**Question Bank**

**Module 1: DWH and Data Mining**

**1. What is DWH? Explain DWH characteristics.**

* A data warehouse is a large collection of business data used to help an organization make decisions.
* A system used for reporting and data analysis ---a core component of business intelligence.
* DWs are central repositories of integrated data from one or more disparate sources. They store current and historical data in one single place.
* According to William H. Inmon, a leading architect in the construction of data warehouse systems, “A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process”

**DWH Characteristics**

**Subject-oriented**:

* A data warehouse is organized around major subjects.
* A data warehouse focuses on the modeling and analysis of data for decision-makers.
* Hence, data warehouses typically provide a simple and concise view of particular subject issues.

**Integrated**:

* A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and online transaction records.
* Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, and so on.

**Time-variant**:

* Data are stored to provide information from a historical perspective (e.g., the past 5–10 years). Every key structure in the data warehouse contains, either implicitly or explicitly, a time element.
* Data in the DWH is mainly meant for data mining and forecasting.

**Nonvolatile**:

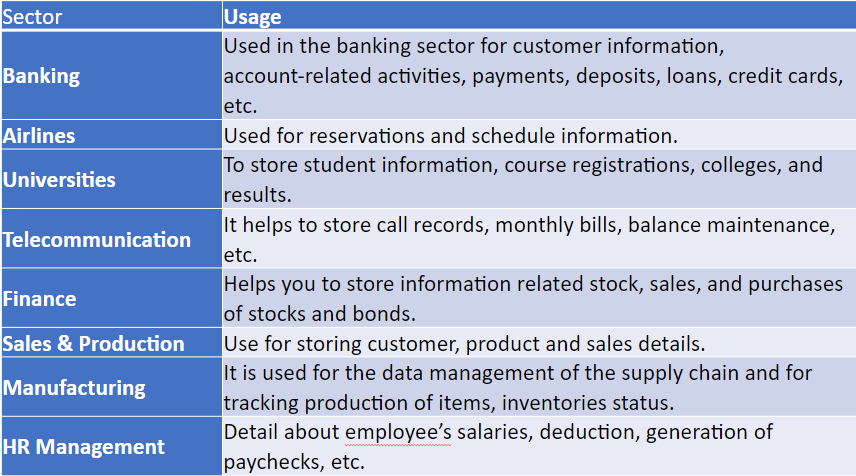
* The data in the data warehouse is read-only.

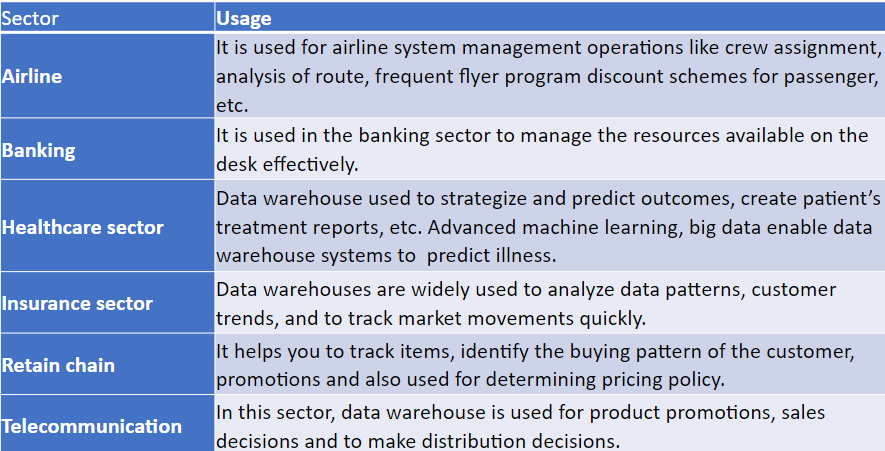
**2. What are the advantages and applications of DWH?**

**Advantages of Data Warehouse (DWH):**

* A data warehouse allows users to access critical data from a number of sources in a single place. Therefore, it saves users time in retrieving data from multiple sources.
* The Data warehouse provides consistent information on various cross-functional activities. It also supports ad-hoc reporting and queries.
* Data Warehouse helps to integrate many sources of data to reduce stress on the production system.
* Data warehouse helps to reduce total turnaround time for analysis and reporting.
* Restructuring and Integration make it easier for the user to use for reporting and analysis.
* The data warehouse stores a large amount of historical data. This helps users to analyze different periods and trends to make future predictions.

**Applications of Data Warehousing**





**3. Why is the ER model not suitable for DWH? What are the steps in dimensional modeling?**

A dimensional model in the data warehouse is designed to read, summarize, and analyze numeric information like values, balances, counts, weights, etc. in a data warehouse.

In contrast, relation models are optimized for addition, updating, and deletion of data in a real-time Online Transaction System.

Dimensional models are used in data warehouse systems and not a good fit for relational systems.

A dimensional model contains the same information as the ER model but packages the data in a symmetric format whose design goals are easy understandability, query performance, and resilience to change.

ER modeling aims to optimize performance for transaction processing. It is also hard to query ER models because of the complexity; Therefore ER models are not suitable for high-performance retrieval of data.

Data warehouses contain huge information. Data can't be fetched by normal technique so it requires special techniques.

The Entity-Relationship (ER) model is primarily designed for conceptual modeling of data within a specific application domain. While it's great for representing entities, their attributes, and the relationships between them in a relational database context, it may not be the best fit for a Data Warehouse (DWH) environment for several reasons:

**Complexity of Data**: Data Warehouses often deal with large volumes of data from various sources. The ER model may not adequately capture the complexity of these data relationships, hierarchies, and aggregations.

**Performance**: ER models may not be optimized for query performance in a Data Warehouse environment. Data Warehouses typically require denormalization, aggregation, and other optimizations to support efficient querying and reporting, which may not be easily represented in an ER model.

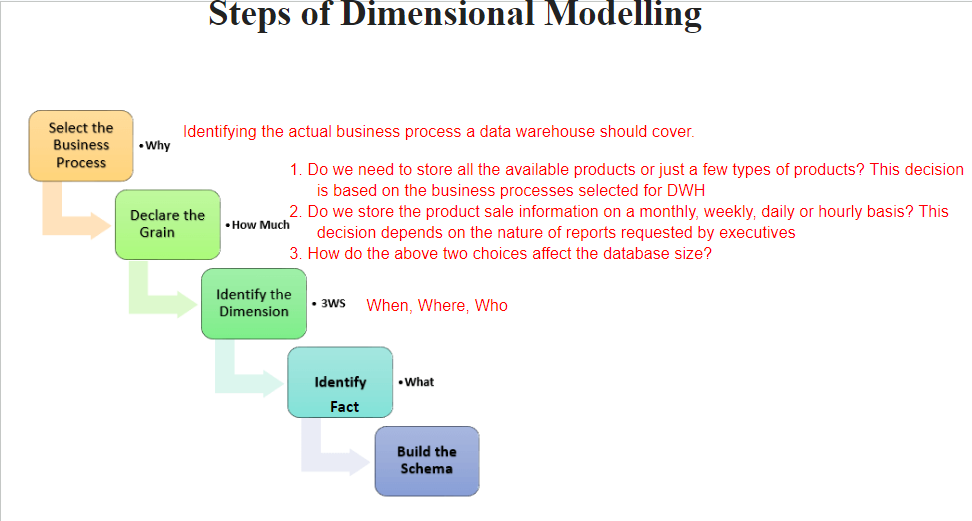
**Historical Data**: Data Warehouses often store historical data for analysis and reporting purposes. The ER model may not have built-in constructs to handle slowly changing dimensions, temporal data, or versioning, which are common requirements in Data Warehousing.

**Multi-dimensional Modeling**: Data Warehouses commonly use multi-dimensional modeling techniques such as star schemas or snowflake schemas to organize data for analytical purposes. These models are more specialized for data analysis and reporting than the generic entity-relationship model.

**Data Integration**: ER models are not inherently designed for integrating data from disparate sources, which is a key requirement in Data Warehousing. Data Warehouse environments often involve data integration, transformation, and cleansing processes, which may not be well-represented in an ER model.

**Aggregation and Summarization**: Data Warehouses often require pre-aggregated or summarized data to support analytical queries efficiently. While ER models can represent relationships between entities, they may not capture the need for pre-computed aggregates that are essential for reporting and analysis in a Data Warehouse.

Because of these limitations, Data Warehouses typically use specialized modeling techniques and architectures, such as dimensional modeling, which are better suited for the specific requirements of analytical querying and reporting on large volumes of data.



**Steps to Create Dimensional Data Modeling**

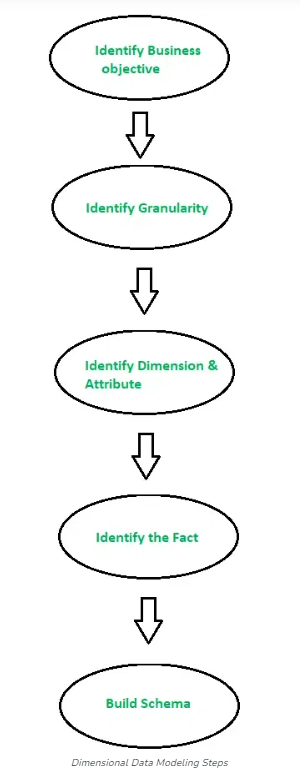
**Step-1**: Identifying the business objective: The first step is to identify the business objective. Sales, HR, Marketing, etc. are some examples of the need of the organization. Since it is the most important step of Data Modelling the selection of business objectives also depends on the quality of data available for that process.

**Step 2**: Identifying Granularity: Granularity is the lowest level of information stored in the table. The level of detail for business problems and its solution is described by Grain.

**Step 3**: Identifying Dimensions and their Attributes: Dimensions are objects or things. Dimensions categorize and describe data warehouse facts and measures in a way that supports meaningful answers to business questions. A data warehouse organizes descriptive attributes as columns in dimension tables. For Example, the data dimension may contain data like a year, month, and weekday.

**Step-4**: Identifying the Fact: The measurable data is held by the fact table. Most of the fact table rows are numerical values like price or cost per unit, etc.

**Step-5**: Building of Schema: We implement the Dimension Model in this step. A schema is a database structure. There are two popular schemes: Star Schema and Snowflake Schema.



**4. Define dimension, fact, fact table, and dimension table with examples.**

**Fact**

* Facts are the measurements/metrics from your business process.
* For a Sales business process, a measurement would be a quarterly sales number

**Dimension**

* A category of information. For example, the time dimension.
* In simple terms, they give who, what, and where of a fact.
* E.g.In the Sales business process, for the fact quarterly sales number, dimensions would be
* Who - Customer Names
* Where - Location
* What - Product Name
* When - Time Dimension
* In other words, a dimension is a window to view information in the facts.

**Attributes**

* The Attributes are the various characteristics of the dimension
* E.g. In the Location dimension, the attributes can be State, Country, Zipcode, etc.
* Attributes are used to search, filter, or classify facts. Dimension Tables contain Attributes.

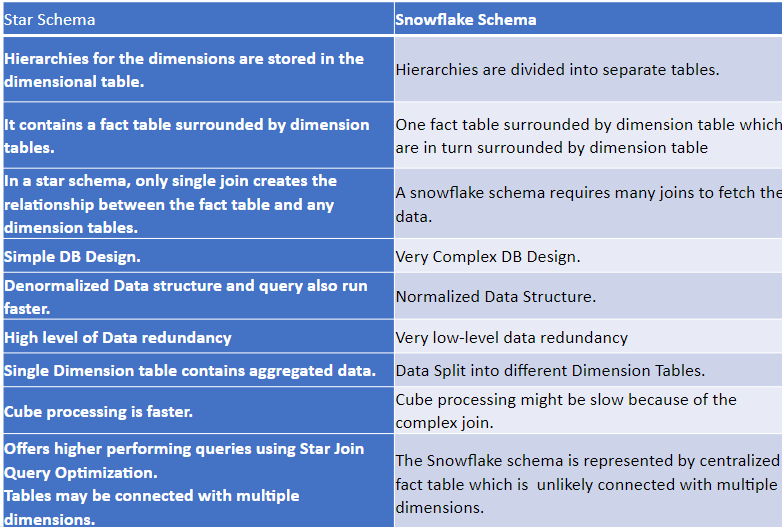
**Fact Table**

* A fact table is a primary table in dimension modeling.
* A fact table consists of the measurements, metrics, or facts of a business process.
* Eg. Monthly sales volume, Average Customer Balance, etc...
* A Fact Table contains
* 1. Measurements/facts
* 2. Foreign key to the dimension table

**Dimension Table**

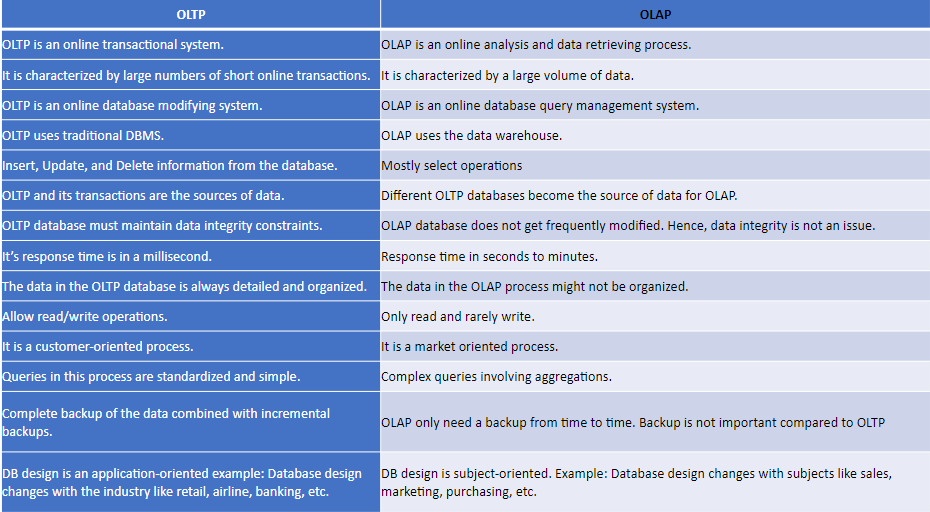
* A dimension table contains the dimensions of a fact.
* They are joined to the fact table via a foreign key.
* Dimension tables are denormalized tables.
* The Dimension Attributes are the various columns in a dimension table
* Dimensions offer descriptive characteristics of the facts with the help of their attributes
* No limit is set for a number of dimensions
* The dimension can also contain one or more hierarchical relationships

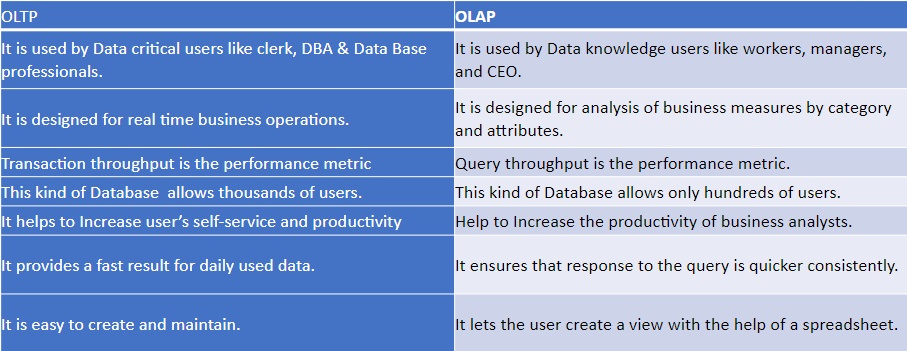
**5. Difference between star and snowflake schema.**

****

**6. Design star and snowflake schema for a given system.**

**7. Difference between OLTP and OLAP.**

****

****

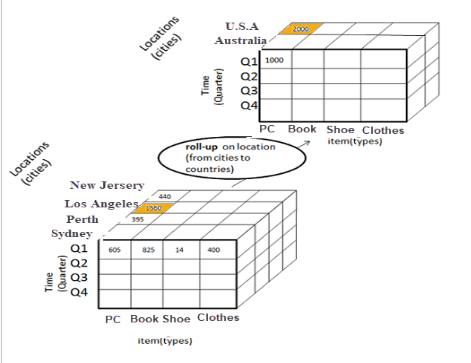
**8. What are different OLAP operations? Explain with an example.**

**Roll-up:**

Roll-up is also known as “consolidation” or “aggregation.” The Roll-up operation can be performed in 2 ways

* Reducing dimensions
* Climbing up concept hierarchy.

In the roll-up process, at least one or more dimensions need to be removed.



In this example, cities New Jersey and Los Angeles are rolled up into country USA

The sales figures of New Jersey and Los Angeles are 440 and 1560 respectively. They become 2000 after roll-up

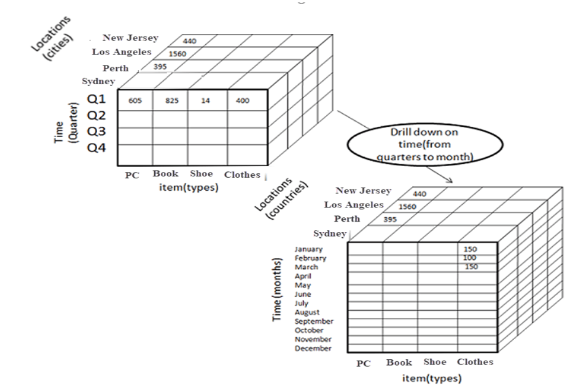
In this aggregation process, data in location hierarchy moves up from the city to the country.

In this example, the Cities dimension is removed.

**Drill-down**

In drill-down, data is fragmented into smaller parts. It is the opposite of the rollup process. It can be done via

* Moving down the concept hierarchy
* Increasing a dimension



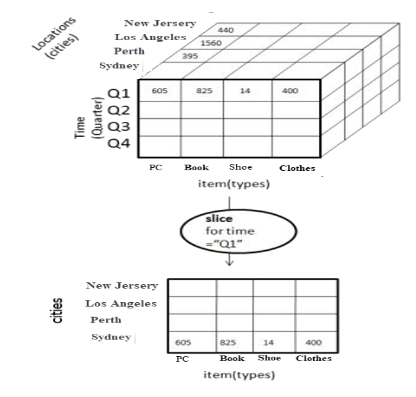
In this Example,

Quarter Q1 is drilled down to months

Here dimension Months is added.

**Slice**

Here, one dimension is selected, and a new sub-cube is created.



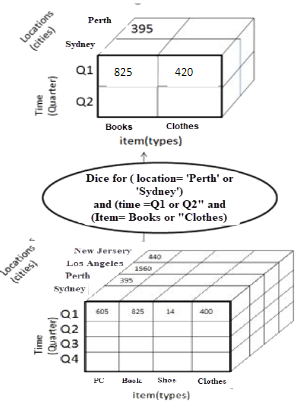
In this example,

Dimension Time is sliced with Q1 as the filter.

A new cube is created altogether.

**Dice**

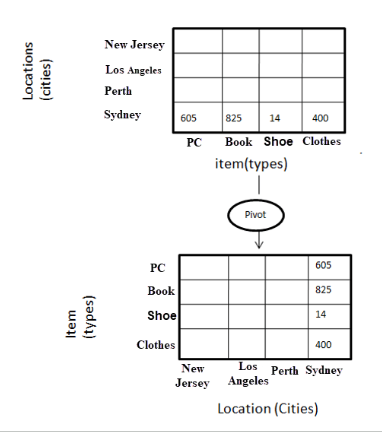
This operation is similar to a slice. The difference in dice is you select 2 or more dimensions that result in the creation of a sub-cube.



Dice operation in OLAP

**Pivot**

In Pivot, you rotate the data axes to provide a substitute presentation of data.

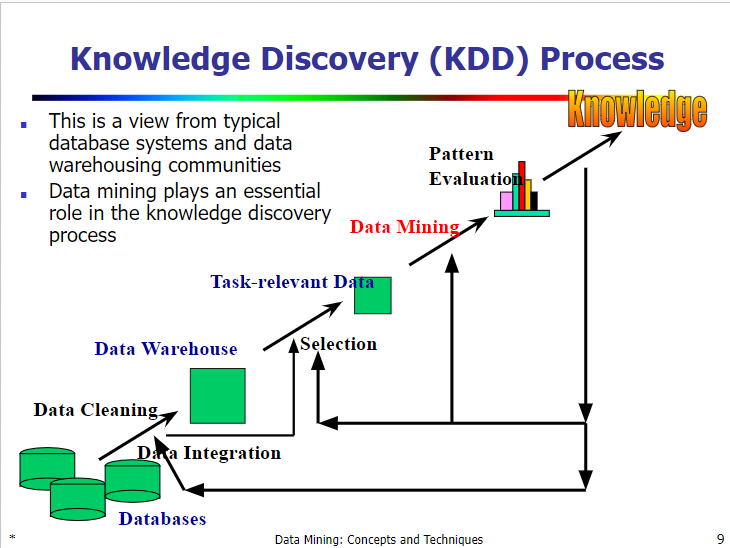


Pivot operation in OLAP

Here the pivot is based on item types

**9. Problems with writing a sequence of OLAP operations for the given query.**

**10. Explain the steps of KDD**

****

**KDD Process**

KDD (Knowledge Discovery in Databases) is a process that involves the extraction of useful, previously unknown, and potentially valuable information from large datasets. The KDD process is an iterative process and it requires multiple iterations of the above steps to extract accurate knowledge from the data.The following steps are included in **KDD process:**

**Data Cleaning**

Data cleaning is defined as removal of noisy and irrelevant data from collection.

1. Cleaning in case of Missing values.
2. Cleaning noisy data, where noise is a random or variance error.
3. Cleaning with Data discrepancy detection and Data transformation tools.

**Data Integration**

Data integration is defined as heterogeneous data from multiple sources combined in a common source(DataWarehouse). Data integration using Data Migration tools, Data Synchronization tools and ETL(Extract-Load-Transformation) process.

**Data Selection**

Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection. For this we can use Neural network, Decision Trees, Naive bayes, Clustering, and Regression methods.

**Data Transformation**

Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure. Data Transformation is a two step process:

1. Data Mapping: Assigning elements from source base to destination to capture transformations.
2. Code generation: Creation of the actual transformation program.

**Data Mining**

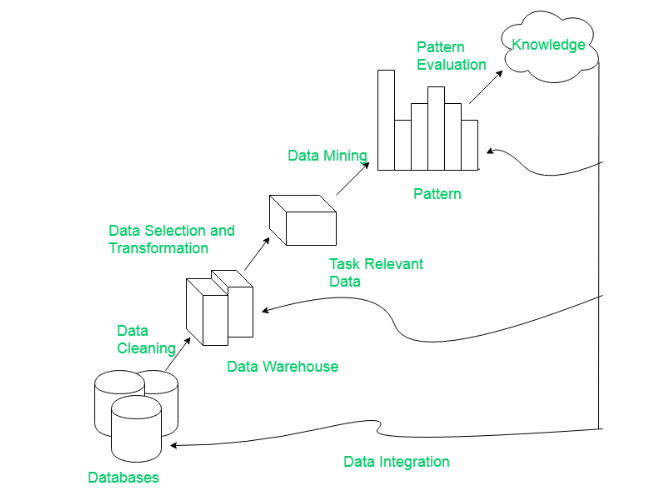
Data mining is defined as techniques that are applied to extract patterns potentially useful. It transforms task relevant data into patterns, and decides purpose of model using classification or characterization.

**Pattern Evaluation**

Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures. It find interestingness score of each pattern, and uses summarization and Visualization to make data understandable by user.

**Knowledge Representation**

This involves presenting the results in a way that is meaningful and can be used to make decisions.



Note: KDD is an iterative process where evaluation measures can be enhanced, mining can be refined, and new data can be integrated and transformed in order to get different and more appropriate results. Preprocessing of databases consists of Data cleaning and Data Integration.

**11. State any 2 decision-making activities for which organizations are using data in DWH.**

Many organizations use this information to support business decision-making activities, including

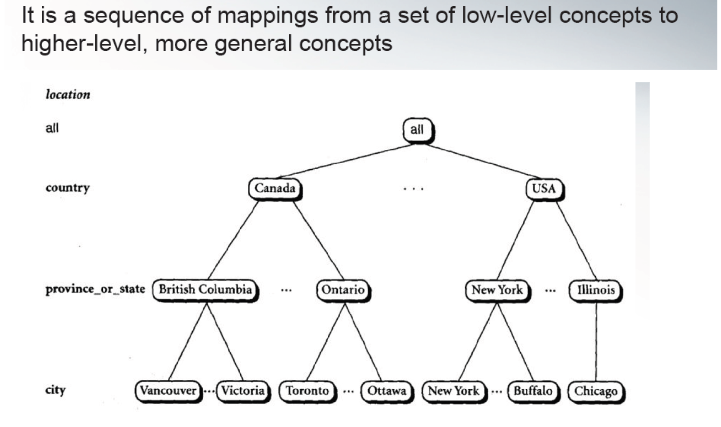
(1) increasing customer focus, which includes the analysis of customer buying patterns ;

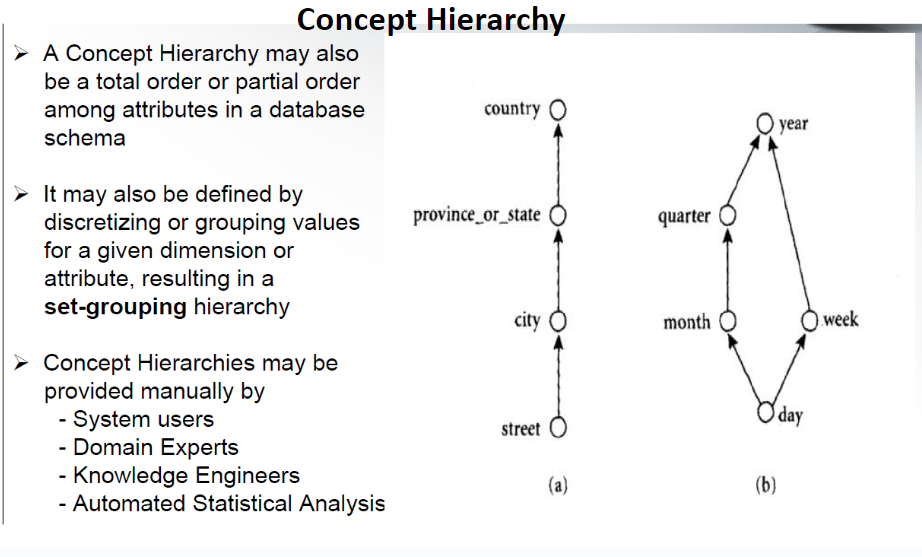
(2) repositioning products and managing product portfolios (by comparing the performance of sales by quarter, by year, and by geographic regions in order to fine-tune production strategies);

(3) analyzing operations and looking for sources of profit; and

(4) managing customer relationships, making environmental corrections, and managing the cost of corporate assets.

**12. What is concept hierarchy, partial and total order concept hierarchy? Ex[plain with an example.**

****

****

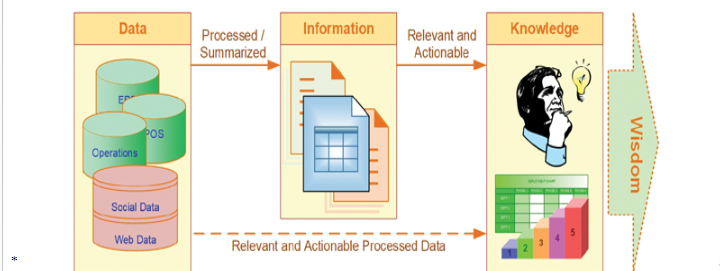
**13. What is data mining? State applications of data mining.**

Data mining is the process of converting data into information and then into knowledge.

Knowledge is very distinct from data and information

Knowledge is information that is contextual, relevant, and actionable.

knowledge has strong experiential and reflective elements that distinguish it from information in a given context.



Data mining is a process that involves using statistical, mathematical, and artificial intelligence techniques and algorithms to extract and identify useful information and subsequent knowledge (or patterns) from large sets of data.



**Marketing and CRM:**

-To identify most likely buyers of new products

-To identify root causes of customer attrition so as to improve customer retention

-To discover time variant associations between products and services to maximize sale and find most profitable customers.

**Banking and Finance:**

-To detect fraudulent credit card and online banking transactions

-To optimize the cash return by forecasting cash flow on banking entities

-To streamline and automate the processing of loan applications by accurately predicting the most probable defaulters.

-To maximize customer value by identifying and selling the products and services that customers are most likely to buy.

**Retailing and Logistics:**

-To identify accurate sales volume at specific retail locations in order to determine correct inventory levels.

-To do an MBA to improve store layout and optimize sales promotions

-To forecast consumption levels for different product types.

-To discover interesting patterns in the movement of products in a supply chain by analyzing sensory and RFID data.

**Manufacturing:**

- To predict machine failures using sensory data

- To discover novel patterns to identify and improve product quality.

**Brokerages and Security Tradings:**

-To predict when and how much certain stock/bond prices will change.

-To forecast the range of market fluctuations, and direction of fluctuations

-To assess the effect of particular issues/events on market movements.

-To identify and prevent fraudulent activities in security trading.

**Insurance:**

- To predict which customers will buy new policies

- Identify fraudulent behavior of customers

- Prevent incorrect claim payments

**Computer Hardware and Software:**

-To predict disk failure

- To identify and filter unwanted web contents and email messages

-To identify potentially unsecured software products

**Government and Defense:**

-To forecast the cost of moving military personnel and equipment.

-To predict resource consumption for better planning and budgeting

**Travel and Lodging:**

- To predict sales of different services to optimally price these services.

- To forecast demand at different locations to better allocate limited organizational resources. .

- To identify the most profitable customers and provide them with personalized services.

- To retain valuable employees by identifying and acting on the root causes for attrition

**Health and Healthcare:**

- To identify successful medical therapies for different illnesses.

- To identify people without health insurance and the reasons behind it.

- To forecast the time of demand at different service locations to optimally allocate organizational resources.

- To retain valuable employees by identifying root causes for attrition

**Entertainment:**

To analyze viewer data to determine which programs to show during prime time.

To decide where to insert advertisements so as to maximize the returns.

To predict the financial success of the movies before they are produced.

Sports:

To improve the performance of NBA teams in the US

To increase the chances of winning.

**14. What are the different types of patterns that can be mined?**

Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks.

Descriptive mining tasks: Deals with the General characteristics and converts them into relevant and useful information

Predictive mining tasks: Predicts future values by analyzing data patterns and their outcomes based on past data.

**Descriptive DM Functionalities**

**1. Class/Concept Description:**

Data entries can be associated with the classes or concepts.

These descriptions can be derived using

(1) data characterization, by summarizing the data of the class under study (often called the target class) in general terms,

Example: At an electronic store a Customer relationship manager asks to Summarize the characteristics of customers who spend more than Rs.10000 a year at the store.

or

(2) data discrimination, by comparison of the target class with one or a set of

comparative classes (often called the contrasting classes),

Example: A customer relationship manager at an Electronics store may want to

compare two groups of customers—those who shop for computer products regularly (e.g., more than twice a month) and those who rarely shop for such products (e.g. less than three times a year).

or

(3) both data characterization and discrimination.

**Mining of frequent patterns:**

Patterns that occur frequently in data.

It includes--

* **Frequent item**: refers to a set of items that often appear together in a transactional data set
* **Frequent subsequences** (also known as sequential patterns): A frequently occurring subsequence like laptop🡪digital camera🡪 memory card
* **Frequent substructures:**
* A substructure can refer to different structural forms (e.g., graphs, trees, or lattices) that may be combined with itemsets or subsequences.
* If a substructure occurs frequently, it is called a (frequent) structured pattern.

Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

**Association Analysis**

Defines relationships between the data and predefined association rules.

Suppose that, as a marketing manager at an Electronics store, you want to know which items are frequently purchased together

A rule, mined from the Electronics store transactional database

Association rules that contain a single predicate are referred to as single-dimensional association rules.

Suppose, instead, that we are given the Electronics relational database related to purchases. A data mining system may find association rules like

Association rules that contain more than one predicate/attribute are referred to as multi-dimensional association rules.

**Clustering**:

* can be used to generate class labels for a group of data. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity.
* i.e. clusters of objects are formed so that objects within a cluster have high similarity, but are rather dissimilar to objects in other clusters.
* Clustering can also facilitate taxonomy formation 🡪 Organization of observations into a hierarchy of classes that group similar events together.

Example: Cluster analysis can be performed on Electronics store customer data to identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing

**Predictive Data mining functionalities**

Predicts future values by analyzing data patterns and their outcomes based on past data.

* Classification
* Regression
* Outlier analysis

**Classification**:

* Is the process of finding a model (or function) that describes and distinguishes data classes or concepts.
* The models are derived based on the analysis of a set of training data (i.e., data objects for which the class labels are known).
* The model is used to predict the class label of objects for which the class label is unknown.
* 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

**Regression:**

* Whereas classification predicts categorical (discrete, unordered) labels, regression models continuous-valued functions. That is, regression is used to predict missing or unavailable numerical data values rather than (discrete) class labels.
* The term prediction refers to both numeric prediction and class label prediction.
* Regression analysis is a statistical methodology that is most often used for numeric prediction.
* Regression also encompasses the identification of distribution trends based on the available data.

**Outlier Analysis:**

* Outlier: A data object that does not comply with the general behavior of the data
* Many data mining methods discard outliers as noise or exceptions. However, in some applications (e.g., fraud detection), the rare events can be more interesting than the more regularly occurring ones.
* Outliers may be detected using statistical tests that assume a distribution or probability model for the data, or using distance measures where objects that are remote from any other cluster are considered outliers.
* Example: Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of unusually large amounts for a given account number in comparison to regular charges incurred by the same account.
* Outlier values may also be detected with respect to the locations and types of purchase, or the purchase frequency.
* It is used in observing the change in trends of buying patterns of a customer.

**Module 2: Preprocessing**

**1. What are the different types of attributes? Explain with examples**

**Attribute (or dimensions, features, variables):**

A data field, representing a characteristic or feature of a data object.

e.g., customer \_ID, name, address

**Categorical/Qualitative Attribute**

**1. Nominal:**

* categories, states, or “names of things”
  + Hair\_color = {auburn, black, blond, brown, grey, red, white}
  + marital status, occupation, ID numbers, zip codes
* In the cases of nominal attributes with numeric values e.g. Cust\_ID, the numbers are not intended to be used quantitatively.
* Also in the case of numeric nominal attributes, values do not have any meaningful order about them.

**2. Binary attributes**

* Nominal attribute with only 2 categories/states (0 or 1)
  + 0: attribute is absent
  + 1: attribute is present
* Symmetric binary: both outcomes equally important
  + e.g., gender
* Asymmetric binary: outcomes not equally important.
  + e.g., medical test (positive vs. negative)
  + Convention: assign 1 to the most important outcome (e.g., HIV positive)
* If two states are True and False,then called as Boolean Attribute

**3. Ordinal Attributes:**

* Values have a meaningful order or a ranking among them but the magnitude between successive values is not known.
  + Ex: Size = {small, medium, large}, grades, professor rankings
* Useful for registering subjective assessments of qualities that cannot be measured objectively; thus often used in surveys for ratings.
  + E.g. Customer satisfaction survey
* We can compute mean and median but not mode for the ordinal attributes.
* Note: nominal, binary, and ordinal attributes are qualitative attributes.

**Numeric Attributes /Quantitative Attributes**

Represent measurable quantity in integer or real values

**1. Interval scaled attributes**

* Measured on a scale of equal-sized units
* Values have order and can be positive, 0, or negative
  + E.g., the temperature in C˚or F˚, calendar dates
* We can obtain a ranking of objects by ordering the values.
* Also allow us to compare and quantify the difference between values.
* No true zero-point -We can not speak of values in terms of ratio.
  + e.g. without a true zero point, we can’t say that 10 C˚ is twice as warm as 5C˚.
* Mean , Median and Mode

**2. Ratio scaled attributes**

* Inherent zero-point
* The values are ordered, and we can also compute the difference between values, as well as the mean, median, and mode
* Examples: Count attributes such as years of experience and number of words attribute for a document
* attributes to measure age, weight, height and monetary quantities (e.g., you are 100 times richer with $100 than with $1).

**Discrete vs. Continuous Attributes**

**Discrete Attribute**

* Has only a finite or countably infinite set of values which may or may not be represented as integers.
  + E.g., zip codes, profession, or the set of words in a collection of documents
* Sometimes, represented as integer variables
* Note: Binary attributes are a special case of discrete attributes

**Continuous Attribute**

* Has real numbers as attribute values
  + E.g., 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

**2. Problems on basic statistical descriptions of data like finding mean, median, midrange standard deviation, variance,modes for given data.Drawing boxplot for**

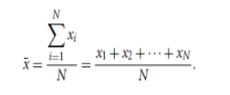
**given data to identify outliers.**

**Mean (algebraic measure) (sample vs. population):**

The most common and effective numeric measure of the “center” of a set of data is the (arithmetic) mean.

Let x1, x2, x3, x4, ….. xN be a set of N values or observations, such as for some numeric attribute X, like salary.

The mean of this set of values is



Weighted mean is :

Problem with mean is its sensitivity to extreme values.

For skewed (asymmetric) data, a better measure of the center of data is the median

**3. What is a five-number summary of data?**

* Five-number summary of a distribution (Minimum, Q1, Median, Q3, Maximum)
* It is more informative to also provide the two quartiles Q1 and Q3, along

with the median

* A common rule of thumb for identifying suspected outliers is to single out values falling at least 1.5IQR above the third quartile or below the first quartile.
* Because Q1, the median, and Q3 together contain no information about the endpoints (e.g., tails) of the data, a fuller summary of the shape of a distribution can be obtained by providing the lowest and highest data values as well. This is known as the five-number summary.
* The five-number summary of distribution consists of the median (Q2), the quartiles Q1 and Q3, and the smallest and largest individual observations, written in the order of Minimum, Q1, Median, Q3, and Maximum.

**4. How can we compute dissimilarity between two binary attributes?**

**5. What is Euclidean distance, Manhattan distance, Minkowski distance? Problems on computing these distances between given objects.**

**6. What is cosine similarity?Problems on finding similarity between given documents.**

**7. Problems based on finding dissimilarity matrices between nominal,binary and ordinal attributes .**

**8. Explain in brief the major tasks in data preprocessing.**

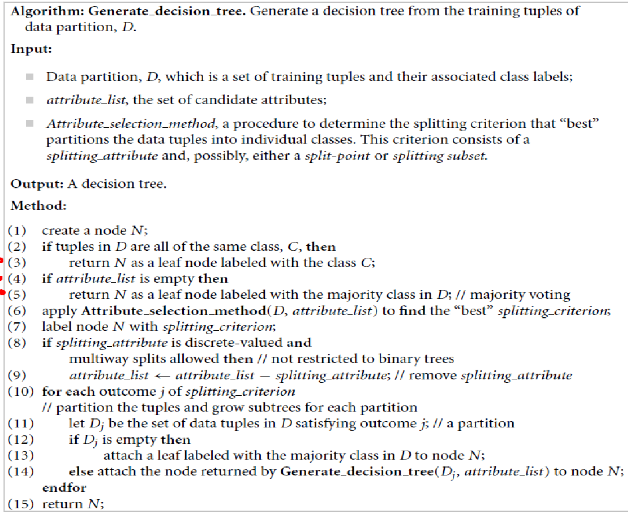
* Data cleaning
  + Fill in missing values, smooth out the noisy data, identify or remove outliers, and resolve inconsistencies
* Data integration
  + Integration of multiple databases, data cubes, or files
* Data transformation
  + Normalization and aggregation
* Data reduction
  + Obtains reduced representation in volume but produces the same or similar analytical results
* Data discretization
  + Part of data reduction but with particular importance, especially for numerical data

**9. Problems based on finding correlation between attributes (Chi Square test, Pearson correlation coefficient, covariance, etc…)**

**Module 3: Classification**

**1. Write Decision Tree algorithm: ID3, C4.5, and CART algorithms**

* ID3 (Iterative Dichotomiser), C4.5, CART(Classification And Regression Trees) Algorithms.
* These algorithms adopt a greedy (i.e., non backtracking) approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner.
* It starts with a training set of tuples and their associated class labels. The training set is recursively partitioned into smaller subsets as the tree is being built.

****

**ID3 (Iterative Dichotomiser 3)**

* An approach for decision trees called ID3 (Iterative Dichotomiser 3) is employed in classification applications. It is one of the first and most used decision tree algorithms, created by Ross Quinlan in 1986. The ID3 algorithm builds a decision tree from a given dataset using a greedy, top-down methodology.
* It works by greedily choosing the feature that maximizes the information gain at each node. ID3 calculates entropy and information gain for each feature and selects the feature with the highest information gain for splitting.
* ID3 uses entropy to measure the uncertainty or disorder in a dataset. Entropy, denoted by H(D) for dataset D, is calculated using the formula:



Information gain quantifies the reduction in entropy achieved by splitting the data based on a particular feature. Features with higher information gain are preferred for splitting. Information gain is calculated as follows:



* Every decision tree node’s dataset is recursively divided using the ID3 algorithm according to the chosen attribute. This method keeps going until either there are no more attributes to divide on, or all the examples in a node belong to the same class.
* The decision tree may be trimmed after it is constructed in order to enhance generalization and lessen overfitting. In order to do this, nodes that do not considerably improve the correctness of the tree must be removed.
* A couple of the ID3 algorithm’s drawbacks are that it tends to overfit the training set and cannot directly handle continuous attributes. Owing to these drawbacks, other decision tree algorithms that address some of these problems have been developed, including C4.5 and CART.
* Entropy, information gain, and recursive partitioning are three key principles in the ID3 algorithm, which is a fundamental technique for creating decision trees. Mastering these ideas is crucial to learning about decision tree algorithms in machine learning.

**C4.5**

* As an enhancement to the ID3 algorithm, Ross Quinlan created the decision tree algorithm C4.5. In machine learning and data mining applications, it is a well-liked approach for creating decision trees. Certain drawbacks of the ID3 algorithm are addressed in C4.5, including its incapacity to deal with continuous characteristics and propensity to overfit the training set.
* A modification of information gain known as the gain ratio is used to address the bias towards qualities with many values. It is computed by dividing the information gain by the intrinsic information, which is a measurement of the quantity of data required to characterize an attribute’s values.



* Where Split Information represents the entropy of the feature itself. The feature with the highest gain ratio is chosen for splitting.
* When dealing with continuous attributes, C4.5 sorts the attribute’s values first, and then chooses the midpoint between each pair of adjacent values as a potential split point. Next, it determines which split point has the largest value by calculating the information gain or gain ratio for each.
* By turning every path from the root to a leaf into a rule, C4.5 can also produce rules from the decision tree. Predictions based on fresh data can be generated using the rules.
* C4.5 is an effective technique for creating decision trees that can produce rules from the tree and handle both discrete and continuous attributes. The model’s accuracy is increased and overfitting is prevented by its utilization of gain ratio and decreased error pruning. Nevertheless, it might still be susceptible to noisy data and might not function effectively on datasets with a lot of features.

**2. Explain Attribute selection measures (Information Gain, Gain Ratio, Gini Index )**

An attribute selection measure is a heuristic for selecting the splitting criterion that “best” separates a given data partition, D , of class-labeled training tuples into individual classes.

The attribute selection measure provides a ranking for each attribute describing the given training tuples.

The three popular attribute selection measures—

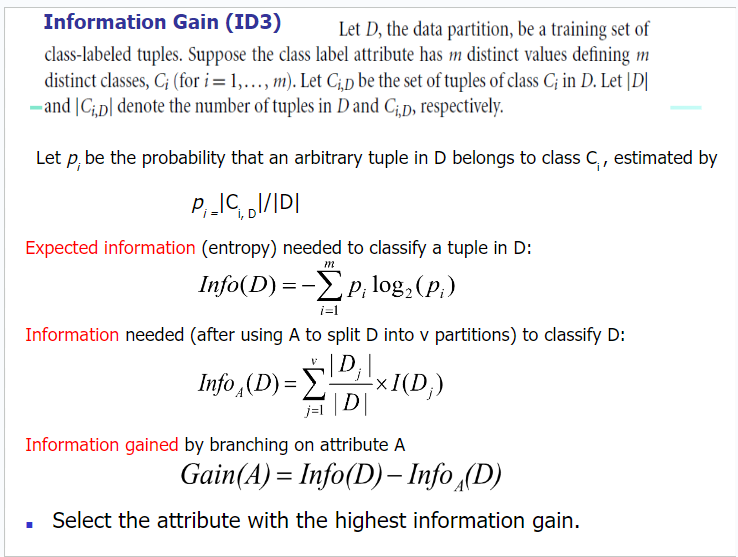
information gain

gain ratio, and

Gini index

**Information Gain:**

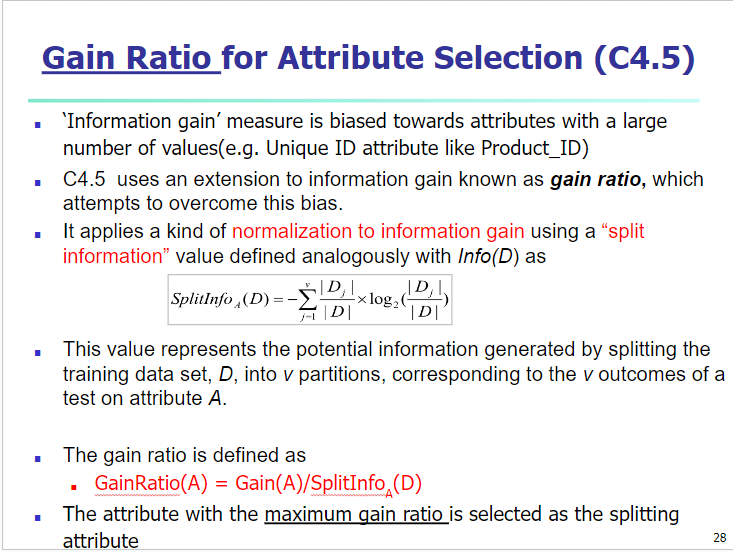
* This measure is based on pioneering work by Claude Shannon on information theory, which studied the value or “information content” of messages.
* The attribute with the highest information gain is chosen as the splitting attribute for node N.
* This attribute minimizes the information needed to classify the tuples in the resulting partitions and reflects the least randomness or “impurity” in these partitions.
* Such an approach minimizes the expected number of tests needed to classify a given tuple and guarantees that a simple (but not necessarily the simplest) tree is found.



Pg-18-24,26,27

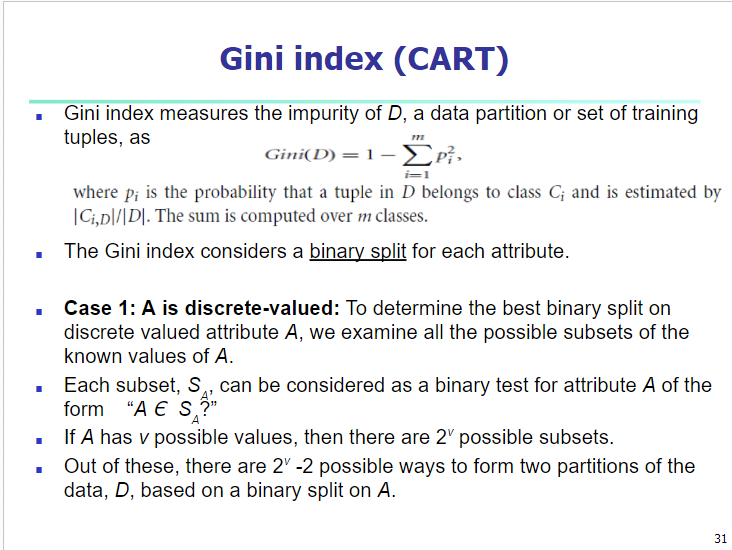
<https://docs.google.com/presentation/d/1R7181g1_QNMPx7wyOm2do2b92fw43_31/edit?usp=sharing&ouid=110281631223246598554&rtpof=true&sd=true>

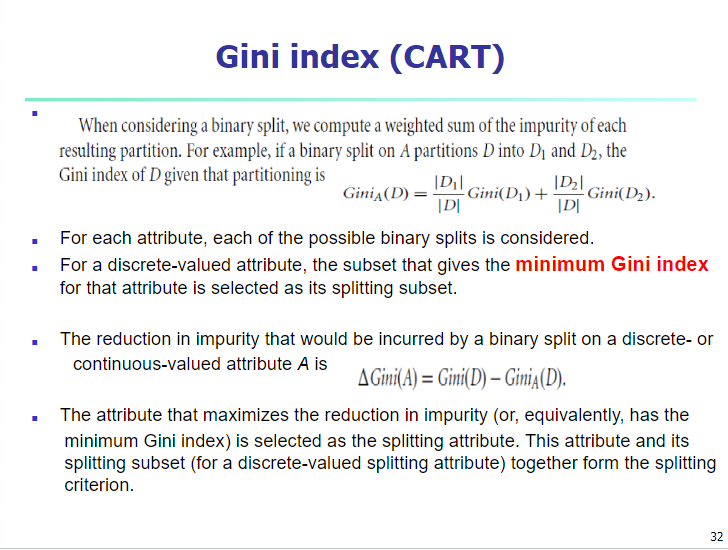
Sums

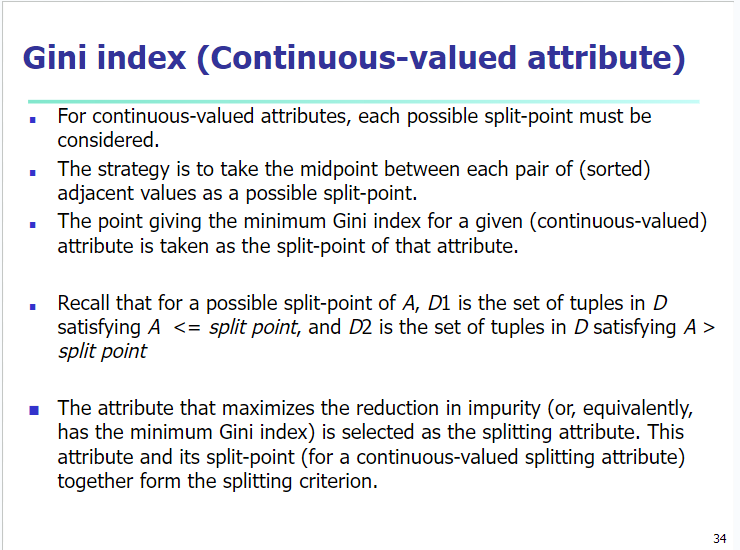
****

Pg-29,30(SUM)

Same Link







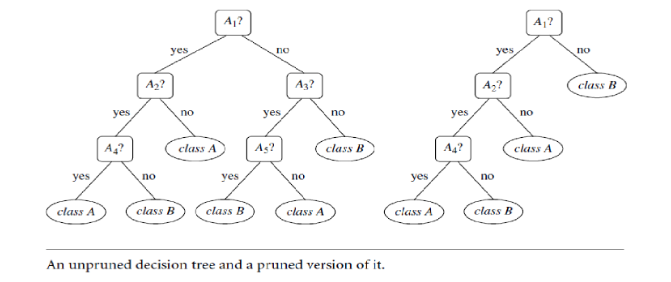
Pg-35,36,37

Same PPT

**3. What is Overfitting and Tree Pruning**

**Overfitting:**

* When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers.
* Tree pruning methods address this problem of overfitting the data.
* Such methods typically use statistical measures to remove the least-reliable branches.

****

Two approaches to avoid overfitting : Prepruning and Postpruning

**Prepruning:**

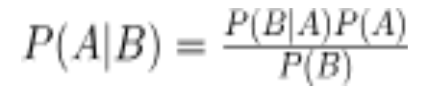
* Halt tree construction early (e.g., by deciding not to further split or partition the subset of training tuples at a given node).
* Upon halting, the node becomes a leaf.
* The leaf may hold the most frequent class among the subset tuples or the probability distribution of those tuples.
* Difficult to choose an appropriate threshold

**Postpruning:**

* Remove branches from a “fully grown” tree—A subtree at a given node is pruned by removing its branches and replacing it with a leaf.
* The leaf is labeled with the most frequent class among the subtree being replaced.

**4. State Bayes Theorem.**

Bayes’ Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes’ theorem is stated mathematically as the following equation:

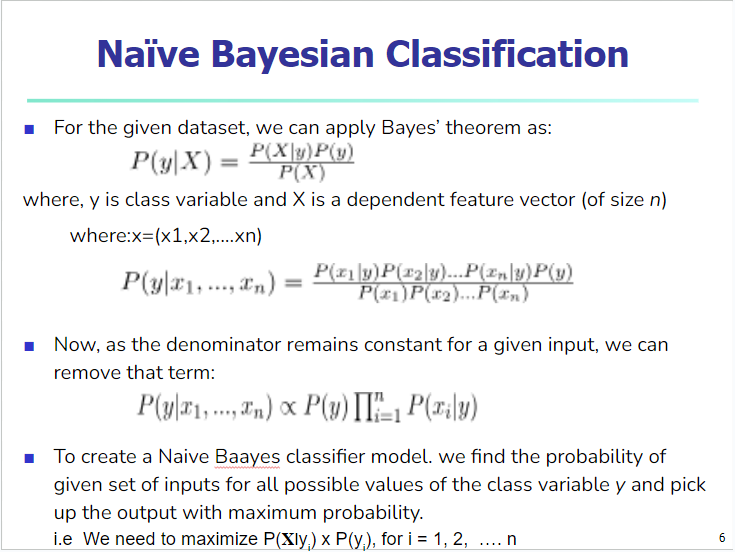


where A and B are events and P(B) ≠ 0

* Basically, we are trying to find the probability of event A, given that event B is true. Event B is also termed as evidence.
* P(A) is the priority of A (the prior probability, i.e. Probability of the event before evidence is seen). The evidence is an attribute value of an unknown instance (here, it is event B).
* P(B) is Marginal Probability: Probability of Evidence.
* P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen.
* P(B|A) is Likelihood probability i.e. the likelihood that a hypothesis will come true based on the evidence.

**5. Explain the Naïve Bayesian Classification Algorithm with an example**

* Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
* Naïve Bayes Classifier is one of the simplest and most effective Classification algorithms that help in building fast machine learning models that can make quick predictions.
* It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
* Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentiment analysis, and classifying articles.
* The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:
* Naïve: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features.
* Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.



Pg-13-18(Sums)

<https://docs.google.com/presentation/d/1OZuHmqYfr6pd0ADHKWtPsmvgpu2hXTYV/edit?usp=sharing&ouid=110281631223246598554&rtpof=true&sd=true>

**6. State advantages and disadvantages of Naive Bayes Algorithm**

Advantages of Naïve Bayes Classifier:

* Naïve Bayes is one of the fastest and easiest ML algorithms to predict a class of datasets.
* It can be used for Binary as well as Multi-class Classifications.
* It performs well in Multi-class predictions as compared to the other Algorithms.
* It is the most popular choice for text classification problems where the dataset is multi-dimensional.

Disadvantages

* Assumption: class conditional independence, therefore loss of accuracy
* Practically, dependencies exist among variables
  + E.g., hospitals: patients: Profile: age, family history, etc.
  + Symptoms: fever, cough etc., Disease: lung cancer, diabetes, etc.
* Dependencies among these cannot be modeled by Naïve Bayesian Classifier

**7. What are the different Metrics for Evaluating Classifier Performance(Accuracy, Precision, Recall, F1 score,Specificity, Sensitivity)**

**8. Problems based on Decision tree and Naive Bayes algorithm.**