1. What is data mining? In your answer, address

the following:

Answer: Data mining refers to the process or method that extracts or \mines interesting knowledge or patterns from large amounts of data..

(a) Is it another hype?

Answer: Data mining is not another hype. Instead, the need for data mining has arisen due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge.

(b) Is it a simple transformation of technology developed from databases ,statistics, and machine learning?

Answer: No. Data mining is more than a simple transformation of technology developed from databases, statistics, and machine learning.

(d) Describe the steps involved in data mining when viewed as a process of knowledge discovery.

Answer: The steps involved in data mining when viewed as a process of knowledge discovery are as follows:

-Data cleaning, a process that removes or transforms noise and inconsistent data.

-Data integration, where multiple data sources may be combined.

-Data selection, where data relevant to the analysis task are retrieved from the database

-Data transformation, where data are transformed or consolidated into forms appropriate for mining.

- Data mining, an essential process where intelligent and efficient methods are applied in order to extract patterns

- Pattern evaluation, a process that identifies the truly interesting patterns representing knowledge based on some interestingness measures.

-Knowledge presentation, where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

2. how is a data warehouse different from a database how are they similar?

Answer:

Operational Database Systems	Data Warehouses
Operational systems are generally designed to support high-volume transaction processing.	Data warehousing systems are generally designed to support high-volume analytical processing. (i.e. OLAP).
Operational systems focuses on Data in.	Data warehousing systems focuses on Information out.
In Operational systems data is stored with a functional or process orientation.	In Data warehousing systems data is stored with a subject orientation.
Performance is low for analysis queries.	Performance is high for analysis queries.
It is used for Online Transactional Processing (OLTP)	It is used for Online Analytical Processing (OLAP).
Operational systems represent current transactions.	Data warehousing systems reads the historical data.
Data within operational systems are generally updated regularly.	Data within a data warehouse is non-volatile, meaning when new data is added old data is not erased so rarely updates.
Complex data structures.	Multi dimensional data structures.

3. Define each of the following data mining functionalities: characterization, discrimination, association and correlation analysis, classification, regression, clustering, and outlier analysis. Give examples of each data mining functionality, using a real-life database that you are familiar with?

Answer:

Data Characterization: This refers to the summary of general characteristics

or features of the class that is under the study. For example. To study the characteristics of a software product whose sales increased by 15% two years ago, anyone can collect these type of data related to such products by running SQL queries.

Data Discrimination: It compares common features of class which is under study. The output of this process can be represented in many forms. Eg., bar charts, curves and pie charts.

5. Explain the difference and similarity between discrimination and classification, between characterization and clustering, and between classification and regression?

Discrimination vs. classification:

Data discrimination is a comparison of the general features of a target class data objects with the general features of objects from one or a set of contrasting classes.

Classification is the process of finding a set of models that describe and distinguish data classes or concepts,

Characterization vs. clustering:

Data characterization is a summarization of the general characteristics or features of a target class of data. In clustering the objects are grouped together based on the principle of maximizing theintraclass similarity and minimizing the interclass similarity

توصيف البيانات هو تلخيص للخصائص أو السمات العامة لفئة البيانات المستهدفة. في التجميع ، يتم تجميع الكائنات معًا بناءً على مبدأ تعظيم التشابه داخل الطبقة وتقليل التشابه بين الفئات

Lecture 2

A decision tree is a flowchart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves
represent classes or class distributions.

Cluster Analysis

- clustering analyzes data objects without consulting class labels
- Clustering can be used to generate class labels for a group of data.
- cluster have high similarity in comparison to one another,

Outlier analysis

Outlier: A data object that does not comply with the general behavior of the data

Rather than using statistical or distance measures, density-based methods may identify outliers in a local region

Outliers may be detected using statistical tests that assume a distribution or probability model for the data

- A statistical model is a set of mathematical functions that describe the behavior of the objects in a target

❑ النموذج الإحصائي هو مجموعة من الوظائف الرياضية التي تصف سلوك كائنات في فئة مستهدفة

Statistical methods can also be used to verify data mining results.

Machine learning : investigates how computers can learn (or improve their performance) based on data

يبحث في كيفية تعلم أجهزة الكمبيوتر (أو تحسينها الأداء) على أساس البيانات

❑ Problems types

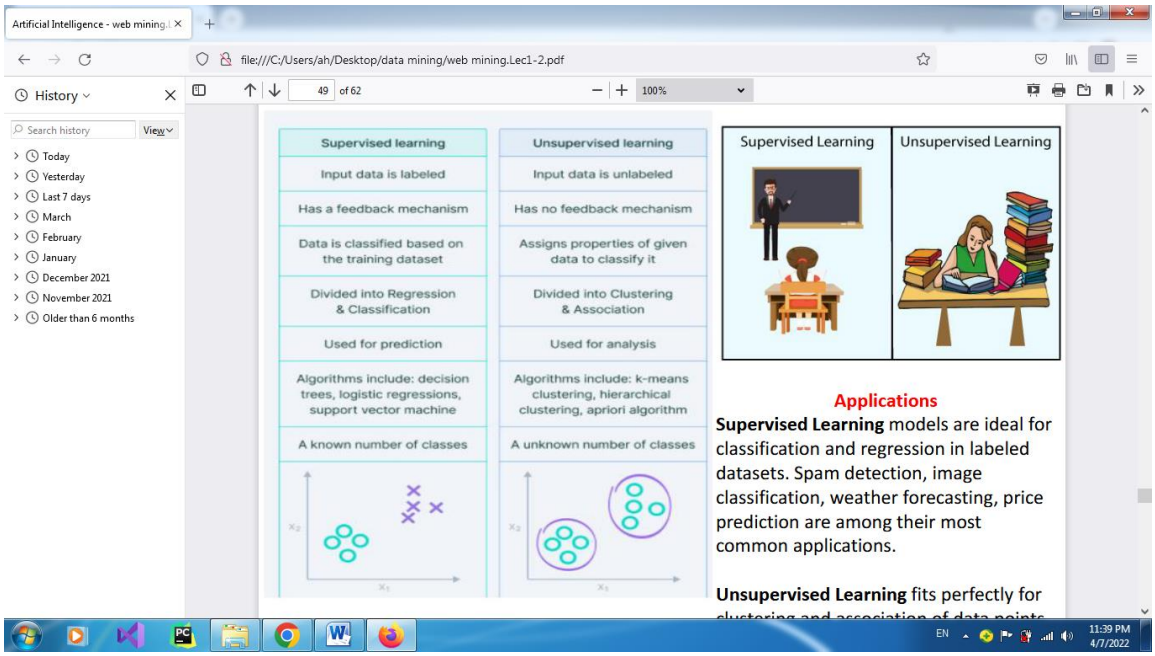
1. Supervised learning is basically a synonym for classification. The supervision in the learning comes from the labeled examples in the training data set

2. Unsupervised learning is essentially a synonym for clustering. The learning process is unsupervised since the input examples are not class labeled.
3. Semi-supervised learning is a class of machine learning techniques that make use of both labeled and unlabeled examples when learning a model.
4. Active learning is a machine learning approach that lets users play an active role in the learning process.

Information retrieval (IR) is the science of searching for documents or information in documents

Applications of Data Mining

- Web page analysis: classification, clustering, ranking
- Collaborative analysis & recommender systems
- Basket data analysis
- Biological and medical data analysis
- Data mining and software engineering
- Data mining and text analysis
- Data mining and social and information network analysis



Lecture 3 and 4

Social Network Analysis (SNA)
- views social relationships in terms of network theory consisting of nodes and ties (also called edges, links or connections).

• A node or vertex is an individual unit in the graph or system.
• A graph or system or network is a set of units that may be connected to each other.

• A neighborhood is the set of its immediately connected nodes.

Degree: The degree of a vertex or node is the number of other nodes in its neighborhood.

• directed graph or network : the edges are reciprocal—so if A is connected to B, B is by definition connected to A. لازم يكونو متبادلين

I can assign the weight

• I undirected graph or network : the edges are not necessarily reciprocal—A may be connected to B, but B may not be connected to A. مش لازم يكونو متبادلين

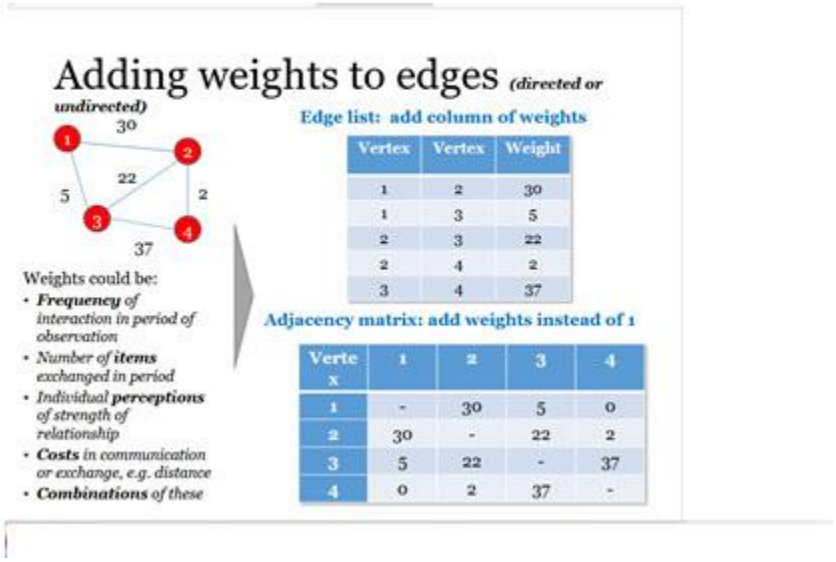
Can not assign the weight

graph is a set of nodes joined by a set of lines or edges.

G is an ordered pair G:=(V, E)
• V is a set of nodes, points, or vertices.
• E is a set, whose elements are known as edges or lines.

A collaboration network (CN) is a partnership of autonomous people and organizations, supported by a computer network, that work together to share resources, such as data and connectivity.

شراكة مستقلة بين الاشخاص و المؤسسات مدعومة بشبكة كمبيوتر لمشاركة البيانات



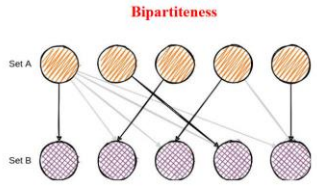
Graph Neural Networks (GNN)

• Machine learning methods are based on data

Three fundamental graph properties:
(1)Connectivity: A graph is connected if there is a path from any vertex to any other vertex in the graph.
(2) Bipartiteness
(3)triangle-free

Bipartiteness : A graph G = (V, E) is bipartite if its set of vertices V can be decomposed into two disjoint sets V1 and V2, i. e., V = V1 ∪ V2, such that every edge e ∈ E connects a vertex in V1 to a vertex in V2

تكون ثنائية & لو عندنا مجموعة نقاط نقسمها الي مجموعتين بحيث كل عنصر من المجموعة الاولي طالع منها اسهم لكل عناصر المجموعة التانية



bipartite matching : is described as a set of edges that are picked in a way to not share an endpoint

المطابقة الثنائية: توصف بأنها مجموعة من الحواف المنتقاة بطريقة لا تسمح بمشاركة نقطة نهاية

triangle-free : if a graph is triangle-free if it does not contain a triangle (a cycle of three vertices)

يكونو متصلين لكن ميكونش في مثلثات

• Path length: number of edges in the shortest path between two nodes
• k-hop neighborhood of a node: the number of nodes that can be reached through paths of length

مجموعة العقد التي يمكن أن تكون تم الوصول إليها من خلال مسارات بطول ك

Graph neural networks (GNNs) : a powerful architecture for learning node and graph representations.

k-hop GNNs algorithm

Algorithm 1: k-hop GNN

Input: Graph $G = (V, E)$, node features $\{h_v : v \in V\}$, number of neighborhood aggregation layers T, number of hops k

Output: Node features $\{h_v^{(T)} : v \in V\}$

for $t \in \{1, \ldots, T\}$ **do**
  for $v \in V$ **do**
    for $u \in R_k(v)$ **do**
      $\mathcal{D} \longleftarrow \mathcal{N}_1(u) \cap R_k(v)$
      $x_u \longleftarrow \text{UPDATE}^{(t)}_{k,within}(u, \mathcal{D})$
    end for
    for $i \in \{k-1, \ldots, 1\}$ **do**
      for $u \in R_i(v)$ **do**
        $\mathcal{B} \longleftarrow \mathcal{N}_1(u) \cap R_{i+1}(v)$
        $x_u \longleftarrow \text{UPDATE}^{(t)}_{i,across}(u, \mathcal{B})$
        $\mathcal{D} \longleftarrow \mathcal{N}_1(u) \cap R_i(v)$
        $x_u \longleftarrow \text{UPDATE}^{(t)}_{i,within}(u, \mathcal{D})$
      end for
    end for
    $h_v^{(t)} = \text{UPDATE}^{(t)}_{0,across}(v, \mathcal{N}_1(v))$
  end for
end for

Main Measures for Social Network are:

1. Degree Centrality: The number of direct connections a node has. node with high degree centrality have the best connection to those around them

الاكثر درجة لها افضل اتصال بمن حولها

Degree centrality is the simplest measure of node connectivity

ابسط مقياس لاتصال نود

When to use it: For finding very connected individuals

2.Betweenness Centrality: the number of times a node lies on the shortest path between other nodes.

عدد المرات التي تقع فيها العقدة على أقصر طريق بين العقد الأخرى.

use it: For finding the individuals who effected around a system.

A node with high betweenness has great influence over what flows in the network .

للعقدة ذات البينية العاليةلها تأثير كبير على ما يتدفق في الشبكة

3.Closeness Centrality: The measure of closeness of a node which are close to everyone else مركزية القرب: مقياس القرب من عقدة قريبة من أي أحد غيره

Help to find the nodes that are closest to the other nodes in a network ,based on their ability to reach them بناء علي قدرتها للوصول اليها

use it: For finding the individuals who are best placed
to effect the entire network

geodesics is : the shortest path between any particular pair of nodes in a network.