`Unit-1: Part-1:

Data Mining:

Data mining is a way of using special computer programs to look through a lot of information at once, and find patterns that might be useful. It can help us make predictions about what might happen in the future based on what has happened in the past. But sometimes it's hard to find useful information because there's just so much of it! And sometimes the information is different from what we're used to, so we have to use different methods to analyze it.

Data Mining and Knowledge Discovery:

Data mining is a way of finding useful information from large amounts of data. It's part of a bigger process called knowledge discovery in databases, which is about turning raw data into useful knowledge.

To do this, data goes through a bunch of steps. First, it's preprocessed - this means cleaning it up and making sure it's in a format that the computer can understand. Then, data mining algorithms are used to find patterns and useful information. Finally, the results are postprocessed - this means summarizing the findings and presenting them in a useful way.

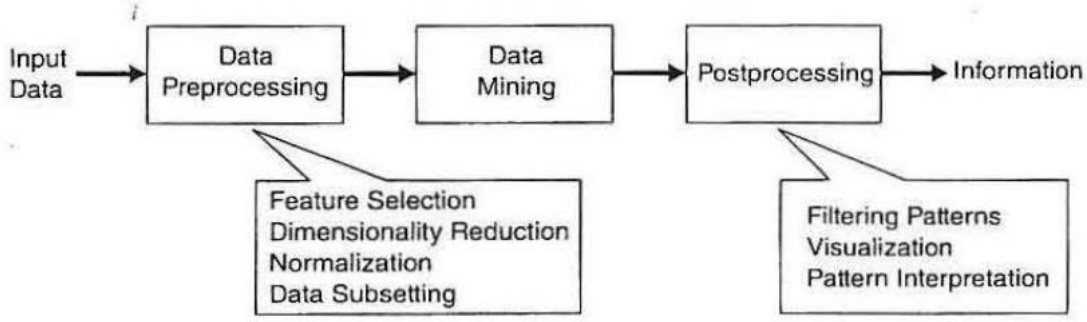


Figure 1.1. The process of knowledge discovery in databases (KDD).

Data can be stored in different formats like spreadsheets, files or tables and it can be in one place or spread out in different locations. Before analyzing the data, it needs to be prepared by a process called pre-processing.

Pre-processing helps to make sure that the data is in a suitable format for further analysis. This involves different steps like combining data from different sources, removing errors or duplicate entries, and selecting only the relevant parts of the data that are needed for the analysis.

Data preprocessing can take a lot of time because there are many ways data can be collected and stored, and it needs to be cleaned and organized before it can be analyzed properly.

Motivating Challenges:

1)Scalability:

Because of advances in data generation and collection, data sets with sizes of gigabytes, terabytes, or even petabytes are becoming common. lf data mining algorithms are to handle these massive data sets, then they must be scalable. Scalability can also be improved by using sampling or developing parallel and distributed algorithms.

2) 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. Data sets with temporal or 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.

Traditional data analysis techniques that were developed for low-dimensional data often do not work well for such highdimensional data.

3) Heterogeneous and Complex Data:

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.

Examples of such non-traditional types of data include collections of Web pages containing semi-structured text and hyperlinks; DNA data and climate data that consists of time series measurements.

4) 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.

5) Non-traditional Analysis:

When people use statistics to analyze data, they usually start by making a guess about what they think might be happening - this is called a hypothesis. Then they design an experiment to gather data and see if their hypothesis is true or not.

But this process can take a long time and it's difficult when there are many hypotheses to test. Nowadays, people need to analyze a lot of data and test thousands of hypotheses.

To help with this, special computer programs called "data mining techniques" have been developed. These programs can help automate the process of coming up with hypotheses and testing them, which makes it easier and faster to analyze a lot of data.

The origins of Data Mining:

To deal with the challenges of analyzing large and diverse data sets, researchers from different fields started working together to create better tools. This led to the development of data mining, which uses ideas from different fields like statistics, artificial intelligence, and machine learning.

Data mining also borrows ideas from other fields like optimization, evolutionary computing, and information retrieval. To help with the storage and processing of large data sets, techniques from database systems and high-performance computing are used. When the data is spread out in different locations, distributed techniques are needed to bring it all together.

By combining ideas from different fields, data mining has become an effective way to analyze complex data sets and discover useful information.

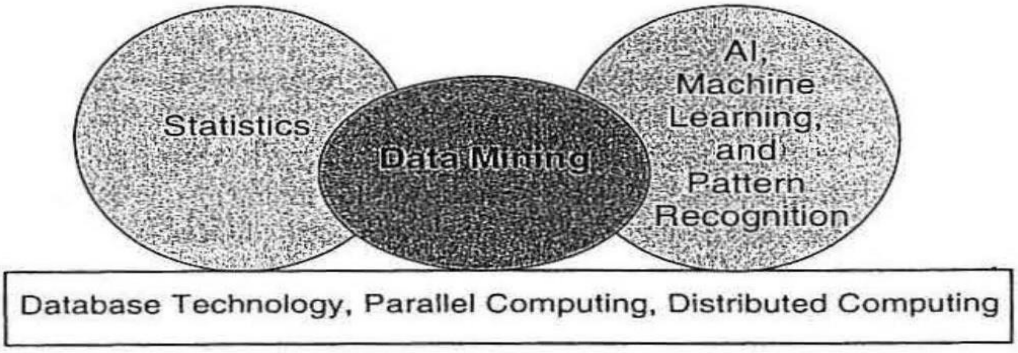


Figure 1.2. Data mining as a confluence of many disciplines.

Data Mining Tasks:

Data mining tasks are generally divided into two major categories:

1) Predictive tasks:

The objective of these tasks 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 explanatory or independent variables.

2) Descriptive tasks:

Here, the objective is to derive patterns (correlations, trends, clusters, trajectories, and anomalies) that summarize the underlying relationships in data. Descriptive data mining tasks are often exploratory in nature and frequently require postprocessing techniques to validate and explain the results.

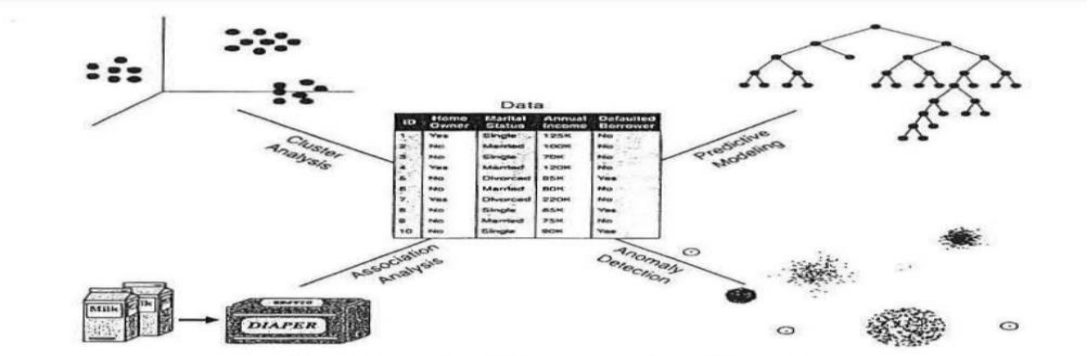


Figure 1.3. Four of the core data mining tasks.

3)Predictive modeling:

It refers to the task of building a model for the target variable as a function of the explanatory 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.Ex: Predicting the Type of a Flower

4)Association analysis:

It 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.Ex: Market Basket Analysis

5)Cluster analysis:

It seeks to find groups of closely related observations so that observations that belong to the same cluster are more similar to each other.Ex: Document Clustering

6)Anomaly detection:

It is the task of identifying observations whose characteristicsare significantly different from the rest of the data. Such observations are known as anomalies or outliers. The goal of an anomaly detection algorithm is to discover the real anomalies and avoid falsely labeling normal objects as anomalous. Ex: Credit CardFraud Detection

Unit-1: Part-2:

What is data:

- Collection of ***data objects*** and their ***attributes***
- An ***attribute*** is a property or characteristic of an object
    - Examples: eye color of a person, temperature, etc.

- Attribute is also known as variable, field, characteristic, dimension, or feature

- A collection of attributes describe an ***object***
  - Object is also known as record, point, case, sample, entity, or instance.

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute values:

- **Attribute values** are numbers or symbols assigned to an attribute for a particular object
- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
  - But properties of attribute can be different than the properties of the values used to represent the attribute

Measurement of Length :

- The way you measure an attribute may not match the attributes properties.

Types of Attributes :

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race)

Properties of Attribute Values:

- The type of an attribute depends on which of the following properties/operations it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Differences are meaningful : + -
  - Ratios are meaningful * /
  - Nominal attribute: distinctness

- Ordinal attribute: distinctness & order
- Interval attribute: distinctness, order & meaningful differences
- Ratio attribute: all 4 properties/operations

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, $\chi 2$ test
	Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
	Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

Discrete and Continuous Attributes :

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.

- Practically, real values can only be measured and represented using a finite number of digits.
- Continuous attributes are typically represented as floating-point variables.

Assymetric attributes:

- Only presence (a non-zero attribute value) is regarded as important
    - Words present in documents
    - Items present in customer transactions
- If we met a friend in the grocery store would we ever say the following?

  "I see our purchases are very similar since we didn't buy most of the same things."

Characteristics of data:

- Dimensionality (number of attributes)
    - High dimensional data brings a number of challenges
- Sparsity
    - Only presence counts
- Resolution
    - Patterns depend on the scale
- Size
    - Type of analysis may depend on size of data

Types of datasets:

Record

- Data Matrix
- Document Data
- Transaction Data

- Graph

- World Wide Web
- Molecular Structures

- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data

Record data:

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data matrix:

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

- Such a data set can be represented by an *m* by *n* matrix, where there are *m* rows, one for each object, and *n* columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document data:

- Each document becomes a 'term' vector
  - Each term is a component (attribute) of the vector
  - The value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

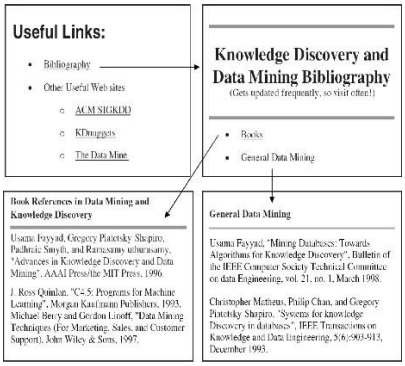
Transaction data:

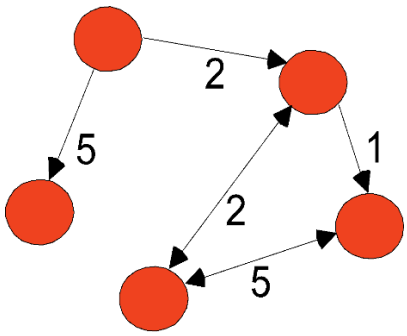
- A special type of data, where
  - Each transaction involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

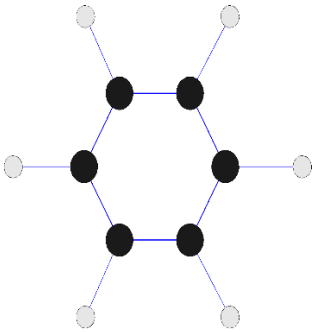
- Can represent transaction data as record data

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

Graph data:







Ordered data:

(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

Data Quality:

Data quality refers to how reliable and accurate the data is. In the real world, there are many situations where the data is incomplete, inconsistent, or contains errors. This is a common problem in big databases.

Data mining is a process that helps to discover useful patterns and insights from data. However, this process can be affected by poor data quality. To overcome this problem, data mining focuses on detecting and correcting data quality issues. The first step in this process is called data cleaning.

Data cleaning involves identifying and correcting errors in the data. For example, if there are missing values in a dataset, data cleaning can involve filling in those missing values based on other information in the dataset. By cleaning the data, we can ensure that the data mining algorithms are more accurate and can provide more meaningful insights.

Measurement and Data Collection Issues:

Measurement and Data Collection Errors:

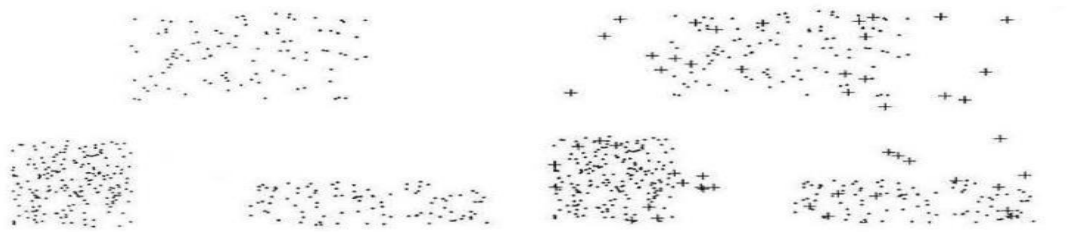
The term measureme nt error refers to any problem resulting from the measurement process. A common problem is that the value recorded differs from the true value to some extent. The term data collection error refers to errors such as omitting data objects or attri bute values, or inappropriately including a data object.

Noise and Artifacts Noise:

It is the random component of a measurement error. It may involve the distortion of a value or the addition of spurious objects.

(a) Time series. (b) Time series with noise.

Figure 2.5. Noise in a time series context.



(a) Three groups of points. (b) With noise points (+) added.

Figure 2.6. Noise in a spatial context.

Precision, Bias, and Accuracy:

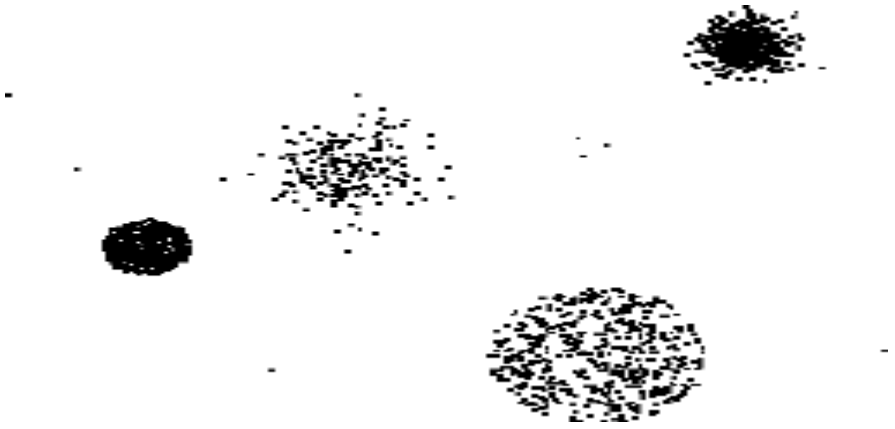
Precision: The closeness of repeated measurements of the same quantity to one another.

Bias: A systematic variation of measurements from the quantity being measured.

Accuracy: The closeness of measurements to t.he true value of the quantity being measured.

Outliers:

Outliers are either data objects that, in some sense, have characteristics that are different from most of the other data objects in the data set, or values of an attribute that are unusual with re spect to the typical values for that attribute.



Missing Values:

It is not unusual for an object to be missing one or more attribute values. In some cases, the information was not collected. The techniques used for dealing missing data:

Eliminate Data Objects or Attributes:

A simple and effective strategy is to eliminate objects with missing values.

Estimate Missing Values:

Sometimes missing data can be reliably estimatedthe missing values can be estimated by using the remai ing values.

Ignore the Missing Value during Analysis:

Many data mining approaches can be modified to ignoremissing values. lf one or both objects of a pair have missing values for some attributes, then the similari ty can be calculated by using only the attributes that do not have missing values.

Inconsistent Values:

Data can contain inconsistent values. Some types of inconsistences are easy to detect and some are not. Once an inconsistency has been detected, it is sometimes possible to correct the data.

Duplicate Data:

A data set may include data objects that are duplicates, or almost duplicates, of one another.

Data Preprocessing:

Data preprocessing is like getting the ingredients ready before cooking a meal. It involves preparing the data in a way that makes it easier for a computer program to understand and analyze. This can involve different techniques like cleaning up the data, removing errors, formatting it in a certain way, and more. The end goal is to make the data as clear and useful as possible for further analysis..

- Aggregation
- Sampling

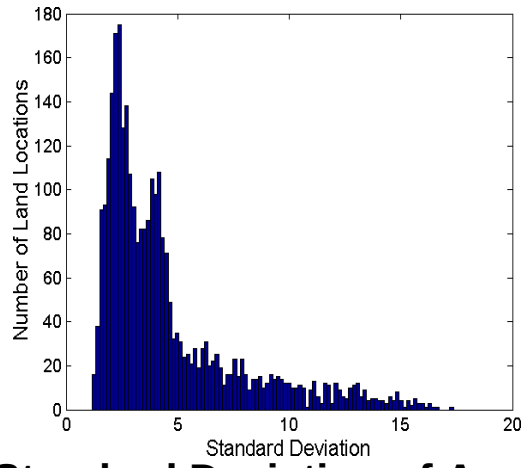
- Dimensionality reduction
- Feature subset selection
- Feature creation
- Discretization and binarization
- Attribute transformation

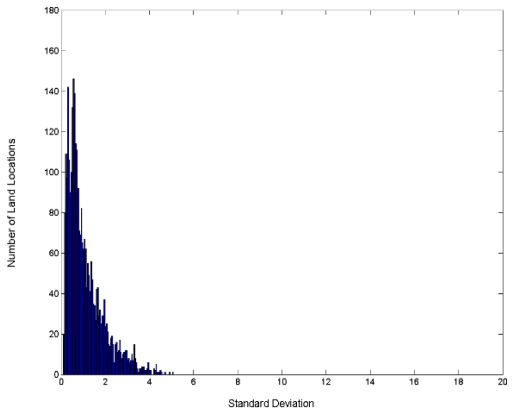
Aggregation:

Combining two or more attributes (or objects) into a single attribute (or object)

Purpose:

- Data reduction - reduce the number of attributes or objects
- Change of scale
  - Cities aggregated into regions, states, countries, etc.
  - Days aggregated into weeks, months, or years
- More “stable” data - aggregated data tends to have less variability.





Standard Deviation of Average Monthly Precipitation
Standard Deviation of Average Yearly Precipitation

Sampling:

Sampling is the main technique employed for data reduction.It is often used for both the preliminary investigation of the data and the final data analysis.

Statisticians often sample because obtaining the entire set of data of interest is too expensive or time consuming.

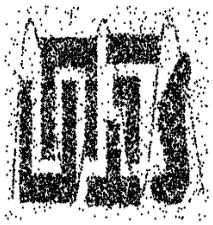
Sampling is typically used in data mining because processing the entire set of data of interest is too expensive or time consuming.

- The key principle for effective sampling is the following:

  Using a sample will work almost as well as using the entire data set,if the sample is representative

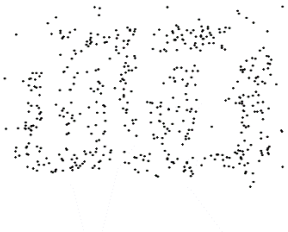
  A sample is representative if it has approximately the same properties (of interest) as the original set of data

Sample size:



8000 points

5000 points



500 points

Types of sampling:

- **Simple Random Sampling**
  - There is an equal probability of selecting any particular item
  - **Sampling without replacement**
    - As each item is selected, it is removed from the population
  - **Sampling with replacement**

- Objects are not removed from the population as they are selected for the sample.
- In sampling with replacement, the same object can be picked up more than once

- **Stratified sampling:**
  - Split the data into several partitions; then draw random samples from each partition.

Dimentionality reduction:

- Avoid curse of dimensionality
- Reduce amount of time and memory required by data mining algorithms
- Allow data to be more easily visualized
- May help to eliminate irrelevant features or reduce noise.

- Techniques
  - Principal Components Analysis (PCA)
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

Cursue of dimentionality:

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies.
- Definitions of density and distance between points, which are critical for clustering and outlier detection, become less meaningful.
- Another way to reduce dimensionality of data.

Feature Subset Selection:

- Redundant features
  - Duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid

- Irrelevant features
    - Contain no information that is useful for the data mining task at hand
    - Example: students' ID is often irrelevant to the task of predicting students' GPA
- Many techniques developed, especially for classification.

Feature Creation:

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes
- Three general methodologies:
    - Feature extraction
        - Example: extracting edges from images
    - Feature construction
        - Example: dividing mass by volume to get density
    - Mapping data to new space

Example: Fourier and wavelet analysis

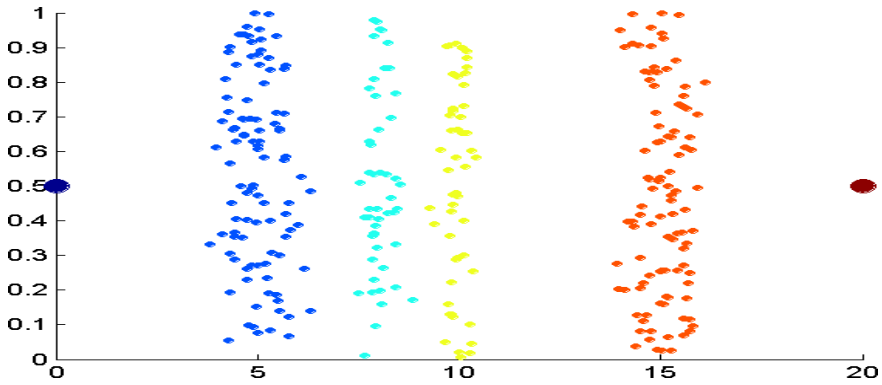
Discretization and binarization:

Discretization:

Discretization is the process of converting a continuous attribute into an ordinal attribute.

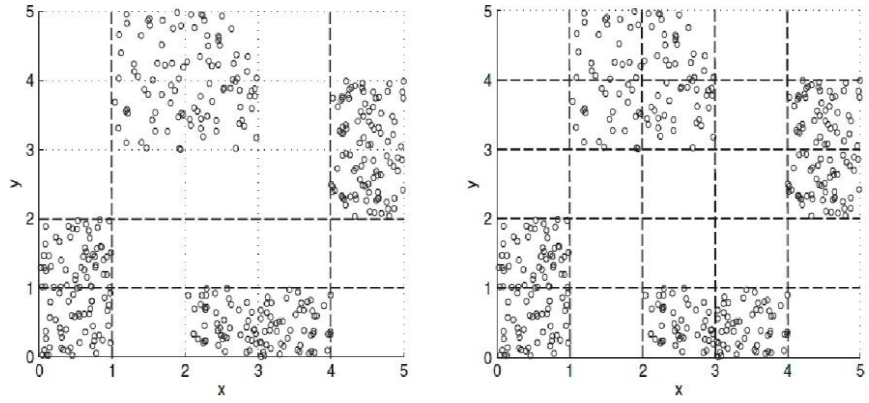
- A potentially infinite number of values are mapped into a small number of categories
- Discretization is used in both unsupervised and supervised settings

- **Unsupervised Discretization**



Data consists of four groups of points and two outliers. Data is one-dimensional, but a random y component is added to reduce overlap.

Supervised Discretization:



(a) Three intervals (b) Five intervals

Figure 2.14. Discretizing x and y attributes for four groups (classes) of points.

Binarization:

It maps a continuous or categorical attribute into one or more binary variables.

Table 2.6. Conversion of a categorical attribute to five asymmetric binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

Attribute transform:

- An attribute transform is a function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
    - Simple functions: x^k, log(x), e^x, |x|
    - Normalization
        - Refers to various techniques to adjust to differences among attributes in terms of frequency of occurrence, mean, variance, range
        - Take out unwanted, common signal, e.g., seasonality
    - In statistics, standardization refers to subtracting off the means and dividing by the standard deviation

Unit-1:Part-3:

MEASURES OF SIMILARITY AND DISSIMILARITY:

The **similarity** between two objects is a numerical measure of the degree to which the two objects are alike.Similarities are higher for pairs of objects that are more alike. Similarities are usually non-negative and are often between 0 (no similarity) and 1 (complete similarity).

The **dissimilarity** between two objects is a numerical measure of the degree to which the two objects are different. Dissimilarities are lower for more similar pairs of objects. Dissimilarities sometimes fall in the interval [0, I], but it is also common for them to range from 0 to oo.

Similarity and Dissimilarity between Simple Attributes:

The proximity of objects with a number of attributes is typically defined by combining the proximities of individual attributes, and thus, we first discuss proximity between objects having a single attribute.

Table 2.7. Similarity and dissimilarity for simple attributes

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x - y \rvert / (n-1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Dissimilarities between Data Objects:

Distances:

Equation is generalized by the Minkowski distance metric:

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r},$$

where r is a parameter. The following are the three most common examples of Minkowski distances.

r = 1 City block (Manhattan) distance

r = 2. Euclidean distance

r = oo. Supremum distance. This is the maximum difference between any attribute of the objects.

Distances, such as the Euclidean distance, have some well-known properties. If d(x, y) is the distance between two points, x and y, then the following properties hold.

1. Positivity

   (a) d(x,x) >=0 for all x and y,

   (b) d(x,y) = 0 only if x = y.

2. Symmetry

   d(x,y) = d(y,x) for all x andy.

3. Triangle Inequality

   d(x, z) <=d{x, y) + d(y , z) for all points x, y, and z

Ex:

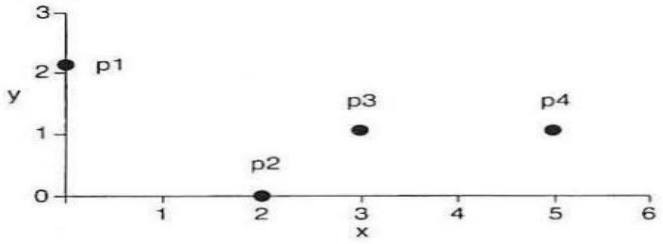


Figure 2.15. Four two-dimensional points.

Table 2.8. x and y coordinates of four points.

point	x coordinate	y coordinate
p1	0	2
p2	2	0
p3	3	1
p4	5	1

Table 2.9. Euclidean distance matrix for Table 2.8.

	p1	p2	p3	p4
p1	0.0	2.8	3.2	5.1
p2	2.8	0.0	1.4	3.2
p3	3.2	1.4	0.0	2.0
p4	5.1	3.2	2.0	0.0

Table 2.10. L_1 distance matrix for Table 2.8.

L_1	p1	p2	p3	p4
p1	0.0	4.0	4.0	6.0
p2	4.0	0.0	2.0	4.0
p3	4.0	2.0	0.0	2.0
p4	6.0	4.0	2.0	0.0

Table 2.11. L_∞ distance matrix for Table 2.8.

L_∞	p1	p2	p3	p4
p1	0.0	2.0	3.0	5.0
p2	2.0	0.0	1.0	3.0
p3	3.0	1.0	0.0	2.0
p4	5.0	3.0	2.0	0.0

Similarities between Data Objects:

For similarities, the triangle inequality (or the analogous property) typically does not hold, but symmetry and positivity typically do.

To be explicit, if s(x, y) is the similarity between points x and y , then the typical properties of similarities are the following:

1. s(x,y) = 1 only if x = y. (0 <=S s <=S 1)
2. s(x,y) = s(y,x) for all x and y. (Symmetry)

Examples of Proximity Measures:

provides specific examples of some similarity and dissimilarity measures.

Similarity Measures for Binary Data:

Similarity measures between objects that contain only binary attributes are called similarity coefficients, and typically have values between 0 and 1. A value of 1 indicates that the two objects are completely similar, while a value of 0 indicates that the objects are not at all similar.

Let x and y be two objects that consist of n binary attributes. The comparison of two such objects, i.e., two binary vectors, leads to the following four quantities (frequencies):

f oo = the number of attributes where x is 0 and y is 0

f01 = the number of attributes where x is 0 and y is 1

f1O = the number of attributes where x is 1 and y is 0

f11 = the number of attributes where x is 1 and y is 1

Simple Matching Coefficient One commonly used similarity coefficient is the simple matching coefficient (SMC) , which is defined as :

$$SMC = \frac{\text{number of matching attribute values}}{\text{number of attributes}} = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}}.$$

Jaccard Coefficient Suppose that x and y are data objects that represent two rows (two transactions) of a transaction matrix.

$\mathbf{x} = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$
$\mathbf{y} = (0, 0, 0, 0, 0, 0, 1, 0, 0, 1)$

$f_{01} = 2$ the number of attributes where x was 0 and y was 1
$f_{10} = 1$ the number of attributes where x was 1 and y was 0
$f_{00} = 7$ the number of attributes where x was 0 and y was 0
$f_{11} = 0$ the number of attributes where x was 1 and y was 1

$$SMC = \frac{f_{11}+f_{00}}{f_{01}+f_{10}+f_{11}+f_{00}} = \frac{0+7}{2+1+0+7} = 0.7$$

$$J = \frac{f_{11}}{f_{01}+f_{10}+f_{11}} = \frac{0}{2+1+0} = 0$$

Cosine Similarity:

Documents are often represented as vectors, where each attribute represents the frequency with which a particular term (word) occurs in the document. Even though documents have thousands or tens of thousands of attributes (terms), each document is sparse since it has relatively few non-zero attributes.

The cosine similarity, defined next, is one of the most common measure of document similarity. If x and y are two document vectors, then

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|}, \tag{2.7}$$

where $\cdot$ indicates the vector dot product, $\mathbf{x} \cdot \mathbf{y} = \sum_{k=1}^{n} x_k y_k$, and $\|\mathbf{x}\|$ is the length of vector $\mathbf{x}$, $\|\mathbf{x}\| = \sqrt{\sum_{k=1}^{n} x_k^2} = \sqrt{\mathbf{x} \cdot \mathbf{x}}$.

Example 2.18 (Cosine Similarity of Two Document Vectors). This example calculates the cosine similarity for the following two data objects, which might represent document vectors:

$\mathbf{x} = (3, 2, 0, 5, 0, 0, 0, 2, 0, 0)$
$\mathbf{y} = (1, 0, 0, 0, 0, 0, 0, 1, 0, 2)$

$\mathbf{x} \cdot \mathbf{y} = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$
$\|\mathbf{x}\| = \sqrt{3*3 + 2*2 + 0*0 + 5*5 + 0*0 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0} = 6.48$
$\|\mathbf{y}\| = \sqrt{1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 2*2} = 2.24$
$\cos(\mathbf{x}, \mathbf{y}) = 0.31$

■

Issues in Proximity Calculation:

(1) how to handle the case in which attributes have different scales and/or are correlated,

(2) how to calculate proximity between objects that are composed of different types of attributes, e.g., quantitative and qualitative,

(3) and how to handle proximity calculation when attributes have different weights; i.e., when not all attributes contribute equally to the proximity of objects.

Standardization and Correlation for Distance Measures:

A related issue is how to compute distance when there is correlation between some of the attributes, perhaps in addition to difrerences in the ranges of values. A generalization of Euclidean distance, the Mahalanobis distance , is useful when attributes are correlated, have different ranges of values.

$$\text{mahalanobis}(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})\Sigma^{-1}(\mathbf{x} - \mathbf{y})^T,$$

Combining Similarities for Heterogeneous Attributes:

The previous definitions of similarity were based on approaches that assumed all the attributes were of the same type. A general approach is needed when the attributes are of different types. One straightforward approach is to compute the similarity between each attribute separately and then combine these similarities using a method that results in a similarity between 0 and l.

Algorithm 2.1 Similarities of heterogeneous objects.

1: For the k^{th} attribute, compute a similarity, $s_k(\mathbf{x}, \mathbf{y})$, in the range $[0, 1]$.
2: Define an indicator variable, δ_k, for the k^{th} attribute as follows:

$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is an asymmetric attribute and} \\ & \text{both objects have a value of 0, or if one of the objects} \\ & \text{has a missing value for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$

3: Compute the overall similarity between the two objects using the following formula:

$$\text{similarity}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^{n} \delta_k} \tag{2.15}$$

the formulas for proximity can be modified by weighting the contribution of each attribute.

If the weights w_k sum to 1, then (2.15) becomes

$$\text{similarity}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^{n} \delta_k}. \tag{2.16}$$

The definition of the Minkowski distance can also be modified as follows:

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} w_k |x_k - y_k|^r \right)^{1/r}. \tag{2.17}$$

Selecting the Right Proximity Measure:

The following are a few general observations that may be helpful. First, the type of proximity measure should fit the type of data. For many types of dense, continuous data, metric distance measures such as Euclidean distance are often used.

Proximity between continuous attributes is most often expressed in terms of differences, and distance measures provide a well-defined way of combining these differences into an overall proximity measure. Although attributes can have different scales and be of differing importance.

For sparse data, which often consists of asymmetric attributes, we typically employ similarity measures that ignore 0- 0 matches.

In some cases, transformation or normalization of t he data is important for obtaining a proper similarity measure since such transformations are not always present in proximity measures.