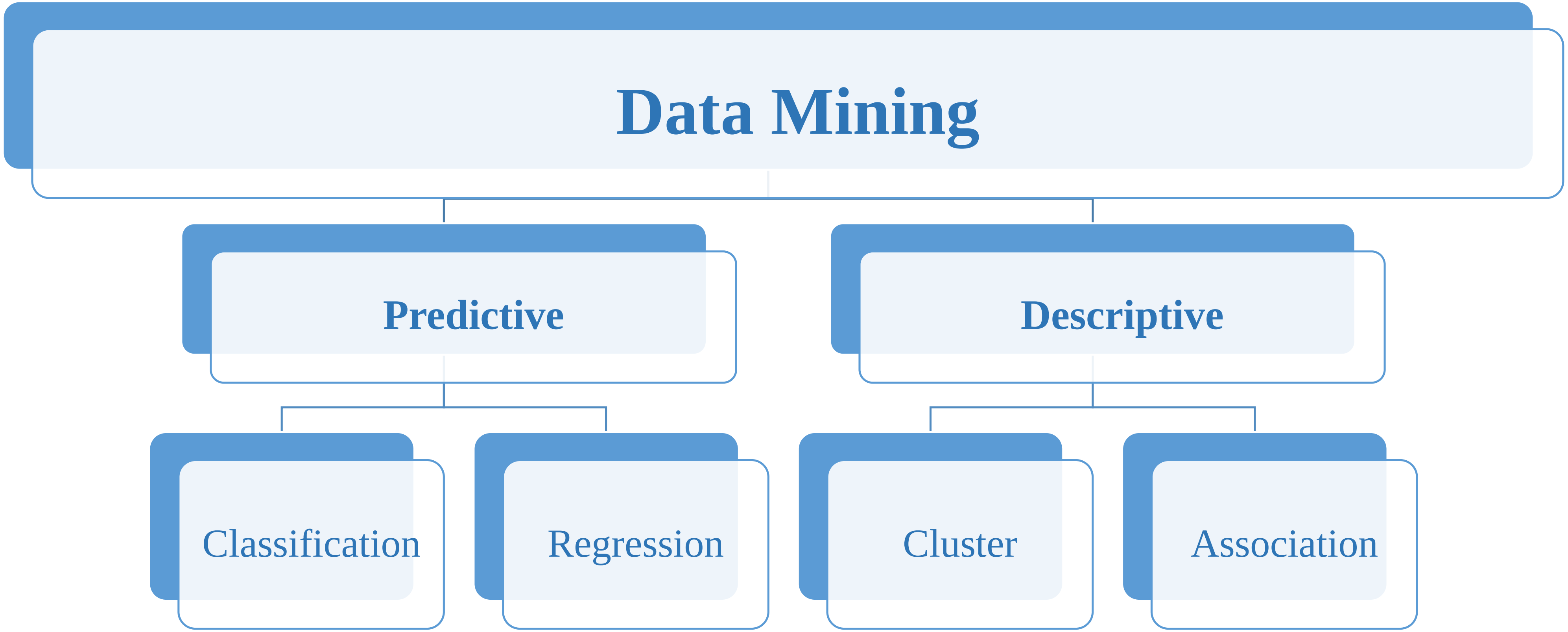
Where visualization & knowledge representation techniques are used to present the mined knowledge to the user.

Data analysis , Data Mining and Data science.

What is the difference?

Data Mining models.

Data mining tasks are generally divided into two major categories:



Data Mining

Predictive

Descriptive

Classification

Regression

Cluster

Association

Data Mining models

Predictive models:

The objective of these models is to **predict the value of a particular attribute based on the values of other attributes**. The attribute to be predicted is commonly known as the target or **dependent** variable, while the attributes used for making the prediction are known as the **independent** variables.

There are **two types of predictive modeling tasks**: classification, which is used for discrete target variables, and regression, which is used for continuous target variables.

Descriptive models

the objective is to find **human-interpretable patterns** (correlations, trends, clusters) that describe the data.

Predictive models

In **classification**, the goal is to assign input data to specific, predefined categories. The output classification is typically a label or a class from a set of predefined options.

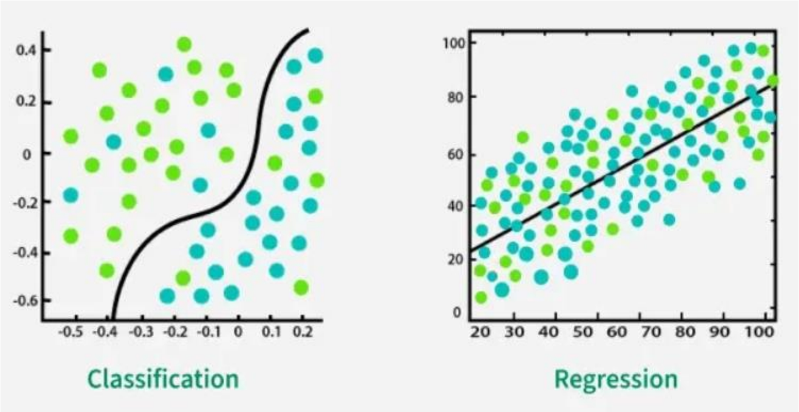
In **regression**, the goal is to establish a relationship between input variables and the output. The output in regression is a real-valued number that can vary within a range.

Both approaches **require labeled data for training but differ in their objectives**

classification aims to find decision boundaries that separate classes, whereas regression focuses on finding the best-fitting line to predict numerical outcomes.

Examples on classification:

It can determine whether an email is spam or not, classify images as “cat” or “dog,” or predict weather conditions like “sunny,” “rainy,” or “cloudy.” with decision boundary , classify IRIS flowers.



Classification

Regression

Examples on regression:

models are used to predict house prices based on features like size and location, or forecast stock prices over time with straight fit line.

Descriptive Model

The descriptive model classifies customers or prospects into groups based on the analyzing relationship between the data as in Clustering, Summarization, Association rule, Sequence discovery, etc.

Clustering:

A cluster is a group of similar objects or data points that share common characteristics, clustering means Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

Examples:

Group genes and proteins that have similar functionality

Descriptive Model

Association:

Association is used to discover patterns that describe strongly associated features in the data. The discovered patterns are typically represented in the form of implication rules or feature subsets. Because of the exponential size of its search space, the goal of association analysis is to extract the most interesting patterns in an efficient manner. Useful applications of association analysis include finding groups of genes that have related functionality, identifying web pages that are accessed together, or understanding the relationships between different elements of Earth’s climate system.

Example (Market Basket Analysis).

For example, we may discover the rule {Diapers}→{Milk}, which suggests that customers who buy diapers also tend to buy milk.

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Frequent Pattern

A **frequent pattern** are patterns that occur frequently in data.

Types of Frequent Patterns:

Frequent Itemsets : **A group of items that appear together** often in transactions.
Example: In a supermarket, **{Milk, Bread}** frequently appear in customer purchases.

Frequent Sequential Patterns : A sequence of **events** that frequently occur in a specific order.
Example: A customer often buys **Laptop** → **Mouse** → **Keyboard** in that order.

Frequent structures –Repeated patterns in chemical and biological data, XML data, software program traces, and Web browsing behaviors , and GPS data.

We will focus on frequent Itemsets

Frequent Pattern mining

Frequent patterns mining: is a Data Mining subject with the objective of **extracting frequent itemsets from a dataset**.

It leads to the discovery of interesting associations and correlations within data.

The uses of frequent pattern mining:

Recommendation Systems (e.g., Amazon’s "Customers also bought...")

Market Basket Analysis (e.g., Which items are bought together?)

Fraud Detection (e.g., Finding unusual spending patterns)

Medical Diagnosis (e.g., Common symptoms in a disease)

Basic concepts

Itemset

A collection of one or more items

Example: {Milk, Bread, Diaper}

k-itemset

An itemset that contains k items

Support count (σ)

Frequency of occurrence of an itemset

σ({Milk, Bread, Diaper}) = 2

Support

Frequency of occurrence of an itemset / all transaction

s({Milk, Bread, Diaper}) = 2/5

TID	*Items*
1	**Bread, Milk**
2	**Bread, Diaper, Beer, Eggs**
3	**Milk, Diaper, Beer, Coke**
4	**Bread, Milk, Diaper, Beer**
5	**Bread, Milk, Diaper, Coke**

Frequent Itemset

Frequent Itemset

An itemset whose support is greater than or equal to a minsup threshold

Example:

As seen in the following transaction table, given the minsup = 60%, which of the following itemset is frequent:

{Bread}, {Eggs}, {Bread, Diaper}, {Bread, Milk}, {Bread, Milk, Diaper}

TID	*Items*
1	**Bread, Milk**
2	**Bread, Diaper, Beer, Eggs**
3	**Milk, Diaper, Beer, Coke**
4	**Bread, Milk, Diaper, Beer**
5	**Bread, Milk, Diaper, Coke**

Solution

Itemset	Support
{Bread}	4/5=80%
{Eggs}	1/5=20%
{Bread, Diaper}	3/5=60%
{Bread, Milk}	3/5=60%
{Bread, Milk, Diaper}	2/5=40%

Since minsup = 60%, then {Bread}, {Bread, Diaper}, {Bread, Milk} are frequent itemsets

, and {Eggs}, {Bread, Milk, Diaper} are not frequent itemsets

Superset

A superset is a set that contains **all elements of another set**, along with possibly more elements.

If an **itemset X** contains all elements of **itemset Y**, then **X is a superset of Y**.

Example. 1:

Given the set {A, B, C, D}, what are the supersets of {C}

Solution

supersets of {C} are:

{C, A}, {C, B} , {C, D}
{C, A, B}, {C, A, D}, {C, B, D}
{C, A, B, D}

Example. 2:

Given the set {A, B, C, D}, what are the supersets of {A, B}

Solution

supersets of {A, B} are:

{A, B, C}, {A, B, D} , {A, B, C, D}

closed frequent itemset

A **closed frequent itemset:** is a frequent **itemset** that doesn’t have a **superset** with the **same support** in the dataset.

Given a frequent itemset, how do you check if it is closed frequent Item set or not, follow the following steps:

1-Find all supersets of this frequent itemset

2-Compute the support of each superset (support=frequency of the set/the number of transactions)

3-

if any superset (larger itemset) has the same support, then the frequent itemset is not closed.

if all supersets (larger itemset) have not the same support, then the frequent itemset is closed.

closed frequent itemset example.1

Given the transactions in the table, determine if the frequent itemset {**Milk**} is closed frequent itemset or not

Solution:

1- find all supersets of the frequent itemset {Milk} :
{Milk, Bread}, {Milk, Butter}, {Milk, Cheese} ,
,{Milk, Bread, Butter} ,{Milk, Bread, Cheese},{Milk, Butter, Cheese},
{Milk, Bread, Butter, Cheese}

Transaction ID	Milk	Bread	Butter	Cheese
1	✓	✓	✓	✗
2	✓	✓	✗	✓
3	✓	✓	✓	✗
4	✗	✓	✓	✓

2- compute Support for each superset:

Superset	support
{Milk, Bread}	3/4=75%
{Milk, Butter}	2/4=50%
{Milk, Cheese}	1/4=25%
{Milk, Bread, Butter}	2/4=50%
{Milk, Bread, Cheese}	1/4=25%
{Milk, Butter, Cheese}	0/4=0%
{Milk, Bread, Butter, Cheese}	0/4=0%

3-the support of S({**Milk**})**= 3/4=75%, it is clear that** S({**Milk**})**=** S({**Milk, Bread**})**=75%, Then the frequent itemset** {**Milk**} **is not closed .**

closed frequent itemset example. 2

Given the transactions in the following table, determine if the frequent itemset **{Bread}** is closed frequent itemset or not

Transaction ID	Milk	Bread	Butter	Cheese
1	✓	✓	✓	✕
2	✓	✓	✕	✓
3	✓	✓	✓	✕
4	✕	✓	✓	✓

Solution:

1- find all supersets of the frequent itemset {Bread} :

{Bread, Milk}, {Bread, Butter}, {Bread, Cheese} ,
,{Bread, Milk, Butter} ,{Bread, Milk, Cheese},{Bread, Butter, Cheese},
{Milk, Bread, Butter, Cheese}

2- compute Support for each superset:

Superset	support
{Bread, Milk}	3/4=75%
{Bread, Butter}	3/4=75%
{Bread, Cheese}	2/4=50%
{Bread, Milk, Butter}	2/4=50%
{Bread, Milk, Cheese}	1/4=25%
{Bread, Butter, Cheese}	1/4=0%
{Milk, Bread, Butter, Cheese}	0/4=0%

3-the support of S({**Bread})= 4/4=100 %, it is clear that all supersets have different support**

Then frequent itemset({ Bread} is closed

closed frequent itemset examples.3

Given the transactions in the table, determine if the frequent itemset {**Milk, Bread}** is closed frequent itemset or not

Transaction ID	Milk	Bread	Butter	Cheese
1	✓	✓	✓	✕
2	✓	✓	✕	✓
3	✓	✓	✓	✕
4	✕	✓	✓	✓

Solution:

1- find all supersets of the frequent itemset {Milk, Bread} :

{Milk, Bread, Butter}, {Milk, Bread, Cheese} , {Milk, Bread, Butter, Cheese}

2- compute Support for each superset:

Superset	support
{Milk, Bread, Butter}	2/4=50%
{Milk, Bread, Cheese}	1/4=25%
{Milk, Bread, Butter, Cheese}	0/4=0%

3-the support of S({**Milk, Bread})= 3/4=75%, it is clear that all supersets have different support**

Then frequent itemset({Milk, Bread} is closed .

Max frequent itemset

A **maximal frequent pattern** (or **max frequent itemset**): is a **frequent itemset that has no superset** that is also frequent.

Given a frequent itemset, how do you check if it is max frequent Item set or not, follow the following steps:

1-Find all supersets of this itemset

2-Compute the support of each superset (support=frequency of the set/the number of transactions)

3-

Checks all superset (larger itemset):

if any superset (larger itemset)is also frequent (i.e. its support >= minsupport), then the frequent itemset is

not max.

if all supersets (larger itemset) are not frequent, then the frequent itemset is max.

Max frequent itemset example. 1

Given the transactions in the table, **and minsupport=50%** determine if the frequent itemset {**Milk, Bread}** is Max frequent itemset or not

Transaction ID	Milk	Bread	Butter	Cheese
1	✓	✓	✓	✗
2	✓	✓	✗	✓
3	✓	✓	✓	✗
4	✗	✓	✓	✓

Solution:

1- find all supersets of the frequent itemset {Milk, Bread} :

{Milk, Bread, Butter}, {Milk, Bread, Cheese} , {Milk, Bread, Butter, Cheese}

2- compute Support for each superset:

Superset	support
{Milk, Bread, Butter}	2/4=50%
{Milk, Bread, Cheese}	1/4=25%
{Milk, Bread, Butter, Cheese}	0/4=0%

3- it is clear that the itemset **{Milk, Bread, Butter} is frequent, (because S({Milk, Bread, Butter}=50%=minsupport)**
Then the frequent itemset **{Milk, Bread}** is not max

Max frequent itemset example. 2

Given the transactions in the table, and **minsupport=60%** determine if the frequent itemset **{Milk, Bread}** is Max frequent itemset or not

Transaction ID	Milk	Bread	Butter	Cheese
1	✓	✓	✓	✗
2	✓	✓	✗	✓
3	✓	✓	✓	✗
4	✗	✓	✓	✓

Solution:

1- find all supersets of the frequent itemset {Milk, Bread} :

{Milk, Bread, Butter}, {Milk, Bread, Cheese} , {Milk, Bread, Butter, Cheese}

2- compute Support for each superset:

Superset	support
{Milk, Bread, Butter}	2/4=50%
{Milk, Bread, Cheese}	1/4=25%
{Milk, Bread, Butter, Cheese}	0/4=0%

3- it is clear that all supersets are not frequent (since all their support is less than the minsupport , then the frequent itemset {Milk, Bread}is max.

Max frequent itemset example. 3

Given the transactions in the table, and minsupport=50% determine if the frequent itemset {**Milk, Bread, Butter**} is Max frequent itemset or not

Transaction ID	Milk	Bread	Butter	Cheese
1	✓	✓	✓	✗
2	✓	✓	✗	✓
3	✓	✓	✓	✗
4	✗	✓	✓	✓

Solution:

1- find all supersets of the frequent itemset {Milk, Bread, Butter} :

{Milk, Bread, Butter, Cheese}

2- compute Support for each superset:

Superset	support
{Milk, Bread, Butter, Cheese}	0/4=0%

3- it is clear that all supersets are not frequent, then the frequent itemset

{**Milk, Bread, Butter**} is max.

Max frequent itemset example.4

Given the transactions in the table, and minsupport=50% determine if the frequent itemset { **Bread, Cheese}** is Max frequent itemset or not

Transaction ID	Milk	Bread	Butter	Cheese
1	✓	✓	✓	✗
2	✓	✓	✗	✓
3	✓	✓	✓	✗
4	✗	✓	✓	✓

Solution:

1- find all supersets of the frequent itemset {Bread, Cheese} :

{ **Bread, Cheese**, Milk}, { **Bread, Cheese** Butter} , {Milk, Bread, Butter, Cheese}

2- compute Support for each superset:

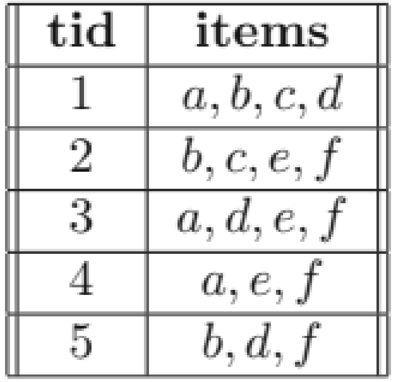
Superset	support
{ Bread, Cheese, Milk}	1/4=25%
{ Bread, Cheese Butter}	1/4=0%
{Milk, Bread, Butter, Cheese}	0/4=0%

3- it is clear that all supersets are not frequent, then the frequent itemset { **Bread, Cheese}** is max.

Assignment #4

1- What is the difference between data analysis, data mining, and data science?

2- Given the following transactions, and the minimum support=60%, determine which of the following itemsets are frequent {a}, {c}, {f}, {e, f}, {a, e, f}



tid	items
1	a, b, c, d
2	b, c, e, f
3	a, d, e, f
4	a, e, f
5	b, d, f

3-Given the transactions in the table, determine if the frequent itemset **{A, B}** is closed frequent itemset or not

Transaction ID	A	B	C	D
1	✓	✓	✓	✗
2	✓	✓	✗	✓
3	✓	✓	✓	✗
4	✗	✓	✓	✓

Assignment #4 cont.

4-Given the transactions in the table, and **minsupport=60%** determine if the frequent itemset **{A, B}** is Max frequent itemset or not

Transaction ID	A	B	C	D
1	✓	✓	✓	✗
2	✓	✓	✗	✓
3	✓	✓	✓	✗
4	✗	✓	✓	✓

References.

[1] P. N. Tan, M. Steinbach , A. Karpatne , V. Kumar, “introduction to data mining”, (2nd edition) , Pearson, 2018.

[2] J. Han, and M. Kamber, “ data mining: concepts and techniques”, 2011.

[3] J. Aguilar-Ruiz, D. Rodríguez -Baena, R. Alves, “frequent pattern mining.” in: W. Ddubitzky, O. Wolkenhauer, K. Hyun Cho, H. Yokota. (eds), encyclopedia of systems biology. springer, 2013.

[4] C. C. Aggarwal, “association pattern mining”, in: “data mining”, springer, 2015.

[5] M. Sharma, “Data Mining Prediction Techniques in Health Care Sector”, Journal of Physics: Conference Series, vol. 2267, pp. 1-9, 2021.