Data Mining and Machine Learning

Lecture Notes – Module 1

Data Mining Introduction: – Data – Types of Data – Data Mining Functionalities – Classification of Data Mining Systems – Issues –Data Pre-processing.
Association Rule Mining Frequent Patterns – Apriori Algorithm Description.

Textbooks:

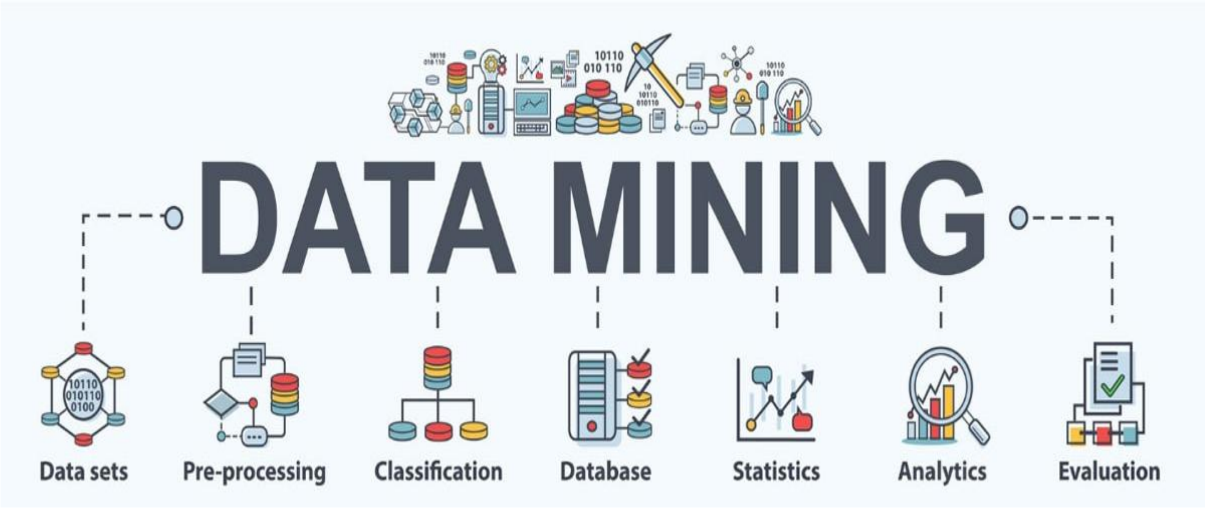
1. Jiawei Han and Micheline Kamber: Data Mining Concepts and Techniques,Elsevier, 2nd Edition, 2009.
2. Stephen Marsland, “Machine Learning - An Algorithmic Perspective”, Second Edition, CRC Press - Taylor and Francis Group, 2015.
3. Ethem Alpaydin, “Introduction to Machine Learning”, Second Edition, MITPress, Prentice Hall of India (PHI) Learning Pvt. Ltd. 2010.
4. Xindong Wu and Vipin Kumar: The top ten Algorithms in Data Mining, Chapman and Hall/CRC press.
5. Pang-Ning Tan, Michael Steinbach and Vipin Kumar, “Introduction to Data Mining”, Pearson Education, 2007.
6. DISCOVERING KNOWLEDGE IN DATA, An Introduction to Data Mining, Second Edition, Daniel T. Larose ,Chantal D. Larose.

Reference Books:

1. K.P. Soman, ShyamDiwakar and V. Aja, “Insight into Data Mining Theory and Practice”, Eastern Economy Edition, Prentice Hall of India, 2006.
2. G. K. Gupta, “Introduction to Data Mining with Case Studies”, Eastern Economy Edition, Prentice Hall of India, 2006.
3. Christopher Bishop, “Pattern Recognition and Machine Learning”, CBS Publishers & Distributors, 2010.
4. Mehryar Mohri, Afshin R, Ameet Talwalkar, "Foundations of Machine Learning”, MIT Press, 2012.

1. What is Data Mining? Why do we use Data Mining?

Data mining is the process of sorting through large data sets to identify patterns and relationships that can help solve business problems through data analysis. Data mining techniques and tools enable enterprises to predict future trends and make more-informed business decisions.



Uses of Data Mining: Clustering, Association, Data Cleaning, Data Visualization, Classification, Machine Learning

2. Briefly Explain Knowledge Discovery (KDD) Process.

The KDD Process involves following steps:

i. **Data cleaning** (to remove noise and inconsistent data)

ii. **Data integration** (where multiple data sources may be combined)

iii. **Data selection** (where data relevant to the analysis task are retrieved from the database)

iv. **Data transformation** (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)

v. **Data mining** (an essential process where intelligent methods are applied to extract data patterns)

vi. **Pattern evaluation** (to identify the truly interesting patterns representing knowledge based on interestingness measures)

vii. **Knowledge presentation** (where visualization and knowledge representation techniques are used to present mined knowledge to users)

Steps 1 through 4 are different forms of data preprocessing, where data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base.

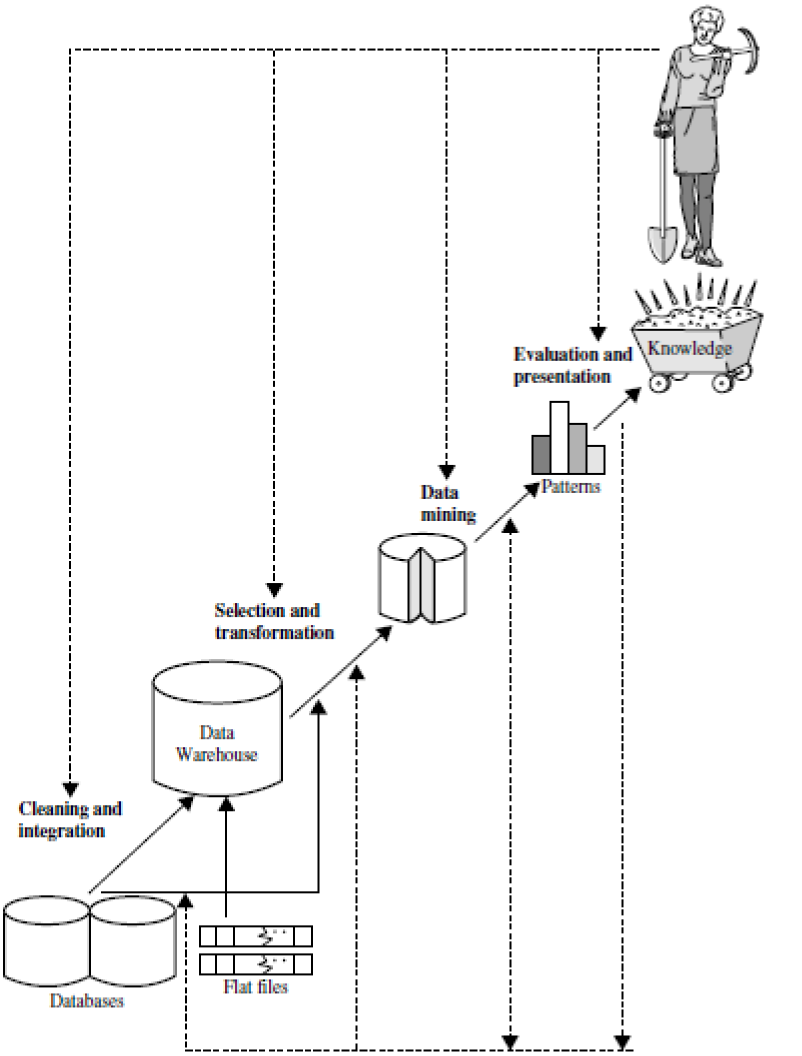


Figure 1.4 Data mining as a step in the process of knowledge discovery.

3. Explain Multi Dimensional View of Data Mining.

- **Data to be mined**
  - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined (or: Data mining functions)**
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Descriptive vs. predictive data mining

    - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
    - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.
- **Applications adapted**
    - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

4. What Kind of Data Can Be Mined?

- Database-oriented data sets and applications
    - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
    - Data streams and sensor data
    - Time-series data, temporal data, sequence data (incl. bio-sequences)
    - Structure data, graphs, social networks and multi-linked data
    - Object-relational databases
    - Heterogeneous databases and legacy databases
    - Spatial data and spatiotemporal data
    - Multimedia database
    - Text databases
    - The World-Wide Web

5. Explain Data Mining Functionalities

(1) Generalization

- Information integration and data warehouse construction
    - Data cleaning, transformation, integration, and multidimensional data model
- Data cube technology
    - Scalable methods for computing (i.e., materializing) multidimensional aggregates
    - OLAP (online analytical processing)
- Multidimensional concept description: Characterization and discrimination
    - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet region

(2) Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)

    - What items are frequently purchased together in your Walmart?
- Association, correlation vs. causality
    - A typical association rule
        - Diaper → Beer [0.5%, 75%] (support, confidence)
    - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

(3) Classification

- Classification and label prediction
    - 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
    - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications:
    - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …

(4) Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications

(5) Outlier Analysis

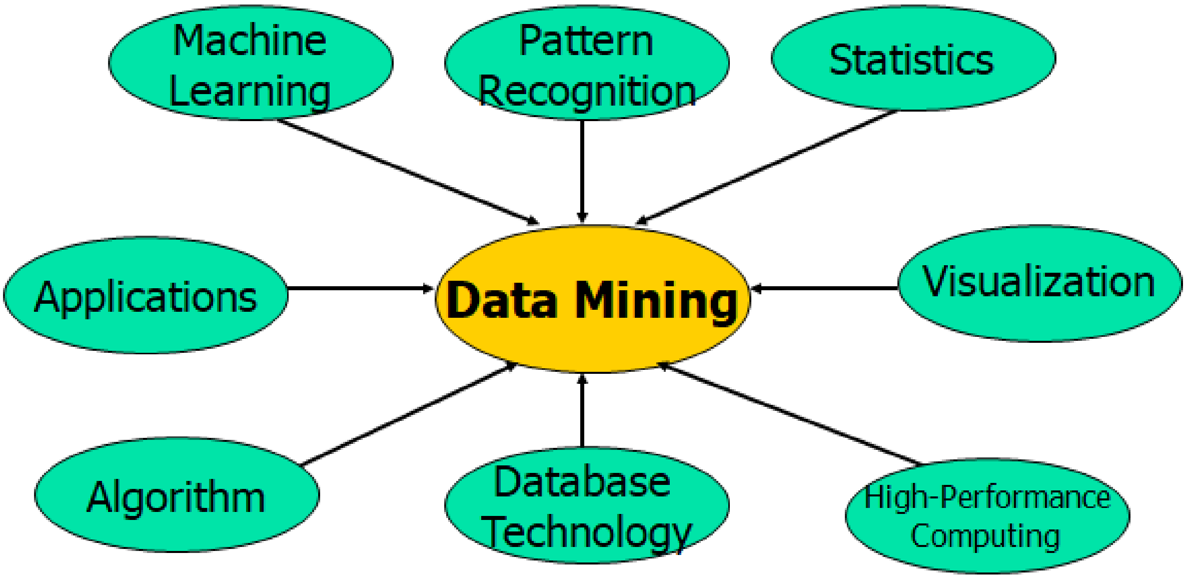
- Outlier: A data object that does not comply with the general behavior of the data
- Noise or exception? — One person’s garbage could be another person’s treasure
- Methods: by product of clustering or regression analysis, …
- Useful in fraud detection, rare events analysis

6. **Explain the classification of Data Mining with a Neat Diagram?**

A. Statistics: Statistics studies the collection, analysis, interpretation or explanation, and presentation of data. Data mining has an inherent connection with statistics. A **statistical model** is a set of mathematical functions that describe the behavior of the objects in a target class in terms of random variables and their associated probability distributions. Statistical models are widely used to model data and data classes. For example, in data mining tasks like data characterization and classification, statistical models of target classes can be built.

B. Machine learning investigates how computers can learn (or improve their performance) based on data. A main research area is for computer programs to *automatically* learn to recognize complex patterns and make intelligent decisions based on data. For example, a typical machine learning problem is to program a computer so that it can automatically recognize handwritten postal codes on mail after learning from a set of examples.

1. **Supervised learning** is basically a synonym for classification. The supervision in the learning comes from the labeled examples in the training data set. For example, in the postal code recognition problem, a set of handwritten postal code images and their corresponding machine-readable translations are used as the training examples, which supervise the learning of the classification model.
2. **Unsupervised learning** is essentially a synonym for clustering. The learning process is unsupervised since the input examples are not class labeled. Typically, we may use clustering to discover classes within the data. For example, an unsupervised learning method can take, as input, a set of images of handwritten digits. Suppose that it finds 10 clusters of data. These clusters may correspond to the 10 distinct digits of 0 to 9, respectively. However, since the training data are not labeled, the learned model cannot tell us the semantic meaning of the clusters found.
3. **Semi-supervised learning** is a class of machine learning techniques that make use of both labeled and unlabeled examples when learning a model. In one approach, labeled examples are used to learn class models and unlabeled examples are used to refine the boundaries between classes. For a two-class problem, we can think of the set of examples belonging to one class as the *positive examples* and those belonging to the other class as the *negative examples*.



4. **Active learning** is a machine learning approach that lets users play an active role in the learning process. An active learning approach can ask a user (e.g., a domain expert) to label an example, which may be from a set of unlabeled examples or synthesized by the learning program. The goal is to optimize the model quality by actively acquiring knowledge from human users, given a constraint on how many examples they can be asked to label.

C. Database Systems and Data Warehouses

Database systems research focuses on the creation, maintenance, and use of databases for organizations and end-users. Particularly, database systems researchers have established highly recognized principles in data models, query languages, query processing and optimization methods, data storage, and indexing and accessing methods. Database systems are often well known for their high scalability in processing very large, relatively structured data sets. Many data mining tasks need to handle large data sets or even real-time, fast streaming data. Therefore, data mining can make good use of scalable database technologies to achieve high efficiency and scalability on large data sets.

D. Information Retrieval

Information retrieval (**IR**) is the science of searching for documents or information in documents. Documents can be text or multimedia, and may reside on the Web. The differences between traditional information retrieval and database systems are twofold:

Information retrieval assumes that (1) the data under search are unstructured; and (2) the queries are formed mainly by keywords, which do not have complex structures (unlike SQL queries in database systems).

7. What Kind of Applications Are Targeted?

- Web page analysis: from web page classification, clustering to PageRank & HITS algorithms
- Collaborative analysis & recommender systems
- Basket data analysis to targeted marketing
- Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis
- Data mining and software engineering (e.g., IEEE Computer, Aug. 2009 issue)
- From major dedicated data mining systems/tools (e.g., SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools) to invisible data mining

8. Explain the Major Issues in Data Mining

- Mining Methodology
  - Mining various and new kinds of knowledge
  - Mining knowledge in multi-dimensional space
  - Data mining: An interdisciplinary effort
  - Boosting the power of discovery in a networked environment
  - Handling noise, uncertainty, and incompleteness of data
  - Pattern evaluation and pattern- or constraint-guided mining
- User Interaction
  - Interactive mining
  - Incorporation of background knowledge
  - Presentation and visualization of data mining results
- Efficiency and Scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed, stream, and incremental mining methods
- Diversity of data types
  - Handling complex types of data
  - Mining dynamic, networked, and global data repositories
- Data mining and society

- Social impacts of data mining
- Privacy-preserving data mining
- Invisible data mining

9. Why Preprocess the Data?

Data preprocessing, a component of data preparation, describes any type of processing performed on raw data to prepare it for another data processing procedure. It has traditionally been an important preliminary step for the data mining process.

We Pre process to improve data quality:

- Accuracy: correct or wrong, accurate or not
- Completeness: not recorded, unavailable, …
- Consistency: some modified but some not, dangling, …
- Timeliness: timely update?
- Believability: how trustable the data are correct?
- Interpretability: how easily the data can be understood?

10. What are the Major Tasks in Data Preprocessing?

Data cleaning

Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Data integration

Integration of multiple databases, data cubes, or files

Data reduction

Dimensionality reduction

Numerosity reduction

Data compression

Data transformation and data discretization

Normalization

Concept hierarchy generation

11. Explain Data Cleaning in detail.

Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error

a. incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    i. e.g., *Occupation*=“ ” (missing data)

b. noisy: containing noise, errors, or outliers
    i. e.g., *Salary*="−10" (an error)

c. inconsistent: containing discrepancies in codes or names, e.g.,
    i. *Age*="42", *Birthday*="03/07/2010"
    ii. Was rating "1, 2, 3", now rating "A, B, C"
    iii. discrepancy between duplicate records

d. Intentional (e.g., *disguised missing* data)

Jan. 1 as everyone's birthday?

The process of Data Cleaning is:

- Data discrepancy detection
    - Use metadata (e.g., domain, range, dependency, distribution)
    - Check field overloading
    - Check uniqueness rule, consecutive rule and null rule
    - Use commercial tools
        - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
        - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- Data migration and integration
    - Data migration tools: allow transformations to be specified
    - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface
- Integration of the two processes

Iterative and interactive (e.g., Potter's Wheels)

12. What is Incomplete (Missing) Data? How to Handle Missing Data?

- Data is not always available
    - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
    - equipment malfunction
    - inconsistent with other recorded data and thus deleted
    - data not entered due to misunderstanding
    - certain data may not be considered important at the time of entry
    - not register history or changes of the data

- Missing data may need to be inferred

Procedure to Handle Missing Data is:

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
    - a global constant : e.g., "unknown", a new class?!
    - the attribute mean
    - the attribute mean for all samples belonging to the same class: smarter
    - the most probable value: inference-based such as Bayesian formula or decision tree

13. What is Noisy Data? How to Handle Noisy Data?

A. Noise: random error or variance in a measured variable

B. Incorrect attribute values may be due to
   a. faulty data collection instruments
   b. data entry problems
   c. data transmission problems
   d. technology limitation
   e. inconsistency in naming convention

C. Other data problems which require data cleaning
   a. duplicate records
   b. incomplete data
   c. inconsistent data

Procedure to Handle Noisy Data is:

a) Binning
   a. first sort data and partition into (equal-frequency) bins
   b. then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

b) Regression
   a. smooth by fitting the data into regression functions

c) Clustering
   a. detect and remove outliers

d) Combined computer and human inspection

a. detect suspicious values and check by human (e.g., deal with possible outliers)

14. What is Data Integration? How to Handle Redundancy in Data Integration?

Data integration:

Combines data from multiple sources into a coherent store

Schema integration: e.g., A.cust-id ≡ B.cust-#

Integrate metadata from different sources

Entity identification problem:

Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton

Detecting and resolving data value conflicts

For the same real world entity, attribute values from different sources are different

Possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by *correlation analysis* and *covariance analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Correlation Analysis (Nominal Data)

- **X^2 (chi-square) test**
- The larger the X^2 value, the more likely the variables are related
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count
- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)
- $$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$
- It shows that like_science_fiction and play_chess are correlated in the group

Correlation Analysis (Numeric Data)

- Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A \sigma_B}$$

where n is the number of tuples, and are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\Sigma(a_i b_i)$ is the sum of the AB cross-product.

- If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.
- $r_{A,B} = 0$: independent; $r_{AB} < 0$: negatively correlated

Covariance (Numeric Data)

- Covariance is similar to correlation

$$Cov(A,B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

where n is the number of tuples, and are the respective mean or **expected values** of A and B, σ_A and σ_B are the respective standard deviation of A and B.

- **Positive covariance**: If $Cov_{A,B} > 0$, then A and B both tend to be larger than their expected values.
- **Negative covariance**: If $Cov_{A,B} < 0$ then if A is larger than its expected value, B is

likely to be smaller than its expected value.

- **Independence**: Cov$_{A,B}$ = 0 but the converse is not true:
    - Some pairs of random variables may have a covariance of 0 but are not independent. Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence

Co-Variance: An Example

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

- It can be simplified in computation as

$$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$$

- Suppose two stocks A and B have the following values in one week: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
    - E(A) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4
    - E(B) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6
    - Cov(A,B) = (2×5+3×8+5×10+4×11+6×14)/5 − 4 × 9.6 = 4
- Thus, A and B rise together since Cov(A, B) > 0.

15. What are Data Reduction Strategies?

Data reduction: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results

Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.

Data reduction strategies

a. Dimensionality reduction, e.g., remove unimportant attributes
    i. Wavelet transforms
    ii. Principal Components Analysis (PCA)
    iii. Feature subset selection, feature creation

b. Numerosity reduction (some simply call it: Data Reduction)
    i. Regression and Log-Linear Models
    ii. Histograms, clustering, sampling
    iii. Data cube aggregation

c. Data compression

Data Reduction 1: Dimensionality Reduction

- **Curse of dimensionality**
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially
- **Dimensionality reduction**
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualization
- **Dimensionality reduction techniques**
  - Wavelet transforms
  - Principal Component Analysis

Supervised and nonlinear techniques (e.g., feature selection)

Data Reduction 2: Numerosity Reduction

- Reduce data volume by choosing alternative, *smaller forms* of data representation
- **Parametric methods** (e.g., regression)
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
  - Ex.: Log-linear models—obtain value at a point in *m*-D space as the product on appropriate marginal subspaces
- **Non-parametric** methods
  - Do not assume models
  - Major families: histograms, clustering, sampling, …

Data Reduction 3: Data Compression

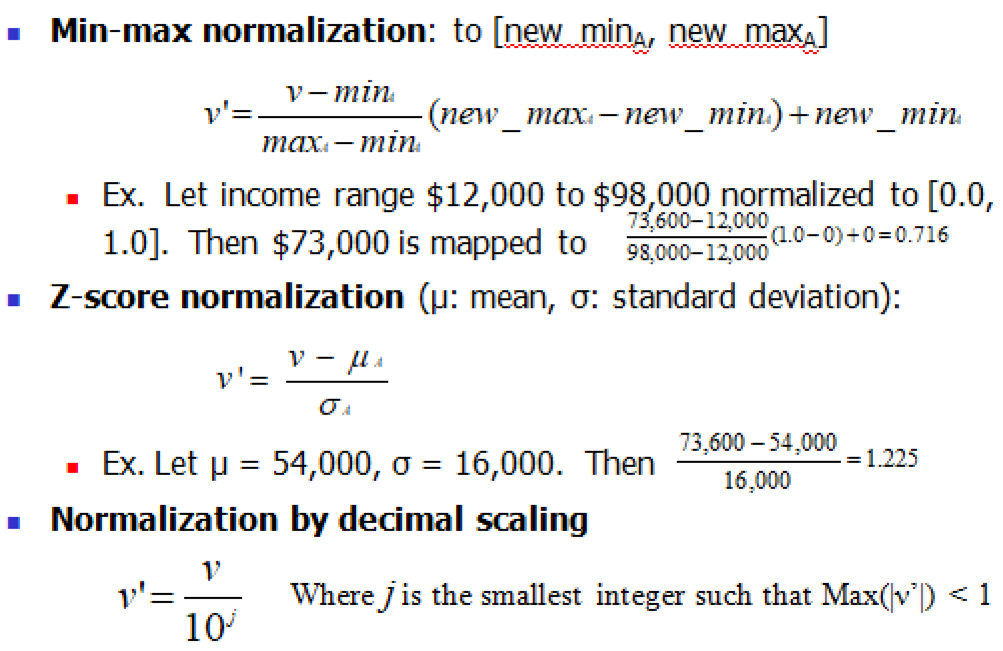
- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless, but only limited manipulation is possible without expansion
- Audio/video compression
  - Typically lossy compression, with progressive refinement
  - Sometimes small fragments of signal can be reconstructed without reconstructing the whole
- Time sequence is not audio

- Typically short and vary slowly with time
- Dimensionality and numerosity reduction may also be considered as forms of data compression

16. What is Data Transformation and Data Discretization ?

- A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values
- Methods
  - Smoothing: Remove noise from data
  - Attribute/feature construction
    - New attributes constructed from the given ones
  - Aggregation: Summarization, data cube construction
  - Normalization: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling
  - Discretization: Concept hierarchy climbing

Normalization



- **Min-max normalization**: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73,600 - 12,000}{98,000 - 12,000}(1.0 - 0) + 0 = 0.716$

- **Z-score normalization** (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73,600 - 54,000}{16,000} = 1.225$

- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max(|v'|) < 1

Discretization

- Three types of attributes
  - Nominal—values from an unordered set, e.g., color, profession
  - Ordinal—values from an ordered set, e.g., military or academic rank

    - Numeric—real numbers, e.g., integer or real numbers
- Discretization: Divide the range of a continuous attribute into intervals
    - Interval labels can then be used to replace actual data values
    - Reduce data size by discretization
    - Supervised vs. unsupervised
    - Split (top-down) vs. merge (bottom-up)
    - Discretization can be performed recursively on an attribute
    - Prepare for further analysis, e.g., classification

Data Discretization Methods

- Typical methods: All the methods can be applied recursively
    - Binning
        - Top-down split, unsupervised
    - Histogram analysis
        - Top-down split, unsupervised
    - Clustering analysis (unsupervised, top-down split or bottom-up merge)
    - Decision-tree analysis (supervised, top-down split)
    - Correlation (e.g., χ^2) analysis (unsupervised, bottom-up merge)

1. Simple Discretization: Binning

Equal-width (distance) partitioning

- Divides the range into *N* intervals of equal size: uniform grid
- if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
- The most straightforward, but outliers may dominate presentation
- Skewed data is not handled well

Equal-depth (frequency) partitioning

- Divides the range into *N* intervals, each containing approximately same number of samples
- Good data scaling
- Managing categorical attributes can be tricky

Binning Methods for Data Smoothing

Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (equi-depth) bins:

- Bin 1: 4, 8, 9, 15

- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

* Smoothing by bin means:

- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

* Smoothing by bin boundaries:

- Bin 1: 4, 4, 4, 15
- Bin 2: 21, 21, 25, 25
- Bin 3: 26, 26, 26, 34

17. What Is Frequent Pattern Analysis? Why Is Freq. Pattern Mining Important?

- ■ Frequent pattern: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set
- ■ First proposed by Agrawal, Imielinski, and Swami [AIS93] in the context of frequent itemsets and association rule mining
- ■ Motivation: Finding inherent regularities in data
  - ■ What products were often purchased together?— Beer and diapers?!
  - ■ What are the subsequent purchases after buying a PC?
  - ■ What kinds of DNA are sensitive to this new drug?
  - ■ Can we automatically classify web documents?
- ■ Applications
  - ■ Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.

Use:

- ■ Freq. pattern: An intrinsic and important property of datasets
- ■ Foundation for many essential data mining tasks
  - ■ Association, correlation, and causality analysis
  - ■ Sequential, structural (e.g., sub-graph) patterns
  - ■ Pattern analysis in spatiotemporal, multimedia, time-series, and stream data
  - ■ Classification: discriminative, frequent pattern analysis
  - ■ Cluster analysis: frequent pattern-based clustering
  - ■ Data warehousing: iceberg cube and cube-gradient
  - ■ Semantic data compression: fascicles
  - ■ Broad applications

18. Explain the terms Itemset, Support, Confidence, Lift, Association Rule.

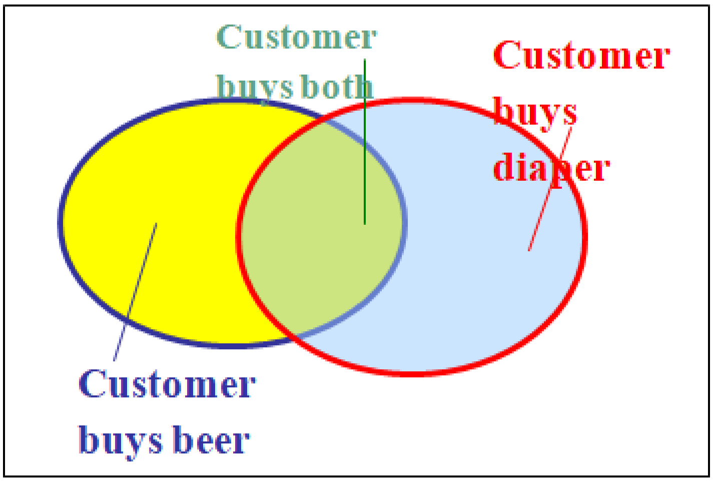
Consider an example:

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

- itemset: A set of one or more items
- k-itemset X = {x_1, …, x_k}
- *(absolute) support*, or, *support count* of X: Frequency or occurrence of an itemset X
- *(relative) support*, *s*, is the fraction of transactions that contains X (i.e., the probability that a transaction contains X)
- An itemset X is *frequent* if X's support is no less than a *minsup* threshold
- Find all the rules *X* → *Y* with minimum support and confidence
  - support, *s*, probability that a transaction contains X ∪ Y
  - confidence, *c*, conditional probability that a transaction having X also contains *Y*

  Let minsup = 50%, minconf = 50%

  Freq. Pat.: Beer:3, Nuts:3, Diaper:4, Eggs:3, {Beer, Diaper}:3

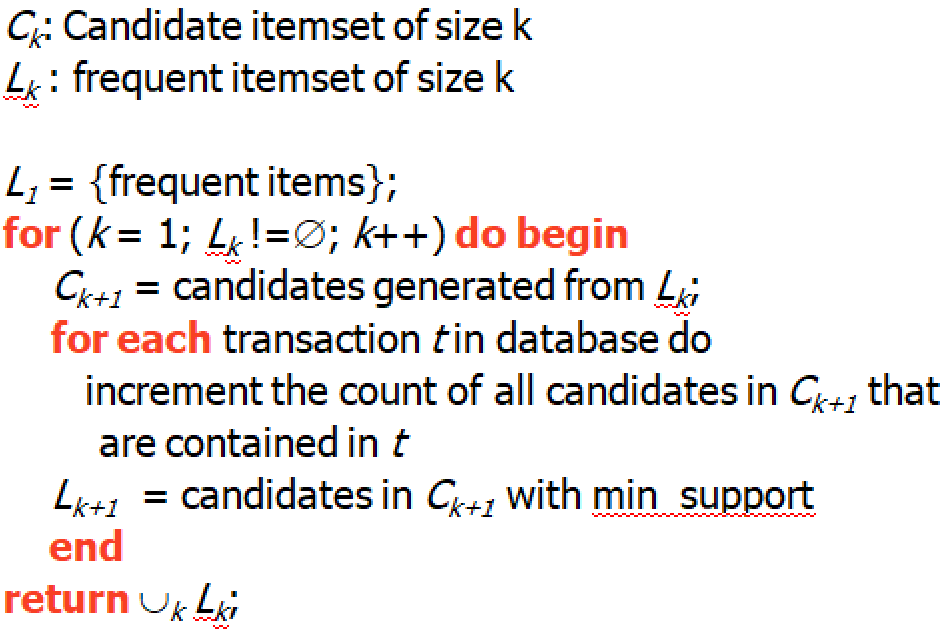


- Association rules: (many more!)
  - *Beer* → *Diaper* (60%, 100%)
  - *Diaper* → *Beer* (60%, 75%)

19. Explain Apriori Algorithm with Pseudo Code.

- Apriori pruning principle: If there is any itemset which is infrequent, its superset should not be generated/tested!
- Method:
  - Initially, scan DB once to get frequent 1-itemset
  - Generate length (k+1) candidate itemsets from length k frequent itemsets
  - Test the candidates against DB
  - Terminate when no frequent or candidate set can be generated

Pseudo Code:

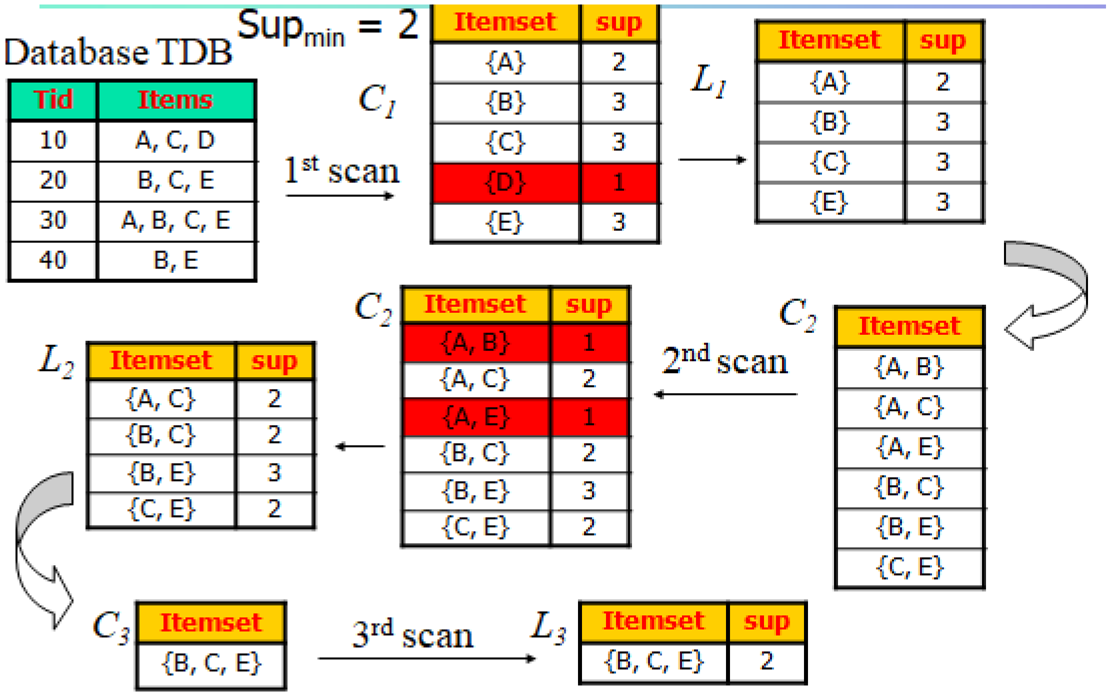


```
$C_k$: Candidate itemset of size k
$L_k$ : frequent itemset of size k

$L_1$ = {frequent items};
for ($k$ = 1; $L_k$ !=∅; $k$++) do begin
    $C_{k+1}$ = candidates generated from $L_k$;
    for each transaction $t$ in database do
        increment the count of all candidates in $C_{k+1}$ that
          are contained in $t$
    $L_{k+1}$ = candidates in $C_{k+1}$ with min_support
end
return $\cup_k L_k$;
```

- **How to generate candidates?**
  - Step 1: self-joining L_k
  - Step 2: pruning
- **Example of Candidate-generation**
  - L_3={*abc, abd, acd, ace, bcd*}
  - Self-joining: L_3*L_3
    - *abcd* from *abc* and *abd*
    - *acde* from *acd* and *ace*
  - Pruning:
    - *acde* is removed because *ade* is not in L_3
  - C_4 = {*abcd*}

Example:



Database TDB

$Sup_{min} = 2$

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

C_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

C_2

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

L_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C_3

Itemset
{B, C, E}

L_3

Itemset	sup
{B, C, E}	2

Possible Questions

1. a) Identify and explain four common methods used for binning numerical predictors.

   b) Use the following stock price data (in dollars)
   20 30 40 60 10 15 21 22 70
   i. Calculate the mean stock price.
   ii. Calculate median and mode stock price.
   iii. Compute the standard deviation of the stock price.
   iv. Find the min-max normalized stock price for the stock worth $105.
   v. Compute the Z – score standardized stock price for the stock worth $105
   vi. Find the decimal scaling stock price for the stock worth $105.
   vii. Identify all possible stock prices that would be outliers, using IQR method.

   c) Explain functionalities of data mining.

2. a) Explain with examples the different transformation techniques used on data to be mined.

   b) i. The data for analysis include the attribute rank. The rank values for the data tuples are (in ascending order)
   X = {1, 1, 1, 1, 1, 2, 2, 11, 11, 12, 12, 44}.
   Apply Equal width binning and Equal frequency binning on the above set.
   ii. The weight of chocolate bars from a particular chocolate factory has a mean of 8 ounces with standard deviation of 0.1 ounce. What is the z-score corresponding to a weight of 8.17 ounces?

   c) Generate 3 frequent itemset using apriori algorithm for the given database. {Consider support count as 2}

Tid	List of items
T100	I1, I2, I5
T200	I2, I4
T300	I1, I2, I3, I5
T400	I1, I2, I3
T500	I1, I2, I4

3. a) Explain KDD process with a neat diagram.

   b) List and explain data mining issues.

   c) Generate three frequent item set for the given database using AprioriTid algorithm. {Consider support count as 3}.

Transaction Id	**Items**
1	R, S
2	U
3	P, R
4	R, S, U
5	Q, R, U
6	S, U, V

7	P
8	P, R, S
9	R, S, U, V
10	S, T
11	Q, S
12	S, V
13	T, U
14	R, S, U

4. a) Explain any five functionalities of data mining.

b) The data for analysis include the attribute rank. The rank values for the data tuples are (in ascending order)
X = {1, 1, 1, 1, 1, 2, 2, 11, 11, 12, 12, 44}.
Apply Equal width binning and Equal frequency binning on the above set.

c) Explain with examples the different transformation techniques used on data to be mined.

5. a) Explain any five data preprocessing techniques.

b) Explain Apriori algorithm with example.

6. a) Describe the basis on which data mining systems are classified. Identify various data mining issues.

b) Identify the different kinds of information that can be collected for data mining.

7. a) Identify the different kinds of information that can be collected for data mining.

b) Use the following stock price data (in dollars)

12 9 22 14 77 17
11 20 6 14 10 16

i. Calculate the mean stock price.
ii. Calculate median and mode stock price.
iii. Compute the standard deviation of the stock price.
iv. Find the min-max normalized stock price for the stock worth $20.
v. Compute the Z – score standardized stock price for the stock worth $20
vi. Find the decimal scaling stock price for the stock worth $20.
vii. Identify all possible stock prices that would be outliers, using Z-score method.
viii. Identify all possible stock prices that would be outliers, using IQR method.

8. a) How do you classify data mining systems? Explain the various data mining issues.

b) Explain apriori algorithm with example.

9. a) List and explain data mining issues.

b) Use the following stock price data (in dollars)

10 7 20 12 75 15
9 18 4 12 8 14

i. Calculate the mean stock price.
ii. Calculate median and mode stock price.

iii. Compute the standard deviation of the stock price.
iv. Find the min-max normalized stock price for the stock worth $20.
v. Compute the Z – score standardized stock price for the stock worth $20.
vi. Find the decimal scaling stock price for the stock worth $20.
vii. Identify all possible stock prices that would be outliers, using IQR method.

c) Explain Apriori algorithm.

10. a) Explain any 5 functionalities of data mining.

b) i. The data for analysis include the attribute rank. The rank values for the data tuples are (in ascending order)
X = {1, 1, 1, 1, 1, 2, 2, 11, 11, 12, 12, 44}.
Apply Equal width binning and Equal frequency binning on the above set.
ii. The weight of chocolate bars from a particular chocolate factory has a mean of 8 ounces with standard deviation of 0.1 ounce. What is the z-score corresponding to a weight of 8.17 ounces?

c) Generate 3 frequent itemset using aprioritid algorithm for the given database. {Consider support count as 2}

Tid	List of items
T100	I1, I2, I5
T200	I2, I4
T300	I1, I2, I3, I5
T400	I1, I2, I3
T500	I1, I2, I4

11. a) Explain KDD process with a neat diagram.

b) List and explain data mining issues.

c) Generate 3 frequent itemset for the given database using AprioriTid algorithm. {Consider support count as 3}.

Transaction Id	**Items**
1	R, S
2	U
3	P, R
4	R, S, U
5	Q, R, U
6	S, U, V
7	P
8	P, R, S
9	R, S, U, V

10	S, T
11	Q, S
12	S, V
13	T, U
14	R, S, U

12. a) Explain any five functionalities of data mining.

b) The data for analysis include the attribute rank. The rank values for the data tuples are (in ascending order)

X = {1, 1, 1, 1, 1, 2, 2, 11, 11, 12, 12, 44}.

Apply Equal width binning and Equal frequency binning on the above set.

c) Explain with examples the different transformation techniques used on data to be mined.

13. a) What is data set? Explain various types of data sets.

b) Discuss various criterias to classify data mining systems.

14. a) Briefly explain any five data preprocessing approaches.

b) Develop the Apriori algorithm for generating frequent item set.

15. a) Identify and explain four common methods used for binning numerical predictors.

b) Use the following stock price data (in dollars)

20 30 40 60 10 15 21 22 70

i. Calculate the mean stock price.

ii. Calculate median and mode stock price.

iii. Compute the standard deviation of the stock price.

iv. Find the min-max normalized stock price for the stock worth $105.

v. Compute the Z – score standardized stock price for the stock worth $105

vi. Find the decimal scaling stock price for the stock worth $105.

vii. Identify all possible stock prices that would be outliers, using IQR method.

c) Explain functionalities of data mining.

16. a) Explain with examples the different transformation techniques used on data to be mined.

b) i. The data for analysis include the attribute rank. The rank values for the data tuples are (in ascending order)
X = {1, 1, 1, 1, 1, 2, 2, 11, 11, 12, 12, 44}.
Apply Equal width binning and Equal frequency binning on the above set.
ii. The weight of chocolate bars from a particular chocolate factory has a mean of 8 ounces with standard deviation of 0.1 ounce. What is the z-score corresponding to a weight of 8.17 ounces?

c) Generate 3 frequent itemset using Apriori-Tid algorithm for the given database. {Consider support count as 2}

Tid	List of items
T100	I1, I2, I5
T200	I2, I4
T300	I1, I2, I3, I5
T400	I1, I2, I3
T500	I1, I2, I4

17. a) Describe how data mining systems are classified and explain its functionalities.

b) Use the following stock price data (in dollars)

10 7 20 12 75 15
9 18 4 12 8 14

i. Calculate the mean stock price.
ii. Calculate median and mode stock price.
iii. Compute the standard deviation of the stock price.
iv. Find the min-max normalized stock price for the stock worth $20.
v. Compute the Z – score standardized stock price for the stock worth $20
vi. Find the decimal scaling stock price for the stock worth $20.
vii. Identify all possible stock prices that would be outliers, using Z-score method.
viii. Identify all possible stock prices that would be outliers, using IQR method.

18. a) Write apriori and association rule generation algorithm.

b) Identify and explain any four types of data that can be mined.

c) Suppose that the data for analysis includes the attribute rank. The rank values for the data tuples are (in increasing order):

X = {1, 1, 1, 1, 1, 2, 2, 11, 11, 12, 12, 44}.

Apply Equal width binning and Equal frequency binning on the above set.