Unit 3

Evolution of Sciences: New Data Science Era

Before 1600: **Empirical science**

1600-1950s: **Theoretical science**

- Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.

1950s-1990s: **Computational science**

- Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
- Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.

1990-now: **Data science**

- The flood of data from new scientific instruments and simulations
- The ability to economically store and manage petabytes of data online
- The Internet and computing Grid that makes all these archives universally accessible
- Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes
- Data mining is a major new challenge!

What is Data Mining?

Data mining refers to :

extracting or "mining" knowledge from large amounts of data.

It is also known as **Knowledge Discovery from Data.**

What Is Data Mining?

Data mining (knowledge discovery from data)
- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Data mining: a misnomer?

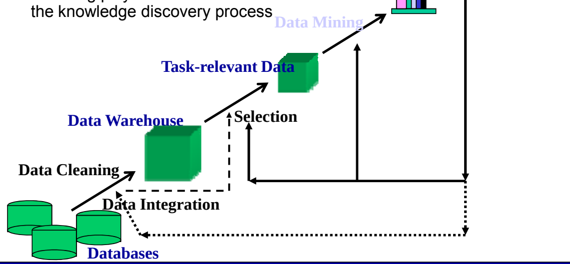
Alternative names
- Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Watch out: Is everything "data mining"?
- Simple search and query processing
- (Deductive) expert systems

Knowledge Discovery (KDD) Process

This is a view from typical database systems and data warehousing communities

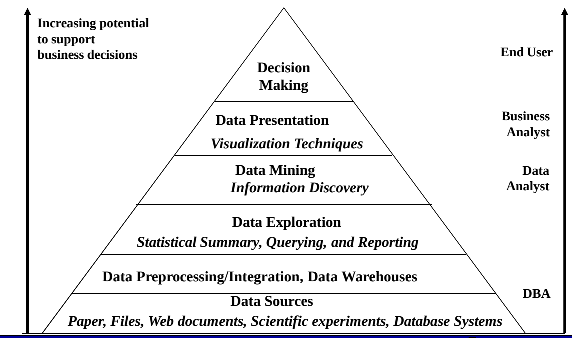
Data mining plays an essential role in the knowledge discovery process



Example: A Web Mining Framework

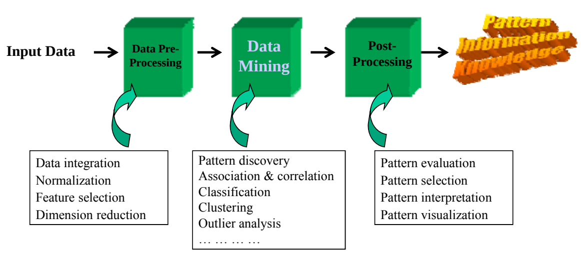
Web mining usually involves
- Data cleaning
- Data integration from multiple sources
- Warehousing the data
- Data cube construction
- Data selection for data mining
- Data mining
- Presentation of the mining results
- Patterns and knowledge to be used or stored into knowledge-base

Data Mining in Business Intelligence



© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, by Imran Khan Asst. Professor. U3

KDD Process: A Typical View from ML and Statistics



This is a view from typical machine learning and statistics communities

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, by Imran Khan Asst. Professor. U3

Example: Medical Data Mining

- Health care & medical data mining – often adopted such a view in statistics and machine learning
- Preprocessing of the data (including feature extraction and dimension reduction)
- Classification or/and clustering processes
- Post-processing for presentation

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, by Imran Khan Asst. Professor. U3

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.

Data Mining: On What Kinds of Data?

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

Data Mining Function: (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

Data Mining Function: (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?

Data Mining Function: (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, …

Data Mining Function: (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

Data Mining Function: (5) Outlier Analysis

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

Relationships

Let us take the case study of a fast food restaurant.

The combo meals that are available are designed after applying data mining to the sales trends' data over some months or years.

Data mining discovers relationships of this type. The relationships may be between two or more different objects along with the time dimension or between the attributes of the same object.

Discovery of knowledge is a key result of data mining.

Case Study

Case Study

The Fast Food industry is highly competitive, one where a very small change in operations can have a significant impact on the bottom line. For this reason, quick access to comprehensive information for both standard and on demand reporting is essential.
Implement the various data mining techniques to address this requirement for ABC Corporation, a fast food franchisee operating approximately 80 outlets at different places. The results should provide strategic and tactical decision support to all levels of management within the Corporation.

Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

Sequence, trend and evolution analysis
- Trend, time-series, and deviation analysis: e.g., regression and value prediction
- Sequential pattern mining
  - ✓e.g., first buy digital camera, then buy large SD memory cards
- Periodicity analysis
- Motifs and biological sequence analysis
  - ✓Approximate and consecutive motifs
- Similarity-based analysis

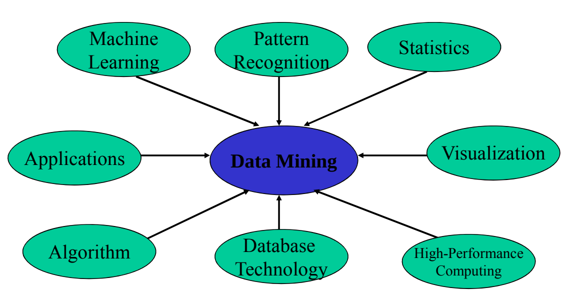
Mining data streams
- Ordered, time-varying, potentially infinite, data streams

Evaluation of Knowledge

Are all mined knowledge interesting?
- One can mine tremendous amount of "patterns" and knowledge
- Some may fit only certain dimension space (time, location, …)
- Some may not be representative, may be transient, …

Evaluation of mined knowledge → directly mine only interesting knowledge?
- Descriptive vs. predictive
- Coverage
- Typicality vs. novelty
- Accuracy
- Timeliness
- …

Data Mining: Confluence of Multiple Disciplines



Why Confluence of Multiple Disciplines?

Tremendous amount of data
- Algorithms must be highly scalable to handle such as tera-bytes of data

High-dimensionality of data
- Micro-array may have tens of thousands of dimensions

High complexity of data
- Data streams and sensor data
- Time-series data, temporal data, sequence data
- Structure data, graphs, social networks and multi-linked data
- Heterogeneous databases and legacy databases
- Spatial, spatiotemporal, multimedia, text and Web data
- Software programs, scientific simulations

New and sophisticated applications

OLAP versus Data Mining

Features	OLAP	DATA MINING
Motivation for Information Request	What is happening in the enterprise?	Predict the future based on why this is happening.
Data granularity	Summary data.	Detailed transaction-level data.
Number of business dimensions	Limited number of dimensions.	Large number of dimensions.
Number of dimension attributes	Small number of attributes.	Many dimension attributes.
Sizes of datasets for the dimensions	Not large for each dimension.	Usually very large for each dimension.
Analysis approach	User-driven interactive analysis.	Data-driven automatic knowledge discovery.
Analysis techniques	Multidimensional, drill-drown, and slice-and-dice.	Prepare data, launch mining tool and sit back.
State of the technology	Mature and widely used.	Still emerging; some parts of the technology more mature.

Steps in Data Mining (contd..)

It includes determining clearly what you want the tool to accomplish. We do not try to predict the knowledge we are going to discover but define the business objectives of the engagement.

Step 1: Define Business Objectives
State why do you need a data mining solution. Define your expectations and express how the final results will be used in the operational system.

Step 2: Prepare Data (Data Preprocessing)
Consists of data selection, pre-processing of data and data transformation. Use the business objectives to determine what data has to be selected. The variables selected are called **active variables.**

Steps in Data Mining (contd..)

Pre-processing is meant to improve the quality of selected data. It involves enriching the selected data with external data, removal of noisy data and missing values.

Step 3: Perform Data Mining
The knowledge discovery engine applies the selected algorithm to the prepared data. The output from this step is a relationship or pattern.

Step 4: Evaluate Results
In this step, you examine all the resulting patterns. You will apply a filtering mechanism and select only the promising patterns to be selected and applied.

Steps in Data Mining (contd..)

Step 5: Present Discoveries
This may be in the form of visual navigation, charts, graphs, or free-form texts. It also includes storing of interesting discoveries in the knowledge base for repeated use.

Step 6: Incorporate Usage of Discoveries
This step is for using the results to create actionable items in the business. The results are assembled in the best way so that they can be exploited to improve the business.

Data Preprocessing

Data Preprocessing: An Overview
- Data Quality
- Major Tasks in Data Preprocessing

Data Cleaning

Data Integration

Data Reduction

Data Transformation and Data Discretization

Summary

Data Quality: Why Preprocess the Data?

Measures for data quality: A multidimensional view

- 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?

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

Data Cleaning

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

- incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - ✓ e.g., *Occupation*=" " (missing data)
- noisy: containing noise, errors, or outliers
  - ✓ e.g., *Salary*="−10" (an error)
- inconsistent: containing discrepancies in codes or names, e.g.,
  - ✓ *Age*="42", *Birthday*="03/07/2010"
  - ✓ Was rating "1, 2, 3", now rating "A, B, C"
  - ✓ discrepancy between duplicate records
- Intentional (e.g., *disguised missing* data)
  - ✓ Jan. 1 as everyone's birthday?

Incomplete (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

How to Handle Missing Data?

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

Noisy Data

Noise: random error or variance in a measured variable

Incorrect attribute values may be due to

- faulty data collection instruments
- data entry problems
- data transmission problems
- technology limitation
- inconsistency in naming convention

Other data problems which require data cleaning

- duplicate records
- incomplete data
- inconsistent data

How to Handle Noisy Data?

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

Regression
- smooth by fitting the data into regression functions

Clustering
- detect and remove outliers

Combined computer and human inspection
- detect suspicious values and check by human (e.g., deal with possible outliers)

Data Integration

Data integration:
- Combines data from multiple sources into a coherent store

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

Entity identification problem:
- Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Metadata of the attributes (that include name, meaning, data type, range of values null rules for handling blanks etc) must be checked before integration data from different sources.
- Integrate metadata from different sources

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

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

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)

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

  - where n is the number of tuples, $\bar{A}$ and $\bar{B}$ 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 (atttribute discourage each other)

Cont..

Mean A

$A=\sum\bar{A}/n$

Standard deviation

$\sigma_A=\text{sqrt}(\sum(A-\bar{A})^2/(n-1)$

Correlation (viewed as linear relationship)

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, A and B, and then take their dot product

$$a'_k = (a_k - mean(A)) / std(A)$$

$$b'_k = (b_k - mean(B)) / std(B)$$

$$correlation(A, B) = A' \bullet B'$$

Covariance (Numeric Data)

- Covariance is similar to correlation

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

Correlation coefficient: $r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$

  - where n is the number of tuples, $\overline{A}$ and $\overline{B}$ 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

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

- Dimensionality reduction, e.g., remove unimportant attributes
  - ✓Wavelet transforms
  - ✓Principal Components Analysis (PCA)
  - ✓Feature subset selection, feature creation
- Numerosity reduction (some simply call it: Data Reduction)
  - ✓Regression and Log-Linear Models
  - ✓Histograms, clustering, sampling
  - ✓Data cube aggregation
- 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

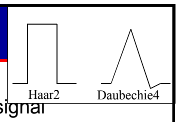
Dimensionality reduction

- 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)

Wavelet Transformation



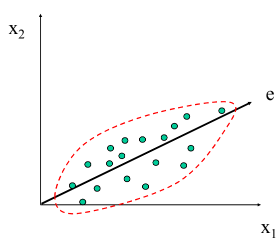
- Discrete wavelet transform (DWT) for linear signal processing, multi-resolution analysis
- Compressed approximation: store only a small fraction of the strongest of the wavelet coefficients
- Similar to discrete Fourier transform (DFT), but better lossy compression, localized in space
- Method:
  - Length, L, must be an integer power of 2 (padding with 0's, when necessary)
  - Each transform has 2 functions: smoothing, difference
  - Applies to pairs of data, resulting in two set of data of length L/2
  - Applies two functions recursively, until reaches the desired length

Wavelet Decomposition

- Wavelets: A math tool for space-efficient hierarchical decomposition of functions
- S = [2, 2, 0, 2, 3, 5, 4, 4] can be transformed to $S_\wedge$ = [$2^3/_4$, $-1^1/_4$, $^1/_2$, 0, 0, -1, -1, 0]
- Compression: many small detail coefficients can be replaced by 0's, and only the significant coefficients are retained

Principal Component Analysis (PCA)

- Find a projection that captures the largest amount of variation in data
- The original data are projected onto a much smaller space, resulting in dimensionality reduction. We find the eigenvectors of the covariance matrix, and these eigenvectors define the new space



Principal Component Analysis (Steps)

Given *N* data vectors from *n*-dimensions, find $k \le n$ orthogonal vectors (*principal components*) that can be best used to represent data

- Normalize input data: Each attribute falls within the same range
- Compute *k* orthonormal (unit) vectors, i.e., *principal components*
- Each input data (vector) is a linear combination of the *k* principal component vectors
- The principal components are sorted in order of decreasing "significance" or strength
- Since the components are sorted, the size of the data can be reduced by eliminating the *weak components*, i.e., those with low variance (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data)

Works for numeric data only

Attribute Subset Selection

Another way to reduce dimensionality of data

Redundant attributes

- Duplicate much or all of the information contained in one or more other attributes
- E.g., purchase price of a product and the amount of sales tax paid

Irrelevant attributes

- Contain no information that is useful for the data mining task at hand
- E.g., students' ID is often irrelevant to the task of predicting students' GPA

- Forward Selection
- Backward selection
- Decision Tree

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)

Non-parametric methods

- Do not assume models
- Major families: histograms, clustering, sampling, …

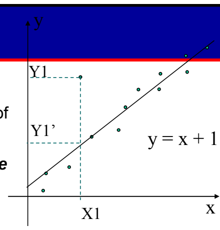
Parametric Data Reduction: Regression and Log-Linear Models

Linear regression
- Data modeled to fit a straight line
- Often uses the least-square method to fit the line

Multiple regression
- Allows a response variable Y to be modeled as a linear function of multidimensional feature vector

Log-linear model
- Approximates discrete multidimensional probability distributions

Regression Analysis



Regression analysis: A collective name for techniques for the modeling and analysis of numerical data consisting of values of a ***dependent variable*** (also called ***response variable*** or *measurement*) and of one or more *independent variables* (aka. ***explanatory variables*** or ***predictors***)

The parameters are estimated so as to give a "**best fit**" of the data

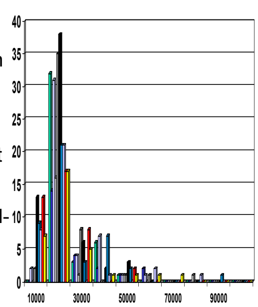
Most commonly the best fit is evaluated by using the ***least squares method***, but other criteria have also been used

Used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships

Histogram Analysis

Divide data into buckets and store average (sum) for each bucket

Partitioning rules:
- Equal-width: equal bucket range
- Equal-frequency (or equal-depth)



Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only
- Can be very effective if data is clustered but not if data is "smeared"
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms

Sampling

- Sampling: obtaining a small sample *s* to represent the whole data set *N*
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Key principle: Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
  - Develop adaptive sampling methods, e.g., stratified sampling:
- Note: Sampling may not reduce database I/Os (page at a time)

Types of Sampling

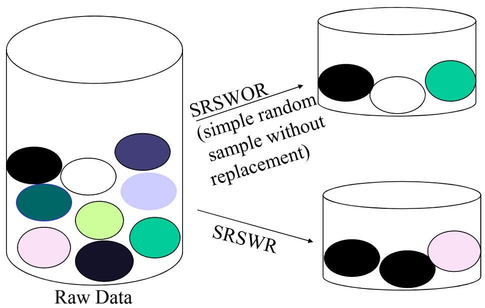
Simple random sampling
- There is an equal probability of selecting any particular item

Sampling without replacement
- Once an object is selected, it is removed from the population

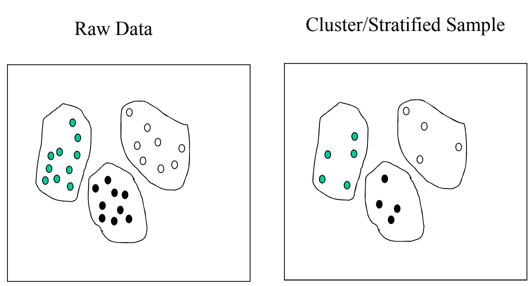
Sampling with replacement
- A selected object is not removed from the population

Stratified sampling:
- Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
- Used in conjunction with skewed data

Sampling: With or without Replacement



Sampling: Cluster or Stratified Sampling



Data Reduction 3: Data Compression

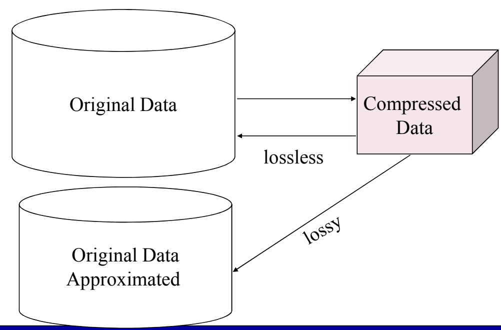
String compression
- There are extensive theories and well-tuned algorithms
- Typically lossless, but only limited manipulation is possible without expansion

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

Data Compression



Data Transformation

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

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)

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

Review Questions

Objective Questions:

1)The types of information that can be garnered from datamining include:

a) sequences, classifications, and clusters.
b) model-driven and data-driven.
c) associations and forecasts.
d) a and c.
e) a, b and c.

2) The term "associations" is associated with:

a) occurrences linked to a single event.
b) classifications when no groups have been defined.
c) pattern recognition describing the group to which an item belongs.
d) a series of existing values used to predict other values.
e) events linked over time.

Review Questions cont..

3)DSS assist management by combining ________ into a single powerful system to support unstructured decision-making.

a) hardware and the Internet
b) data, analytical models and tools, and user-friendly software
c) analytical models and tools and data from the Internet
d) group decision processes and electronics
e) data and people

4)DSS, GDSS, and ESS are part of a special category of information systems that are explicitly designed to:

a) make decisions for managers.
b) enhance Web performance.
c) gather data and build data warehouses.
d) enhance managerial decision-making.
e) interpret data for management.

Review Questions cont..

5)The term "sequences" is associated with:
a) occurrences linked to a single event.
b) classifications when no groups have been defined.
c) pattern recognition describing the group to which an item belongs.
d) a series of existing values used to predict other values.
e) events linked over time.

6)The earliest DSS tended to:
a) rely on Internet data.
b) draw on small subsets of corporate data.
c) be heavily model-driven.
d) b and c.
e) a and c.

Review Questions cont..

7)The term "classifications" is associated with:
a) occurrences linked to a single event.
b) classifications when no groups have been defined.
c) pattern recognition describing the group to which an item belongs.
d) a series of existing values used to predict other values.
e) events linked over time.

8)Model-driven DSS:
a) analyze large pools of data.
b) are an outgrowth of data mining.
c) use TPS and OLAP.
d) begin with a given group of data and change variables.
e) use events linked over time.

Review Questions cont..

9)The term "forecasting" is associated with:
a) Occurrences linked to a single event.
b) Classifications when no groups have been defined.
c) Pattern recognition describing the group to which an item belongs.
d) A series of existing values used to predict other values.
e) Events linked over time.

10)A goal of data mining includes which of the following?
a) To explain some observed event or condition
b) To confirm that data exists
c) To analyze data for expected relationships
d) To create a new data warehouse
e) None of these

Review Questions cont..

Short answer type Questions

1. Define data mining in two or three sentences
2. How is data mining different from OLAP?
3. Is the data warehouse prerequisite for data mining? Does the data warehouse help data mining? If so, in what ways?
4. Name the three common problems of link analysis technique?
5. What is market basket analysis? Give two examples of this application in business.
6. Give three broad reasons why you think data mining is being used in today's businesses.
7. What business problems can data mining help solve?
8. What is Predictive Analytics?
9. What is the difference between data mining, online analytical processing (OLAP) ?
10. State various benefits of Data mining.

Review Questions cont..

Long answer type Questions

1. Describe how decision trees work. Explain with the help of an example.
2. What do you mean by KDD? Explain all the steps of KDD in detail.
3. What are the basic principles of genetic algorithms? Use the example to describe how this technique works
4. Describe cluster detection technique?
5. Discuss Data mining Application in the field of Banking and finance.
6. Do neural networks and genetic algorithms have anything in common? Point out differences.
7. How does the memory-based reasoning technique work? What is the underlying principle?
8. Explain Neural Network in detail?
9. What are the golden rules for data mining?
10. Discuss Data mining Application in the field of Retail Industry.

Suggested Reading/References

- Kamber and Han, "Data Mining Concepts and Techniques", Hartcourt India P. Ltd.,2001
- Paul Raj Poonia, "Fundamentals of Data Warehousing", John Wiley & Sons, 2003.
- Sam Anahony, "Data Warehousing in the real world: A practical guide for building decision support systems", John Wiley, 2004
- W. H. Inmon, "Building the operational data store", 2nd Ed., John Wiley, 1999.
- E. Alpaydin. Introduction to Machine Learning, 2nd ed., MIT Press, 2011
- S. Chakrabarti. Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data. Morgan Kaufmann, 2002
- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996

References

- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2nd ed., 2006 (3ed. 2011)
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer-Verlag, 2009
- B. Liu, Web Data Mining, Springer 2006.
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005
- S. M. Weiss and N. Indurkhya, Predictive Data Mining, Morgan Kaufmann, 1998
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, by Imran Khan