Parul® University

Data Mining and Warehousing (03105430)

Prof. Dheeraj Kumar Singh, Assistant Professor
Information Technology Department

Parul®
University

DIGITAL LEARNING CONTENT

CHAPTER-4

Data Pre-processing

No quality data, results in no quality mining !

What is Data Preprocessing?

- Allows to analyse and summarize main characteristics of data sets.
- Refers to steps applied to make data more suitable for data mining.

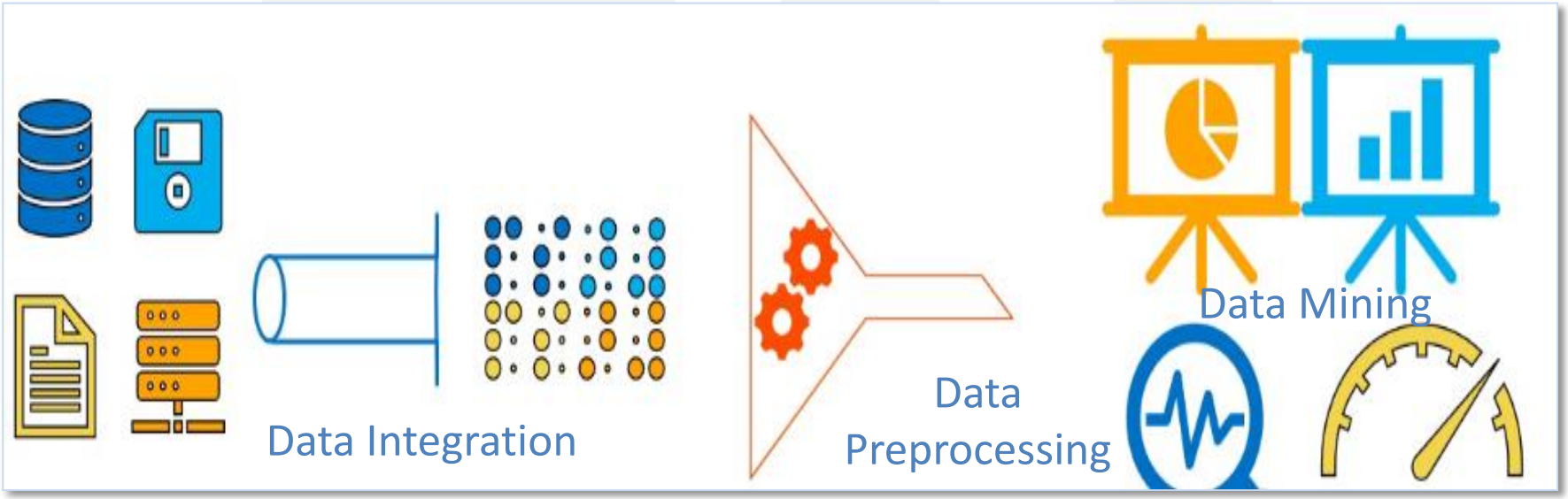


Image source : Google

What is Data Preprocessing?

- **Used to deal with:**

  Incomplete data : missing attribute values, unavailable data properties, or containing only summary data.

  Noisy data: random errors or incorrect attribute values

  Inconsistent data: containing discrepancies in codes or names, outliers.

Significance of Preprocessing

- **Data in the real world is generally incomplete, noisy and inconsistent.**
- **Quality decisions must be based on quality data**

Index	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	5.1	3.5	1.4	0.2	Iris-set…
2	4.9	nan	1.4	0.2	nan
3	4.7	3.2	1.3	0.2	Iris-set…
4	??	3.1	1.5	0.2	Iris-set…
5	5	3.6	###	0.2	Iris-set…

Image source : NPTEL

Statistical Descriptions of Data

- Better understanding of data is important for successful data preprocessing.
- Provides a way to characterize the central tendency and dispersion of data.
- Helps in understanding distribution of data.
- Data quality can be assessed in terms of
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Value added
  - Interpretability
  - Accessibility

Statistical Description Measure

- **Measures of central tendency:**
  - Mean
  - Median
  - Mode
  - Midrange
- **Measures of Data Dispersion:**
  - Range
  - Quartiles (five number summary)
  - Interquartile range (IQR)
  - Variance
  - Standard deviation

Statistical Descriptions of Data

- **Distributive Measure:** sum(), count(), max(), min() etc.
  - **Partition the data into subsets and merge values obtained for each subsets.**
- **Algebraic Measure:** average() or mean()
  - **Computed as sum() /count().**
- **Holistic Measure:** median()
  - **Computed on entire dataset as a whole.**

Measuring the Central Tendency: Mean

- **Arithmetic Mean or Average**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{or,} \quad \mu = \frac{\sum x}{N}$$

- **Weighted Arithmetic Mean on Weighted Average**

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- **Trimmed Mean**
  - problem with the mean is its sensitivity to extreme (e.g., outlier)
  - offset this effect, mean is obtained after chopping off values at high and low extremes.

Measuring the Central Tendency: Median

- **Median is a better measure of center of data for skewed (asymmetric) data.**
- **Given dataset of N distinct values is sorted in numerical order.**
  - **If N is odd, then the median is the middle value,**
  - **otherwise median is the average of the two middle values.**

- **Suppose that a given data grouped in intervals according to frequency.**

$$median = L_1 + (\frac{N/2 - (\sum f)l}{f_{median}})c$$

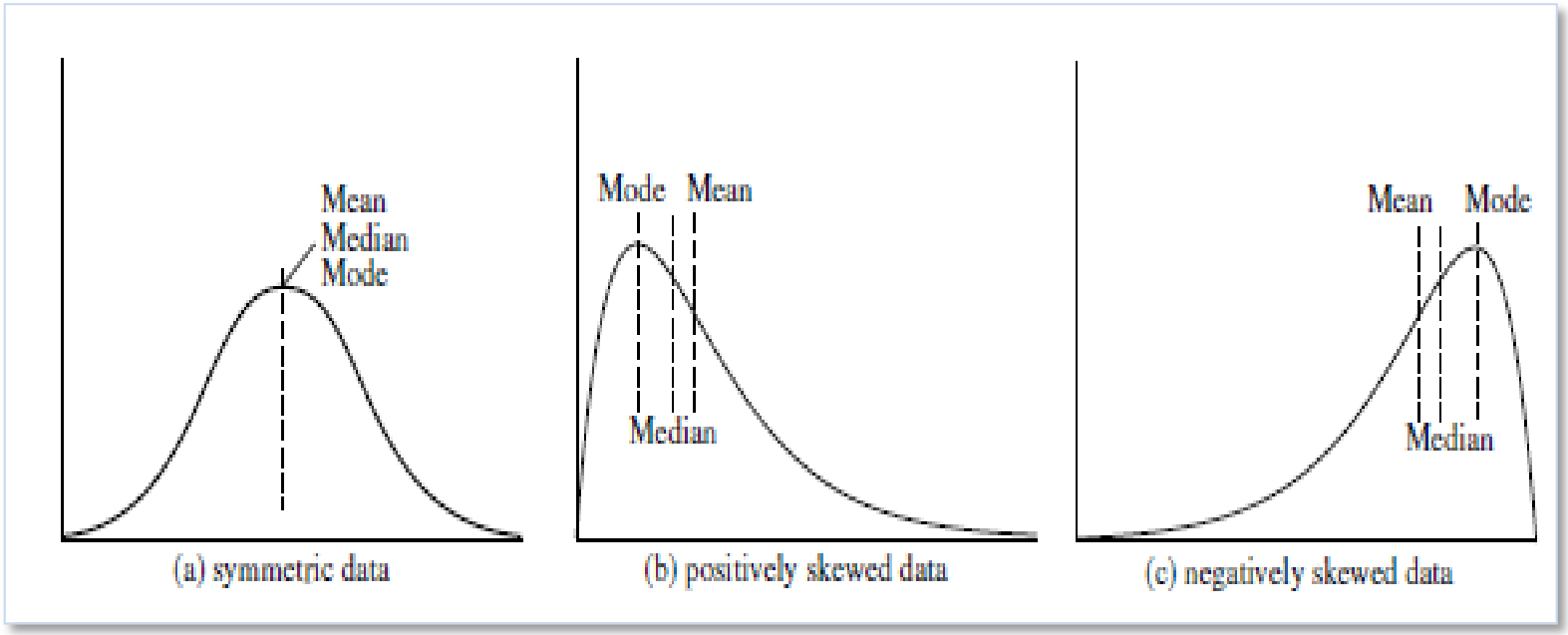
where L1 is lower bound of median interval, N is number of values in entire data set, $(\sum f)l$ is sum of frequencies of all of intervals that are lower than median interval. f_{median} is frequency of median interval, and c is the width of median interval.

Measuring the Central Tendency: Mode and Midrange

- **Value that occurs most frequently in the data set.**
- **It is possible for the greatest frequency to correspond to several different values, which results in more than one mode.**
- **Data sets with two or more modes is multimodal.**
  - **Unimodal**: with only one mode.
  - **Bimodal**: with two modes.
  - **Trimodal**: with three modes.
- **If each data value occurs only once, then there is no mode.**
- **Midrange is average of the largest and smallest values in the set.**

Empirical relation: Mean, Median and Mode

- **Empirical formula:** $mean - mode = 3 \times (mean - median)$



Mean, median, and mode of symmetric versus positively and negatively skewed data

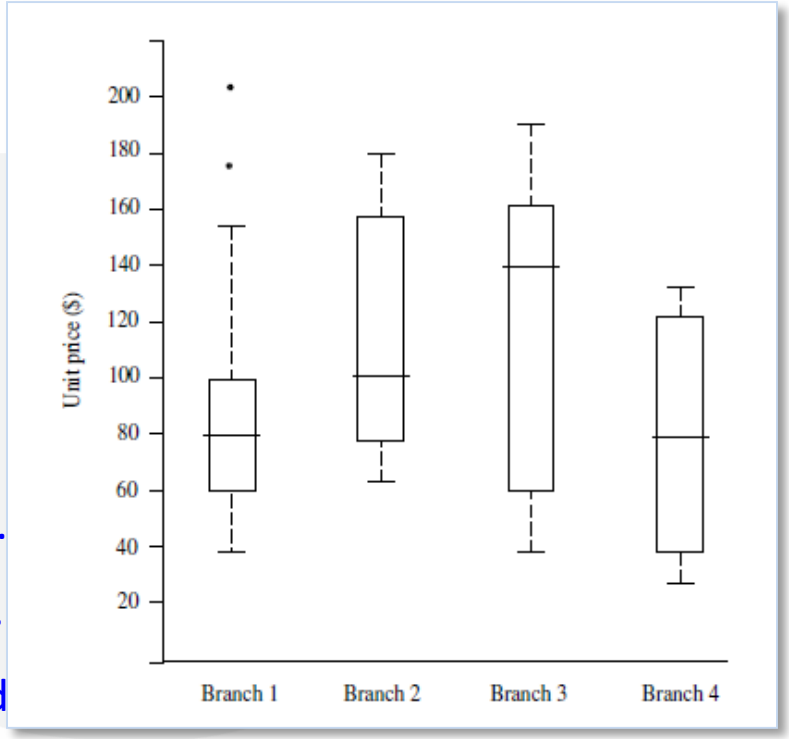
Image source : Data Mining Concepts and Techniques Book

Measuring the Dispersion of Data

- The degree to which numerical data tend to spread is called the dispersion, or variance of the data.
- The k_{th} percentile of a set of data in numerical order is the value x_i having property that k percent of the data entries lie at or below x_i.
- Range, Quartiles, and Inter-quartile range(IQR):
  - Range: Difference between the largest (max()) and smallest (min()) values.
  - Quartiles: First quartile: Q_1 (25^{th} percentile), Third quartile: Q_3 (75^{th} percentile)
  - Inter-quartile range: Distance between first and third quartiles, IQR = $Q_3 - Q_1$
  - Five-number summary: min, Q_1, M, Q_3, max
  - Outlier: usually, a value higher/lower than 1.5 x IQR

Measuring the Dispersion of Data: Boxplot Analysis

- **way of visualizing a distribution.**
- **A boxplot incorporates the five-number summary:**
  - Data is represented with a box.
  - The ends of the box are at the first and third quartiles, i.e.,height of the box is IRQ.
  - The median is marked by a line within box.
  - Whiskers: two lines outside the box extend to Minimum and Maximum.



Boxplot of unit price data for items sold at four branches of *a shop*

Image source : Data Mining Concepts and Techniques Book

Variance and Standard Deviation

- The variance of N observations, $x_1, x_2, \ldots x_N$, is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n}x_i^{\,2} - \mu^2$$

- where μ is the mean value of the observations.
- Standard deviation, σ is the square root of variance σ^2.
- σ measures spread about mean and should be used only when mean is chosen as the measure of center.
- σ =0 only when there is no spread, i.e. , when all observations have same value. Otherwise σ > 0.

Data Visualization: Histogram Analysis

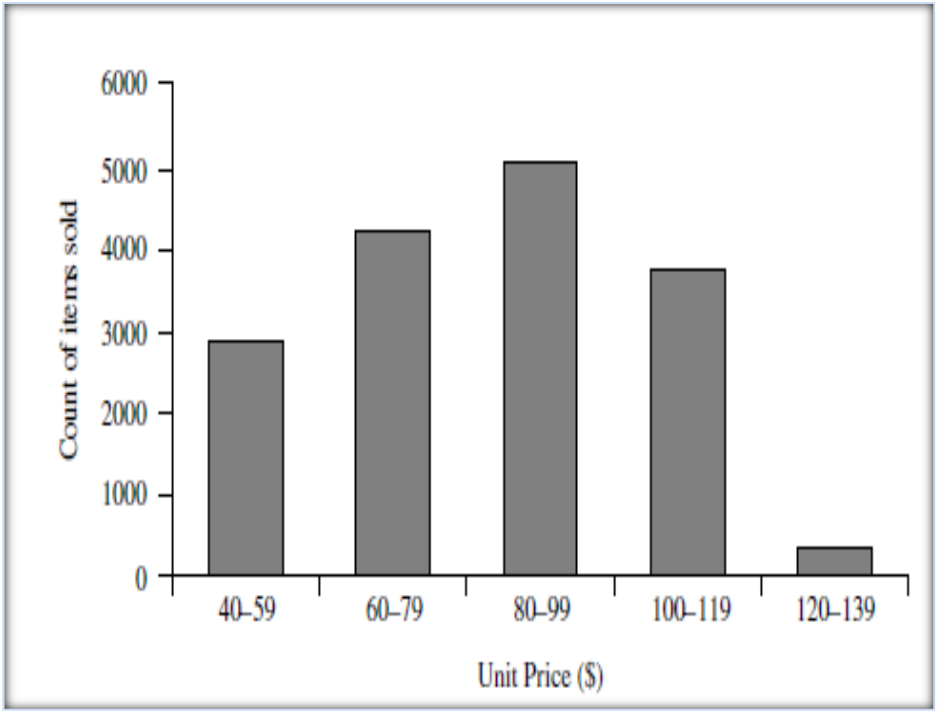
- Plotting histograms, or frequency histograms, is a graphical method for summarizing distribution of a given attribute.
- Categorical attributes histograms are referred as a bar chart, where as for numeric attribute, the term histogram is preferred.
- A histogram for partitions data distribution of given attribute into disjoint subsets, called buckets.
- The width of each bucket is uniform.
- Each bucket is represented by a rectangle whose height is equal to the relative frequency of values at the bucket.

Data Visualization: Histogram Analysis

- **Example**

Table of unit price data for items sold

Unit price ($)	*Count of items sold*
40	275
43	300
47	250
.. ..	
74	360
75	515
78	540
.. ..	
115	320
117	270
120	350



Histogram for the data set of Table

Image source : Data Mining Concepts and Techniques Book

Data Visualization: Quantile Plot

- Displays all of data for given attribute, allowing user to assess both overall behavior and unusual occurrences.
- Allows us to compare different distributions based on their quantiles.
- It plots quantile information.
- For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i.

$$f_i = (i - 0.5) / N$$

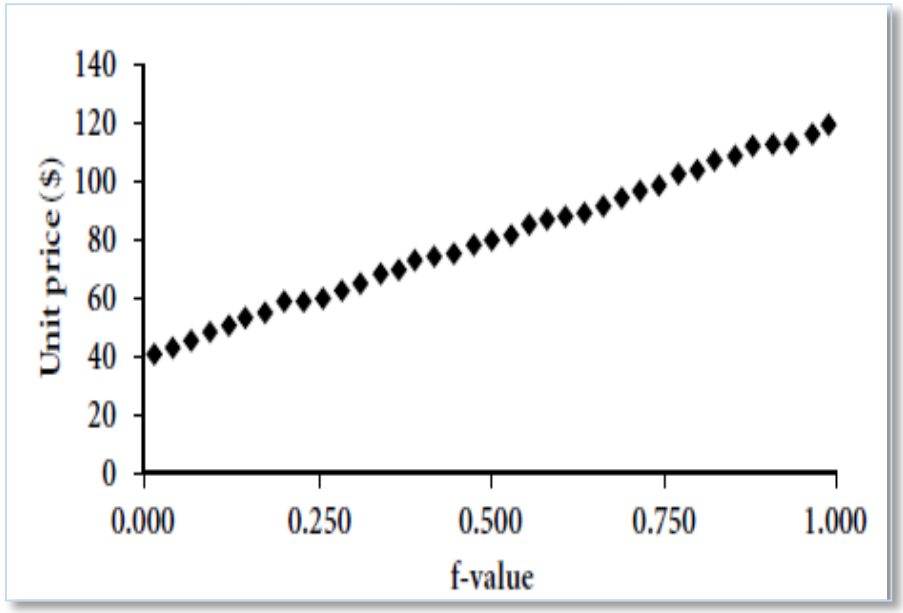
These numbers increase in equal steps of 1/N, ranging from 1/2N to 1- 1/2N.

Data Visualization: Quantile Plot

- **Example**

Table of unit price data for items sold

Unit price ($)	*Count of items sold*
40	275
43	300
47	250
.. ..	
74	360
75	515
78	540
.. ..	
115	320
117	270
120	350



Quantile Plot for unit price of Table

Image source : Data Mining Concepts and Techniques Book

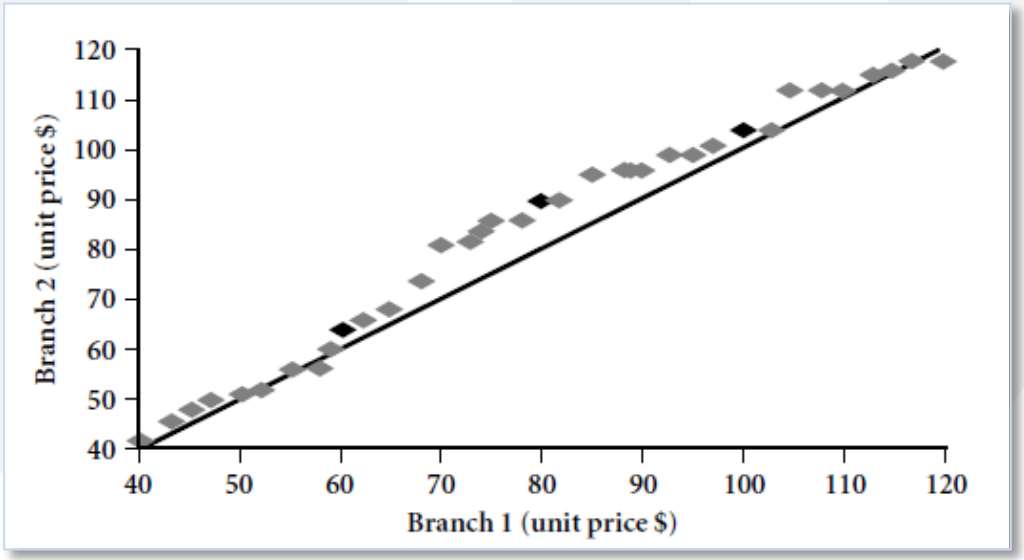
Data Visualization: Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against corresponding quantiles of another.
- Allows user to view whether there is a shift in going from one distribution to another.
- Let $x_1, \ldots x_N$, and $y_1, \ldots y_M$ be two sets of observations, where each data set is sorted in increasing order.
  - If $M = N$ *then* $f_i = (i - 0.5) / N$
  - If $M < N$ *then* $f_i = (i - 0.5) / M$
- This computation typically involves interpolation.

Data Visualization: Quantile-Quantile (Q-Q) Plot

- **Example**

Each point corresponds to same quantile for each data set and shows the unit price of items sold at branch 1 versus branch 2 for that quantile.



Q-Q Plot of unit price data from two different branches

Image source : Data Mining Concepts and Techniques Book

Data Visualization: Scatter plot

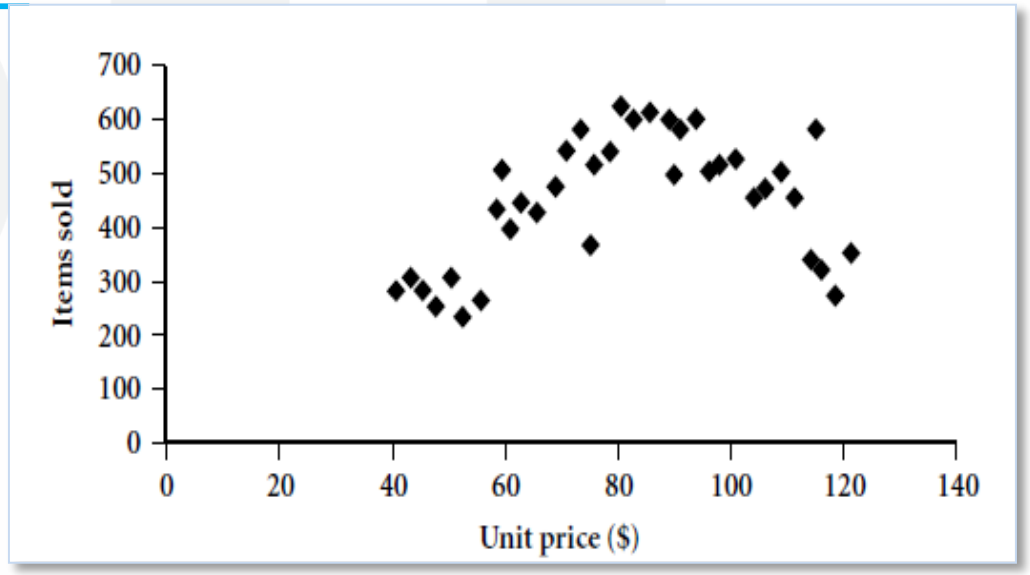
- Used to find a relationship, pattern, or trend between two numerical attributes.
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane.
- Useful method for providing a first look at bivariate data to see clusters of points and outliers.
- When dealing with several attributes, the scatter-plot matrix is a useful extension to the scatter plot.
- Given n attributes, a scatter-plot matrix is an nxn grid of scatter plotsthat provides a visualization of each attribute with every other attribute.

Data Visualization: Scatter plot

- **Example**

Table of unit price data for items sold

Unit price ($)	*Count of items sold*
40	275
43	300
47	250
.. ..	
74	360
75	515
78	540
.. ..	
115	320
117	270
120	350



Scatter Plot for unit price of Table

Image source : Data Mining Concepts and Techniques Book

Data Visualization: Loess Curve

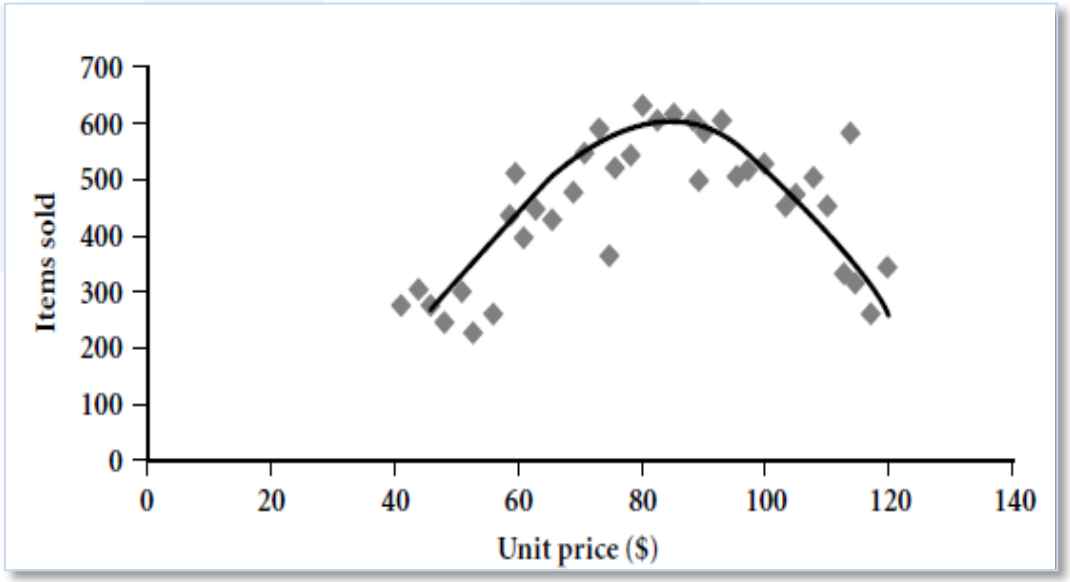
- Adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence.
- The word loess is short for “local regression”.
- Loess curve is fitted by setting two parameters: α, smoothing parameter, and λ, the degree of the polynomials that are fitted by the regression.
- While α can be any positive number (typical values are between 1=4 and 1), λ can be 1 or 2.
- The goal in choosing α is to produce a fit that is as smooth as possible without unduly distorting the underlying pattern in the data.

Data Visualization: Loess Curve

- **Example**

Table of unit price data for items sold

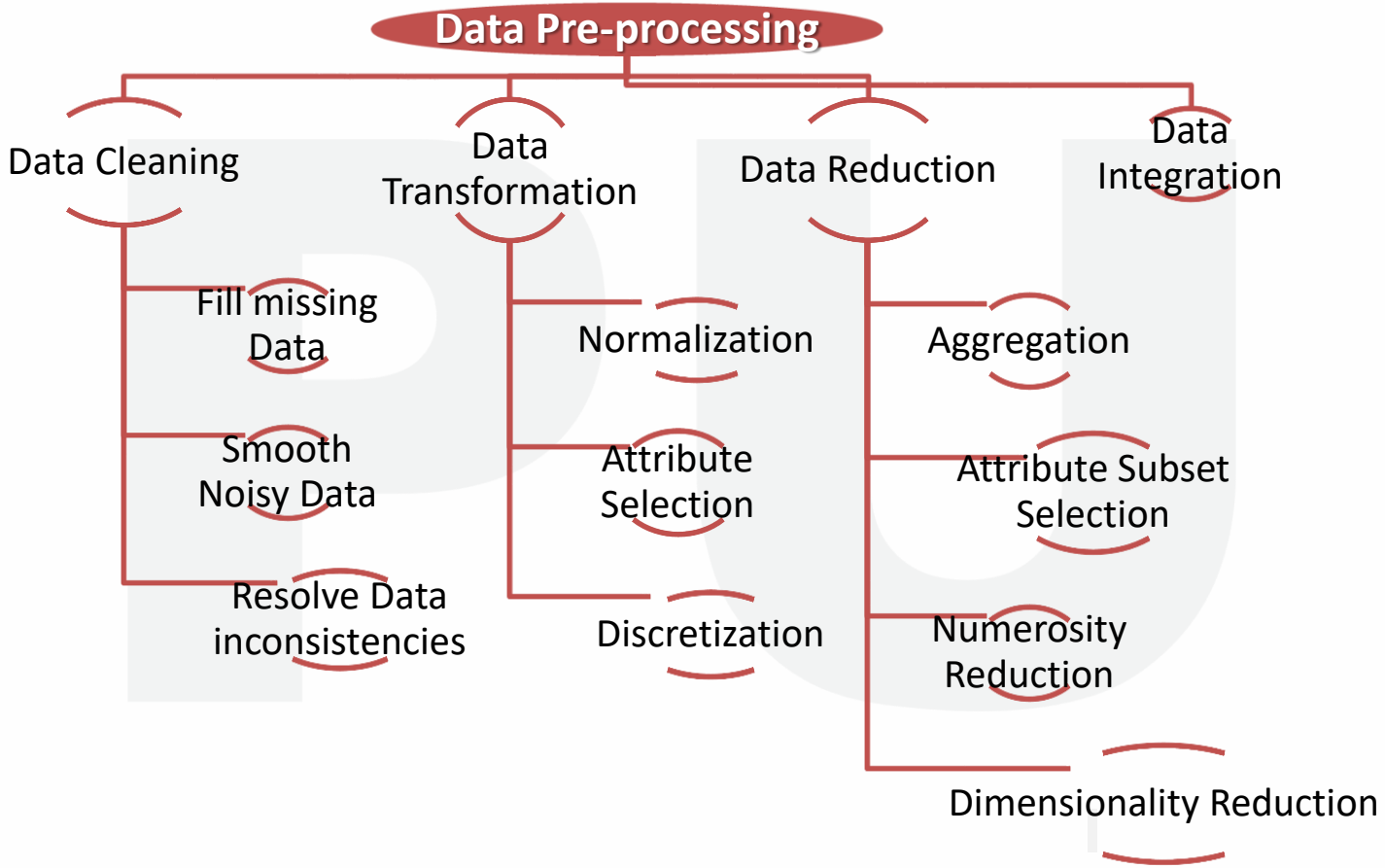
Unit price ($)	*Count of items sold*
40	275
43	300
47	250
.. ..	
74	360
75	515
78	540
.. ..	
115	320
117	270
120	350



Loess Curve for unit price of Table

Image source : Data Mining Concepts and Techniques Book

Tasks involved in Data Pre-processing



Data Cleaning : Fill missing Data

- **Ignore the tuple:** usually done when most of attribute values are missing
- **Fill in the missing value manually:** Tedious & infeasible sometime
- **Use a global constant to fill in the missing value**: e.g., unknown, a new class
- **Use the attribute mean to fill in the missing value:** calculate mean for all samples of the same class
- **Use the most probable value to fill in the missing value:** based on inference

Data Cleaning : Noisy Data

- Meaningless data or unknown values
- Cannot interpreted by machine
- Error in a measured variable
- Incorrect attribute values
- Sources of such data:
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention

Data Cleaning : Handling Noisy Data

- **Binning Method:** sort the data and partition into segments of equal size called bins. Then binning methods are performed on each partition.
- **Clustering:** groups the similar data in a cluster and detect outliers.
- **Regression:** made smooth by fitting in to a regression function.
- **Semi-automated method:** human detect suspicious values and update manually.

Data Cleaning : Binning Method

- Sort the input data
- Create bins according to given bin size
- Partition data in to equal segment and arrange bins
- Apply smoothing by
  - bin mean
  - bin boundaries
  - bin medians

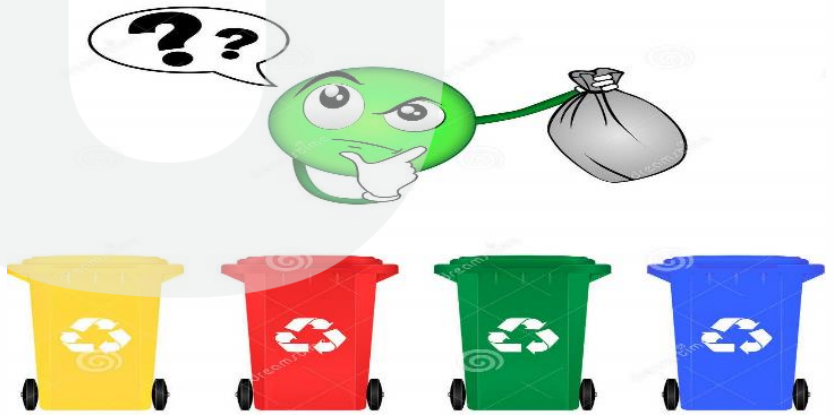


Image source : Google

Data Cleaning : Binning Method

- **Example: Sorted data for price (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34**

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15

Bin 2: 21, 21, 24

Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9

Bin 2: 22, 22, 22

Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15

Bin 2: 21, 21, 24

Bin 3: 25, 25, 34

Smoothing by bin median:

Bin 1: 8, 8, 8

Bin 2: 21, 21, 21

Bin 3: 28, 28, 28

Data Cleaning : Clustering

- Outliers may be detected by clustering
- In clustering, similar values are organized into groups.
- Intuitively, values that fall outside of the set of clusters may be considered outliers

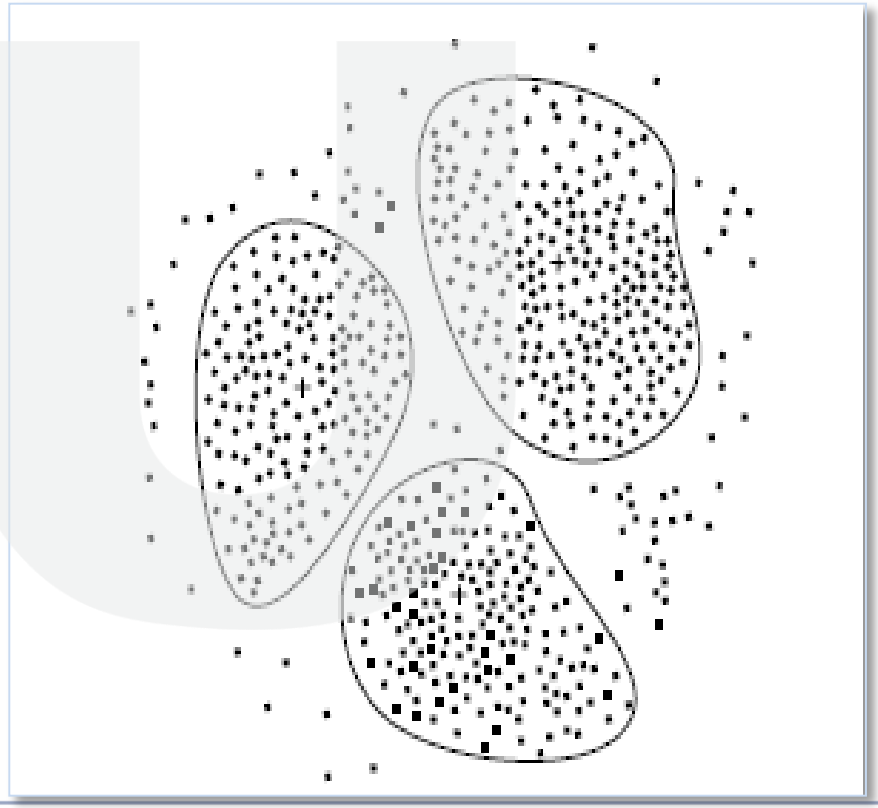


Image source : Data Mining Concepts and Techniques Book

Data Cleaning : Regression

- **Regression fits the data to a function.**
  - Linear regression
  - Multiple linear regression

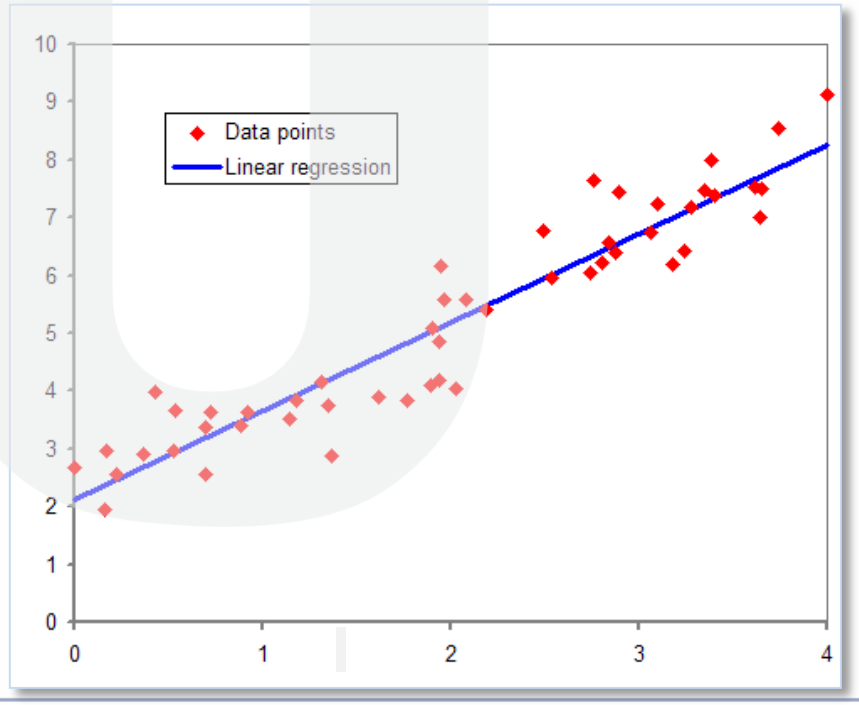


Image source : Data Mining Concepts and Techniques Book

Data Cleaning : Handling Inconsistent Data

- **Manual correction using external references**
- **Semi-automatic using various tools**
  - To detect violation of known functional dependencies and data constraints
  - To correct redundant data

Image source : Google

Data Transformation

- **Data transformed or consolidated into forms appropriate for mining.**
  - **To detect violation of functional dependencies and data constraints**
  - **To correct redundant data**

•Transformation can involve:

- **Smoothing: remove noise from data (binning, clustering, regression)**
- **Aggregation: summarization, data cube construction**
- **Generalization: uses concept hierarchy**
- **Normalization: scaled to fall within a small, specified range**
- **Attribute or feature construction: New attributes constructed, added from the given set of attributes**

Data Transformation : Normalization

- **Solve issue of features with different scales for comparing features of data.**
  - **min-max normalization**

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

***Where minA* and *maxA* are minimum and maximum values of an attribute A.**

  - **z-score normalization** $v' = \frac{v - mean_A}{stand\_dev_A}$

***Where meanA and stand_devA* are mean & standard deviation of attribute A.**

  - **normalization by decimal scaling**

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

Data Transformation : Normalization

- **min-max normalization example**
  - **Suppose that the minimum and maximum values for an attribute income are $12,000 and $98,000, respectively. We would like to map income to range [0,1]. By min-max normalization, a value of $73,600 for income is transformed as following:**

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

=(73600-12000)/(98000-12000) x (1-0) +0 = 0.716

Data Transformation : Normalization

- **min-max normalization example**
  - **Suppose that the mean and standard deviation of values for the attribute income are $54,000 and $16,000, respectively. With z-score normalization, a value of $73,600 for income is transformed as :**

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

=(73600-54000)/16000 = 1.225

Data Transformation : Normalization

- **min-max normalization example**
  - Suppose that the recorded values of A range from -986 to 917. The maximum absolute value of A is 986.

To normalize by decimal scaling, we divide each value by 1,000 (i.e., j = 3) $v' = \frac{v}{10^j}$

So, -986 normalizes to -986/1000= - 0.986

917 normalizes to 917/1000 =0.917

Data Reduction

- Mining on huge amounts of data can take a long time, making such analysis infeasible.
- Data reduction applied to obtain a reduced representation of data set that is much smaller in volume, yet closely maintains integrity of the original data.
- Strategies for data reduction:
  - Data cube aggregation
  - Attribute subset selection
  - Dimensionality reduction
  - Numerosity reduction
  - Discretization and concept hierarchy generation

Data Reduction: Data cube aggregation

- **Applied to the data set in the construction of a data cube.**
- **Supports multiple level of aggregation in data cube.**
- **Use the smallest representation capable to solve the task.**

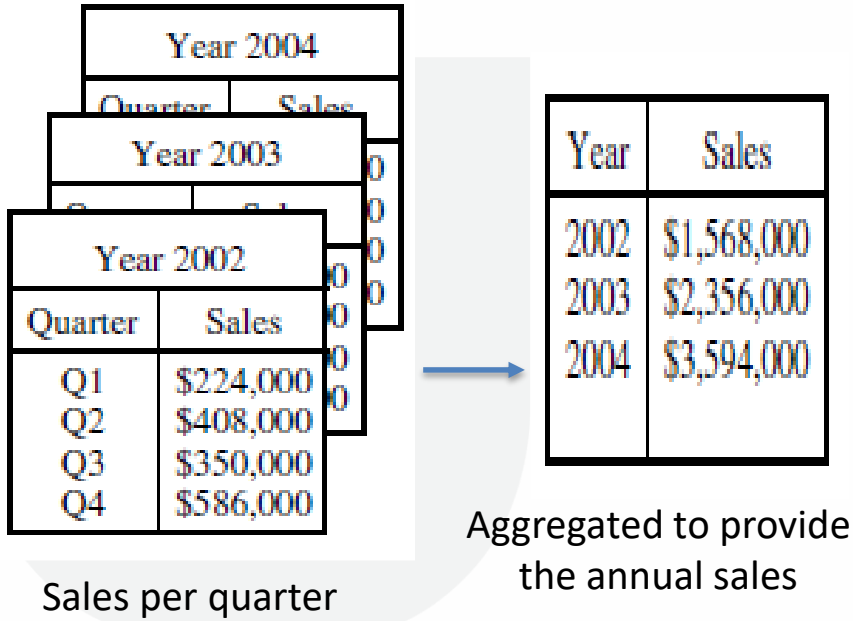


Image source : Data Mining Concepts and Techniques Book

Data Reduction: Attribute Subset Selection

- Data sets for analysis may contain hundreds of attributes, many of which may be irrelevant to the mining task or redundant.
- Attribute subset selection removes irrelevant attributes or dimensions.
- For n attributes, there are 2^n possible subsets. An exhaustive search for the optimal subset of attributes can be prohibitively expensive.
- Basic heuristic methods
  - Stepwise forward selection
  - Stepwise backward elimination
  - Combination of forward selection and backward elimination
  - Decision tree induction

Data Reduction: Forward Attribute Subset Selection

- **Forward selection starts with an empty set of attributes as reduced set.**
- **The best of original attributes is determined and added to reduced set.**
- **At each iteration, best of remaining original attributes is added to the set.**
- **Example:**

  Initial attribute set: {A1, A2, A3, A4, A5, A6}

  Initial reduced set: {}

  => {A1}

  => {A1, A4}

  => Reduced attribute set: {A1, A4, A6}

Data Reduction: Backward elimination

- **Backward selection starts with the full set of attributes.**
- **At each step, it removes the worst attribute remaining in the set.**
- **Example:**

  Initial attribute set: {A1, A2, A3, A4, A5, A6}

  => {A1, A3, A4, A5, A6}

  => {A1, A4, A5, A6}

  => Reduced attribute set: {A1, A4, A6}

Combination of forward selection and backward elimination

- It combines forward selection and backward elimination.
- At each step, the procedure selects the best attribute and removes the worst from the remaining attributes.

Data Reduction: Decision tree induction

- **Decision tree classification uses attribute selection measures, such as information gain for selecting best attributes based on given task .**
- **Example:**

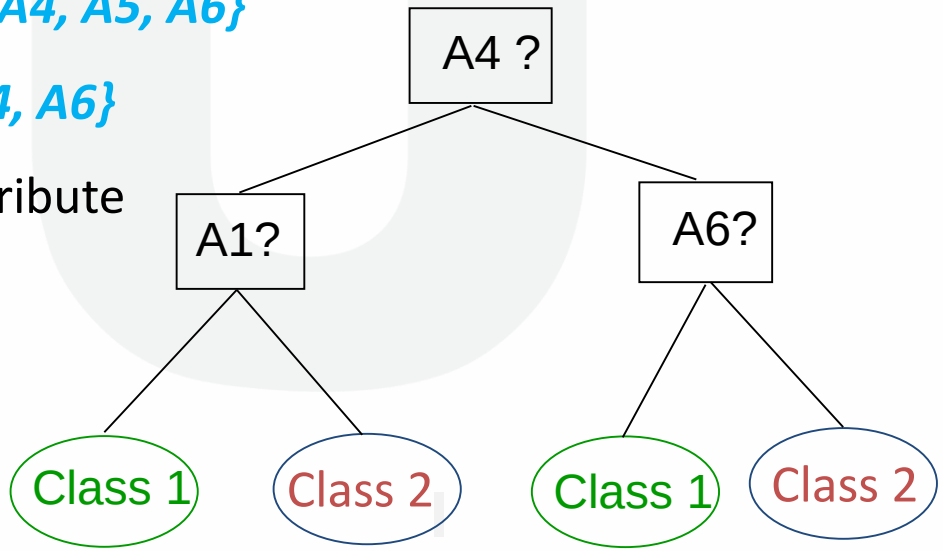
  Initial attribute set: {A1, A2, A3, A4, A5, A6}

  => Reduced attribute set: {A1, A4, A6}

Non leaf nodes: tests on values of attribute

Branches: outcomes of tests

Leaf nodes: class prediction



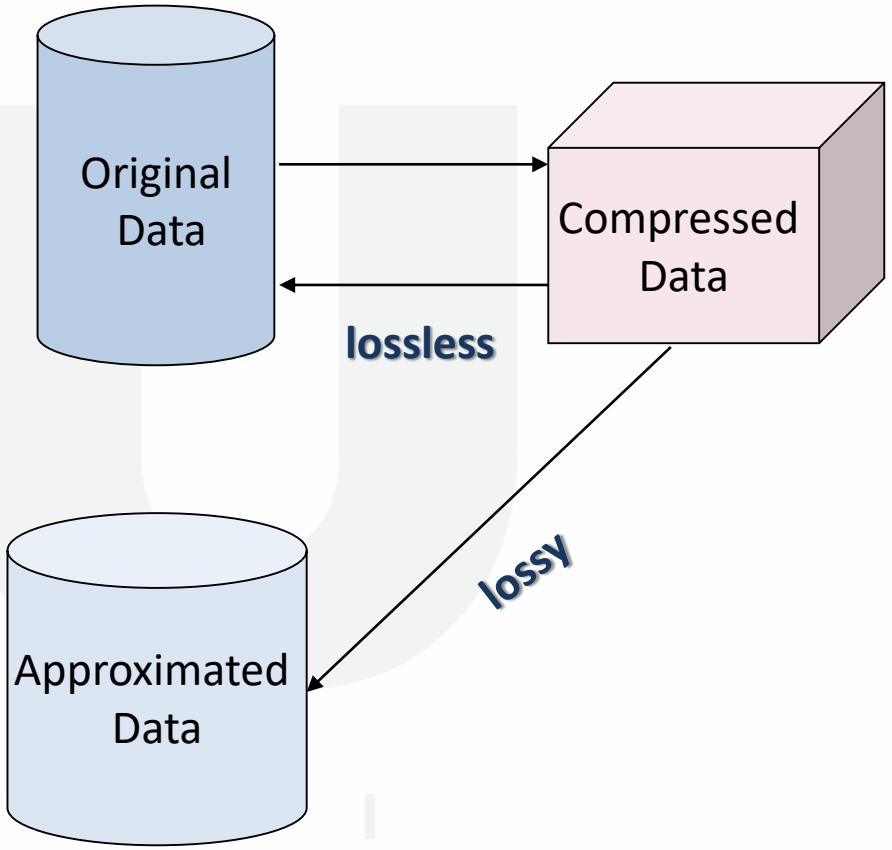
Data Reduction: Dimensionality reduction

- **Data encoding or data compression mechanisms are used to reduce the original data.**
- **Data reduction can be**
  - **Lossless**

    String compression algorithms
  - **Lossy**

    Wavelet transforms

    Principal components analysis.



String Compression Algorithms

- Several algorithms for string compression are available.
- Typically lossless
- Allow only limited manipulation of the data.

Discrete wavelet transform (DWT):

- Linear signal processing technique that, when applied to a n-dimensional data vector X, transforms it to a numerically different vector X' of same length , of wavelet coefficients.
- Compressed approximation of data can be retained by storing only a small fraction of strongest wavelet coefficients.
- Similar to discrete Fourier transform (DFT), however, the DWT achieves better lossy compression.
- Wavelet transforms have many real-world applications, such as compression of fingerprint images, computer vision, time-series data analysis.

DWT: Hierarchical Pyramid Algorithm

- Length L, input data vector must be an integer power of 2.
  - This condition can be met by padding data vector with zeros as(L >= n).
- Each transform involves two functions applied to pairs of data points in X:
  - Applies data smoothing (e.g., sum or weighted avg.)
  - Then performs a weighted difference to bring out detailed features of data. This results in two sets of data of length L/2.
- Applies these two functions recursively, until resulting data sets obtained are of length 2.
- Selected values from data sets obtained in above iterations are designated wavelet coefficients of the transformed data.

Principal Components Analysis

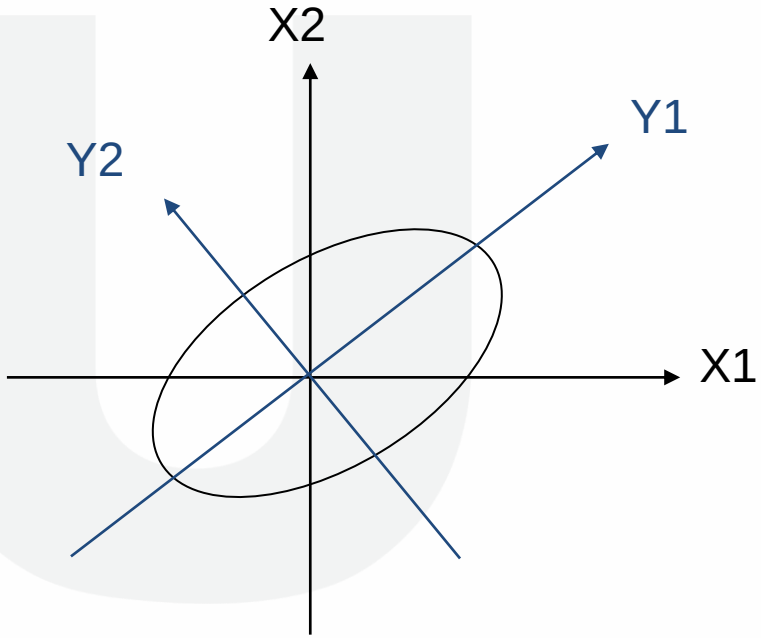
- **Given N data vectors from k-dimensions, Principal components analysis, or PCA (**also called the Karhunen-Loeve, or K-L, method**), searches for k n-dimensional orthogonal vectors that can best be used to represent the data, where k >n.**
- **PCA combines essence of attributes by creating an alternative, smaller set of variables.**
- **PCA can be applied to ordered and unordered attributes.**
- **It can handle sparse data and skewed data.**
- **PCA tends to be better at handling sparse data, whereas wavelet transforms are more suitable for data of high dimensionality**.

Principal Components Analysis Procedure

- The input data are normalized, so that each attribute falls within same range.
- PCA computes k orthonormal vectors that provide a basis for normalized data.
- The input data are a linear combination of the principal components.
- The principal components are sorted in order of decreasing “significance”.
- The principal components essentially serve as a new set of sorted axis, such that the first axis shows data having most variance, the second axis shows the next highest variance, and so on.
- The size of data can be reduced by eliminating weaker components with low variance.

Principal Components Analysis Example

- **Figure shows, first two principal components, Y1 and Y2, for the given data set originally mapped to the axes X1 and X2.**



Numerosity Reduction

- **Data are replaced or estimated by alternative, smaller data representations using**
  - **parametric models**

    Instead of the actual data, only the data parameters need to be stored.

    Example: Regression and Log-linear models
  - **nonparametric methods**

    Used to store reduced representations of the data.

    Example: histograms, clustering, and sampling.

Parametric models: Regression

- **Linear regression: The data are modeled to fit a straight line.**

  For example, a random variable, y can be modeled as a linear function of another random variable, x using following equation:

  $y = w\,x + b$,

  where w and b are regression coefficients used to specify the slope of the line and the Y-intercept, respectively.

- **Multiple linear regression: Allows a response variable, y to be modeled as a linear function of two or more predictor variables.**

  – **For example, a random variable, Y can be modeled as a linear function of two random variable, X_1, X_2 using following equation:**

  $Y = b_0 + b_1 X_1 + b_2 X_2$,

  where w, b_1 and b_2 are regression coefficients.

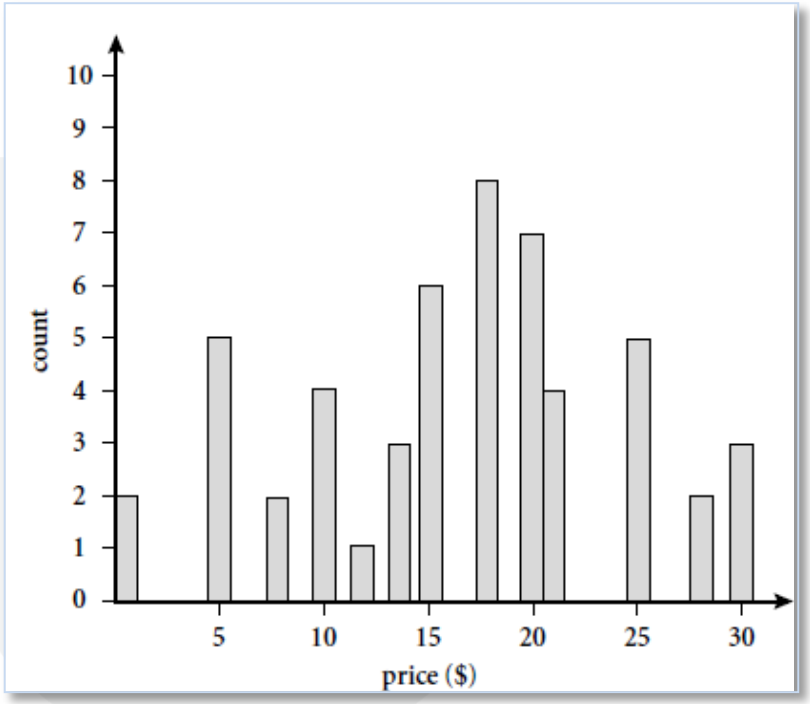
Parametric models: Log-Linear Models

- approximates discrete multidimensional joint probability distributions.
- Useful in dimensionality reduction and data smoothing.
- Used to handle skewed data.
- Good scalability for up to 10 or so dimensions.

Nonparametric models: Histogram

- **Histograms use binning to approximate data distributions.**
- **Partitions the data distribution into disjoint subsets also called buckets.**
- **Example:** *Following data are a sorted list of prices of commonly sold items:*

1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.



Histogram using **singleton buckets** (each bucket represents one price value/frequency pair)

Image source : Data Mining Concepts and Techniques Book

Nonparametric models: Clustering

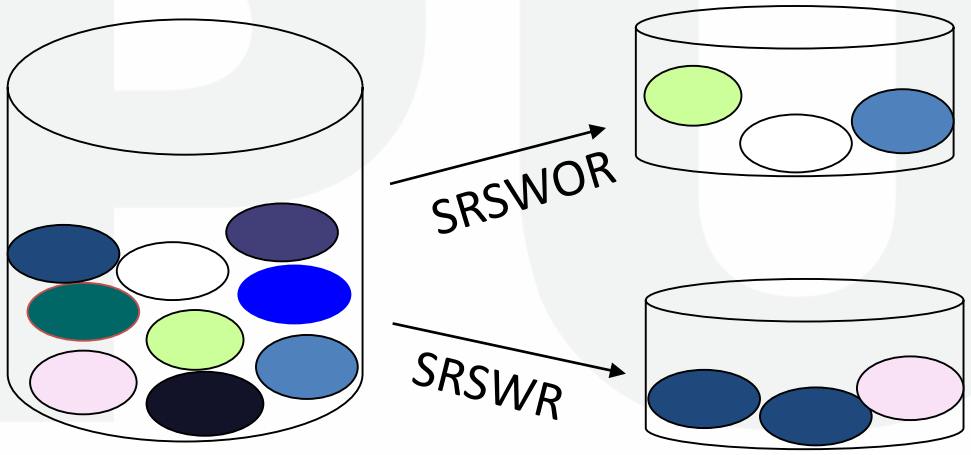
- Partition data set into groups called clusters, such that objects within a cluster are similar to one another and dissimilar to objects in other clusters.
- For data reduction, the cluster representations of data are used to replace the actual data.
- It is more effective for data that can be organized into distinct clusters than for smeared data.
- Quality of clusters measured by their diameter (max distance between any two objects in cluster) or centroid distance (average distance of each cluster object from its centroid).

Nonparametric models: Sampling

- Choose a representative subset of the data.
- Sampling complexity is potentially sublinear to the size of the data.
- Other data reduction techniques can require at least one complete pass through *D*.
- For a fixed sample size, sampling complexity increases only linearly as the number of data dimensions increases, whereas techniques using histograms, for example, increase exponentially in *n*.
- Common ways of Sampling:
  - Simple random sample without replacement (SRSWOR)
  - Simple random sample with replacement (SRSWR)
  - Cluster sample
  - Stratified sample

SRSWOR and SRSWR Sampling

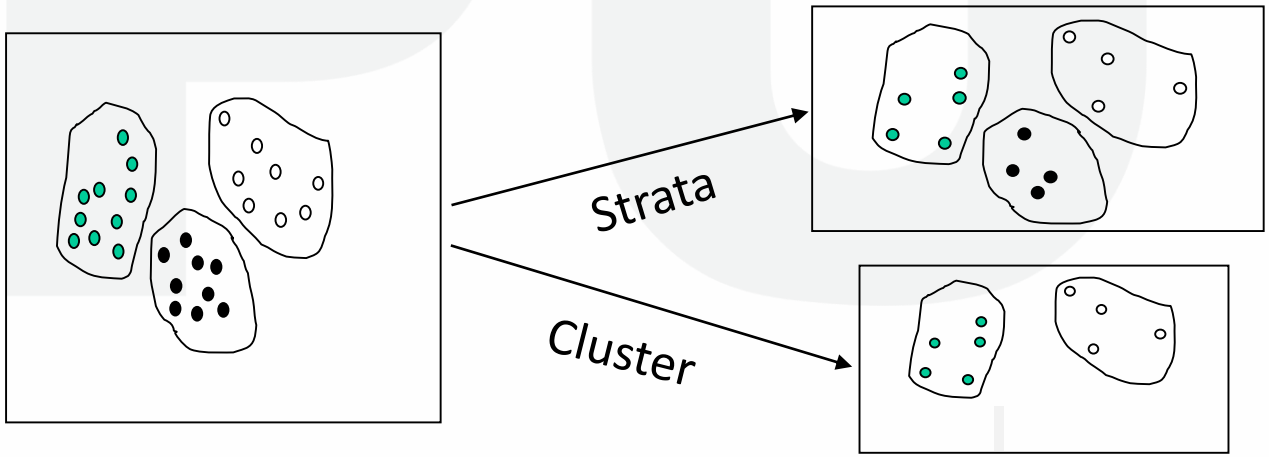
- **SRSWOR of size s is created by drawing s of the N tuples from dataset D (s < N), such that all tuples are equally likely to be sampled.**
- **SRSWR is similar to SRSWOR, except that each time a tuple is drawn from D, it is recorded and then placed back in D so that it may be drawn again.**



Data Object from **D**

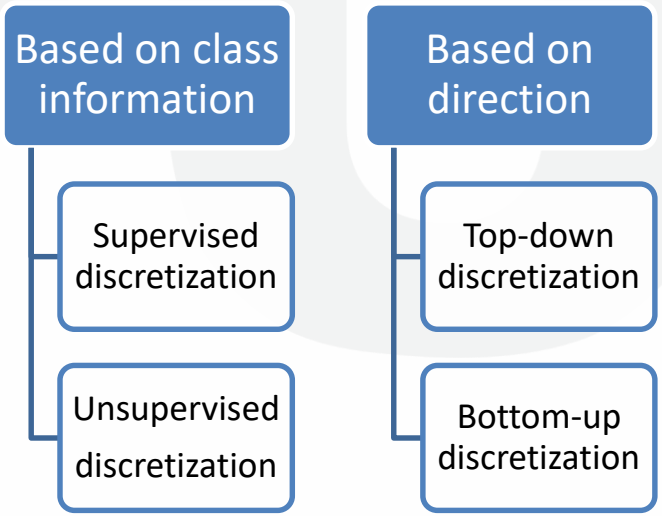
Cluster and Stratified Sampling

- For Cluster sampling, the tuples in D are grouped into M mutually disjoint clusters, then an SRS of s clusters can be obtained, where s < M.
- In Stratified sampling, data set D is divided into mutually disjoint parts called strata, then a stratified sample of D is generated by obtaining an SRS at each stratum.
- This helps ensure a representative sample, especially when the data are skewed.



Data Discretization (Quantization)

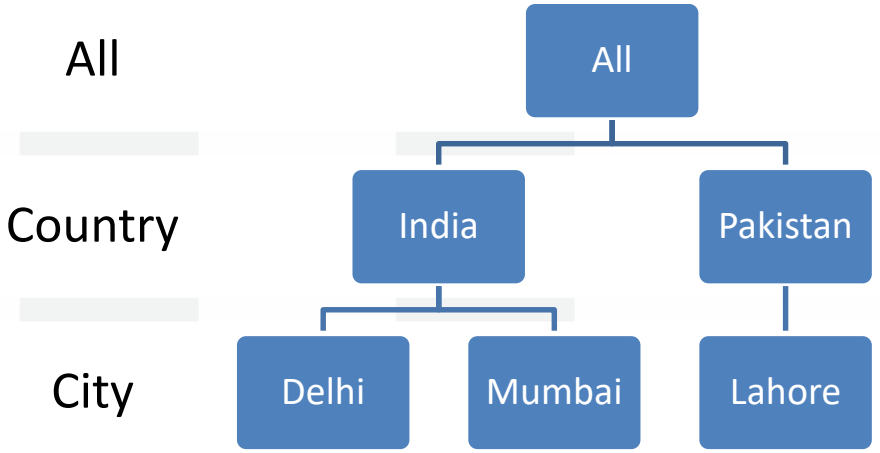
- **Reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals.**
- **Interval labels can then be used to replace actual data values.**
- **Replacing numerous values of a continuous attribute by a small number of interval labels reduces the original data.**
- **Types of discretization:**



Concept Hierarchy Generation

- **Reduce the data by collecting and replacing low level concepts (such as numeric values for the attribute age) by higher level concepts (such as young, middle-aged, or senior).**

- **Although detail is lost by such data generalization, the generalized data may be more meaningful and easier to interpret.**



Concept Hierarchy of Location

Discretization and concept hierarchy generation for numeric data

- **Wide diversity of possible data ranges and the frequent updates of data values for numerical attributes create difficulty for concept hierarchy.**
- **Concept hierarchies for numerical attributes can be constructed automatically based on data discretization.**
- **Methods:**
  - Binning
  - histogram analysis
  - cluster analysis
  - **entropy-based discretization**
  - **χ^2-merging,**
  - **discretization by intuitive partitioning**

Entropy-based discretization

- **Entropy-based discretization is a supervised, top-down splitting technique.**
- **Given a set of samples S, if S is partitioned into two intervals S1 and S2 using threshold T on the value of attribute A, the information gain resulting from the partitioning is:**

$$I(S,T) = \frac{|S_1|}{|S|} E(S_1) + \frac{|S_2|}{|S|} E(S_2)$$

- **where entropy E, of S1 given m classes is:** $E(S_1) = -\sum_{i=1}^{m} p_i \log_2(p_i)$

  p_i is the probability of class i in S1.

- **The process is recursively applied to partitions obtained until some stopping criterion is met, e.g.,** $E(S) - I(S,T) < \delta$

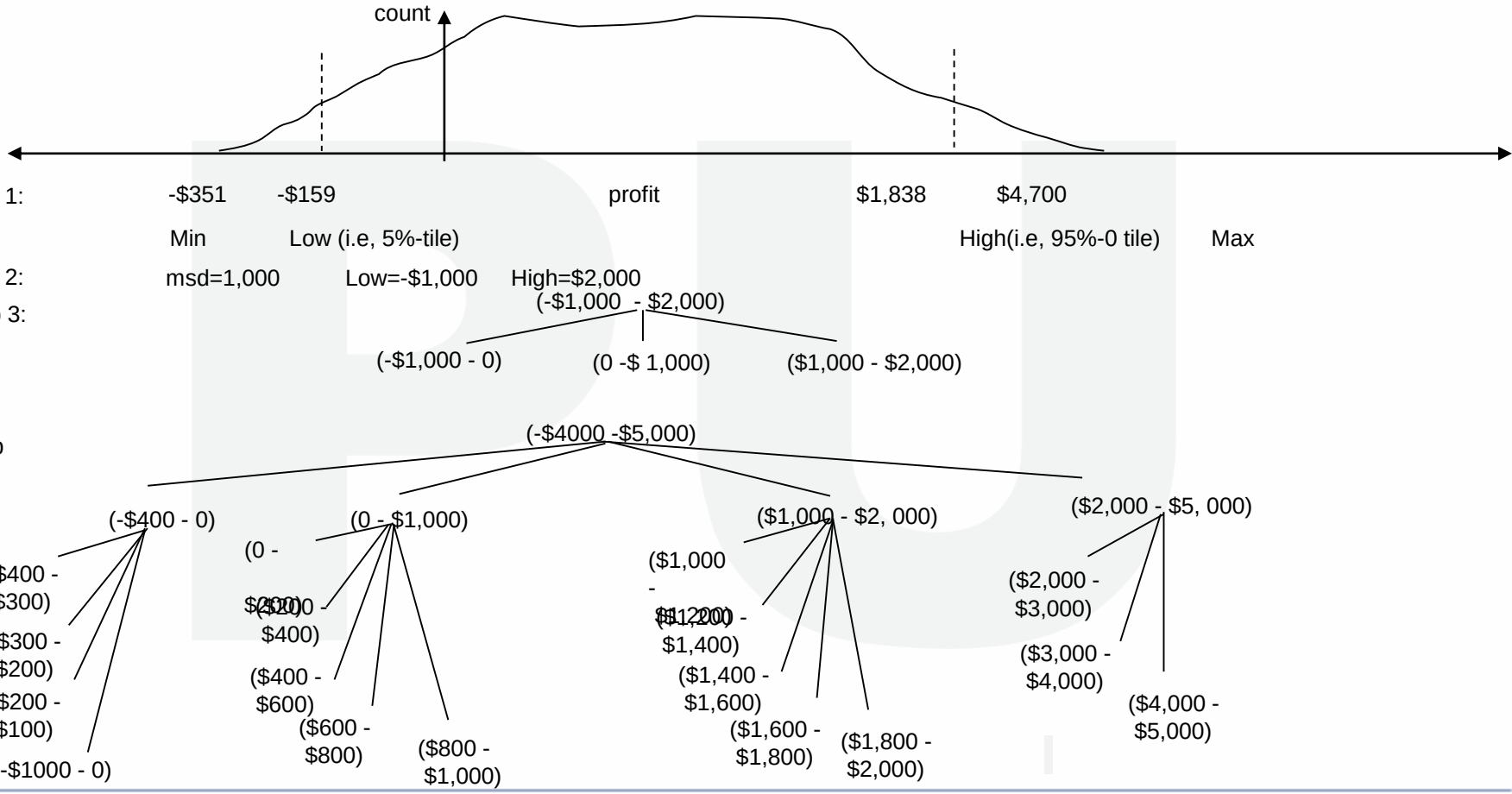
χ^2-merging

- ChiMerge is a $\chi 2$ based discretization method
- Applies supervised, bottom-up approach by finding the best neighboring intervals and then merging these to form larger intervals, recursively.
- If two adjacent intervals have very similar class distribution (low $\chi 2$ values), then intervals can be merged. Otherwise, they should remain separate.
- Steps:
  - each distinct value is considered to be one interval.
  - $\chi 2$ tests are performed for every pair of adjacent intervals.
  - Adjacent intervals with the least $\chi 2$ values are merged together,
  - This merging process proceeds recursively until a predefined stopping criterion is met.

Discretization by intuitive partitioning

- 3-4-5 rule can be used to segment numeric data into relatively uniform, “natural” intervals.
- It partitions a given range into 3,4, or 5 equiwidth intervals recursively level-by-level based on the value range of the most significant digit.
  - If an interval covers 3, 6, 7 or 9 distinct values at the most significant digit, partition the range into 3 equi-width intervals.
  - If it covers 2, 4, or 8 distinct values at the most significant digit, partition the range into 4 intervals.
  - If it covers 1, 5, or 10 distinct values at the most significant digit, partition the range into 5 intervals.
- The rule can be recursively applied to each interval, for creating a concept hierarchy for given numerical attribute.

Discretization by intuitive partitioning: Example of 3-4-5 rule



Concept hierarchy generation for categorical data

- **Categorical attributes have a finite number of distinct values, with no ordering among the values.**
- **Methods:**
  - **Specification of a partial ordering of attributes explicitly at the schema level by users or experts.**
  - **Specification of a portion of a hierarchy by explicit data grouping.**
  - **Specification of a set of attributes, but not of their partial ordering.**
  - **Specification of only a partial set of attributes.**

Data Integration

- **Combines data from multiple sources into a coherent store.**
- **These sources may include multiple databases, data cubes, or flat files.**
- **Issues:**
  - **How can equivalent real-world entities from multiple data sources be matched up? This is referred to as the entity identification problem.**

    Solution: Schema integration - integrate metadata from different sources
  - **How to deal with Redundancy?**

    Solution: Some redundancies can be detected by **correlation analysis**.
  - **How to detect and resolve data value conflicts?**

    Solution: Careful integration with special attention to structure of data.

DIGITAL LEARNING CONTENT

Parul® University

www.paruluniversity.ac.in