# Chapter 4:
# Data Mining Process

Lecturer: Dr. *Nguyen Thi Ngoc Anh*
*Email: ngocanhnt@ude.edu.vn*



## How to Choose a Data Mining System?

- Commercial data mining systems have little in common
  - Different data mining functionality or methodology
  - May even work with completely different kinds of data sets
- Need multiple dimensional view in selection
- Data types: relational, transactional, text, time sequence, spatial?
- System issues
  - running on only one or on several operating systems?
  - a client/server architecture?
  - Provide Web-based interfaces and allow XML data as input and/or output?

## How to Choose a Data Mining System? (2)

- Data sources
  - ASCII text files, multiple relational data sources
  - support ODBC connections (OLE DB, JDBC)?
- Data mining functions and methodologies
  - One vs. multiple data mining functions
  - One vs. variety of methods per function
    - More data mining functions and methods per function provide the user with greater flexibility and analysis power
- Coupling with DB and/or data warehouse systems
  - Four forms of coupling: no coupling, loose coupling, semitight coupling, and tight coupling
    - Ideally, a data mining system should be tightly coupled with a database system



## How to Choose a Data Mining System? (3)

- Scalability
  - Row (or database size) scalability
  - Column (or dimension) scalability
  - Curse of dimensionality: it is much more challenging to make a system column scalable that row scalable
- Visualization tools
  - "A picture is worth a thousand words"
  - Visualization categories: data visualization, mining result visualization, mining process visualization, and visual data mining
- Data mining query language and graphical user interface
  - Easy-to-use and high-quality graphical user interface
  - Essential for user-guided, highly interactive data mining

## Examples of Data Mining Systems (1)

- IBM Intelligent Miner
  - A wide range of data mining algorithms
  - Scalable mining algorithms
  - Toolkits: neural network algorithms, statistical methods, data preparation, and data visualization tools
  - Tight integration with IBM's DB2 relational database system
- SAS Enterprise Miner
  - A variety of statistical analysis tools
  - Data warehouse tools and multiple data mining algorithms
- Mirosoft SQLServer 2000
  - Integrate DB and OLAP with mining
  - Support OLEDB for DM standard

5

## Examples of Data Mining Systems (2)

- SGI MineSet
  - Multiple data mining algorithms and advanced statistics
  - Advanced visualization tools
- Clementine (SPSS)
  - An integrated data mining development environment for end-users and developers
  - Multiple data mining algorithms and visualization tools
- DBMiner (DBMiner Technology Inc.)
  - Multiple data mining modules: discovery-driven OLAP analysis, association, classification, and clustering
  - Efficient, association and sequential-pattern mining functions, and visual classification tool
  - Mining both relational databases and data warehouses

6

## Data Mining Process

- Cross-Industry Standard Process for Data Mining (CRISP-DM)
- European Community funded effort to develop framework for data mining tasks
- Goals:
  - Encourage interoperable tools across entire data mining process
  - Take the mystery/high-priced expertise out of simple data mining tasks

7

## Why Should There be a Standard Process?

*The data mining process must be reliable and repeatable by people with little data mining background.*

- Framework for recording experience
  - Allows projects to be replicated
- Aid to project planning and management
- "Comfort factor" for new adopters
  - Demonstrates maturity of Data Mining
  - Reduces dependency on "stars"

8

## Process Standardization

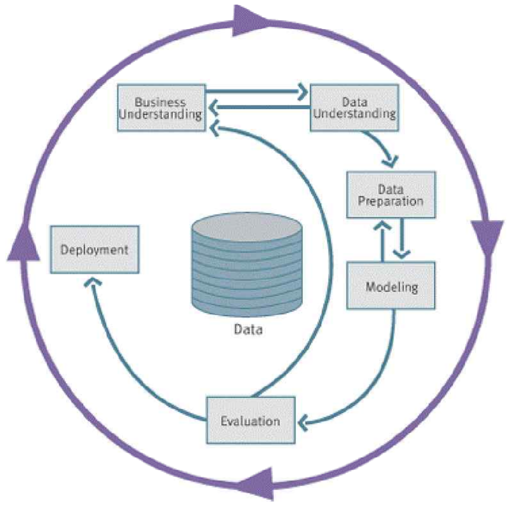
- CRoss Industry Standard Process for Data Mining
- Initiative launched Sept.1996
- SPSS/ISL, NCR, Daimler-Benz, OHRA
- Funding from European commission
- Over 200 members of the CRISP-DM SIG worldwide
  - DM Vendors - SPSS, NCR, IBM, SAS, SGI, Data Distilleries, Syllogic, Magnify, ..
  - System Suppliers / consultants - Cap Gemini, ICL Retail, Deloitte & Touche, …
  - End Users - BT, ABB, Lloyds Bank, AirTouch, Experian, ...



## CRISP-DM

- Non-proprietary
- Application/Industry neutral
- Tool neutral
- Focus on business issues
  - As well as technical analysis
- Framework for guidance
- Experience base
  - Templates for Analysis

# CRISP-DM:  Overview




11

# CRISP-DM:  Phases

- **Business Understanding**
  - Understanding project objectives and requirements
  - Data mining problem definition
- **Data Understanding**
  - Initial data collection and familiarization
  - Identify data quality issues
  - Initial, obvious results
- **Data Preparation**
  - Record and attribute selection
  - Data cleansing
- **Modeling**
  - Run the data mining tools
- **Evaluation**
  - Determine if results meet business objectives
  - Identify business issues that should have been addressed earlier
- **Deployment**
  - Put the resulting models into practice
  - Set up for repeated/continuous mining of the data

12

# Phases and Tasks

**Business Understanding**

**Determine Business Objectives**
*Background*
*Business Objectives*
*Business Success Criteria*

**Situation Assessment**
*Inventory of Resources*
*Requirements, Assumptions, and Constraints*
*Risks and Contingencies*
*Terminology*
*Costs and Benefits*

**Determine Data Mining Goal**
*Data Mining Goals*
*Data Mining Success Criteria*

**Produce Project Plan**
*Project Plan*
*Initial Assessment of Tools and Techniques*

**Data Understanding**

**Collect Initial Data**
*Initial Data Collection Report*

**Describe Data**
*Data Description Report*

**Explore Data**
*Data Exploration Report*

**Verify Data Quality**
*Data Quality Report*

**Data Preparation**

*Data Set*
*Data Set Description*

**Select Data**
*Rationale for Inclusion / Exclusion*

**Clean Data**
*Data Cleaning Report*

**Construct Data**
*Derived Attributes*
*Generated Records*

**Integrate Data**
*Merged Data*

**Format Data**
*Reformatted Data*

**Modeling**

**Select Modeling Technique**
*Modeling Technique*
*Modeling Assumptions*

**Generate Test Design**
*Test Design*

**Build Model**
*Parameter Settings*
*Models*
*Model Description*

**Assess Model**
*Model Assessment*
*Revised Parameter Settings*

**Evaluation**

**Evaluate Results**
*Assessment of Data Mining Results w.r.t. Business Success Criteria*
*Approved Models*

**Review Process**
*Review of Process*

**Determine Next Steps**
*List of Possible Actions*
*Decision*

**Deployment**

**Plan Deployment**
*Deployment Plan*

**Plan Monitoring and Maintenance**
*Monitoring and Maintenance Plan*

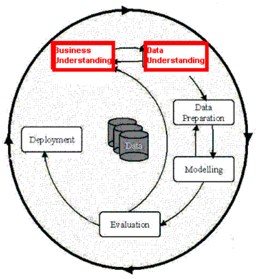
**Produce Final Report**
*Final Report*
*Final Presentation*

**Review Project**
*Experience Documentation*

13

# Phases in the DM Process (1 & 2)

- Business Understanding:
  - Statement of Business Objective
  - Statement of Data Mining objective
  - Statement of Success Criteria



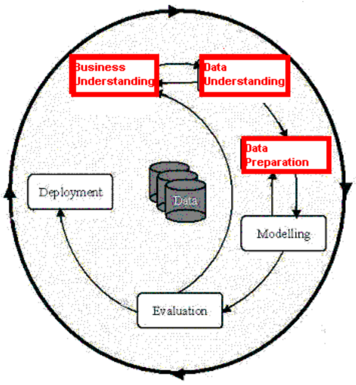

- Data Understanding
  - Explore the data and verify the quality
  - Find outliers

14

# Phases in the DM Process (3)

Data preparation:

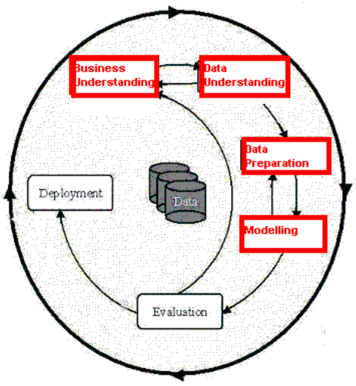
- Takes usually over 90% of the time
  - Collection
  - Assessment
  - Consolidation and Cleaning
    - table links, aggregation level, missing values, etc
  - Data selection
    - active role in ignoring non-contributory data?
    - outliers?
    - Use of samples
    - visualization tools
  - Transformations - create new variables





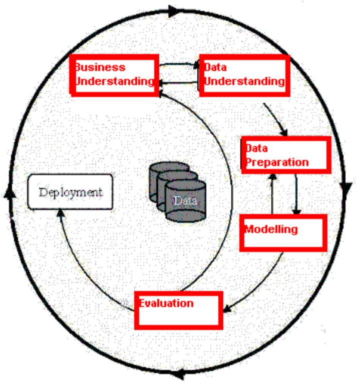

# Phases in the DM Process (4)

- Model building
  - Selection of the modeling techniques is based upon the data mining objective
  - Modeling is an iterative process - different for *supervised* and *unsupervised learning*
    - May model for either description or prediction



## Phases in the DM Process (5)

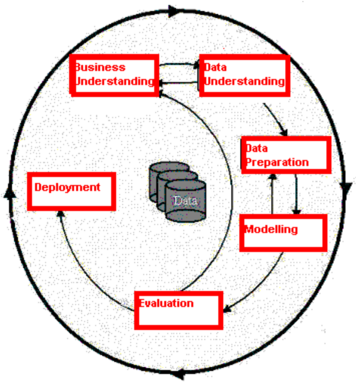
- Model Evaluation
  - Evaluation of model: how well it performed on test data
  - Methods and criteria depend on model type:
    - e.g., coincidence matrix with classification models, mean error rate with regression models
  - Interpretation of model: important or not, easy or hard depends on algorithm




18

## Phases in the DM Process (6)

- Deployment
  - Determine how the results need to be utilized
  - Who needs to use them?
  - How often do they need to be used
- Deploy Data Mining results by:
  - Scoring a database
  - Utilizing results as business rules
  - interactive scoring on-line




19

## Why CRISP-DM?

- The data mining process must be reliable and repeatable by people with little data mining skills
- CRISP-DM provides a uniform framework for
  - guidelines
  - experience documentation
- CRISP-DM is flexible to account for differences
  - Different business/agency problems
  - Different data



# Attribute-Oriented Induction

## Attribute-Oriented Induction

- Proposed in 1989 (KDD '89 workshop)
- Not confined to categorical data nor particular measures.
- How it is done?
  - Collect the task-relevant data (*initial relation*) using a relational database query
  - Perform generalization by attribute removal or attribute generalization.
  - Apply aggregation by merging identical, generalized tuples and accumulating their respective counts
  - Interactive presentation with users



## Basic Principles of Attribute-Oriented Induction

- Data focusing: task-relevant data, including dimensions, and the result is the *initial relation*.
- Attribute-removal: remove attribute *A* if there is a large set of distinct values for *A* but (1) there is no generalization operator on *A*, or (2) *A*'s higher level concepts are expressed in terms of other attributes.
- Attribute-generalization: If there is a large set of distinct values for *A*, and there exists a set of generalization operators on *A*, then select an operator and generalize *A*.
- Attribute-threshold control: typical 2-8, specified/default.
- Generalized relation threshold control: control the final relation/rule size. see example

## Attribute-Oriented Induction: Basic Algorithm

- InitialRel: Query processing of task-relevant data, deriving the *initial relation*.
- PreGen: Based on the analysis of the number of distinct values in each attribute, determine generalization plan for each attribute: removal? or how high to generalize?
- PrimeGen: Based on the PreGen plan, perform generalization to the right level to derive a "prime generalized relation", accumulating the counts.
- Presentation: User interaction: (1) adjust levels by drilling, (2) pivoting, (3) mapping into rules, cross tabs, visualization presentations.



## Example

- DMQL: Describe general characteristics of graduate students in the Big-University database

```
use Big_University_DB
mine characteristics as "Science_Students"
in relevance to name, gender, major, birth_place, birth_date,
   residence, phone#, gpa
from student
where  status in "graduate"
```

- Corresponding SQL statement:

```
Select name, gender, major, birth_place, birth_date, residence,
   phone#, gpa
from student
where  status in {"Msc", "MBA", "PhD" }
```

## Class Characterization: An Example

**Initial Relation**

| Name | Gender | Major | Birth-Place | Birth_date | Residence | Phone # | GPA |
|---|---|---|---|---|---|---|---|
| Jim Woodman | M | CS | Vancouver,BC, Canada | 8-12-76 | 3511 Main St., Richmond | 687-4598 | 3.67 |
| Scott Lachance | M | CS | Montreal, Que, Canada | 28-7-75 | 345 1st Ave., Richmond | 253-9106 | 3.70 |
| Laura Lee | F | Physics | Seattle, WA, USA | 25-8-70 | 125 Austin Ave., Burnaby | 420-5232 | 3.83 |
| … | … | … | … | … | … | … | … |
| Removed | Retained | Sci,Eng, Bus | Country | Age range | City | Removed | Excl, VG,.. |

**Prime Generalized Relation**

| Gender | Major | Birth_region | Age_range | Residence | GPA | Count |
|---|---|---|---|---|---|---|
| M | Science | Canada | 20-25 | Richmond | Very-good | 16 |
| F | Science | Foreign | 25-30 | Burnaby | Excellent | 22 |
| … | … | … | … | … | … | … |

| Gender \ Birth_Region | Canada | Foreign | Total |
|---|---|---|---|
| M | 16 | 14 | 30 |
| F | 10 | 22 | 32 |
| Total | 26 | 36 | 62 |

33

## Presentation of Generalized Results

- Generalized relation:
  - Relations where some or all attributes are generalized, with counts or other aggregation values accumulated.
- Cross tabulation:
  - Mapping results into cross tabulation form (similar to contingency tables).
  - Visualization techniques:
  - Pie charts, bar charts, curves, cubes, and other visual forms.
- Quantitative characteristic rules:
  - Mapping generalized result into characteristic rules with quantitative information associated with it, e.g.,

$$grad(x) \land male(x) \Rightarrow$$
$$birth\_region(x) = "Canada"[t:53\%] \lor birth\_region(x) = "foreign"[t:47\%].$$

34

## Presentation—Generalized Relation

| location | item | sales (in million dollars) | count (in thousands) |
|---|---|---|---|
| Asia | TV | 15 | 300 |
| Europe | TV | 12 | 250 |
| North_America | TV | 28 | 450 |
| Asia | computer | 120 | 1000 |
| Europe | computer | 150 | 1200 |
| North_America | computer | 200 | 1800 |

Table 5.3: A generalized relation for the sales in 1997.

35

## Presentation—Crosstab

| location \ item | TV | | computer | | *both_items* | |
|---|---|---|---|---|---|---|
| | sales | count | sales | count | sales | count |
| Asia | 15 | 300 | 120 | 1000 | 135 | 1300 |
| Europe | 12 | 250 | 150 | 1200 | 162 | 1450 |
| North_America | 28 | 450 | 200 | 1800 | 228 | 2250 |
| *all_regions* | 45 | 1000 | 470 | 4000 | 525 | 5000 |

Table 5.4: A crosstab for the sales in 1997.

36

## Concept Description: Characterization and Comparison

- What is concept description?
- Data generalization and summarization-based characterization
- Analytical characterization: Analysis of attribute relevance
- Mining class comparisons: Discriminating between different classes
- Mining descriptive statistical measures in large databases
- Discussion
- Summary



## Characterization vs. OLAP

- Similarity:
  - Presentation of data summarization at multiple levels of abstraction.
  - Interactive drilling, pivoting, slicing and dicing.
- Differences:
  - Automated desired level allocation.
  - Dimension relevance analysis and ranking when there are many relevant dimensions.
  - Sophisticated typing on dimensions and measures.
  - Analytical characterization: data dispersion analysis.

## Attribute Relevance Analysis

- Why?
  - Which dimensions should be included?
  - How high level of generalization?
  - Automatic VS. Interactive
  - Reduce # attributes; Easy to understand patterns
- What?
  - statistical method for preprocessing data
    - filter out irrelevant or weakly relevant attributes
    - retain or rank the relevant attributes
  - relevance related to dimensions and levels
  - analytical characterization, analytical comparison



## Attribute relevance analysis (cont'd)

- How?
  - Data Collection
  - Analytical Generalization
    - Use information gain analysis (e.g., entropy or other measures) to identify highly relevant dimensions and levels.
  - Relevance Analysis
    - Sort and select the most relevant dimensions and levels.
  - Attribute-oriented Induction for class description
    - On selected dimension/level
  - OLAP operations (e.g. drilling, slicing) on relevance rules

## Relevance Measures

- Quantitative relevance measure determines the classifying power of an attribute within a set of data.
- Methods
  - information gain (ID3)
  - gain ratio (C4.5)
  - gini index
  - $\chi^2$ contingency table statistics
  - uncertainty coefficient



## Information-Theoretic Approach

- Decision tree
  - each internal node tests an attribute
  - each branch corresponds to attribute value
  - each leaf node assigns a classification
- ID3 algorithm
  - build decision tree based on training objects with known class labels to classify testing objects
  - rank attributes with information gain measure
  - minimal height
    - the least number of tests to classify an object

# Example: Analytical Characterization

- Task
  - Mine general characteristics describing graduate students using analytical characterization
- Given
  - attributes *name, gender, major, birth_place, birth_date, phone#*, and *gpa*
  - $Gen(a_i)$ = concept hierarchies on $a_i$
  - $U_i$ = attribute analytical thresholds for $a_i$
  - $T_i$ = attribute generalization thresholds for $a_i$
  - $R$ = attribute relevance threshold

46

# Example: Analytical Characterization (cont'd)

- 1. Data collection
  - target class: graduate student
  - contrasting class: undergraduate student
- 2. Analytical generalization using $U_i$
  - attribute removal
    - remove *name* and *phone#*
  - attribute generalization
    - generalize *major*, *birth_place*, *birth_date* and *gpa*
    - accumulate counts
  - candidate relation: *gender*, *major*, *birth_country*, *age_range* and *gpa*

47

## Example: Analytical characterization (2)

| gender | major | birth_country | age_range | gpa | count |
|---|---|---|---|---|---|
| M | Science | Canada | 20-25 | Very_good | 16 |
| F | Science | Foreign | 25-30 | Excellent | 22 |
| M | Engineering | Foreign | 25-30 | Excellent | 18 |
| F | Science | Foreign | 25-30 | Excellent | 25 |
| M | Science | Canada | 20-25 | Excellent | 21 |
| F | Engineering | Canada | 20-25 | Excellent | 18 |

***Candidate relation for Target class: Graduate students (Σ=120)***

| gender | major | birth_country | age_range | gpa | count |
|---|---|---|---|---|---|
| M | Science | Foreign | <20 | Very_good | 18 |
| F | Business | Canada | <20 | Fair | 20 |
| M | Business | Canada | <20 | Fair | 22 |
| F | Science | Canada | 20-25 | Fair | 24 |
| M | Engineering | Foreign | 20-25 | Very_good | 22 |
| F | Engineering | Canada | <20 | Excellent | 24 |

***Candidate relation for Contrasting class: Undergraduate students (Σ=130)***

48

## Example: Analytical characterization (3)

- 3. Relevance analysis
  - Calculate expected info required to classify an arbitrary tuple

$$I(s_1, s_2) = I(120,130) = -\frac{120}{250}\log_2\frac{120}{250} - \frac{130}{250}\log_2\frac{130}{250} = 0.9988$$

  - Calculate entropy of each attribute: e.g. *major*

For *major*="*Science*": $s_{11}=84$ $s_{21}=42$ $I(s_{11},s_{21})=0.9183$

For *major*="*Engineering*": $s_{12}=36$ $s_{22}=46$ $I(s_{12},s_{22})=0.9892$

For *major*="*Business*": $s_{13}=0$ $s_{23}=42$ $I(s_{13},s_{23})=0$

Number of grad students in "Science"

Number of undergrad students in "Science"

49

# Example: Analytical Characterization (4)

- Calculate expected info required to classify a given sample if S is partitioned according to the attribute

$$E(major) = \frac{126}{250} I(s_{11}, s_{21}) + \frac{82}{250} I(s_{12}, s_{22}) + \frac{42}{250} I(s_{13}, s_{23}) = 0.7873$$

- Calculate information gain for each attribute

$$Gain(major) = I(s_1, s_2) - E(major) = 0.2115$$

  - Information gain for all attributes

| | |
|---|---|
| Gain(gender) | = 0.0003 |
| Gain(birth_country) | = 0.0407 |
| Gain(major) | = 0.2115 |
| Gain(gpa) | = 0.4490 |
| Gain(age_range) | = 0.5971 |

50

# Example: Analytical characterization (5)

- 4. Initial working relation ($W_0$) derivation
  - R = 0.1
  - remove irrelevant/weakly relevant attributes from candidate relation => drop *gender*, *birth_country*
  - remove contrasting class candidate relation

| major | age_range | gpa | count |
|---|---|---|---|
| Science | 20-25 | Very_good | 16 |
| Science | 25-30 | Excellent | 47 |
| Science | 20-25 | Excellent | 21 |
| Engineering | 20-25 | Excellent | 18 |
| Engineering | 25-30 | Excellent | 18 |

**Initial target class working relation $W_0$: Graduate students**

- 5. Perform attribute-oriented induction on $W_0$ using $T_i$

51

## Concept Description: Characterization and Comparison

- What is concept description?
- Data generalization and summarization-based characterization
- Analytical characterization: Analysis of attribute relevance
- Mining class comparisons: Discriminating between different classes
- Mining descriptive statistical measures in large databases
- Discussion
- Summary

52

## Mining Class Comparisons

- Comparison: Comparing two or more classes
- Method:
  - Partition the set of relevant data into the target class and the contrasting class(es)
  - Generalize both classes to the same high level concepts
  - Compare tuples with the same high level descriptions
  - Present for every tuple its description and two measures
    - support - distribution within single class
    - comparison - distribution between classes
  - Highlight the tuples with strong discriminant features
- Relevance Analysis:
  - Find attributes (features) which best distinguish different classes

53

## Example: Analytical comparison

- Task
  - Compare graduate and undergraduate students using discriminant rule.
  - DMQL query

**use** Big_University_DB
**mine comparison as** “grad_vs_undergrad_students”
**in relevance to** *name, gender, major, birth_place, birth_date, residence, phone#, gpa*
**for** “graduate_students”
**where** status in graduate”
**versus** “undergraduate_students”
**where** status in “undergraduate”
**analyze** count%
**from** student



## Example: Analytical comparison (2)

- Given
  - attributes *name, gender, major, birth_place, birth_date, residence, phone#* and *gpa*
  - $Gen(a_i)$ = concept hierarchies on attributes $a_i$
  - $U_i$ = attribute analytical thresholds for attributes $a_i$
  - $T_i$ = attribute generalization thresholds for attributes $a_i$
  - $R$ = attribute relevance threshold

# Example: Analytical comparison (3)

- 1. Data collection
  - target and contrasting classes
- 2. Attribute relevance analysis
  - remove attributes *name, gender, major, phone#*
- 3. Synchronous generalization
  - controlled by user-specified dimension thresholds
  - prime target and contrasting class(es) relations/cuboids

56

# Example: Analytical comparison (4)

| Birth_country | Age_range | Gpa | Count% |
|---|---|---|---|
| Canada | 20-25 | Good | 5.53% |
| Canada | 25-30 | Good | 2.32% |
| Canada | Over_30 | Very_good | 5.86% |
| … | … | … | … |
| Other | Over_30 | Excellent | 4.68% |

**Prime generalized relation for the target class: Graduate students**

| Birth_country | Age_range | Gpa | Count% |
|---|---|---|---|
| Canada | 15-20 | Fair | 5.53% |
| Canada | 15-20 | Good | 4.53% |
| … | … | … | … |
| Canada | 25-30 | Good | 5.02% |
| … | … | … | … |
| Other | Over_30 | Excellent | 0.68% |

**Prime generalized relation for the contrasting class: Undergraduate students**

57

## Example: Analytical comparison (5)

- 4. Drill down, roll up and other OLAP operations on target and contrasting classes to adjust levels of abstractions of resulting description
- 5. Presentation
  - as generalized relations, crosstabs, bar charts, pie charts, or rules
  - contrasting measures to reflect comparison between target and contrasting classes
    - e.g. count%



## Quantitative Discriminant Rules

- Cj = target class
- $q_a$ = a generalized tuple covers some tuples of class
  - but can also cover some tuples of contrasting class
- d-weight
  - range: [0, 1] $d\_weight = \frac{count(q_a \in C_j)}{\sum_{i=1}^{m} count(q_a \in C_i)}$
- quantitative discriminant rule form

$$\forall X,\ target\_class(X) \Leftarrow condition(X)\quad [d : d\_weight]$$

## Example: Quantitative Discriminant Rule

| Status | Birth_country | Age_range | Gpa | Count |
|---|---|---|---|---|
| Graduate | Canada | 25-30 | Good | 90 |
| Undergraduate | Canada | 25-30 | Good | 120 |

Count distribution between graduate and undergraduate students for a generalized tuple

- Quantitative discriminant rule

$$\forall X,\ graduate\_student(X) \Leftarrow$$
$$birth\_country(X) = \text{"Canada"} \wedge age\_range(X) = \text{"25-30"} \wedge gpa(X) = \text{"good"}\ \ [d: 30\%]$$

  - where 90/(90+120) = 30%



## Class Description

- Quantitative characteristic rule

$$\forall\ X,\ \ target\_class(X) \Rightarrow condition(X)\ \ \ [t : t\_weight]$$

  - necessary
- Quantitative discriminant rule

$$\forall\ X,\ \ target\_class(X) \Leftarrow condition(X)\ \ \ [d : d\_weight]$$

  - sufficient
- Quantitative description rule

$$\forall\ X,\ \ target\_class(X) \Leftrightarrow$$
$$condition_1(X)[t : w_1, d : w'_1] \vee \ldots \vee condition_n(X)[t : w_n, d : w'_n]$$

  - necessary and sufficient

## Example: Quantitative Description Rule

| Location/item | TV | | | Computer | | | Both_items | | |
|---|---|---|---|---|---|---|---|---|---|
| | *Count* | *t-wt* | *d-wt* | *Count* | *t-wt* | *d-wt* | *Count* | *t-wt* | *d-wt* |
| **Europe** | 80 | 25% | 40% | 240 | 75% | 30% | 320 | 100% | 32% |
| **N_Am** | 120 | 17.65% | 60% | 560 | 82.35% | 70% | 680 | 100% | 68% |
| **Both_ regions** | 200 | 20% | 100% | 800 | 80% | 100% | 1000 | 100% | 100% |

**Crosstab showing associated t-weight, d-weight values and total number (in thousands) of TVs and computers sold at AllElectronics in 1998**

- Quantitative description rule for target class *Europe*

$$\forall X, Europe(X) \Leftrightarrow$$
$$(item(X) = "TV")[\mathrm{t}:25\%, \mathrm{d}:40\%] \vee (item(X) = "computer")[\mathrm{t}:75\%, \mathrm{d}:30\%]$$

62

## Concept Description: Characterization and Comparison

- What is concept description?
- Data generalization and summarization-based characterization
- Analytical characterization: Analysis of attribute relevance
- Mining class comparisons: Discriminating between different classes
- Mining descriptive statistical measures in large databases
- Discussion
- Summary

72

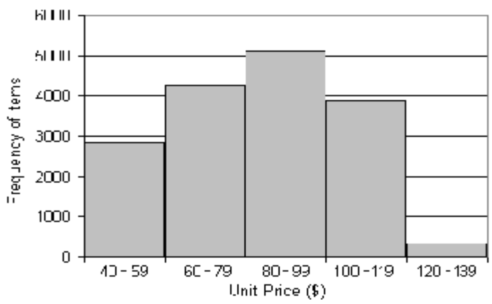
# Mining Data Dispersion Characteristics

- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals
- Dispersion analysis on computed measures
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube

73

# Histogram Analysis

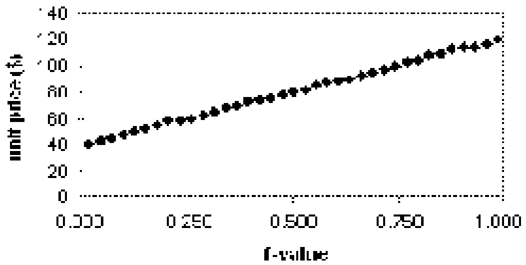
- Graph displays of basic statistical class descriptions
  - Frequency histograms
    - A univariate graphical method
    - Consists of a set of rectangles that reflect the counts or frequencies of the classes present in the given data




79

# Quantile Plot

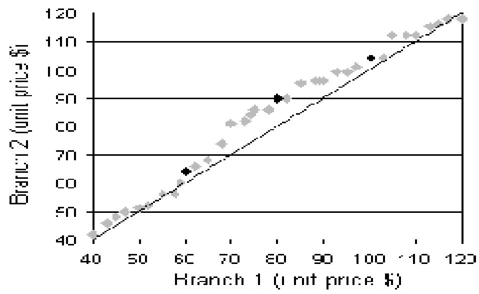
- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots quantile information
  - For a data $x_i$ data sorted in increasing order, $f_i$ indicates that approximately 100 $f_i$% of the data are below or equal to the value $x_i$





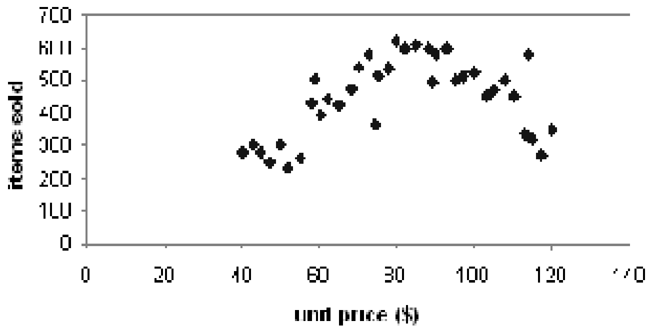

# Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- Allows the user to view whether there is a shift in going from one distribution to another



# Scatter plot

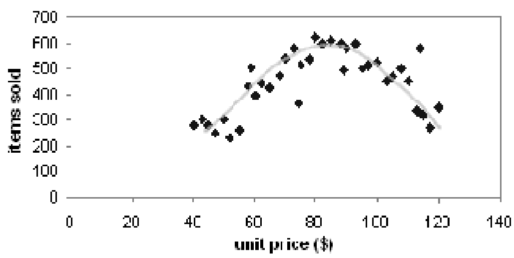
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane




82

# Loess Curve

- Adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence
- Loess curve is fitted by setting two parameters: a smoothing parameter, and the degree of the polynomials that are fitted by the regression




83

## Summary

- Concept description: characterization and discrimination
- OLAP-based vs. attribute-oriented induction
- Efficient implementation of AOI
- Analytical characterization and comparison
- Mining descriptive statistical measures in large databases
- Discussion
  - Incremental and parallel mining of description
  - Descriptive mining of complex types of data

90

## References


- Y. Cai, N. Cercone, and J. Han. Attribute-oriented induction in relational databases. In G. Piatetsky-Shapiro and W. J. Frawley, editors, Knowledge Discovery in Databases, pages 213-228. AAAI/MIT Press, 1991.
- S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26:65-74, 1997
- C. Carter and H. Hamilton. Efficient attribute-oriented generalization for knowledge discovery from large databases. IEEE Trans. Knowledge and Data Engineering, 10:193-208, 1998.
- W. Cleveland. Visualizing Data. Hobart Press, Summit NJ, 1993.
- J. L. Devore. Probability and Statistics for Engineering and the Science, 4th ed. Duxbury Press, 1995.
- T. G. Dietterich and R. S. Michalski. A comparative review of selected methods for learning from examples. In Michalski et al., editor, Machine Learning: An Artificial Intelligence Approach, Vol. 1, pages 41-82. Morgan Kaufmann, 1983.
- J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. Data Mining and Knowledge Discovery, 1:29-54, 1997.
- J. Han, Y. Cai, and N. Cercone. Data-driven discovery of quantitative rules in relational databases. IEEE Trans. Knowledge and Data Engineering, 5:29-40, 1993.


91

# References (cont.)


- J. Han and Y. Fu. Exploration of the power of attribute-oriented induction in data mining. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, pages 399-421. AAAI/MIT Press, 1996.
- R. A. Johnson and D. A. Wichern. Applied Multivariate Statistical Analysis, 3rd ed. Prentice Hall, 1992.
- E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. VLDB'98, New York, NY, Aug. 1998.
- H. Liu and H. Motoda. Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, 1998.
- R. S. Michalski. A theory and methodology of inductive learning. In Michalski et al., editor, Machine Learning: An Artificial Intelligence Approach, Vol. 1, Morgan Kaufmann, 1983.
- T. M. Mitchell. Version spaces: A candidate elimination approach to rule learning. IJCAI'97, Cambridge, MA.
- T. M. Mitchell. Generalization as search. Artificial Intelligence, 18:203-226, 1982.
- T. M. Mitchell. Machine Learning. McGraw Hill, 1997.
- J. R. Quinlan. Induction of decision trees. Machine Learning, 1:81-106, 1986.
- D. Subramanian and J. Feigenbaum. Factorization in experiment generation. AAAI'86, Philadelphia, PA, Aug. 1986.