Unit II

Data Pre-Processing

- Why Pre-process the Data?
- Descriptive Data Summarization
- Data Cleaning
- Data Integration and Transformation
- Data Reduction
- Data Discretization and Concept Hierarchy Generation

Why Pre-process the data?

Sometimes data are:

Incomplete: lacking attribute values or certain attributes of interest or containing only aggregate data.

Noisy:

Containing errors (or) outlier values that deviate from the expected.

Inconsistent:

Containing discrepancies in the department codes used to categorize items.

Descriptive Data summarization:

It is used to identify the typical properties of data.

Measuring the central Tendency:

Center of set of data is arithmetic mean i.e; the mean (average) of set of values is

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Distributive measure:

Partitioning data into similar subsets and compute sum() & count()->max()&min()

Algebraic measure=mean() i.e; sum()/count().

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1x_1 + w_2x_2 + \cdots + w_Nx_N}{w_1 + w_2 + \cdots + w_N}$$

called as weighted arithmetic mean.

Use - <u>trimmed mean</u> because of the disadvantage of using mean is its avoid 2% of high and low sensitivity to extreme values.

For skewed (asymmetric) data, (enter of data is median).

(for calculating median) N values are in sorted order,

if N is even-> avg of the middle two numbers

N is odd-> center value

A holistic measure:

Computed on the entire data set as a whole.

Eg: **Median**

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width$$

L1->lower boundary of the median interval

N-> No. of values

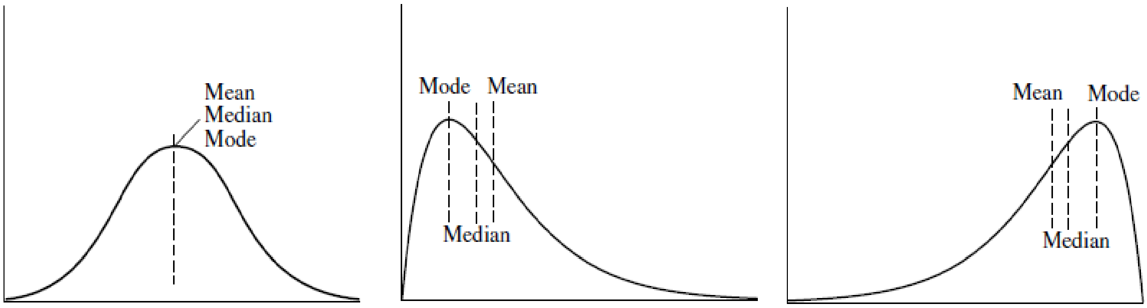
(∑freq)l->sum of all the intervals that are lower than the median intervals that are lower than the median interval.

Mode:

i.e; that occurs most frequently in the set.

Data sets with one, two (or) three modes are called unimodal, bimodal and trimodal & multimodal

mean-mod = 3*(mean-median)



(a) symmetric data (b) positively skewed data (c) negatively skewed data

Measuring the Dispersion of Data:

The degree to which numerical data tend to spread is called the dispersion (or) variance of data

Measures of data dispersion are

- Range
- Five-number summary
- Inter quartile range
- Standard deviation

Box-plots can be plotted based on 5-number summary and are a useful tool for identifying outliers.

Range:

Difference between max() & min() values.

Q_1=25% percentile Q_3=75%

Inter Quartile Range:

IQR= Q_3- Q_1

Outliers are =1.5*IQR i.e; above Q_3 or below Q_1

Five-number Summary:

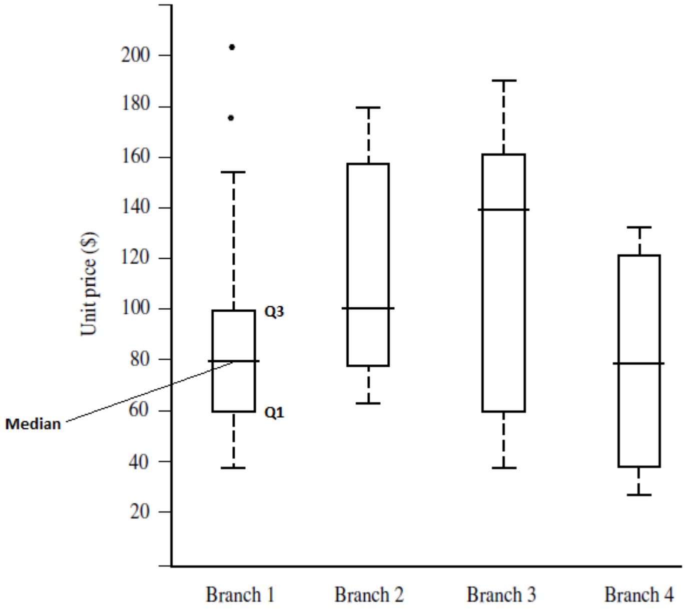
min1, Q_1,Q_3

Box-plots (visualizing a distribution)

->ends are Quartiles

1->median

-Two lines(whiskey) outside the box extend to the smallest and longest observations.



Unit price data for item sold at 4 branches during a given time period.

Variance & Standard deviation:

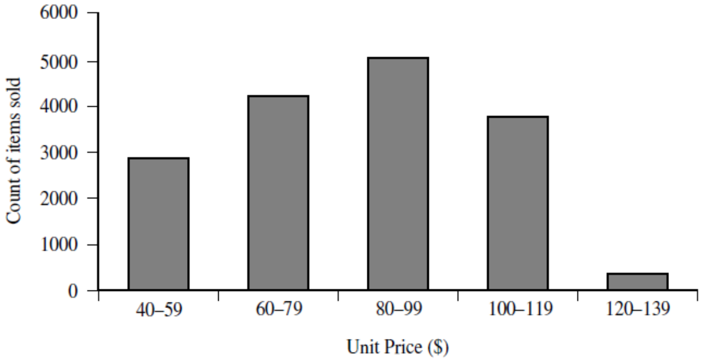
Variance of N observations, x_1,x_2,………,x_N is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\left[\sum x_i^2 - \frac{1}{N}(\sum x_i)^2\right]$$

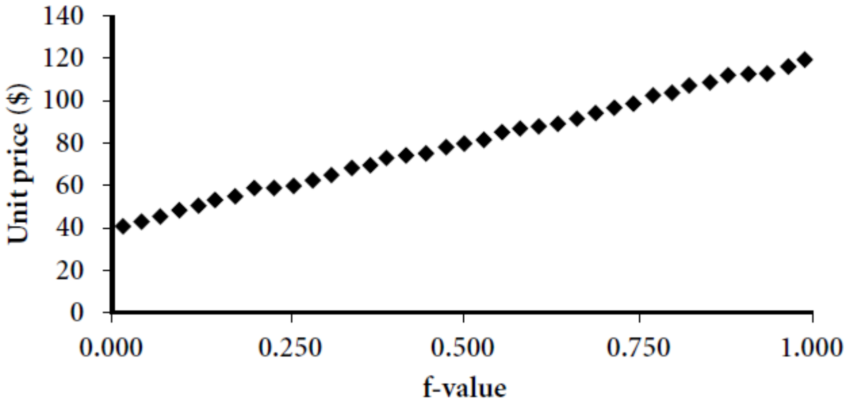
Standard deviation=σ

Graphic Displays of Basic Descriptive Data summaries:

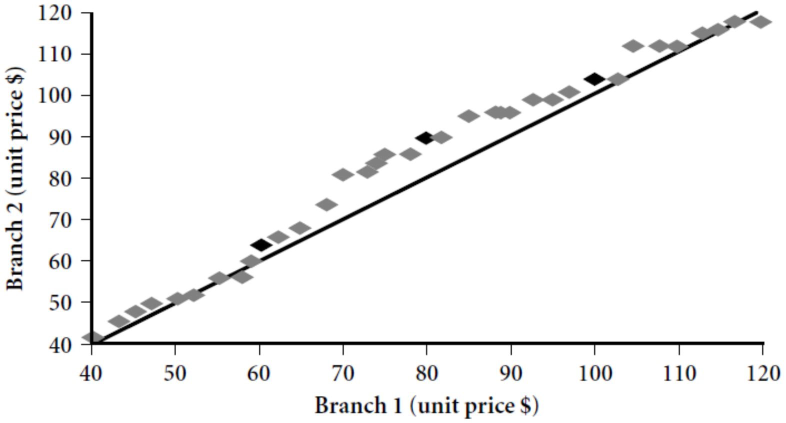
- Histograms, quantile plots, q-q plots, scatter plots and less curves
- Frequency histograms->Data buckets & Distribute data information
- Quantile plot-> q-q plot.
- Scatter plot



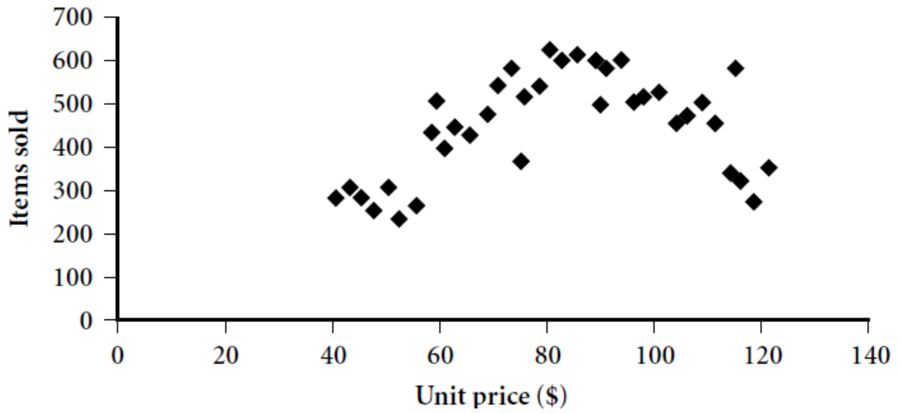
Histogram



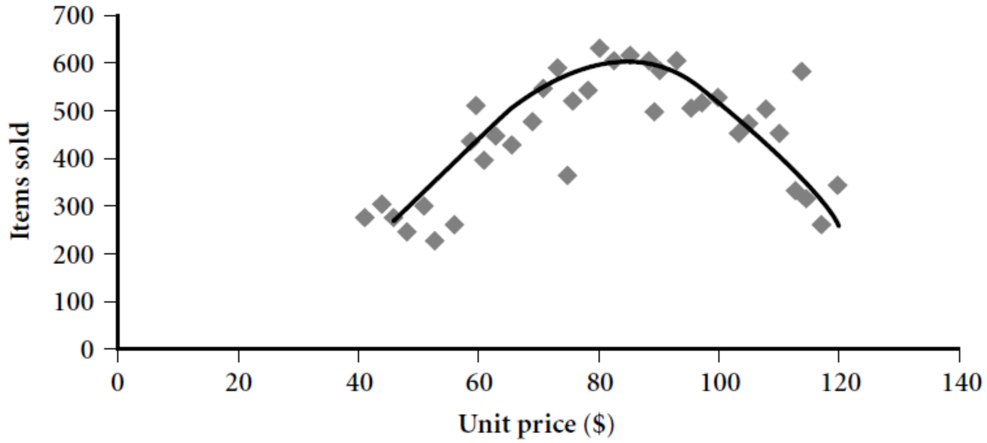
A quantile plot



A q-q plot



A scatter plot



A loose curve

Data Cleaning:

Real-world data tend to be incomplete, noisy & inconsistent. Data cleaning attempt to fill in missing values, smooth out noise while identifying Outliers, and correct inconsistencies in the data.

Missing values:

Ex: customer income is not recorded.

Methods used for handling missing values:

1. Ignore the tuple:
   - This is usually done when the class label is missing.
   - It is not very effective, unless the tuple contains several attributes with missing values.
2. Filling in the missing value manually:
   - It is time consuming and not feasible when many missing values.
3. Use a global constant to fill in the missing value:
   - i.e; with 'unknown or α'
4. Use the attribute mean to fill in the missing value:
   - Suppose average value for income is 56,000$, use this value to till missing values.
5. Use the attribute mean for all samples belonging to the same class as the given tuple

Ex: credit – risk: A-> then fill average income:56,000$

:B-> then fill average income:30,000$

6. Use the most probable value to fill in the missing value:
   - Probable value may be determined by regression, Decision-true- It is a popular strategy.

Noisy Data:

Noise is a random error or variance in a measured variable.

Techniques used to remove the noise:

1) Binning:

Binning methods smooth a sorted data value by consulting its "neighbour hood" values around it.

Stored values are distributed into a number of buckets (or) bins.

Sorted data for price (in dollars): 48,15,21,21,24,25,28,34

Partition into equal-frequency bins of size 3:

Bin1: 4,8,15

Bin2: 21,21,24

Bin3: 25,28,34

Smoothing by bin means:

Bin1: 9,9,9

Bin2: 22,22,22

Bin3: 29,29,29

Smoothing by bin medians:

Bin1: 8,8,8

Bin2: 21,21,21

Bin3: 28,28,28

Smoothing by bin boundaries:

Bin1: 4,4,4

Bin2: 21,21,24

Bin3: 25,25,34

2) Regression:

Data can be smoothed by fitting the data to a function, such as with linear regression involves finding the best line to fit two attributes.

Multiple linear regression is an extension, where more than two attributes are involved and the data are fit to a multidimensional surface.

3) Clustering:

Outliers may be detected by clustering, where similar values are organized into groups or clusters. The values fall outside of the set of clusters may be considered outliers.

Ex: Data cleaning needs for poorly designed data entry forms, human error in data entry, data decary (out dated addresses), system errors.

Data Integration:

Data Integration means which combines data from multiple sources into a coherent data store.

➔ Shcema integration and object matching

How can equivalent real-world entities from multiple data sources be matched up? -> this is called as " entity – identification problem".

Redundancy: (annual revenue)

An attribute may be redundant if it can be derived from another attribute (or) set of attributes.

➔ Redundancies can be detected by correlation analysis -> how strongly one attribute implies another.

For numerical attribute A&B, compute correlation coefficient

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

N-> No. of tuples

a_i and b_i-> respective values of Attributes A and B in tuple i,

$\bar{A}$ and $\bar{B}$-> respective mean values of A & B

σ_A ans σ_B -> respective standard deviations of A & B

S(*aibi*)-> sum of the AB cross-product i.e; for each tuple, the value for *A* is multiplied by the value for *B* in that tuple

Different types of attributes:

$\left.\begin{array}{l}\text{Nominal}\\ \\ \text{Ordinal}\end{array}\right\}$ categorical (qualitative) (symbols)

$\left.\begin{array}{l}\text{Inter val}\\ \\ \text{Ratio}\end{array}\right\}$ Numeric (quantitative) (numbers)

Nominal -> just different names(only provide information to distinguish one object from Another)(=,≠) Distinctness

Ex: Zipcodes, e ID's, eye color, gender.

Ordinal -> Provide enough information to order objects(>,<)

Ex: Hardness of materials {good,better,best}

Grades, street numbers

Interval -> Differences between values are meaning ful(+,-)

Ex: calendar dates, temperature in Celsius

Ratio -> both differences and ratios are meaning ful(*,/)

If $r_{A,B}>0$ then A&B are positively correlated.

If A increases, B also increases

Either A or B may be remoted as redundancy.

If $r_{A,B}=0$ then A&B are independent and there is no correlation between them.

If $r_{A,B}<0$ then A&B are negatively related.

For categorical(discrete) data, correlation can be discovered by (chi-square) test.

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

A has 'c' distinct values, B has 'r' distinct values.

$a_1,a_2,a_3,\ldots\ldots\ldots.a_c$ $\qquad\qquad$ $b_1,b_2,b_3,\ldots\ldots\ldots.b_r$

o_{ij}=observed frequency (actual count) of the joint event(A_i,B_j) and

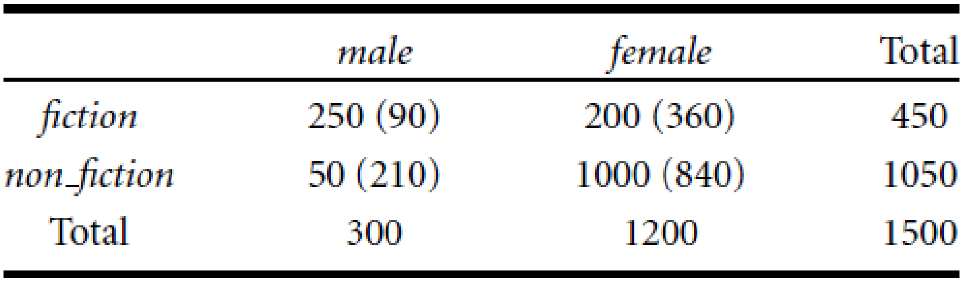
$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N}$$

N is the number of data tuples,
count(*A*=*ai*) is the number of tuples having value *ai* for *A*, and
count(*B* = *bj*) is the number of tuples having value *bj* for *B*.

χ2 tests the hypothesis that A&B are independent

Correlation analysis of categorical attributes using **χ2** :

Contingency table are "gender" and "preferred-Reading".



	male	*female*	Total
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

$$e_{11} = \frac{count(male) \times count(fiction)}{N} = \frac{300 \times 450}{1500} = 90$$

$$\begin{aligned}\chi^2 &= \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} \\ &= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.\end{aligned}$$

A,B are independent.

The test is based on a significance level, with (1-1)*(c-1) degrees of freedom.

If the hypothesis can be rejected, then A&B are related.

(2*2) table (2-1)*(2-1)=1
r-1 c-1

for 1 degree of freedom $\chi 2$ value needed to reject the hypothesis at the 0.001 significance level is 10.828 .

➔ Then we can reject the hypothesis, and the two attributes are (strongly) correlated for the given group of people

3rd one is

<u>The detection and resolution of data value conflicts:</u>

For the same entity, attribute values from different sources may differ. This may be due to differences in representations, scaling or encoding.

Ex: weight-in metric units,

Price of rooms in different cities -> in different services(taxes)

An attribute in one system may be recorded at a lower level of abstraction

➔ When matching attributes from one data base to anther during integration, special attention must be paid to the structure of the data. (to ensure fd's & referential Integrity constraints must be there is target system.)

➔ Careful integration of data from multiple sources can help reduce and avoid redundancies and inconsistencies in the resulting data set. This can improve the resulting data set. This can improve the accuracy and speed of the subsequent mining process.

Data Transformation:

Data are transformed or consolidated into forms appropriate for mining.

1. Smoothing: which works to remove noise from the data.
   Ex: regression, binning, clustering
2. Aggregation: where summary or aggregation operations are applied to the data
3. Generalization: of the data, where low-level data are replaced by higher-level concepts by using concept hierarchies.
4. Normalization: where the attribute data are scaled. So as to fall within small specified range, such as -1.0 to 1.0 or 0.0 to 1.0
   - Useful for classification also.

1. Min-max normalization
2. Z-score normalization
3. Normalization by decimal scaling

Min-max Normalization:

Performs a linear transformation on the original data. min_A&max_A -> values of attribute A
It maps a value *v*, of A to v' in the range [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: the minimum and maximum values for the attribute *income* are $12,000 and $98,000, respectively. We would like to map *income* to the range [0:0;1:0]. By min-max normalization, a value of $73,600 for *income* is transformed to 0:716.

Z-score Normalization: (or) (zero-mean)

The value of an attribute A. are normalized based on the mean and standard deviation of A.

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

Ex: mean=54,000
SD=16,000
Value=73,600
Obtained value=1.225

Decimal scaling:

Normalization by decimal scaling normalizes by moving the decimal point of values of attribute *A*. The number of decimal points moved depends on the maximum absolute value of *A*. A value, *v*, of *A* is normalized to *v*0 by computing

$$v' = \frac{v}{10^j}$$

Suppose recorded values by A in the range -986 to 917
The maximum absolute value of A is 986
So decide each value by 1000 (i.e j=3) so that -986 to -0.986 and 917 to 0.917

Normalizes by moving the decimal point the decimal point of value of attribute A
No. of decimals points moved depends on the maximum absolute value of A

5. Attribute Construction:

   New attributes are constructed from the given attributes and added in order to help improve the accuracy and understanding of structure in high-dimensional data.

   Area(attribute) can be added based on height & width.

Data Reduction:

Complex data analysis and mining on huge amounts of data can take a long time, making such analysis impractical or infeasible.

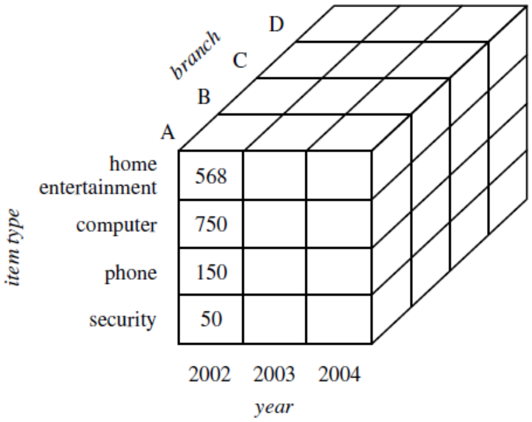
Data Reduction techniques can be applied to obtain a reduced representation of the data set and mining yields efficient results.

Techniques are:

1) Data cube aggregation-aggregation operations are applied to Data.
2) Attribute subset selection- irrelevant or redundant attributes may be detected and removed.
3) Dimensionality reduction- where encoding mechanisms are used.
4) Numerosity reduction-data are replaced or estimated through parametric , non-parametric methods.
   ->clustering, sampling and histograms
5) Discretization and concept hierarchy generation.

1) Data cube Aggregation:

   Data can be aggregated from Quarters to years in data cubes.



Concept hierarchies used for performing the analysis at multiple levels of abstraction(base avoid, apex cuboid)

Data cubes created → cuboids.

2) Attribute subset selection:

   Data sets may contain hundreds of attributes, which may be irrelevant to the mining task or redundant.

   To purchase CD or not "phno" of customer-is irrelevant age, music-taste-are relevant.

   →Attribute subset selection reduces the data set size by removing irrelevant or redundant attributes.

   But for 'n' attributes 2^n subsets are true.

   →"Optional subset" can be selected through heuristic greedy approach.

   →"best attributes" are determined using tests of statistical significance.

   Ex: Information gain , entropy.

   Basic heuristic methods of attribute subset selection includes.

i) Stepwise forward selection:

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

Initial reduced set= { }

- ⇨ {A1}=> {A1,A4}

Reduced attribute set= {A1,A4,A6}

Procedure starts with an empty set of attributes, best of original attributes is determined and added to the reduced set.

ii) Stepwise backward elimination:

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

- ⇨ {A1,A4,A5,A6}

Reduced attribute set: {A1,A4,A6}

The procedure starts with the full set of attributes. At each step, it removes the worst attribute remaining in the set.

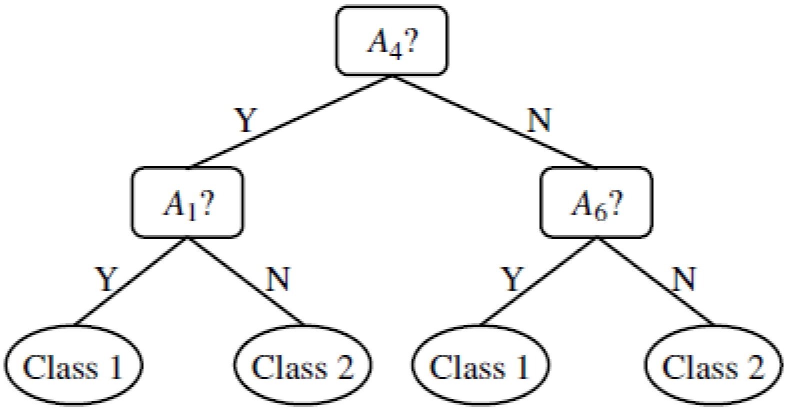
iii) Combination of forward selection and backward elimination:

At each step, the procedure selects the best attributes, and removes the worst from the set.

iv) Decision tree induction:

ID3,C4.5 &CART used for classification. It constructs flowchart like structure, where each internal node denotes a test on an attribute, each branch corresponds to an outcome of the test and each external (leaf) node denotes a class prediction. At each node algorithm chooses best node.

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



Therefore Reduced attributeset={A1,A4,A6}

3. Dimensionality Reduction: (Data compression)

Data encoding or transformations are applied to obtain a reduced or compressed representation of the original data.

Looseless: If the original data can be reconstructed from the compressed data without any loose of information.

Loosy: We reconstruct only an approximation of the original data.

Loosy dimensionality reduction techniques are: wavelet transforms, and principle components analysis.

Wavelet transforms: Discrete wavelet transform(DWT) is a linear signal processing technique.

→Each tuple is an n-dimensional data vector. i.e; X=(X1,X2,……Xn)

X→ can be transformed to X^1, of wavelet coefficients.

- ➔ Given a set of coefficients, approximation of original data can be constructed by applying the inverse of DWT.

DWT uses a hierarichal pyramid algorithm that halves the data at each iteration.

1)Length L, of input data vector must be an integer power of 2 or by padding with zeros as necessary (L>=n)

2)Two functions used.

First applies data smoothing such as weighted average.

Second performs a weighted difference.

3)Two functions are applied on two halfs of data.

i.e; Two sets L/2

4)Two functions are recursively applied to sets of data until the resulting dataset are of length 2.

5)selected values are designated as wavelet coefficients of the transformed data.

So, strength lies wavelet coefficients.

Principle components Analysis:

PCA procedure is

1)The input data are normalized.

2)PCA computers K orthogonal vectors perpendicular to the other called as principal components.

3)Principle components are stored in order of decreasing "significance" or "strength" of variance.

4)The size of the data can be reduced by eliminating the weaker components i.e; of low variance.

→ PCA is computationally inexpensive.

Pc's used as i/p to multiple regression and cluster analysis PCA suitable for handling sparse data.

DWT suitable for data of high dimensionality.

4. Numerosity Reduction:

It reduce the data volume by choosing 'smaller' forms of data representation.

Parametric methods: A model is used to estimate the data, the data parameters need to be stored, instead of the actual data. i.e; log-linear models.

Non parametric methods: For string reduced representations of the data include histograms, clustering and sampling.

Regression & log-linear Models:

Linear regression → the data are modeled to bit a straight line.

Y→response variable

X→predictor variable

➔Y=wX+b

X,Y are numerical database attributes

W,b are regression coefficients.

Multiple linear regression, which allows a response variable u, to be modeled as a linear function of two or more predictor variables.

Log-linear Models:

Approximate discrete multidimensional probability distributions.

It is used to estimate the probability of each point in a multidimensional space for a set of discretized attributes, based on smaller of dimensional combinations.

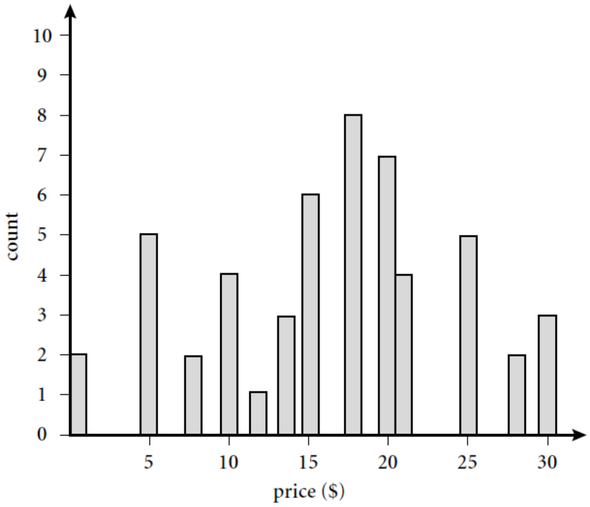
Histograms:

A histogram for an attribute, A, partitions the data distribution of A into disjoint subsets, or buckets.

If each bucket represents only a single attribute values, the buckets are called singleton buckets.

Price:

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



Equal-width:

The width of each bucket range is uniform (10$)

Equal Frequency: or (Equi depth)

The frequency of each bucket is constant.

V-optimal:

For all of the possible histograms, v-optimal histogram is the one with least variance is weighted sum of the original value that each bucket represents, where bucket weight is equal to the no. of values in the bucket.

Max-diff:

Consider difference between each pair of adjacent values. V-optimal & max-diff→ tend to be the most accurate and practical.

Clustering:

It takes data types as objects. They partition the objects into groups or clusters. Object within a cluster are "similar" to one another and "dissimilar" to objects in other clusters.

→Quality of a cluster may be repreented by its diameter.

Diameter means maximum distance between any two objects in cluster.

Centroid distance means average distance of each object from the cluster centroid.

→Cluster representations of the data are used to replace the actual data.

Sampling:

Used for data reduction.

Sampling allows a large data set to be represented by much smaller random sample of the data.

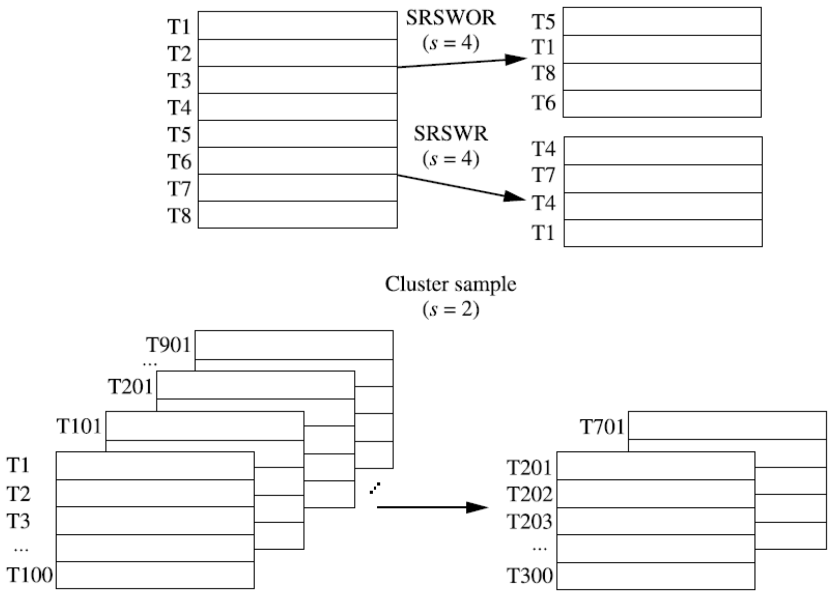
Methods:

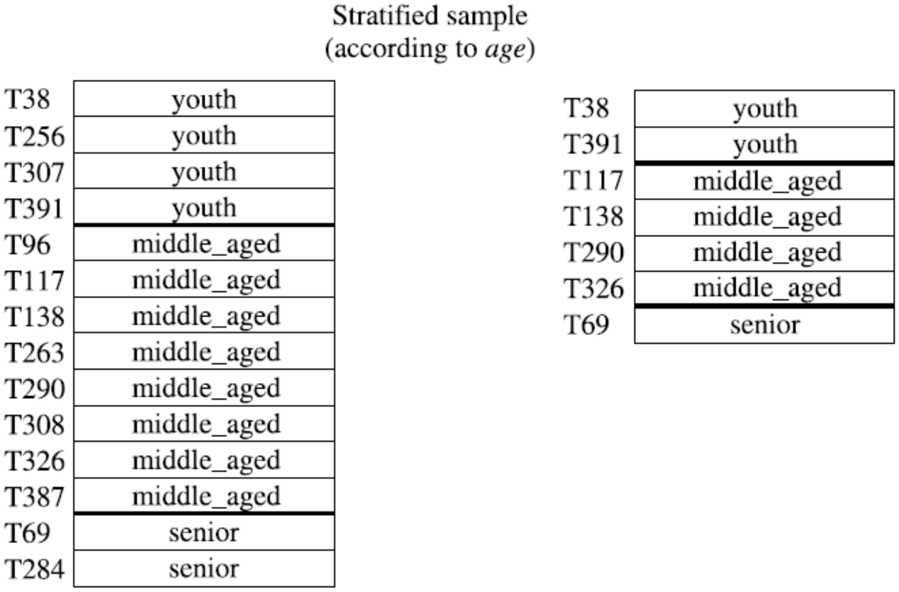
1) Simple random sample without replacement (SRSWOR) of size s:
   By taking s of the N tuples from D(s<N), where the probability of drawing any tuple in D is 1/N, i.e; all tuples are equally likely to be sampled.
2) Simple random sample with replacement(SRSWR) of size s:
   Similar to SRSWOR, except that each time a tuple is drawn from D, it is recorded and then replace.
3) Cluster Sample:

If tuples in D are grouped into M mutually disjoint "clusters", then an SRS of s clusters can be obtained i.e;s<M

4) Stratified Sample:

If 'D' is divided into manually disjoint parats called strata, a stratified sample of D is generated by obtaining an SRS at each stratum.





Stratified sample
(according to *age*)

T38	youth
T256	youth
T307	youth
T391	youth
T96	middle_aged
T117	middle_aged
T138	middle_aged
T263	middle_aged
T290	middle_aged
T308	middle_aged
T326	middle_aged
T387	middle_aged
T69	senior
T284	senior

T38	youth
T391	youth
T117	middle_aged
T138	middle_aged
T290	middle_aged
T326	middle_aged
T69	senior

Advantage of sampling is the cost of obtaining a sample is proportional to the size of the sample s.

→It is possible to determine a sufficient sample size for estimating a given function within a specified degree of error.

S is small in comparision to N.

Data Discretization and concept hierarchy generation:

Data discretization can be used to reduce the no. of values for a given continuous attribute by dividing the range of the attribute into intervals. This leads to a concise, easy-to-use, knowledge level representation of mining results.

Data Discretization→i) top-down approach
ii) bottom-up approach

Discretization and concept hierarchy generation for Numerical Data:

→Because of the wide diversity of possible data ranges, and frequent updates of data values, it is difficult to specify concept hierarchies for numerical attributes.

→These are automatically constructed basd on data discretization.

Methods used are→i) binning
ii) histogram analysis
iii) entropy based discretization
iv) X^2-merging &
v) Discretization by intuitive partitioning.

1) <u>Binning:</u>

It is top-down splitting Technique based on specified no. of bins. Attribute values can be discretized by applying equiwidth, equidepth binning and then replacing each bin value by the mean (or) median.

These techniques can be recursively applied to resulting partitions in order to generate concept hierarchies.

→It does not use class information so it is unsupervised technique.

2) <u>Histogram analysis:</u>

It is also unsupervised discretization technique. It partition the values into disjoint ranges called buckets.

The histogram analysis algorithm can be recursively applied to each partition in order to automatically generate a multilevel concept hierarchy.

→A minimum interval size can also used per level to control the recursive procedure.

3) <u>Entropy based Discretization:</u>

→ It is one of the most commonly used discretization measures.

➔ It is supervised discretization, & top-down splitting technique.

➔ To discretize 'A', it selects value of A that has the minimum entropy as a split-point and recursively partitions the resulting intervals to arrive at a hierarchical discretization.

D-Dat set of tuples with attributes & class level attributes method.

i)Each value of A can be boundary or split-point to partition the range of A.

→Split point can partition the tuples in D ito two subsets, satisfying the conditions A<= Split-point and A>split-point, (binary discretization).

ii) Class c1 c2

$$Info_A(D) = \frac{|D_1|}{|D|}Entropy(D_1) + \frac{|D_2|}{|D|}Entropy(D_2)$$

$$Entropy(D_1) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

<u>Discretization by intuitive partitioning:</u>

Many users would like to see numerical ranges partitioned into relatively uniform, easy-to-read intervals, that appear "natural" or "intuitive"

Ex: income –($50,000 , 60,000) instead of (51,263.98,$60,872.34)

→3-4-5-rule can be used to segment numerical data into uniform, natural seeming intervals.

→The rule partitions a given range of data into 3,4 or 5 relatively equal-width intervals, recursively and level by level, based on the value range at the most significant digit.

→If an interval covers 3,6,7 or 9 distinct values at the most significant digit, then partition the range into 3 intervals.

→If it covers 2,4,(or)8 distinct values at the most significant digit, then partition the range into 4 equal-width intervals.

→If it covers 1,5, (or) 10 distinct values at the MSD , then partition the range into 5-equal width intervals.

→The top-level discretization can be performed based on the range of data values representing the majority.

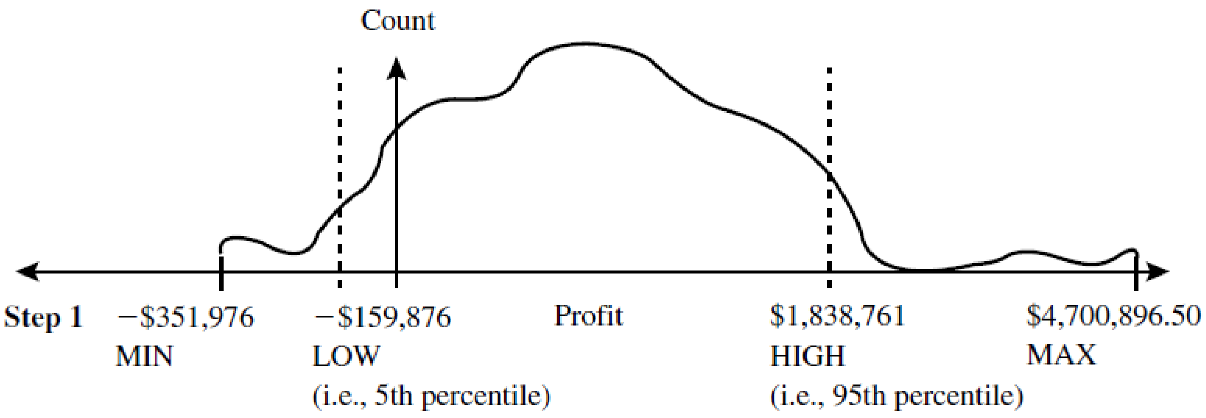
(5th percentile to 95th percentile)

Ex:

Profits at different branches for the year 2004.

-$351,976.00 to $4,700,896,50.

Step1:



Step2: Most significant digit =1,000,000

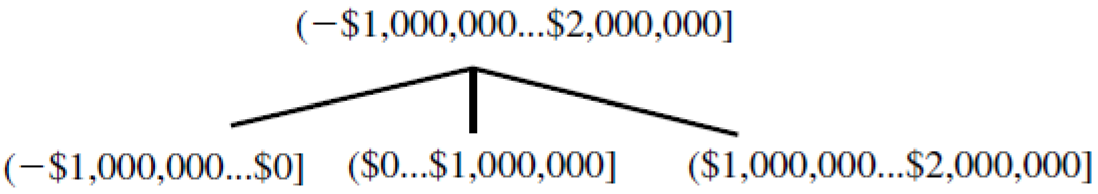
Low=1,000,000$ and high[1]=$2,000,000

Step3: Since this interval ranges over three distinct values at the MSD

i.e; (2,000,000-(-1,000,000))/1,000,000 =3

Partitioned into 3 equal-width subsequent according to 3-4-5 rule.

Step3:



Step4:

Examine MIN&MAX values to see how they “fit” into the first –level partitions.

i.e; first interval covers MIN value, i.e; Low^1<MIN, so we can adjust the left boundary of this interval to make the interval smaller.

i.e; MSD of MIN is the hundred thousand digit position. So Rounding MIN down to this partition.

We get MIN^1=-$4,00,000

i.e; first interval is (-$4,00,000……..0]

→Since the last interval ($1,000,000…..$2,000,000] does not cover the max value. So ($2,000,000…..$5,000,000] is new interval.

Step4:

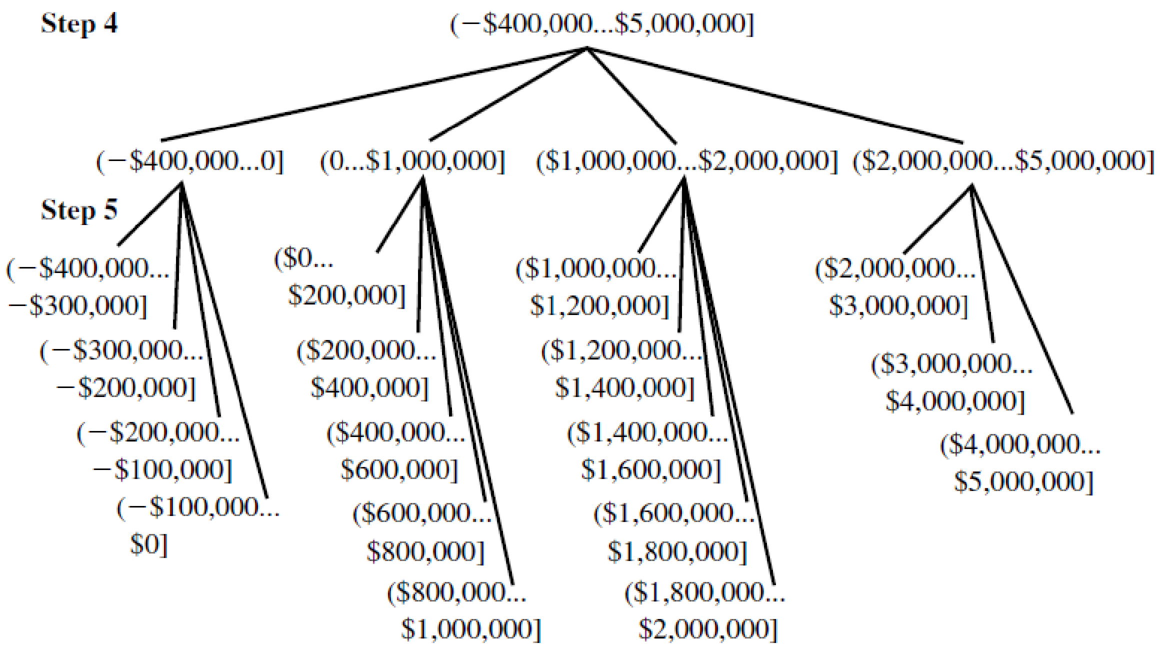


Fig: Automatic generation of concept hierarchy for profit based on the 3-4-5 rule.

Step5:Recursively, each interval can be further partitioned according to the 3-4-5 rule to from the next lower level of the hierarchy. Similarly, the 3-4-5 rule can be carried on iteratively at deeper levels, as necessary.

Concept hierarchy generation for categorical Data:

Categorical attribute have a finite no. of distinct values, with no ordering among the values.

Ex: Geographic location, job category and item type to generate concept hierarchies.

1) Specification of a partial ordering of attributes explicitly at the schema level by users (or) experts:

   A user or except can easily define a concept hierarchy by specifying a partial or total ordering of the attributes at the schema level.

   Ex: location, total ordering is street<city<state<country.

2) <u>Specification of a partition of a hierarchy by explicit Data grouping:</u>

This is manual definition of a portion of a concept hierarchy.

→In large database, it is unrealistic to define an entire concept hierarchy by explicit value enumeration. So we can easily specify explicit grouping for a small portion of intermediate level data.

Ex: After specifying that state and country from a hierarchy at the schema level, a user could define some intermediate levels manually such as.

{ Alberta, suskatchowan, Manitoba}< prairks-canada" and {British, Columbia, prairks-canada} < western – canada.

3) <u>Specification of a set of attribute, but not of their partial ordering:</u>

A user may specify a set of attributes forming a concept hierarchy, but omit to explicitly state their partial ordering.

→The system automatically generate the attribute ordering so as to construct a meaningful concept hierarchy.

- An attribute with most distinct values is placed at the lowest level of the hierarchy.
- Some adjustments may be applied by user or experts afterwards.

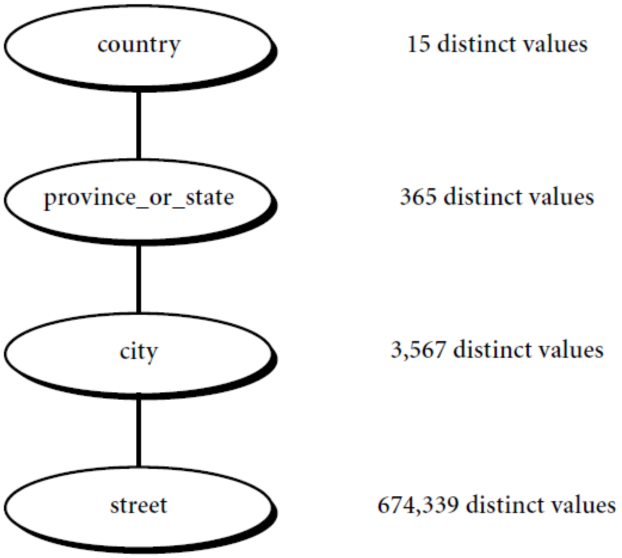
  Ex: location: street, country, state & city

  (It does not specify the hierarchical ordering among the attributes)

a. Sort the attributes in the ascending order based on the roof distinct values in each attributes.

   country(15),State(365),city(3567),street(6,74,339)

b. Generate concept hierarchy from the top down according to sorted order, with the 1st attribute at the top level and last attribute at the bottom level.



4) <u>Specification of only a partial set of attributes</u>

So, A user may have include only a small subset of the relevant attribute in the hirarchy specification.

Ex: location → {Street, City} →

To handle partially specified hierarchies , it is important to embed data sematics in the database schema so that attributes with tight sematic connections can be pinned together.

In their way the specifications of an attribute may trigger a whole group of semantically tightly linked attributes to be " drags in" to form a complete hierarchy users have the option to override their feature as necessary.

Ex: Location: {number, street, city, state & country}

If a user were to specify only the attribute "city" '5' semantically relate attributes to form a hierarchy and he may choose to drop any attribute like number and street.