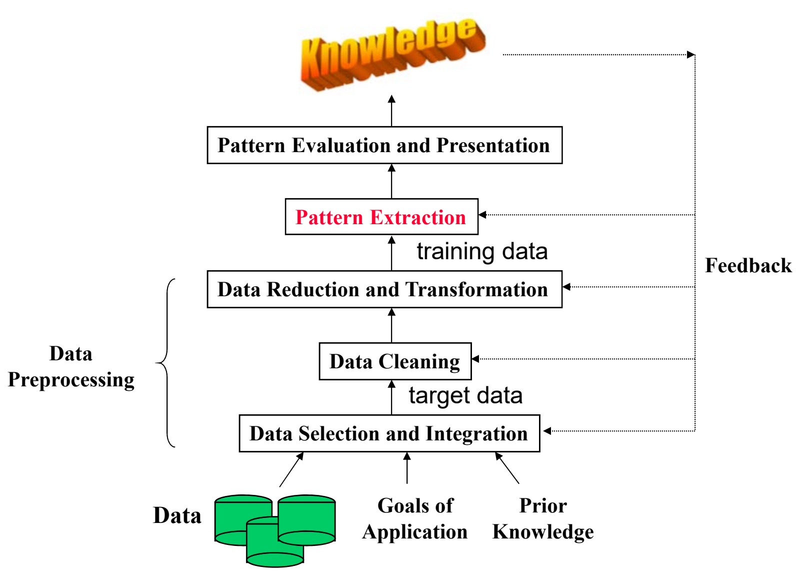
Data Mining
(EECS 6412)

Data Preprocessing

Aijun An
Department of Electrical Engineering and Computer Science
York University

Process of Data Mining and KDD



Outline

- Why preprocess the data?
- Data integration
- Data cleaning
- Data transformation
- Data reduction
- Discretization

3

Why Data Preprocessing?

- Heterogeneous data – data integration
  - From various departments, in various forms
- Dirty data – data cleaning
  - Incomplete data: missing attribute values
    - e.g., occupation="”
  - Noisy data: containing errors
    - e.g., Salary="-10”
  - Discrepancies in codes or names
    - e.g., US=USA
- Data not in the right format – data transformation
  - Normalization, discretization, etc.
- A huge amount of data – data reduction
  - Speed up mining

No quality data, no quality mining results!

4

Major Tasks in Data Preprocessing

- Data integration
  - Integration of multiple databases or files
- Data cleaning
  - Fill in missing values, identify outliers and smooth out noisy data, and resolve discrepancies
- Data transformation
  - Feed right data to the mining algorithm
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
  - Part of data reduction and data transformation but with particular importance, transform numerical data into symbolic (discrete) data

5

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., hotel price in different currencies, metric vs. British units
    - e.g., Age="42" Birthday="03/07/1997"
    - e.g., Was rating "1,2,3", now rating "A, B, C"

6

Data Cleaning

- Why is data dirty?
  - Incomplete data come from
    - human/hardware/software problems (e.g. equipment malfunction)
    - different consideration between the time when the data was collected and when it is analyzed.
      - certain data may not be considered important at the time of entry
  - Noisy data come from the process of
    - data collection
    - data entry
    - data transmission
- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data

7

How to Handle Missing Values?

- Fill in the missing value manually: tedious + infeasible?
- Ignore the tuple containing missing values:

Cust-id	Age	Gender	Income	Credit
1	36	M	$54K	good
2	24	M	$20K	bad
~~3~~	~~37~~	~~M~~	~~$50K~~	~~?~~
4	23	F	$30K	good
5	55	F	$25K	good
~~6~~	~~35~~	~~?~~	~~$16K~~	~~bad~~
7	33	F	$10K	bad

  - usually done when class label is missing (assuming the task is classification)
  - not effective when missing values in attributes spread in many different tuples.
- Fill it in with a value “unknown”
  - patterns containing “unknown” is ugly

8

How to Handle Missing Values? *(Contd.)*

- Global estimation
  - the attribute mean/median for numeric attributes
  - the most probable value for symbolic (i.e. categorical) attributes
- Local estimation: smarter
  - the attribute mean/median for all the tuples belonging to the same class (for numeric attributes)
  - the most probable value within the same class (for symbolic attributes)

Cust-id	Age	Gender	Income	Credit
1	36	F	$55K	good
2	24	?	$20K	bad
3	37	F	$50K	good
4	23	F	$30K	good
5	55	F	$25K	good
6	35	M	?	bad
7	33	M	$10K	bad

- Use inference-based prediction techniques, such as
  - Nearest-neighbor estimator, decision tree, regression, neural network, etc.
  - good method with overhead

9

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

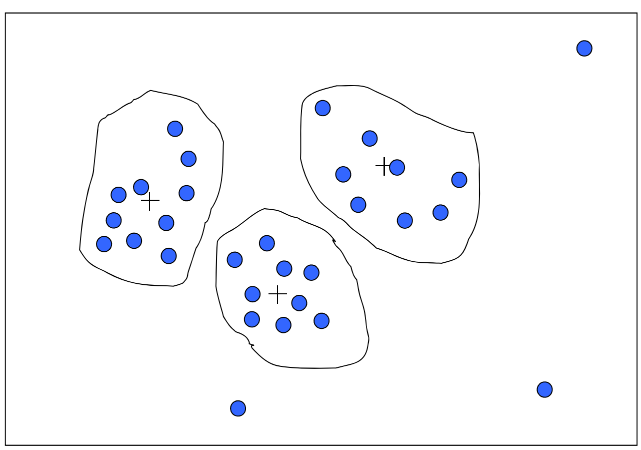
10

How to Handle Noisy Data?

- Clustering
  - detect and remove outliers (An outlier is a value that does not follow the general pattern of the rest)
- Regression
  - smooth by fitting the data into regression functions
- Binning method:
  - first sort data and partition into bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Moving average
  - Use the arithmetic mean of neighborhood examples
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

11

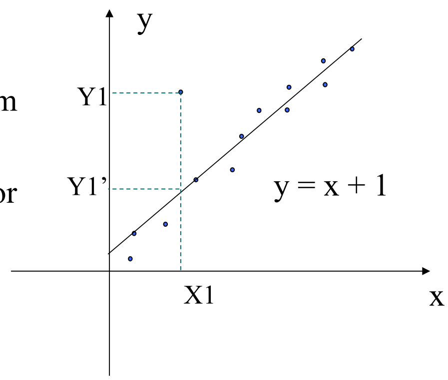
Cluster Analysis



12

Regression

- Fit the data to a function.
- Data points too far away from the function are outliers.
- A *single linear regression*, for instance, finds the line to fit data with 2 variables so that one variable can predict the other.
- More variables can be involved in *multiple linear regression*.



13

Binning:

- Equal-width (distance) partitioning:
  - It divides the range of an attribute 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 values
  - Good data scaling; better handle skewed data

14

Equal-width Binning Methods for Smoothing Data

* Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Partition into (equi-width) bins: 3 intervals of equal size
  - Bin 1: 4, 8, 9
  - Bin 2: 15, 21, 21, 24
  - Bin 3: 25, 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 7, 7, 7
  - Bin 2: 20, 20, 20, 20
  - Bin 3: 28, 28, 28, 28, 28
* Smoothing by bin boundaries:
  - Bin 1: 4, 9, 9
  - Bin 2: 15, 24, 24, 24
  - Bin 3: 25, 25, 25, 25, 34

15

Equal-depth Binning Methods for Smoothing Data

* Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34 (12 points in total)
* Partition into 3 (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

16

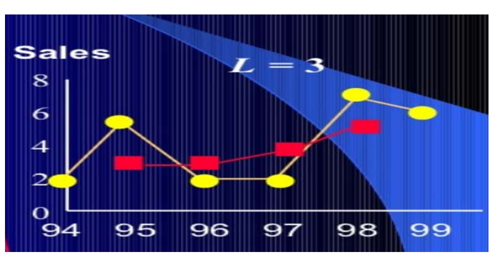
Moving Average

- Use neighborhood values to smooth out noise
- Typically used for time-series data
  - Use series of arithmetic means over time
  - Result depends on choice of length L for computing mean.
- Can also be used on spatial data such as images

17

Moving Average Example

Year	Sales	Moving Average
1994	2	NA
1995	5	3
1996	2	3
1997	2	3.67
1998	7	5
1999	6	NA



18

Outline

- Why preprocess the data?
- Data integration
- Data cleaning
- Data transformation
- Data reduction
- Discretization

Data Transformation

- Transform the data into appropriate form for mining
- Attribute/feature construction
  - New attributes constructed from the given ones
    - e.g., compute average sale amount using total sale amount divided by units sold.
- Normalization: scale attribute values to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Discretization
  - Transform numeric attributes into symbolic attributes

Data Transformation: Normalization

- min-max normalization

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

where min_A and max_A are the minimum and maximum values of attribute A, and $[new\_min_A, new\_max_A]$ is the new range

- Example: Attribute *income* has values
  - $12,000, $20,000, $25,000, $30,000, $45,000, $60,000, $73,600, $98,000
  - normalized into values in range [0, 1]:

    0, 0.093, 0.151, 0.209, 0.384, 0.558, 0.716, 1
- Problems:
  - "Out of bounds" error occurs if a future input case falls outside the original range for A
  - A too big or too small value could be noise. If they are used as min or max value for normalization, the results are not reliable.

21

Data Transformation: Normalization (Contd.)

- z-score normalization

$$v' = \frac{v - mean_A}{s_A}$$

where $mean_A$ is the mean of attribute A and s_A is the standard deviation of A (suppose values are : $v_1, v_2, \ldots, v_n$):

$$s_A = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(v_i - mean_A)^2}$$

- Example:
  - The mean and standard deviation of the attribute *income* are 45,450 and 29735
  - With z-score normalization, the values are transformed into:

    -1.12, -0.86, -0.69, -0.52, -0.02, 0.49, 0.95, 1.77
- Advantages:
  - useful when the actual min and max are unknown
  - better deal with outliers than min-max normalization

22

Data Transformation: Normalization (Contd.)

- Normalization by decimal scaling

$$v_i' = \frac{v_i}{10^k}$$

where k is the smallest integer such that Max(| v_i'|)≤1

- Example:
  - Suppose the recorded values of A range from -986 to 97
  - The maximum absolute value of A is 986.
  - Then k= 3
  - -986 is normalized to -0.986 and 97 is normalized to 0.097

23

Outline

- Why preprocess the data?
- Data integration
- Data cleaning
- Data transformation
- Data reduction
- Discretization

24

Data Reduction

- What is data reduction?
  - A preprocessing step before applying learning or mining techniques to the data
  - Purpose: reduce the size of data.
- Why data reduction?
  - A data set may be too large for a learning program. The dimensions exceed the processing capacity of the program.
  - The expected time for inducing a solution may be too long. Trade off accuracy for speed-up.
  - Sometimes, better answers are found by using a reduced subset of the available data. Too large data may cause the program to fit too many exceptions.

25

Data Reduction Operations

- Standard data form

Case/Example	*feature$_1$*	…	*feature$_k$*	*Class*
e_1	$V_{1,1}$	…	$V_{1,k}$	c_1
…	…	…	…	…
e_i	$V_{i,1}$	…	$V_{i,k}$	c_i
…	…	…	...	…
e_n	$V_{n,1}$	…	$V_{n,k}$	c_n

- Data reduction operations
  - Feature reduction (reduce the number of columns)
  - Case reduction (reduce the number of rows)
  - Value reduction (reduce the number of distinct values in a column)

26

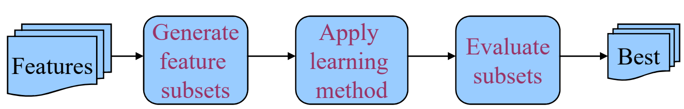
Types of Attributes (Features)

- Three types of attributes:
  - Nominal (symbolic, categorical) — values from an unordered set
    - Eg: {red, yellow, blue, ….}
  - Ordinal — values from an ordered set
    - Eg: {low, medium, high}
  - Continuous — real numbers
    - Eg: {-9.8, 3.9, 8.7, 19.1}
- Next: feature selection for classification tasks.

27

Feature Selection

- Objective
  - Find a subset of features with predictive performance comparable to the full set of features.
  - An optimal subset selection



  - A *practical objective* is to remove clearly extraneous features - leaving the table reduced to manageable dimensions - not necessarily to select the optimal subset.

28

Feature Selection Methods

- Filter Methods: select a subset of original features.
  - Feature Selection from Means and Variances (√)
  - Feature Selection by Mutual Information (√)
  - Feature Selection by Decision Trees (√)
  - Feature Selection by Rough Sets, etc.
- Merger Methods: merge features, resulting in a new set of fewer columns with new values.
  - Principal component analysis (PCA)
- Wrapper Methods: feature selection is being "wrapper around" a learning algorithm.
  - This is the optimal method in the last slide.
  - Running time is long; infeasible in practice if there are many features.

29