Data Pre-processing

1

Learning Outcomes

- Data cleaning
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
- Data Integration
- Data Transformation
- Data Reduction and Discretisation

2

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
  - no register history or changes of the data
- Missing data may need to be inferred

3

How to Handle Missing Data?

- Ignore the tuple: not effective when the percentage of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- 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
- Use the attribute mean for all samples belonging to the same class to fill in the missing value: smarter
- Use the most probable value to fill in the missing value: inference-based such as Bayesian formula or decision tree

4

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may 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

5

How to Handle Noisy Data?

- Binning method
  - first sort data and partition into bins (buckets)
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human
- Regression
  - smooth by fitting the data into regression functions

6

Simple Discretisation Methods: Binning

- Equal-width (distance) partitioning
  - It divides the range into *N* intervals of equal size: uniform grid
  - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$
  - The most straightforward
  - But outliers may dominate presentation
  - Skewed data is not handled well
- Equal-depth (frequency) partitioning
  - It divides the range into *N* intervals, each containing approximately the same number of objects
  - Good data scaling
  - Managing categorical attributes can be tricky

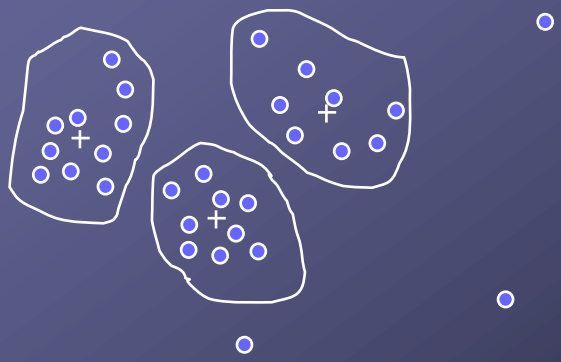
7

Binning Methods for Data Smoothing

- Sorted data for price (in €)
  - 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Partition into (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

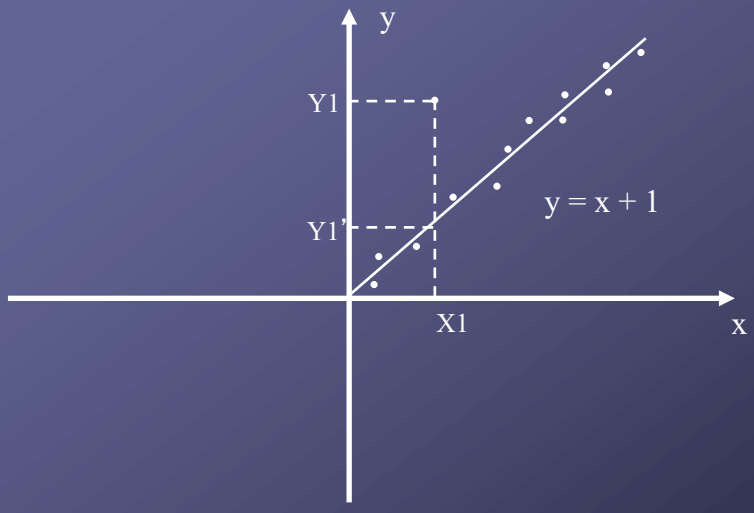
8

Cluster Analysis



9

Regression



10

Data Integration

- Data integration
  - combines data from multiple sources into a coherent store
- Schema integration
  - integrate metadata from different sources
  - Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id ≡ B.cust-#
- 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. imperial units

11

Handling Redundant Data in Data Integration

- Redundant data occur often when integrating multiple datasets
  - The same attribute may have different names in different datasets, (e.g., customer_ID, customer_No)
  - One attribute may be a "derived" attribute in another table, (e.g., annual revenue)
- Redundant data may be detected by correlation-based analysis
  - Correlation-based analysis measures how strongly one attribute implies the other
  - Correlation does not mean causality
- Careful integration of the data from multiple sources
  - May help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

12

Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarisation, data cube construction
- Generalisation: concept hierarchy climbing
- Normalisation: scaled to fall within a small, specified range
  - min-max normalisation
  - z-score normalisation
  - normalisation by decimal scaling
- Attribute/feature construction
  - New attributes constructed from the given ones

13

Data Transformation: Normalisation

- min-max normalisation

$$v' = \frac{v - minA}{max_A - minA}(Nmax_A - NminA) + NminA$$

- z-score normalisation

$$v' = \frac{v - MeanA}{std_A}$$

- normalisation by decimal scaling

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

14

Data Reduction Strategies

- Data Warehouse may store exabytes of data
  - Complex data analysis/mining may take a very long time to run on the complete data set
- Data reduction
  - Obtains a reduced representation of the dataset that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Data reduction strategies
  - Data cube aggregation
  - Dimensionality reduction
  - Numerosity reduction
  - Discretisation and concept hierarchy generation

15

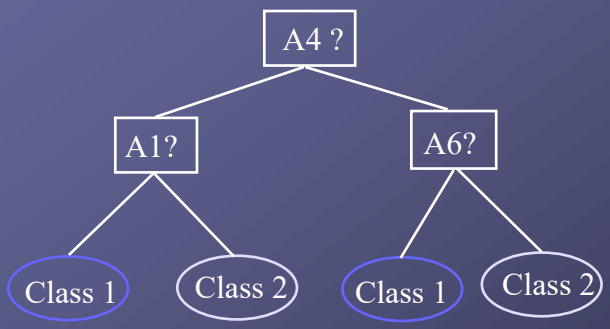
Dimensionality Reduction

- 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 the number of patterns ==> easier to understand
- Heuristic methods (due to exponential # of choices)
  - step-wise forward selection
  - step-wise backward elimination
  - combining forward selection and backward elimination
  - decision-tree induction

17

Example

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



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

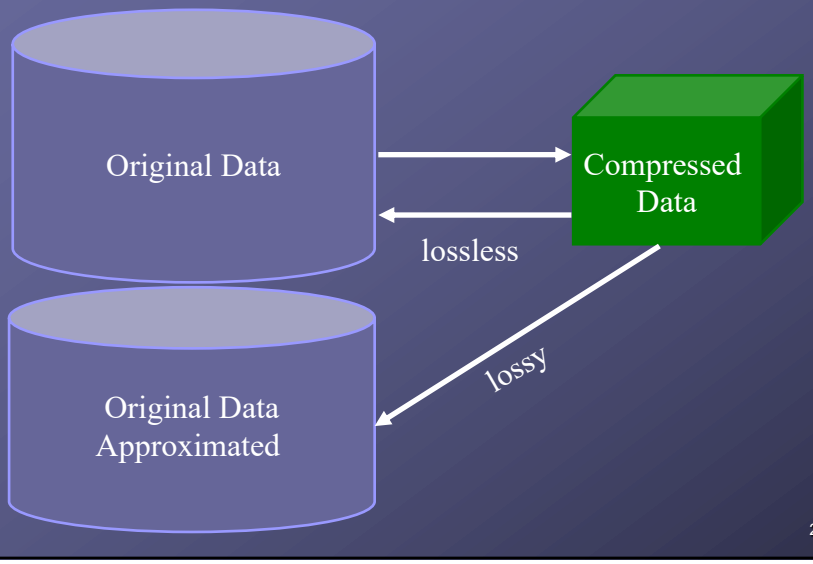
18

Data Compression

- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless
  - But only limited manipulation is possible without expansion
- Audio/video compression
  - Typically lossy compression, with progressive refinement
  - Sometimes small fragments of signal can be reconstructed without reconstructing the whole
- Time sequence is not audio
  - Typically short and vary slowly with time

19

Data Compression



20

Wavelet Transforms

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

21

Principal Component Analysis (PCA)

- Given M data vectors from n-dimensions, find $k \leq n$ orthogonal vectors that can be best used to represent data
  - The original data set is reduced to one consisting of N data vectors on x principal components (reduced dimensions)
- Each data vector is a linear combination of the *x* principal component vectors
- Works for numerical data only
- Used when the number of dimensions is large

22

Numerosity Reduction

- Parametric methods
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
  - Log-linear models: obtain value at a point in m-D space as the product on appropriate marginal subspaces
- Non-parametric methods
  - Do not assume models
  - Major families: histograms, clustering, sampling

23

Regression and Log-Linear Models

- Linear regression
  - Data are modeled to fit a straight line
  - Often uses the least-square method to fit the line
- Multiple regression
  - allows a response variable Y to be modeled as a linear function of multidimensional feature vector
- Log-linear model
  - approximates discrete multidimensional probability distributions

24

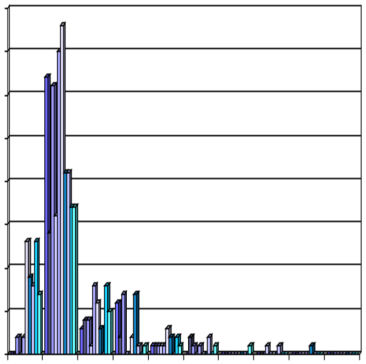
Regression Analysis and Log-Linear Models

- Linear regression: $Y = a X + b$
  - Two parameters , a and b specify the line and are to be estimated by using the data at hand
  - using the least squares criterion to the known values of $Y_1, Y_2, \ldots, X_1, X_2, \ldots$.
- Multiple regression: $Y = a_0 + a_1 X_1 + a_2 X_2$.
  - Many nonlinear functions can be transformed into the above
- Log-linear models:
  - The multi-way table of joint probabilities is approximated by a product of lower-order tables.
  - Probability: $p(a, b, c, d)$

25

Histograms

- A popular data reduction technique
- Divide data into buckets and store average (sum) for each bucket
- Can be constructed optimally in one dimension using dynamic programming
- Related to quantization problems



26

Clustering

- Partition data set into clusters, and one can store cluster representation only
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms

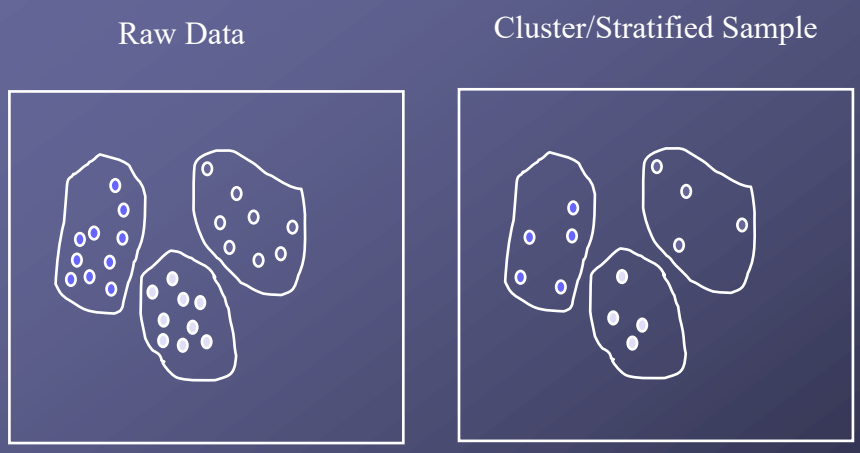
27

Sampling

- Complexity
  - Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
- Develop adaptive sampling methods
  - Stratified sampling:
    - Approximate the percentage of each class (or subpopulation of interest) in the overall database
    - Used in conjunction with skewed data

28

Sampling



30

Hierarchical Reduction

- Use multi-resolution structure with different degrees of reduction
- Hierarchical clustering is often performed but tends to define partitions of data sets rather than “clusters”
- Parametric methods are usually not amenable to hierarchical representation
- Hierarchical aggregation
  - An index tree hierarchically divides a data set into partitions by value range of some attributes
  - Each partition can be considered as a bucket
  - Thus an index tree with aggregates stored at each node is a hierarchical histogram

31

Discretisation

- Three types of attributes
  - Nominal — values from an unordered set
  - Ordinal — values from an ordered set
  - Continuous — real numbers
- Discretisation
  - divide the range of a continuous attribute into intervals
  - Some classification algorithms only accept categorical attributes
  - Reduce data size by discretisation
  - Prepare for further analysis

32

Discretisation and Concept hierarchy

- Discretisation
  - 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
- Concept hierarchies
  - 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).

33

Discretisation and concept hierarchy generation for numeric data

- Binning
- Histogram analysis
- Clustering analysis
- Entropy-based discretisation
- Segmentation by natural partitioning

34

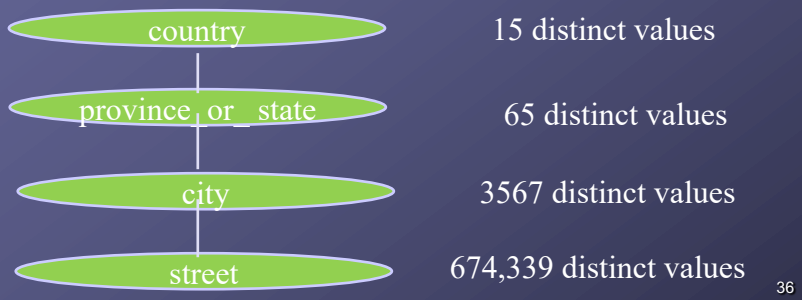
Concept hierarchy generation for categorical data

- 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

35

Specification of a set of attributes

- Concept hierarchy can be automatically generated based on the number of distinct values per attribute in the given attribute set. The attribute with the most distinct values is placed at the lowest level of the hierarchy.



36