

## Data Preprocessing

---

- Lecture 3: Overview of data preprocessing/ Descriptive data summarization
- Lecture 4: Data cleaning / Data integration/transformation
- **Lecture 5: Data reduction /Discretization and concept hierarchy generation and summary**

Lecture 5

## Data Reduction Strategies

---

- 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
- Data reduction
  - 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**

1

Lecture 5

2

## Data Reduction Strategies

---

- **Data reduction strategies**
  - Data cube aggregation
  - Dimensionality reduction — e.g., remove unimportant attributes
  - Data Compression
  - Numerosity reduction — e.g., fit data into models
  - Discretization and concept hierarchy generation

Lecture 5

3

## Data Cube Aggregation

---

- The lowest level of a data cube (base cuboid)
  - The aggregated data for an **individual entity of interest**
  - E.g., a customer in a phone calling data warehouse
- Multiple levels of aggregation in data cubes
  - Further reduce the size of data to deal with
- 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

Lecture 5

4

## Attribute Subset Selection

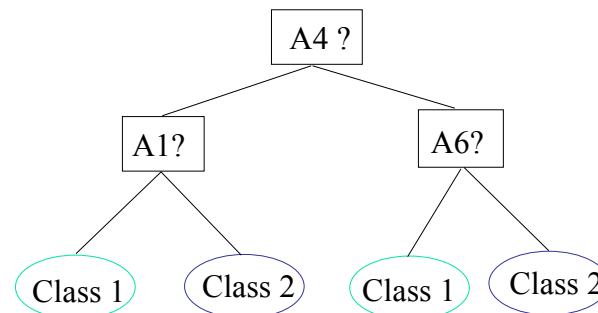
- Feature selection (i.e., attribute subset selection):
  - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
  - reduce # of patterns in the 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

Lecture 5

5

## Example of Decision Tree Induction

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



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

Lecture 5

6

## Heuristic Feature Selection Methods

- There are  $2^d$  possible sub-features of  $d$  features
- Several heuristic feature selection methods:
  - Best single features under the feature independence assumption: choose by significance tests
  - Best step-wise feature selection:
    - The best single-feature is picked first
    - Then next best feature condition to the first, ...
  - Step-wise feature elimination:
    - Repeatedly eliminate the worst feature
  - Best combined feature selection and elimination
  - Optimal branch and bound:
    - Use feature elimination and backtracking

Lecture 5

7

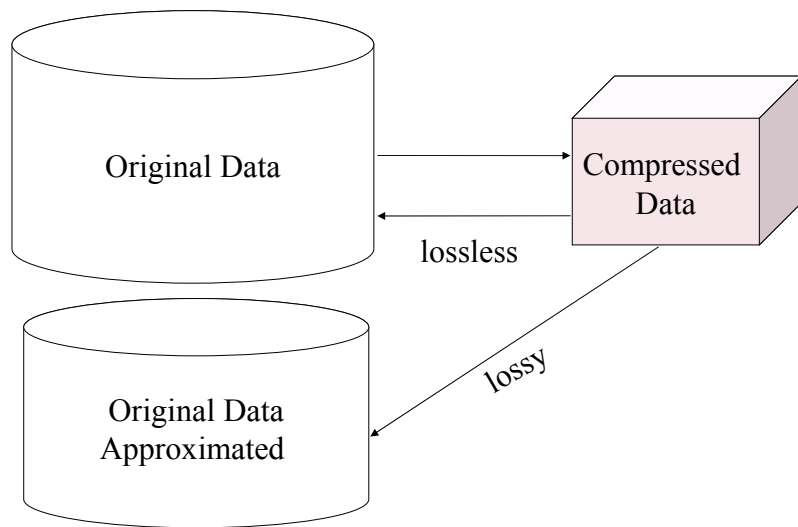
## Data Compression

- 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

Lecture 5

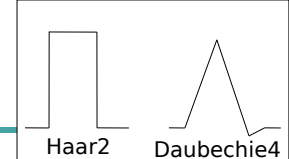
8

# Data Compression



Lecture 5

# Dimensionality Reduction: Wavelet Transformation



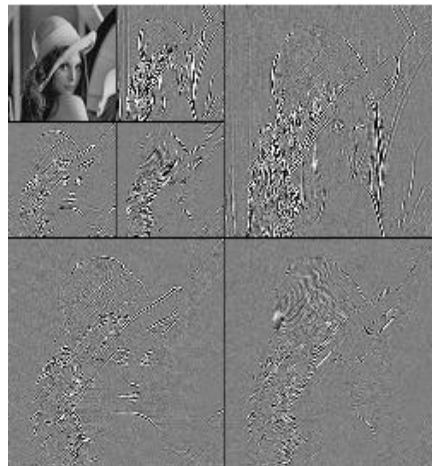
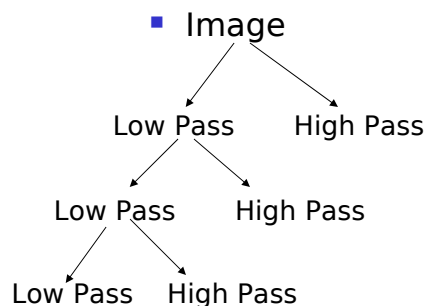
- Discrete wavelet transform (DWT): linear signal processing, multi-resolutional analysis
- Compressed approximation: store only a small fraction of the strongest of the wavelet coefficients
- Similar to discrete Fourier transform (DFT), but better lossy compression, localized in space
- Method:
  -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

9

Lecture 5

10

# DWT for Image Compression



Lecture 5

11

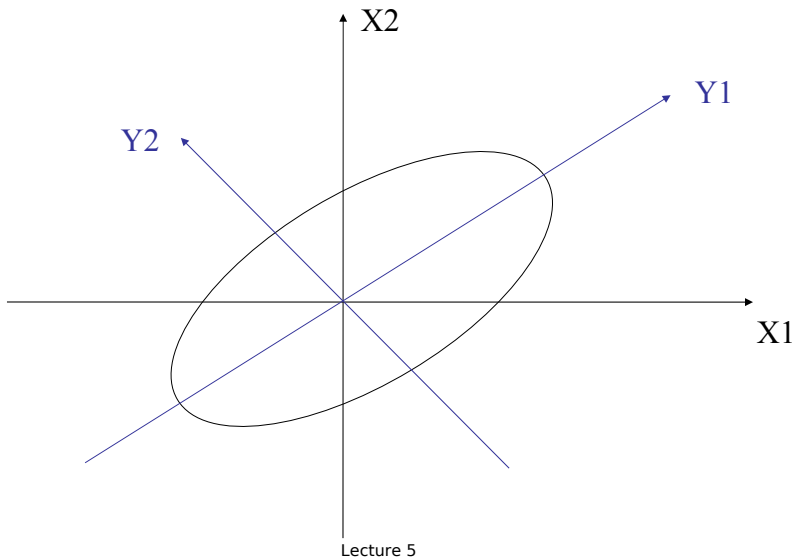
# Dimensionality Reduction: Principal Component Analysis (PCA)

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

Lecture 5

12

# Principal Component Analysis



13

## Data Reduction Method (1): 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

Lecture 5

15

# 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)
  - Example: 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

Lecture 5

14

## Regress Analysis and Log-Linear Models

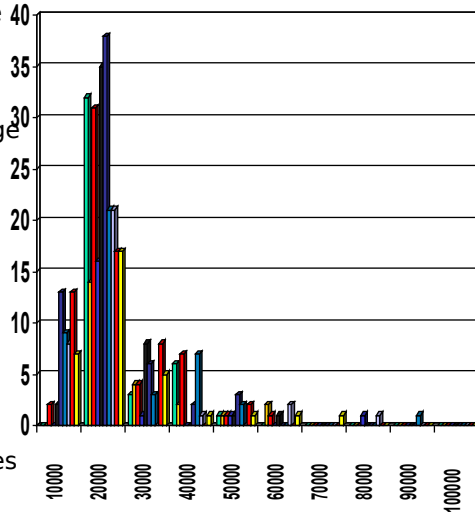
- Linear regression:  $Y = w X + b$ 
  - Two regression coefficients,  $w$  and  $b$ , specify the line and are to be estimated by using the data at hand
  - Using the least squares criterion to the known values of  $Y_1, Y_2, \dots, X_1, X_2, \dots$
- Multiple regression:  $Y = b_0 + b_1 X_1 + b_2 X_2$ .
  - 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:  $p(a, b, c, d) = \alpha \beta \gamma \delta$

Lecture 5

16

## Data Reduction Method (2): Histograms

- Divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
  - Equal-width:** equal bucket range
  - Equal-frequency** (or equal-depth)
  - V-optimal:** with the least *histogram variance* (weighted sum of the original values that each bucket represents)
  - MaxDiff:** set bucket boundary between each pair for pairs have the  $\beta-1$  largest differences



Lecture 5

17

## Data Reduction Method (3): Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only
- Can be very effective if data is clustered but not if data is "smeared"
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms
- Cluster analysis will be studied in depth in Chapter 7

Lecture 5

18

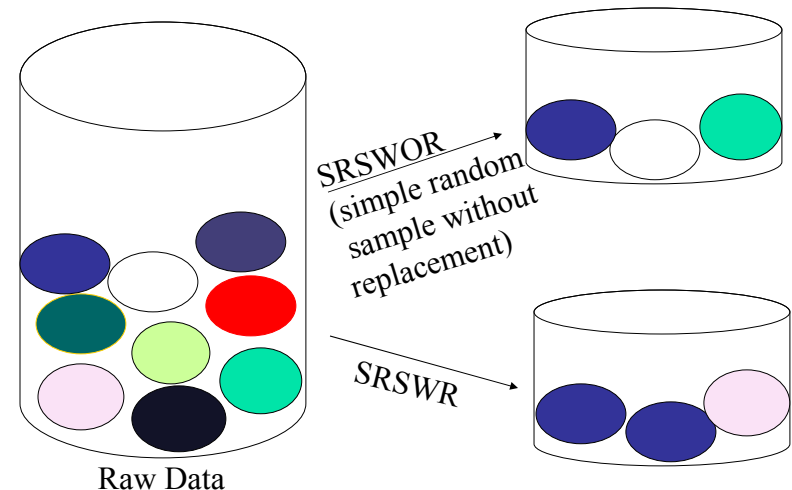
## Data Reduction Method (4): Sampling

- Sampling: obtaining a small sample  $s$  to represent the whole data set  $N$
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- 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
- Note: Sampling may not reduce database I/Os (page at a time)

Lecture 5

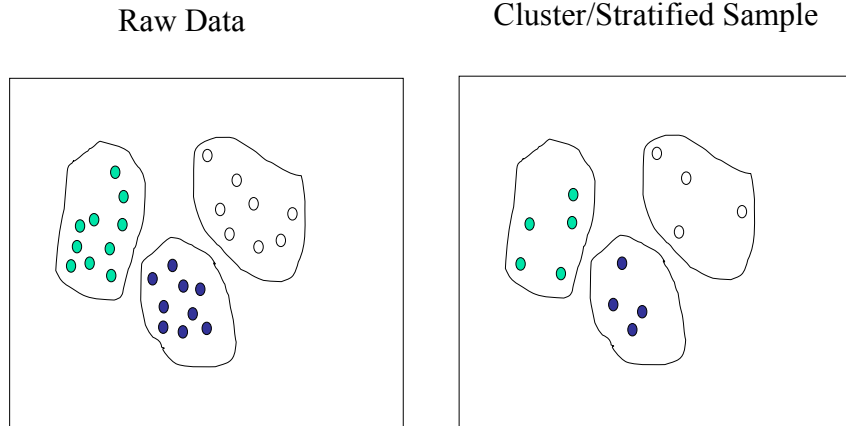
19

## Sampling: with or without Replacement



Lecture 5

20



Lecture 5

21

## 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

Lecture 5

22

# 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)

Lecture 5

23

# Discretization and Concept Hierarchy Generation for Numeric Data

## Typical methods: All the methods can be applied recursively

### Binning

### Top-down split, unsupervised

### Histogram analysis

#### Top-down split, unsupervised

### Clustering analysis

- Either top-down split or bottom-up merge, unsupervised

### Entropy-based discretization: supervised, top-down split

### Interval merging by $\chi^2$ Analysis: unsupervised, bottom-up merge

### Segmentation by natural partitioning: top-down split, unsupervised

Lecture 5

24

# 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
 
$$I(S, T) = \frac{|S_1|}{|S|} \text{Entropy}(S_1) + \frac{|S_2|}{|S|} \text{Entropy}(S_2)$$
- Entropy is calculated based on class distribution of the samples in the set. Given  $m$  classes, the entropy of  $S_1$  is
 
$$\text{Entropy}(S_1) = - \sum_{i=1}^m p_i \log_2(p_i)$$
 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

Lecture 5

25

# Interval Merge by $\chi^2$ Analysis

- Merging-based (bottom-up) vs. splitting-based methods
- Merge: Find the best neighboring intervals and merge them to form larger intervals recursively
- ChiMerge [Kerber AAAI 1992, See also Liu et al. DMKD 2002]
  - Initially, each distinct value of a numerical attr.  $A$  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, since low  $\chi^2$  values for a pair indicate similar class distributions
  - This merge process proceeds recursively until a predefined stopping criterion is met (such as significance level, max-interval, max inconsistency, etc.)

Lecture 5

26

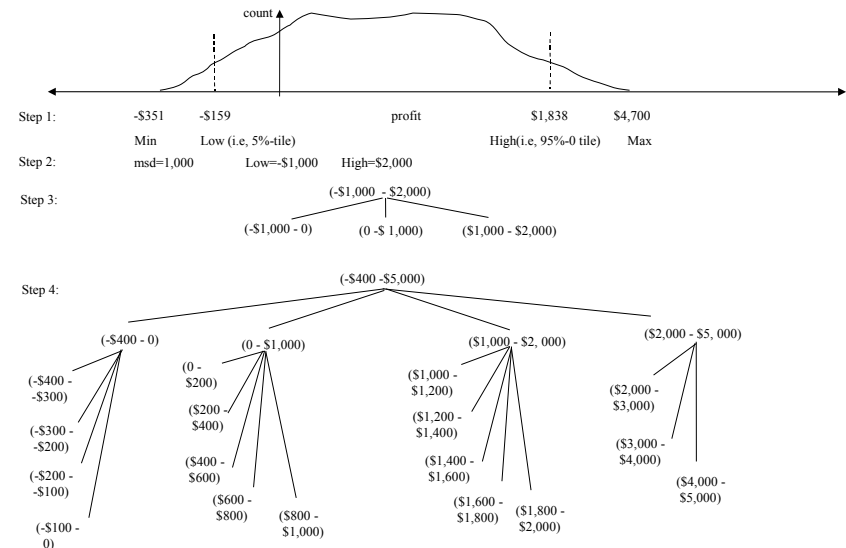
# Segmentation by Natural Partitioning

- A simply 3-4-5 rule can be used to segment numeric data into relatively uniform, "natural" intervals.
  - 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

Lecture 5

27

# Example of 3-4-5 Rule



Lecture 5

28

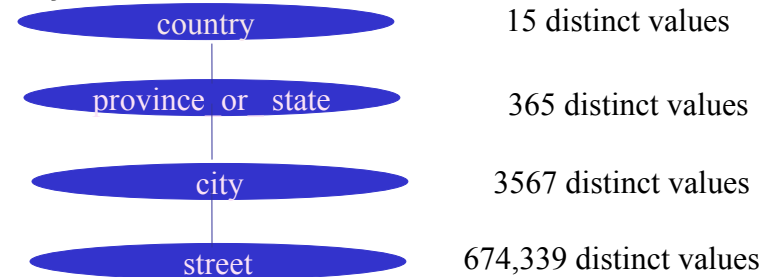
# Concept Hierarchy Generation for Categorical Data

- Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
  - street < city < state < country
- Specification of a hierarchy for a set of values by explicit data grouping
  - {Urbana, Champaign, Chicago} < Illinois
- Specification of only a partial set of attributes
  - E.g., only street < city, not others
- Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values
  - E.g., for a set of attributes: {street, city, state, country}

Lecture 5

# Automatic Concept Hierarchy Generation

- Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy
  - Exceptions, e.g., weekday, month, quarter, year



Lecture 5

## 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 a methods have been developed but data preprocessing still an active area of research

Lecture 5

## References

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. *Communications of ACM*, 42:73-78, 1999
- T. Dasu and T. Johnson. *Exploratory Data Mining and Data Cleaning*. John Wiley & Sons, 2003
- T. Dasu, T. Johnson, S. Muthukrishnan, V. Shkapenyuk. [Mining Database Structure: Or, How to Build a Data Quality Browser](#). SIGMOD'02.
- H.V. Jagadish et al., *Special Issue on Data Reduction Techniques*. *Bulletin of the Technical Committee on Data Engineering*, 20(4), December 1997
- D. Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann, 1999
- E. Rahm and H. H. Do. Data Cleaning: Problems and Current Approaches. *IEEE Bulletin of the Technical Committee on Data Engineering*. Vol.23, No.4
- V. Raman and J. Hellerstein. *Potters Wheel: An Interactive Framework for Data Cleaning and Transformation*, VLDB'2001
- T. Redman. *Data Quality: Management and Technology*. Bantam Books, 1992
- Y. Wand and R. Wang. Anchoring data quality dimensions ontological foundations. *Communications of ACM*, 39:86-95, 1996
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. *IEEE Trans. Knowledge and Data Engineering*, 7:623-640, 1995

29

30

31

Lecture 5

32