MUST

Wisdom & Virtue

MIRPUR UNIVERSITY OF SCIENCE AND TECHNOLOGY (MUST), MIRPUR
DEPARTMENT OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY

DATA MINING

BCS-3605

Lecture 04

Dr Yasir Mehmood

(Assistant Professor)

Agenda of Today’s Lecture

- ***What Kinds of Pattern Can Be Mined?***
  - 3- Classification and Regression
  - 4- Cluster Analysis
  - 5- Outlier Analysis
- ***What Technologies are used?***
- ***Why Confluence of Multiple Disciplines?***
- ***Applications of Data Mining***

3- Classification and Regression

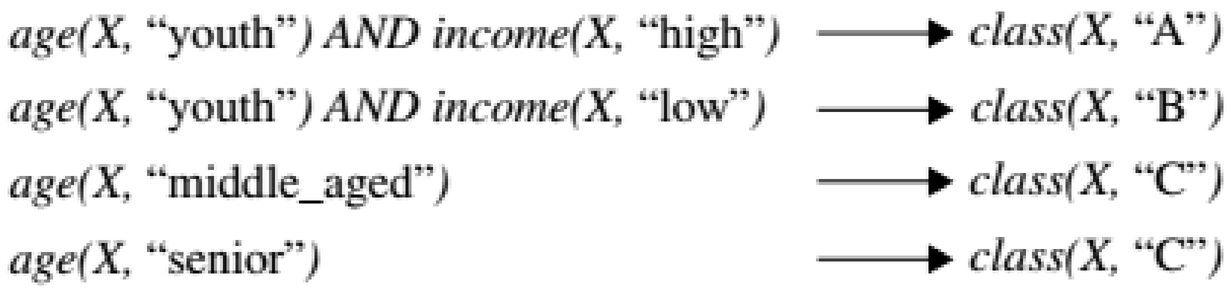
- **Classification** is the process of finding a **model** (or function) that describes and distinguishes data classes or concepts.
- Classification predict the categorical (discrete, ordered) labels
  - Construct models (functions) based on some training examples.
  - Describe and distinguish classes or concepts for future prediction

    e.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown class labels

- Typical methods
  - Rule-based classification, decision trees, neural networks, naïve Bayesian classification, support vector machines, *K*-nn, pattern-based classification, logistic regression, …

3- Classification and Regression: Cont’d

- The derived model may be represented in various forms, such as *classification rules* (i.e., *IF-THEN rules*), *decision trees, mathematical formulae,* or *neural networks*



age(*X*, “youth”) *AND income*(*X*, “high”) → *class*(*X*, “A”)

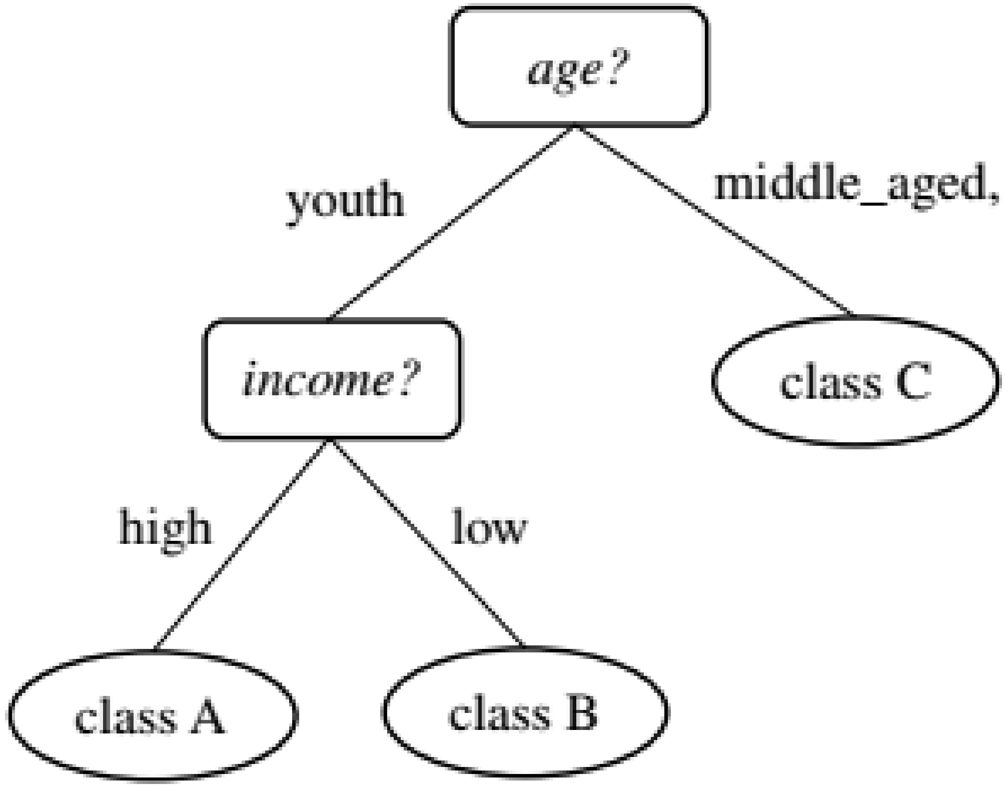
age(*X*, “youth”) *AND income*(*X*, “low”) → *class*(*X*, “B”)

age(*X*, “middle_aged”) → *class*(*X*, “C”)

age(*X*, “senior”) → *class*(*X*, “C”)

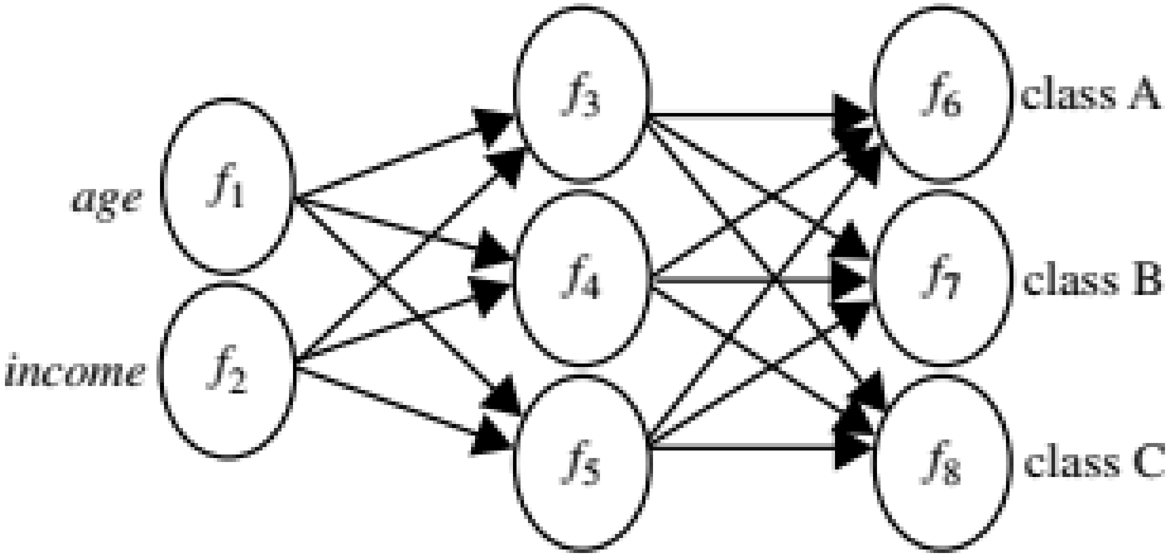
3- Classification and Regression: Cont'd

- **Decision Tree**: Flowchart-like tree structure
- Node denotes a test on an attribute value
- Branch represents an outcome of the test
- Leaves represent classes or class distributions.
- Decision tree classify the person as Class A, B, or C using age and income attribute.



3- Classification and Regression: Cont'd

- A **neural network**, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units.

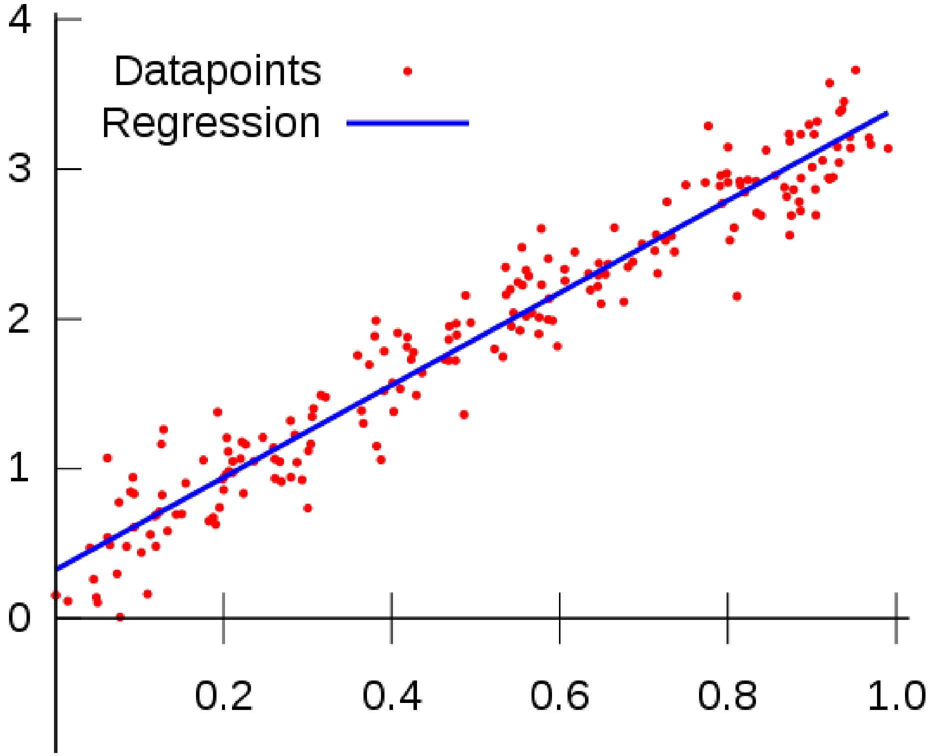


-

3- Classification and Regression: Cont’d

- **Regression** predicts the continuous-valued function
  - Predict missing or unavailable data rather than class labels
  - Statistical method for numeric prediction
  - Identify the distribution trends of available data
    - Regression predicts the amount of revenue that each item will generate.

- E.g Pg #56 [Data Mining Concepts and Techniques 3rd Edition]

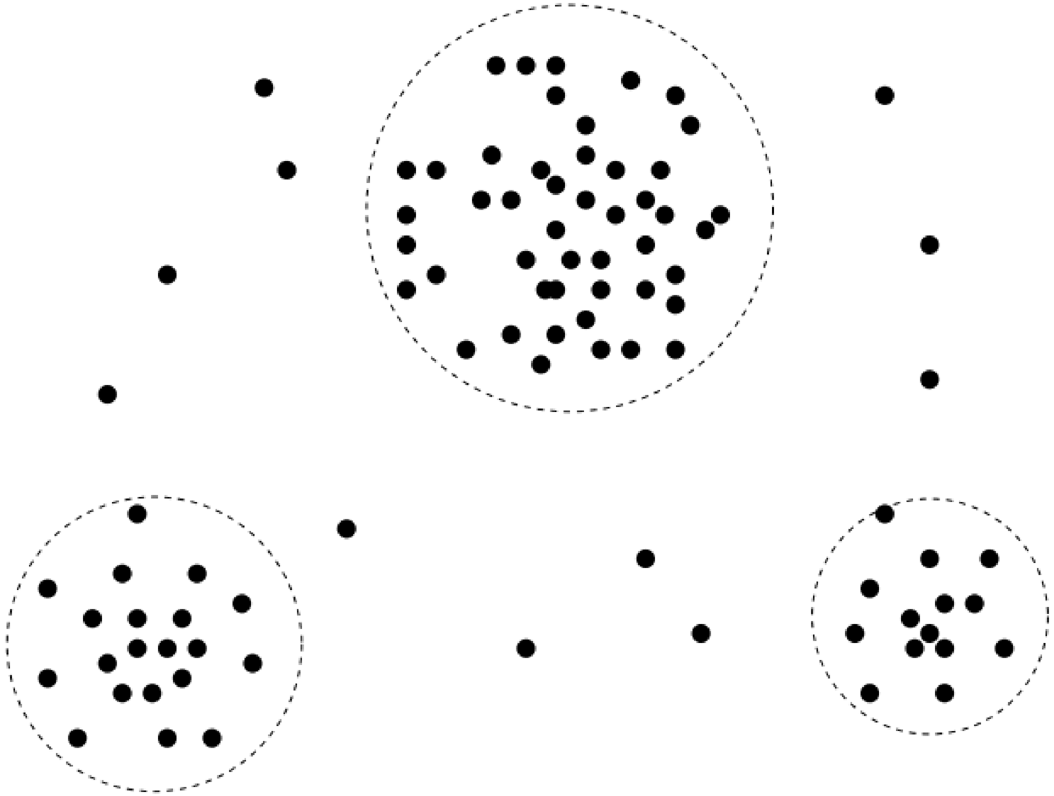


3- Classification and Regression: Cont’d

- Classify the descriptive features of the items, such as *price*, *brand*, *place made*, *type*, and *category* into three classes *good response*, *mild response* and *no response*.
- Suppose classification is expressed as a decision tree which identify *price* as being the single factor that best distinguishes the three classes. In addition to *price*, other features include *brand* and *place made* help to further distinguish objects of each class from one another.
- To predict the amount of revenue that each item will generate during an upcoming sale at some store, based on the previous sales data. This is an example of regression analysis because the regression model constructed will predict a continuous function

4- Cluster Analysis

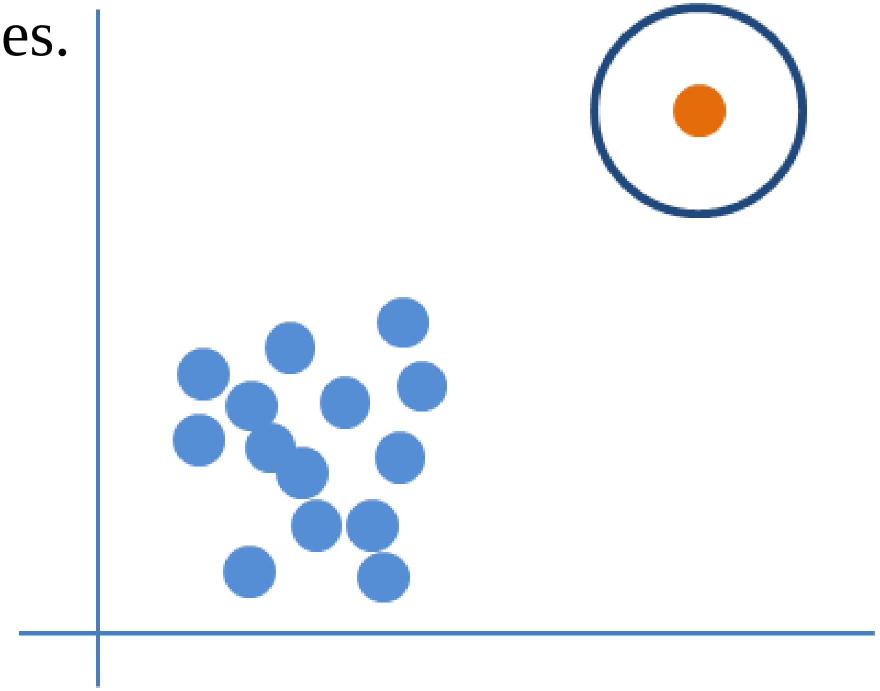
- Unsupervised learning
- Analyzes data objects without consulting class labels
- Group data to form new categories
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Objects within a cluster have high similarity and dissimilar in other clusters
- Taxonomy formulation



5- Outlier Analysis

- Outlier analysis or anomaly mining
  - A data object that does not comply with the general behavior of the data. e.g, noise and exception
  - Rare events are more interesting than the regular ones.

    e.g, in credit card fraud detection unusual purchase
  - Outlier values may also be detected with respect to the locations and types of purchase, or the purchase frequency
  - Methods: Statically measures, Distance measures, Density-based, by product of clustering or regression analysis, …



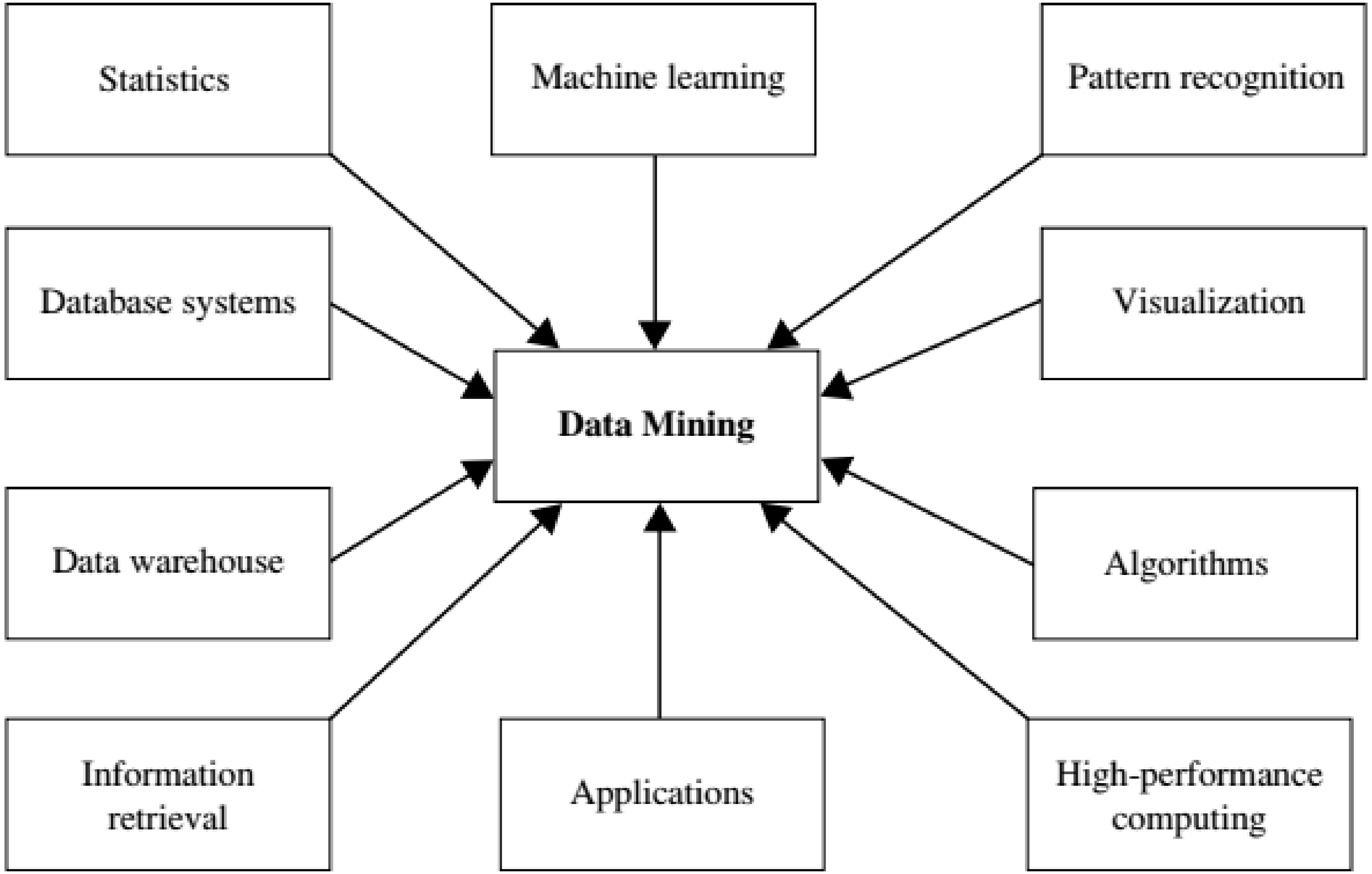
6- Are all pattern interesting?

- Are all mined knowledge interesting?
- What makes a pattern interesting?
  - Understandable by human.
  - Valid for new and test data
  - Useful
  - Novel
  - Validates a hypothesis
  - Represents knowledge
- Objective measures
  - Support and confidence for association rules
  - Accuracy and coverage for classification rules

6- Are all pattern interesting?

- Can a DM system generate all of the interesting patterns?
  - Unrealistic and inefficient to generate all patterns
  - Constraints and interestingness measures to focus the search
  - Association rule mining ensures the completeness.
- Can system generate only interesting ones?
  - Highly desirable task of DM system
  - Challenging optimization problem

What Technologies are used?



Why Confluence of Multiple Disciplines?

- DM is highly application-driven domain
- Tremendous amount of data
  - Algorithms must be highly scalable to handle tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

Applications of Data Mining

- *Where there is data, there is data mining applications*
- Knowledge-intensive application-driven domain
  - Business intelligence:
    - To understand customers, market, supply and competitors
    - Provides historical, current, and predictive view of business
  - BI help to compare customer feedback, analyze competitors, retain valuable customers, smart business decisions
- DM is a core of BI
  - **Predictive analysis** to analyzes markets, suppliers, and sales
  - **Clustering** to groups customers based on similarities
  - **Characterization** to understand features of customers group

Applications of Data Mining: Cont’d

- Web search engine: algorithmically maintain list of pages, images, files, adopts crawling, indexing, searching (rank, adds, personalized)
- Challenges:
  - How to handle huge and ever-growing data (clouds)
  - Deal with online data (query classifier)
  - Maintain and update model
  - Rear queries

Reference

Data Mining Concepts and Techniques Third Edition

1.4 What Kinds of Patterns Can Be Mined?

1.5 Which Technologies Are Used?

1.6 Which Kinds of Applications Are Targeted?

THANKS