Lecture Slide - 2

Data Preprocessing

1

Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

2

Why Data Preprocessing?

- Data in the real world is dirty
  - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., Telephone no, Mother's name, etc.
  - Noisy: containing errors or outliers
    - e.g., Salary="-10"
  - Inconsistent: containing discrepancies in codes or names
    - e.g., Age="42" Birthday="03/07/1997"
    - e.g., Was rating "1,2,3", now rating "A, B, C"
    - e.g., discrepancy between duplicate records

3

Why Is Data Dirty?

- Incomplete data may come from
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify linked data)

4

Why Is Data Preprocessing Important?

- No quality data, no quality mining results! (GIGO)
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
  - Data warehouse needs consistent integration of quality data
- Data extraction, Cleaning, and Transformation comprise the majority of the work of building a data warehouse

5

Multi-Dimensional Measure of Data Quality

- A well-accepted multidimensional view:
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Value added
  - Interpretability
  - Accessibility

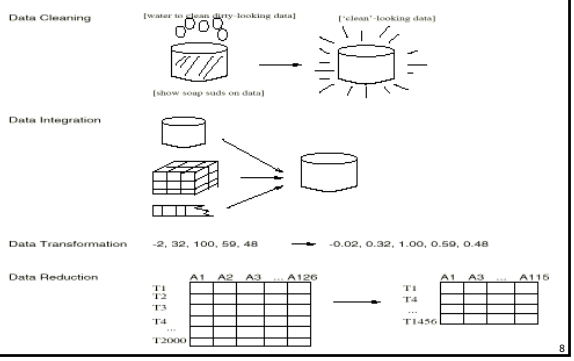
6

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 data sources (databases, flat files, etc.) into a coherent data store (data warehouse)
- Data transformation: Transformation of data from one form to another for accuracy and efficiency purpose (Normalization, Aggregation, etc.)
- Data reduction: Obtains reduced representation in volume but produces the same or similar analytical results (aggregation, duplicate elimination, clustering, attribute subset selection, etc.)

7

Visualization of Data Preprocessing Tasks



8

Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

9

Descriptive Data Summarization

- A technique used to identify the typical properties (central tendency & dispersion) of data and highlight which value should be treated as noise or outliers.
- Measure of Central tendency
  - Mean, Median, Mode, and Midrange
- Measure of dispersion
  - Quartiles, Inter-Quartile Range (IQR), Variance

10

Central Tendency Measure: Mean

- An algebraic measure
- Arithmetic Mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Weighted Arithmetic Mean $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
- Trimmed Arithmetic Mean
  - chopping extreme values (n%) $\bar{x} = \frac{\sum_{i=m}^{n-m} w_i x_i}{\sum_{i=m}^{n-m} w_i}$
- Sensitive to outliers

11

Central Tendency Measure: Median

- A holistic measure
- Middle value if odd number of values, or average of the middle two values, otherwise
- Estimated by interpolation (for *grouped data*):

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

L_1=lower boundary of the median interval
n=Number of values in the entire data set
$(\Sigma f)_l$=sum of frequencies of all the intervals lower than median interval
$freq_{median}$=frequency of the median interval
c=width of the median interval

12

Central Tendency Measure: Mode

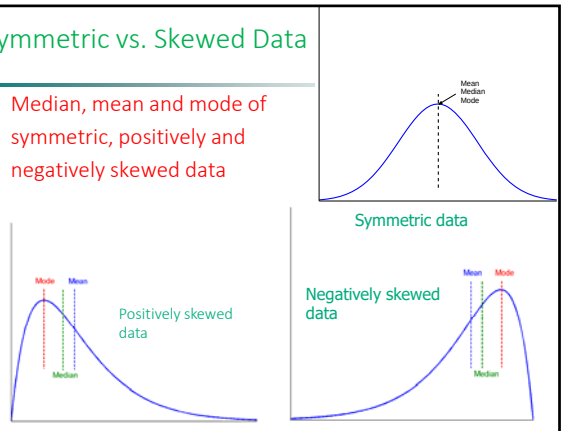
- A value that occurs most frequently in the data set
- Unimodal vs Multimodal (Bimodal, Trimodal, etc.)
- No mode: If each data occurs only once
- For unimodal moderately skewed (asymmetrical) data set

$$mean - mode = 3 \times (mean - median)$$

13

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data



Central Tendency Measure: Midrange

- Average of largest and smallest values in the data set
- Can be calculated using SQL aggregate functions max() and min()

15

Measuring Dispersion of Data

- The degree to which numerical data tend to spread is called dispersion or variance of data
- **Common measures are:**
  - Range
  - Interquartile range (IQR)
  - Standard deviation & Variance
  - The five-number summary (based on quartiles)
    - Can be used to draw boxplots (used for outlier analysis)

16

Dispersion Measure: Range, Quartiles & IQR

- Difference between largest and smallest data values
- The k^{th} percentile of a set of data in numerical order is the value x_i having the property that k% of the data entries lie at or below x_i
- The median is the 50^{th} percentile
- Percentiles other than median are quartiles
  - First quartile (Q_1) is the 25^{th} percentile
  - Third quartile (Q_3) is the 75^{th} percentile
- IQR = $Q_3 - Q_1$
  - Outlier: Usually, a value higher/lower than 1.5 x IQR

17

Dispersion Measure: Variance & SD

- Variance (σ^2): (algebraic measure)
- The variance is the average squared distance of each point from the mean

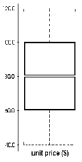
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{...} \quad \sigma^2 = \frac{1}{n}\left(\sum_{i=1}^{n}(x_i)^2\right) - \left(\bar{x}\right)^2$$

- *The variance is thus the difference between the average of the squared magnitude of the data points and the squared magnitude of the mean.*
- Standard Deviation (σ): (algebraic measure)
  - square root of variance σ^2

18

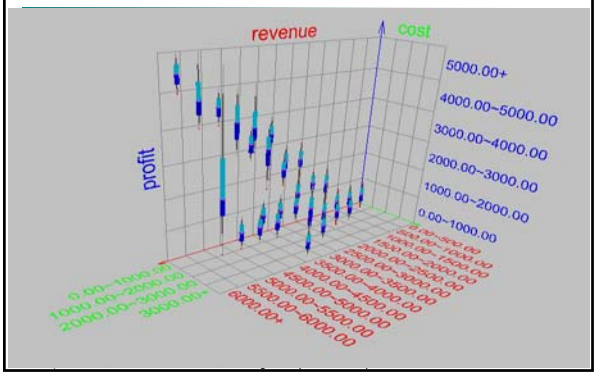
Boxplot Analysis

- Five-number summary of a distribution:
  Minimum, Q1, M, Q3, Maximum
- Boxplot
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extend to Minimum and Maximum



19

Visualization of Data Dispersion: Boxplot Analysis

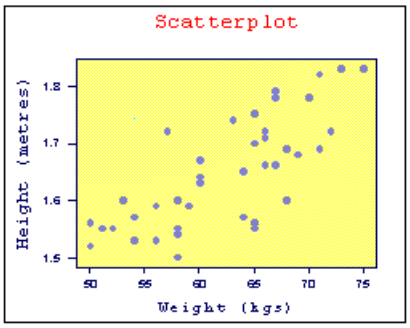


Scatter plot

- Provides a first look at bivariate data (two variables) to see clusters of points, outliers, etc.
- Gives a good visual picture of the relationship between the two variables
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane
- Points are plotted but not joined
- The resulting pattern indicates the type and strength of the relationship between the two variables.

21

Scatter plot (cont…)



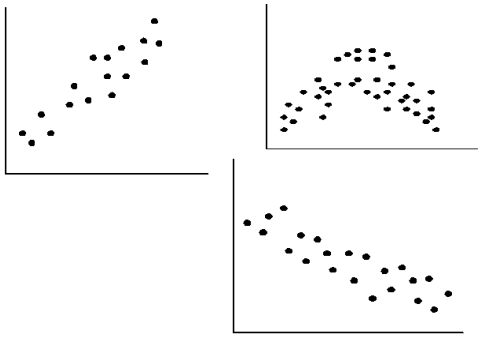
22

Scatter plot (cont…)

- The more the points tend to cluster around a straight line, the stronger the linear relationship between the two variables (the higher the correlation).
- Line runs from lower left to upper right → positive relationship (direct)
- Line runs from upper left to lower right → negative relationship (inverse).
- Random scatter of points → no relationship (very low or zero correlation).
- Points clustering around a curve → non-linear relationship (the correlation coefficient will not be a good measure of the strength)

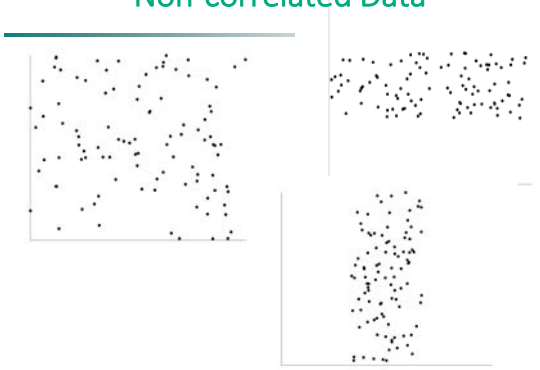
23

Positively and Negatively Correlated and Non-linear Data



24

Non-correlated Data



25

Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

26

Data Cleaning

- Importance:
  - "Data cleaning is one of the three biggest problems in data warehousing"—Ralph Kimball
  - "Data cleaning is the number one problem in data warehousing"—DCI survey
- Data cleaning tasks:
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

27

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
- Missing data may need to be inferred

28

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage 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
  - The most probable value: inference-based such as Bayesian formula or decision tree

29

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 requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

30

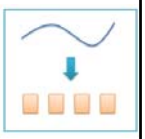
How to Handle Noisy Data?

- Binning
  - First sort data and partition into 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)

31

Binning (aka Discretization)

- A process of transforming numerical variables into categorical counterparts.
  - Example: Bin values for Age into categories such as 20-39, 40-59, and 60-79.
  - Numerical variables are usually discretized in the modeling methods based on frequency tables (e.g., decision trees)



- Binning may be supervised or unsupervised
- Advantages:
  - Binning may improve accuracy of the predictive models by reducing the noise or non-linearity.
  - Binning allows easy identification of outliers, invalid and missing values of numerical variables.

32

Unsupervised Binning

- Transform numerical variables into categorical counterparts without using the target (class) information.
- **Categories:** Equal width and Equal frequency
- Equal Width Binning:
  - The algorithm divides data into k intervals of equal size.
  - The width of intervals is: w = (max-min)/k
  - The interval boundaries are: min+w, min+2w, ... , min+(k-1)w

Example: Data: 0, 4, 12, 16, 16, 18, 24, 26, 28

- Bin-1: 0, 4 [-, 10]
- Bin-2: 12, 16, 16, 18 [10, 20]
- Bin-3: 24, 26, 28 [20, +]

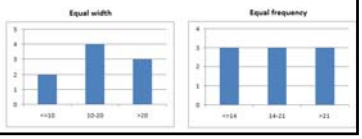
33

Unsupervised Binning (cont…)

- Equal Frequency (aka Equal Depth) Binning:
  - The algorithm divides data into k groups where each group contains approximately same number of values.
  - For both methods, the best way of determining k is by looking at the histogram and try different intervals or groups.

Example: Data: 0, 4, 12, 16, 16, 18, 24, 26, 28

- Bin-1: 0, 4, 12 [-, 14]
- Bin-2: 16, 16, 18 [14, 21]
- Bin-3: 24, 26, 28 [21, +]



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:

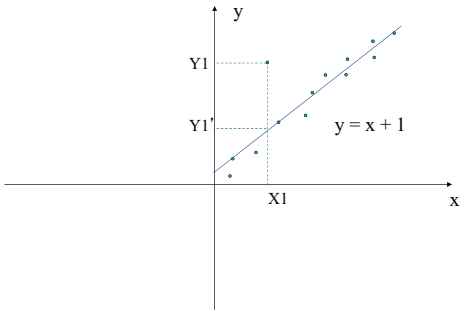
- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

* Smoothing by bin boundaries:

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

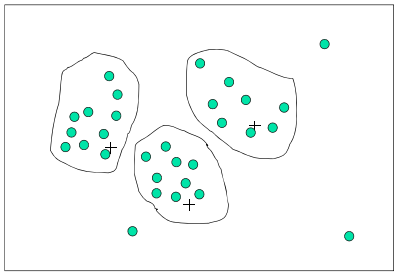
35

Regression



36

Cluster Analysis



37

Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

38

Data Integration

- A process to combine data from multiple sources into a coherent store
- Schema integration: e.g., Scholarship ≡ Fellowship
  - Integrate metadata from different sources
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts:
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

39

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, age, etc.
- Redundant attributes may be able to be detected by correlation analysis
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

40

Correlation Analysis (Numerical Data)

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

$$r_{A,B} = \frac{\sum (a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B}$$

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.

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

41

Correlation Analysis (Categorical Data)

- X^2 (chi-square) test

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad E_{ij} = \frac{count(A = ai) \times count(B = bj)}{N}$$

- 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

42

X^2 (chi-square) test (cont…)

- X^2 tests the hypothesis that A and B are independent
- The test is based on a significance level, with (r-1)×(c-1) degrees of freedom
- If hypothesis can be rejected then A and B are said to be statistically related or associated
- Example:
  - Suppose that a group of 1,500 people was surveyed. The gender of each person was noted. Each person was polled as to whether their preferred type of reading material was fiction or nonfiction. Thus, we have two attributes, gender and preferred reading.

  Are gender and preferred Reading correlated?

43

Example: Chi-Square Calculation

	Male	Female	Sum (row)
Fiction	250 (90)	200 (360)	450
Non-fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

2 x 2 Contingency table

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- Degrees of freedom = (2-1)×(2-1)=1
- For 1 degree freedom, the X^2 value needed to reject the hypothesis at 0.001 significance level is 10.828

44

Upper critical values of chi-square distribution with ν degrees of freedom

	Probability of exceeding the critical value				
ν	0.10	0.05	0.025	0.01	0.001
1	2.706	3.841	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266
4	7.779	9.488	11.143	13.277	18.467
5	9.236	11.070	12.833	15.086	20.515
6	10.645	12.592	14.449	16.812	22.458
7	12.017	14.067	16.013	18.475	24.322
8	13.362	15.507	17.535	20.090	26.125
9	14.684	16.919	19.023	21.666	27.877
10	15.987	18.307	20.483	23.209	29.588
11	17.275	19.675	21.920	24.725	31.264
12	18.549	21.026	23.337	26.217	32.910
13	19.812	22.362	24.736	27.688	34.528
14	21.064	23.685	26.119	29.141	36.123
15	22.307	24.996	27.488	30.578	37.697
16	23.542	26.296	28.845	32.000	39.252
17	24.769	27.587	30.191	33.409	40.790
18	25.989	28.869	31.526	34.805	42.312
19	27.204	30.144	32.852	36.191	43.820
20	28.412	31.410	34.170	37.566	45.315
21	29.615	32.671	35.479	38.932	46.797
22	30.813	33.924	36.781	40.289	48.268
23	32.007	35.172	38.076	41.638	49.728
24	33.196	36.415	39.364	42.980	51.179
25	34.382	37.652	40.646	44.314	52.620

45

Data Transformation

- Smoothing: Removing noise from data
- Aggregation: Summarization, data cube construction, etc.
- Generalization: Concept hierarchy climbing
- Normalization: Scaling to fall within a small, specified range
  - Min-Max normalization
  - Z-Score normalization
  - Normalization by decimal scaling
- Attribute/Feature construction:
  - New attributes constructed from the given ones
  - E.g., *area* attribute from *height* and *width* attributes

46

Data Transformation: Normalization

- **Min-max normalization:** to [new_min_A, new_max_A]
- Performs linear transformation on the original data

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

  - Example: Let income range Rs.12,000 to Rs. 98,000 normalized to [0, 1]. Then Rs. 73,600 is mapped to

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

- Preserves relationship among the original data values
- Encounters an "out-of-bounds" error if the new value falls outside of the original data range

47

Normalization (cont…)

- Z-score normalization (or Zero-mean normalization)
- Let μ: mean and σ: standard deviation

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

  - Example: Let μ = 54,000, σ = 16,000. Then 73,600 is mapped to

$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

- Useful when the actual min and max values of the attribute is unknown or when there are outliers that dominate the min-max normalization.

48

Normalization (cont…)

- Normalization by decimal scaling
  - Normalizes data values by moving the decimal point at extreme left position
  - Number of decimal points moved depends on the max. absolute value
  - A value v is normalized to v' by computing

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

  - Example: A ranges from -986 to 97
    Max absolute value = 986
    So, j=3 (i.e., divide each value by 1000)

49

Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

50

Data Reduction Techniques

- Why data reduction?
  - A database/data warehouse may store terabytes of data
  - Complex data analysis/mining may take a very long time to run on the complete data set
- What is data reduction?
  - A process to obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
- Data reduction techniques:
  - Data cube aggregation
  - Numerosity reduction: e.g., fit data into models
  - Dimensionality reduction: e.g., remove unimportant attributes
  - Discretization and concept hierarchy generation

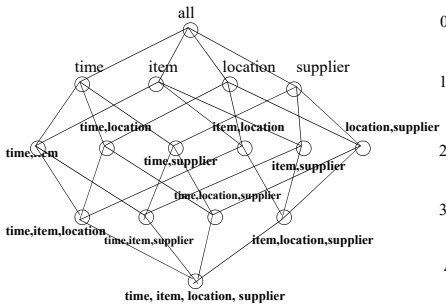
51

Data Cube Aggregation

- An Data Cube is a multidimensional database that is optimized for data warehouse and online analytical processing (OLAP) applications.
- A method to store data in a multidimensional form, generally for reporting purposes.
- In Data Cubes, data (measures) are categorized by dimensions.
- Provides multiple levels of aggregation
- Reference appropriate levels
  - Use the smallest representation which is enough to solve the task
- Queries regarding aggregated information should be answered using data cube, when possible

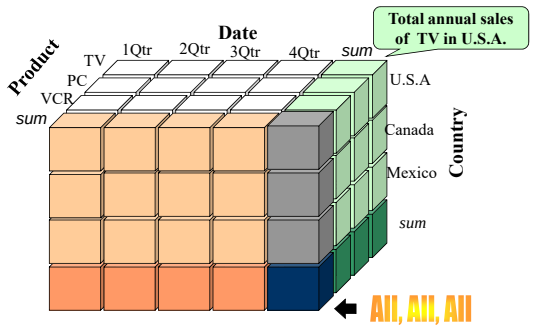
52

Data Cube: A Lattice of Cuboids



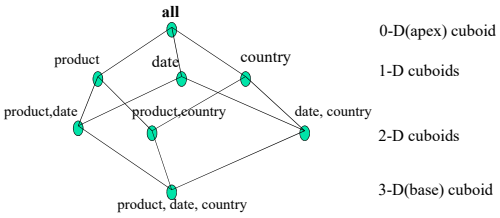
53

A Sample Data Cube



54

Cuboids Corresponding to the Cube



55

Numerosity Reduction

- Reduce data volume by choosing alternative, smaller forms of data representation
- **Parametric methods**
  - 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

56

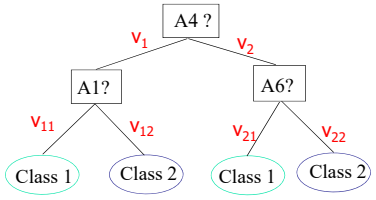
Dimensionality Reduction: Attribute Subset Selection

- Feature selection (i.e., attribute subset selection):
  - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
  - Reduce #of patterns - easier to understand
- Heuristic methods (due to exponential #of choices):
  - Step-wise forward selection
  - Step-wise backward elimination
  - Decision-tree induction

57

Example of Decision Tree Induction

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



Reduced attribute set: {A1, A4, A6}

58

Dimensionality Reduction: Principal Component Analysis (PCA)

- Given *N* data vectors from *n*-dimensions, find $k \le n$ orthogonal vectors (*principal components*) that can best represent the data
- **Steps:**
  - 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
- Used when the number of dimensions is large

Read given tutorial for technical details of PCA

59

Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

60

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
  - Continuous — real numbers, e.g., integer or real numbers
- Discretization:
  - Divide the range of a continuous attribute into intervals
  - Some classification algorithms only accept categorical attributes.
  - Reduce data size by discretization
  - Prepare for further analysis

61

Discretization and Concept Hierarchy

- Discretization
  - 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
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute
- Concept hierarchy formation
  - Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for age) by higher level concepts (such as young, middle-aged, or senior)

62

Discretization and Concept Hierarchy Generation for Numeric Data

- Typical methods: All the methods can be applied recursively
  - Binning (covered above)
    - Top-down split, unsupervised,
  - Clustering analysis (covered above)
    - Either top-down split or bottom-up merge, unsupervised
  - Entropy-based discretization: supervised, top-down split
  - Segmentation by natural partitioning: top-down split, unsupervised
  - Segmentation by X^2 test (ChiMerge): Bottom-up merge, supervised

63

Entropy-Based Discretization

- Given a set of samples S, if S is partitioned into two intervals S_1 and S_2 using boundary T, the information gain after partitioning is

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

- Entropy is calculated based on class distribution of the samples in the set. Given *m* classes, the entropy of S_1 is

$$Entropy(S_1) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

where p_i is the probability of class *i* in S_1

- The boundary that minimizes the entropy function over all possible boundaries is selected as a binary discretization
- The process is recursively applied to partitions obtained until some stopping criterion is met
- Such a boundary may reduce data size and improve classification accuracy

64

Segmentation by Natural Partitioning: 3-4-5 Rule

- Segments numeric data into relatively uniform, "natural" intervals.
- Partitions a given range of data into either 3, 4, or 5 relatively equi-length intervals, recursively and level by level, based on the value range at the most significant digit.

65

3-4-5 Rule for Natural Partitioning

Distinct values of the value range at MSD	Number of intervals Segmented
3, 6, 9	3 equiwidth intervals
7	3 intervals in the grouping of 2-3-2
2, 4, 8	4 equiwidth intervals
1, 5, 10	5 equiwidth intervals

66

3-4-5 Partitioning Process

- Let E_{min}, and E_{max}, be the minimum and maximum data value
- Choose appropriate positive integer numbers e_1 and e_2 such that $L=(E_{max}+ e_2) - (E_{min} - e_1)$ is divisible by the digital position of the most significant digit of L.
- Define universe of discourse $U = [E_{min} – e_1, E_{max} + e_2]$
- Divide U into sub-intervals $u_1, u_2, …, u_m$ using 3-4-5 rules. This represents the top tier of the hierarchy
- Divide each sub-interval, u_i, recursively using 3-4-5 rules to get next tier of the hierarchy.

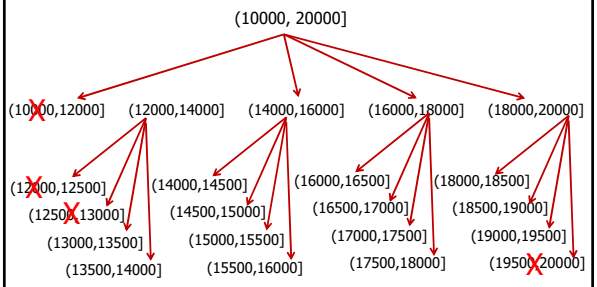
67

Example-1: Discretization using 3-4-5 Rule

- Let E_{min} = 13055 and E_{max} = 19337, then e1=3055 and e2=663, Thus U = [l0000, 20000]
  Now, (20000 - 10000) / 10000 = 1, So, as per the 3-4-5 rule U is partitioned into five equiwidth intervals as
  [10000, 12000], [12000, 14000], [14000, 16000], [16000, 18000], [18000, 20000]
- Remove [10000, 12000] since it does not contain any data value
- Recursively, Divide [12000, 14000] into 4 sub-intervals [12000, 12500], [12500, 13000], [13000, 13500], [13500, 1400]
- Do the same for all sub-intervals

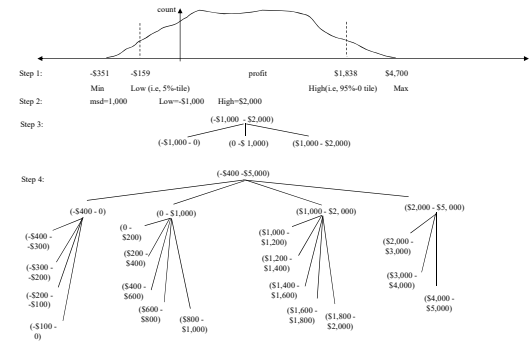
68

Example-1 (cont…)



69

Example-2: Discretization using 3-4-5 Rule



70

ChiMerge Algorithm for Discretization

- A bottom-up supervised Discretization method
- Consists of an initialization step and a bottom-up merging process, where intervals are continuously merged until a termination condition is met.
- Initialized by first sorting the training examples according to their value for the attribute being discretized and then constructing the initial discretization, in which each example is put into its own interval

71

Cont…

- Interval merging process contains two steps, repeated continuously:
  - Compute the Chi-square value for each pair of adjacent intervals
  - Merge (combine) the pair of adjacent intervals with the lowest Chi-square value
- Merging continues until all pairs of intervals have Chi-square values exceeding the parameter Chi-square threshold
  - That is, all adjacent intervals are considered significantly different by the Chi-square independence test.

72

Cont…

- The formula for computing χ^2 value is

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(A_{ij}-E_{ij})^2}{E_{ij}}$$

Where:

$m = 2$ (the 2 intervals being compared)
$k =$ number of classes
$A_{ij} =$ number of examples in ith interval, jth class.
$R_i =$ number of examples in ith interval $= \sum_{j=1}^{k} A_{ij}$
$C_j =$ number of examples in jth class $= \sum_{i=1}^{m} A_{ij}$
$N =$ total number of examples $= \sum_{j=1}^{k} C_j$
$E_{ij} =$ expected frequency of $A_{ij} = \frac{R_i * C_j}{N}$

For algorithmic detail and Example, see the research paper given in the class.

73

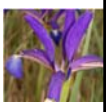
Example: Iris Dataset

- Contains 150 examples belonging to three classes
- 50 examples each of the classes Iris setosa, Iris versicolor, and Iris virginica (species of iris).
- Each example is described using four numeric attributes: petal-length, petal-width, sepal-length, and sepal-width.

74

Description of Iris Dataset

Data Set Characteristics:	Multivariate	Number of Instances:	150	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	4	Date Donated	1988-07-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	143101



Attribute Information:

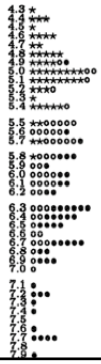
1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica

Original source of data:

http://archive.ics.uci.edu/ml/datasets/Iris

75

Class histogram for sepal-length



'*' → Setosa, o → Versicolor,
• → Virginica

76

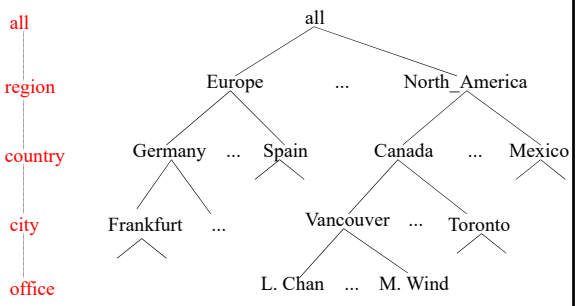
ChiMerge discretizations for sepal-length at the .50 and .90 significance levels (x2 = 1.4 and 4.6)

Int	Class frequency			χ^2
4.3	16	0	0	4.1
4.9	4	1	1	2.4
5.0	25	5	0	8.6
5.5	2	5	0	2.9
5.6	0	5	1	1.7
5.7	2	5	1	1.8
5.8	1	3	3	2.2
5.9	0	12	7	4.8
6.3	0	6	15	4.1
6.6	0	2	0	3.2
6.7	0	5	10	1.5
7.0	0	1	0	3.6
7.1	0	0	12	

Int	Class frequency			χ^2
4.3	45	6	1	30.9
5.5	4	15	2	6.7
5.8	1	15	10	4.9
6.3	0	14	25	5.9
7.1	0	0	12	

77

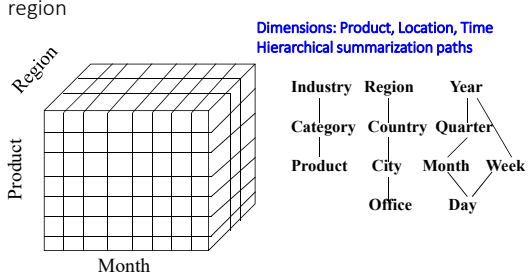
A Concept Hierarchy: Dimension (location)



78

Multidimensional Data

- Sales volume as a function of product, month, and region



79

Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

80

Summary

- Data preparation or preprocessing is a big issue for both data warehousing and data mining
- Descriptive data summarization is need for quality data preprocessing
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- A lot of methods have been developed but data preprocessing still an active area of research

81

Assignment

- Text Book
  - Exercises: 2.2 (Page: 98), 2.9(Page: 99)
- Discretize and generate a concept hierarchy using 3-4-5 rule for the following data values:

  -519, -456, 453, 12, 345, 123, 23, 56, 78, 232, 624, -45, -2, -67, 33, 44, 56, 123, 789, 88, 219
- Using ChiMerge algorithm to discretize all attributes of the iris dataset available on the Web.

82