new, see p. 20

Quantitative Characteristic Rules

- *Typicality weight* *(t_weight)* of the disjuncts in a rule
  - *n:* number of tuples in the initial generalized relation *R*
  - *t_weight:* fraction of tuples in R that represent target class
  - q_a: generalized tuple describing the target class
  - definition $$t\_weight\,(q_a) = \frac{\text{count}\,(q_a)}{\sum_{i=1}^{n} \text{count}\,(q_i)}$$
  - range is [0…1]
- Form of a *Quantitative Characteristic Rule:* (cf. crosstab)

$$\forall X, \text{target\_class}(X) \;\Rightarrow\; \text{condition}_1(X)[t : w_1] \;\vee\; \ldots \;\vee\; \text{condition}_m(X)[t : w_m]$$

  - Disjunction represents a *necessary* condition of the target class
  - *Not sufficient*: a tuple that meets the conditions could belong to another class

Chapter 5: 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
- Descriptive statistical measures in large databases
- Summary

Characterization vs. OLAP

- Shared concepts:
  - 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.

Streuung

Attribute Relevance Analysis

- Why?—Support for specifying generalization parameters
  - Which dimensions should be included?
  - How high level of generalization?
  - Automatic vs. interactive
  - Reduce number of attributes
    → easy to understand patterns / rules
- What?—Purpose of the method
  - 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?—Steps of the algorithm:
  - 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

Relevance Measures

- Quantitative relevance measure determines the classifying power of an attribute within a set of data.

- Competing methods
  - information gain (ID3)—discussed here
  - gain ratio (C4.5)
  - gini index (IBM Intelligent Miner)
  - χ^2 contingency table statistics
  - uncertainty coefficient

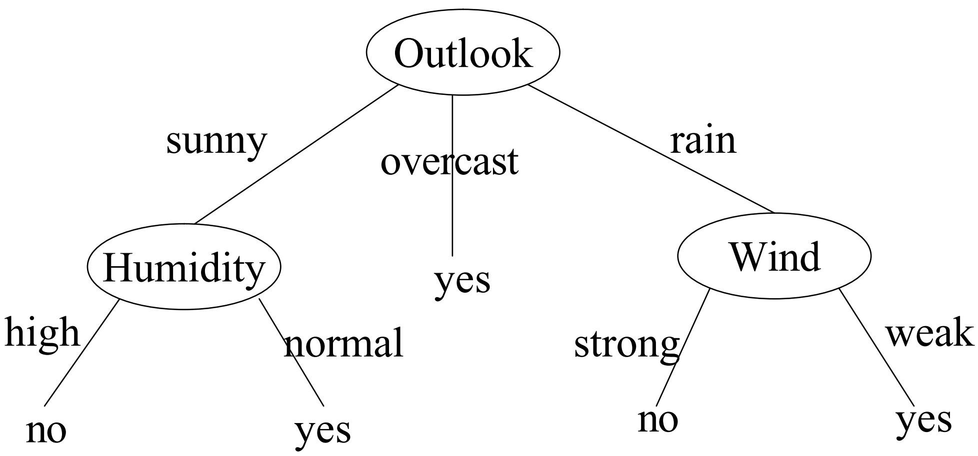
Information-Theoretic Approach

- 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

Top-Down Induction of Decision Tree

Attributes = {Outlook, Temperature, Humidity, Wind}

PlayTennis = {yes, no}



Entropy and Information Gain

- S contains s_i tuples of class C_i for $i = \{1, \ldots, m\}$
- Information measures info required to classify any arbitrary tuple

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} \log_2 \frac{s_i}{s}$$

- Entropy of attribute A with values $\{a_1, a_2, \ldots, a_v\}$

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} I(s_{1j}, \ldots, s_{mj})$$

- Information gained by branching on attribute A

$$\text{Gain}(A) = I(s_1, s_2, \ldots, s_m) - E(A)$$

Example: Analytical Characterization

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

Example: Analytical Characterization (2)

- Step 1: Data collection
  - target class: graduate student
  - contrasting class: undergraduate student

- Step 2: Analytical generalization using thresholds U_i
  - attribute removal
    - remove *name* and *phone#*
  - attribute generalization
    - generalize *major, birth_place, birth_date, gpa*
    - accumulate counts
  - candidate relation
    - *gender, major, birth_country, age_range, gpa*

Example: Analytical characterization (3)

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)

Example: Analytical Characterization (4)

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

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

  - 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"

Example: Analytical Characterization (5)

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

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

- Calculate information gain for each attribute

$$\mathrm{Gain(major)} = \mathrm{I}(s_1, s_2) - \mathrm{E(major)} = 0.2115$$

  - 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

Example: Analytical Characterization (6)

- Step 4a: Derive initial working relation W_0
  - Use attribute relevance threshold *R*, e.g., *R* = 0.1
  - remove irrelevant/weakly relevant attributes (*gain* < *R*) from candidate relation, i.e., 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

- Step 4b: Perform attribute-oriented induction using thresholds T_i

Chapter 5: 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
- Descriptive statistical measures in large databases
- Summary

Mining Class Comparisons

- Comparison
  - Comparing two or more classes.
- Relevance Analysis
  - Find attributes (features) which best distinguish different classes.
- Method
  - Partition the set of relevant data into the target class and the contrasting class(es)
  - Analyze the attribute's relevances
  - Generalize both classes to the same high level concepts
  - Compare tuples with the same high level descriptions
  - Present the results and highlight the tuples with strong discriminant features

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#, gpa*
  - *generalization*(a_i) = concept hierarchies on attributes a_i
  - U_i = attribute analytical thresholds for attributes a_i
  - R = attribute relevance threshold
  - T_i = attribute generalization thresholds for attributes a_i

Example: Analytical comparison (3)

- Step1: Data collection
  - target and contrasting classes
- Step 2: Attribute relevance analysis
  - remove attributes *name, gender, major, phone#*
- Step 3: Synchronous generalization
  - controlled by user-specified dimension thresholds
  - prime target and contrasting class(es) relations/cuboids

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

Example: Analytical comparison (5)

- Step 4: Compare tuples; drill down, roll up and other OLAP operations on target and contrasting classes to adjust levels of abstractions of resulting description.
- Step 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

- C_j = target class
- q_a = a generalized tuple covers some tuples of class
  - but can also cover some tuples of contrasting class
- *Discrimination weight* (d_weight)
  - m classes C_i
  - definition:

$$d\_weight(q_a, C_j) = \frac{\text{count}(q_a \in C_j)}{\sum_{i=1}^{m} \text{count}(q_a \in C_i)}$$

  - range: [0, 1]
  - high *d*_weight: q_a primarily represents a target class concept
  - low *d*_weight: q_a is primarily derived from contrasting classes

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

Example: Quantitative Discriminant Rule

Status	Birth_country	Age_range	Gpa	Count
Graduate	Canada	25-30	Good	90
Undergraduate	Canada	25-30	Good	210

Count distribution between graduate and undergraduate students for a generalized tuple

- Quantitative discriminant rule

$$\begin{aligned}\forall X, \text{graduate\_student}(X) \Leftarrow\ & \text{birth\_country}(X) = \text{‘Canada’} \wedge \\ & \text{age\_range}(X) = \text{’25-30’} \wedge \\ & \text{gpa}(X) = \text{‘good’} \quad [d{:}\ 30\%]\end{aligned}$$

  - *d_weight* = 90/(90+210) = 30%
  - Rule is *sufficient* (but not *necessary*):
    - if X fulfills the condition, the probability that X is a graduate student is 30%, but not vice versa, i.e., there are other grad studs, too.

Class Description

- Quantitative *characteristic* rule (*necessary*)

$$\forall X, target\_class(X) \Rightarrow condition_1(X)[t{:}w_1] \vee \ldots \vee condition_m(X)[t{:}w_m]$$

- Quantitative *discriminant* rule (*sufficient*)

$$\forall X, target\_class(X) \Leftarrow condition_1(X)[d{:}w_1'] \vee \ldots \vee condition_m(X)[d{:}w_m']$$

- Quantitative *description* rule (*necessary and sufficient*)

$$\forall X, target\_class(X) \Rightarrow$$
$$condition_1(X)[t{:}w_1, d{:}w_1'] \vee \ldots \vee condition_m(X)[t{:}w_m, d{:}w_m']$$

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*

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

Chapter 5: 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
- Descriptive statistical measures in large databases
- Summary

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

Measuring the Central Tendency (1)

- *Mean* — (weighted) arithmetic mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- *Median* — a holistic measure
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimate the median for grouped data by interpolation:

$$median \approx L_1 + \left(\frac{n/2 - \left(\sum f\right)_{lower}}{f_{median}}\right) \cdot c$$

L_1 — lowest value of the class containing the median
n — overall number of data values
$(\Sigma f)_{lower}$ — sum of the frequencies of all classes that are lower than the median
f_{median} — frequency of the median class
c — size of the median class interval

Measuring the Central Tendency (2)

- *Mode*
  - Value that occurs most frequently in the data
  - Well suited for categorial (i.e., non-numeric) data
  - Unimodal, bimodal, trimodal, …: there are 1, 2, 3, … modes in the data (multimodal in general)
  - There is no mode if each data value occurs only once
  - Empirical formula for unimodal frequency curves that are moderately skewed:

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

- *Midrange*
  - Average of the largest and the smallest values in a data set:

$$(max - min) / 2$$

Measuring the Dispersion of Data

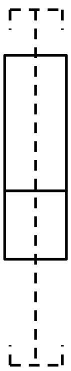
- Quartiles, outliers and boxplots
  - Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)
  - Inter-quartile range: IQR = $Q_3 - Q_1$
  - Five number summary: min, Q_1, M, Q_3, max
  - Boxplot: ends of the box are the quartiles, median is marked, whiskers (Barthaare, Backenbart), and plot outlier individually
  - Outlier: usually, values that are more than 1.5 x IQR below Q_1 or above Q_3
- Variance and standard deviation
  - Variance s^2: (algebraic, scalable computation) $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
  - Standard deviation s is the square root of variance s^2

Boxplot Analysis

- Five-number summary of a distribution:

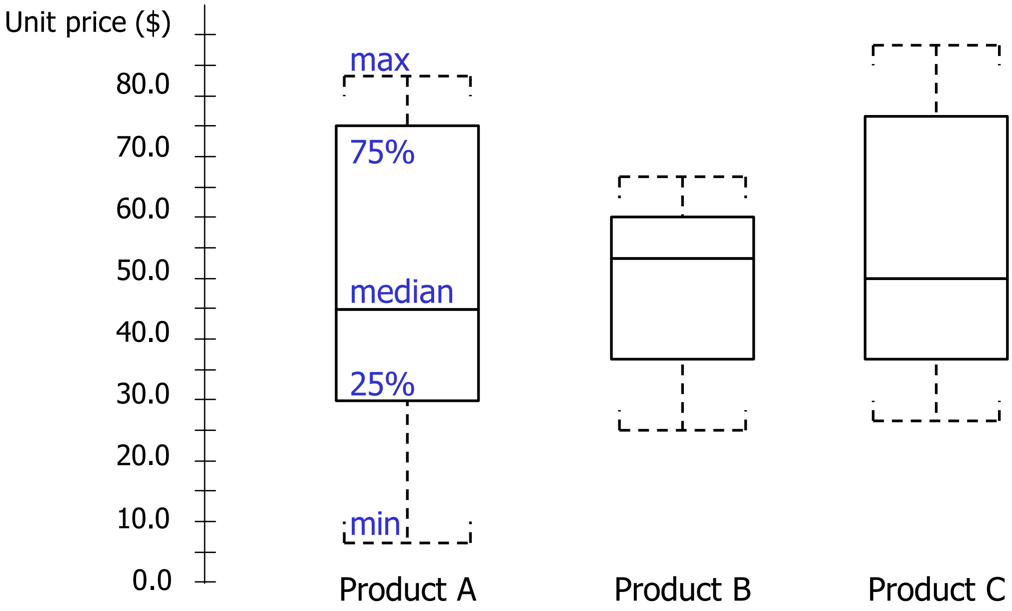
  Minimum, Q1, M, Q3, Maximum

  = (0%, 25%, 50%, 75%, 100%-quantiles)
- Boxplot

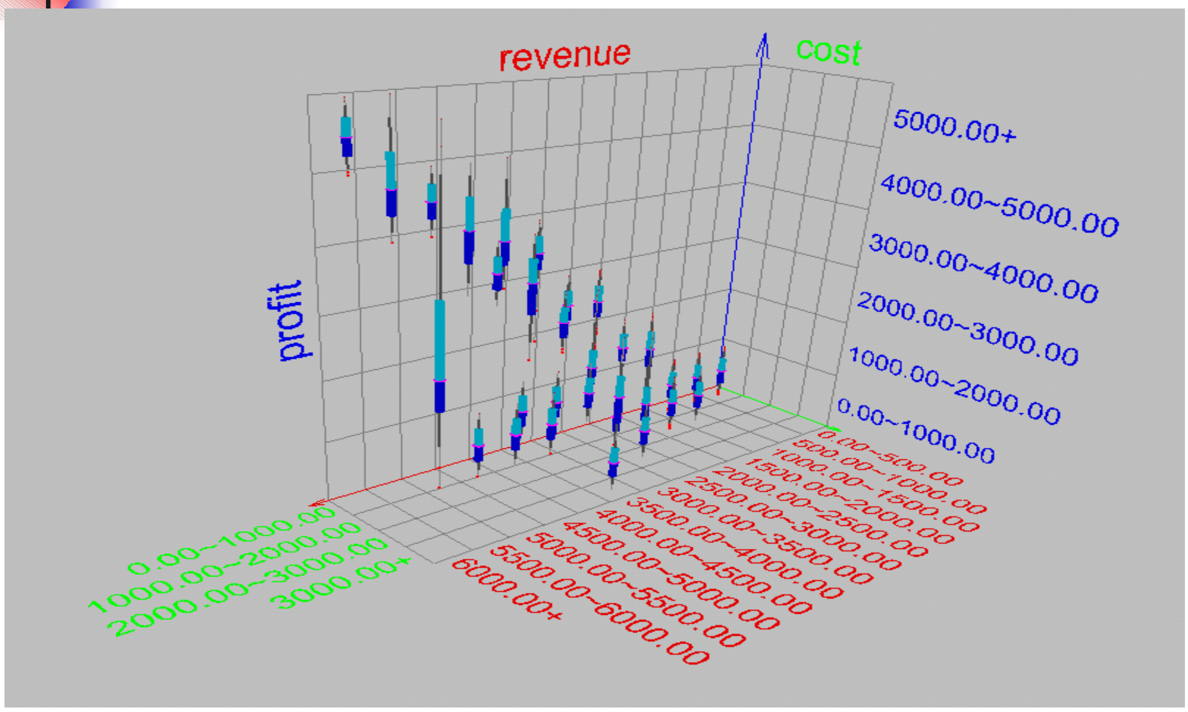


  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extend to Minimum and Maximum

Boxplot Examples



Visualization of Data Dispersion: Boxplot Analysis



Mining Descriptive Statistical Measures in Large Databases

- Variance

alternatives: $\frac{1}{n-1}, \quad \frac{1}{n}$

May be computed in a single pass!

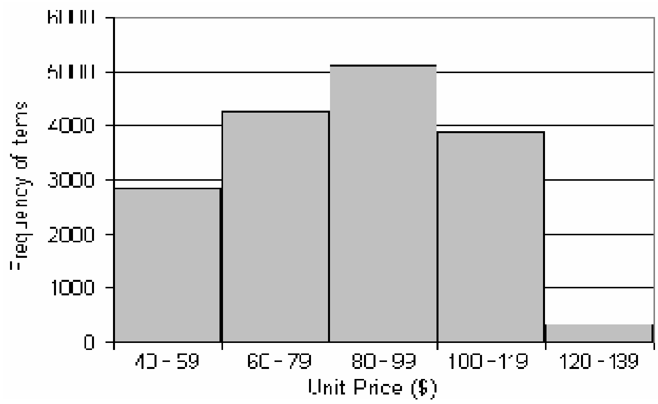
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right]$$

Requires two passes but is numerically much more stable

- Standard deviation: the square root of the variance
  - Measures the spread around the mean
  - It is zero if and only if all the values are equal
  - Both the deviation and the variance are algebraic

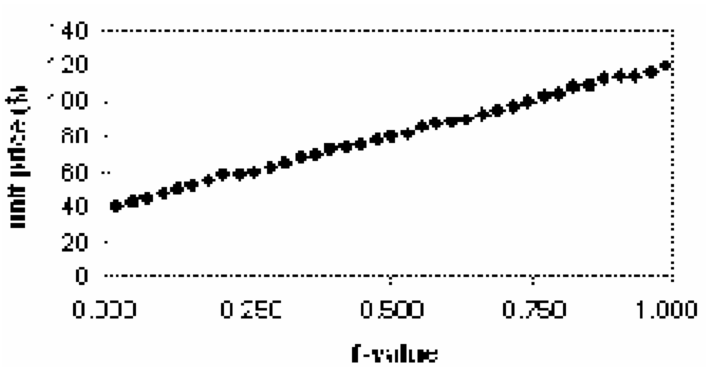
Histogram Analysis

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



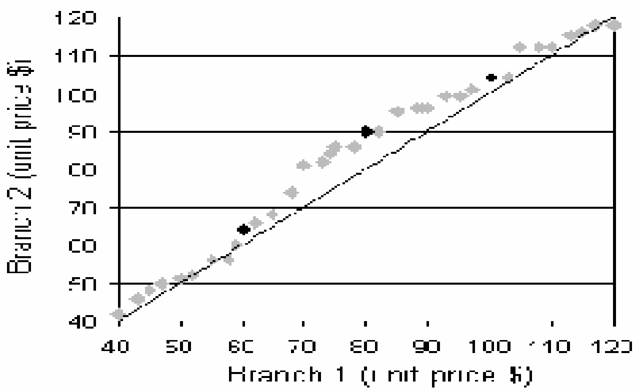
Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots quantile information
  - The q-quantile x_q indicates the value x_q for which the fraction q of all data is less than or equal to x_q (called percentile if q is a percentage); e.g., median = 50%-quantile or 50th percentile.



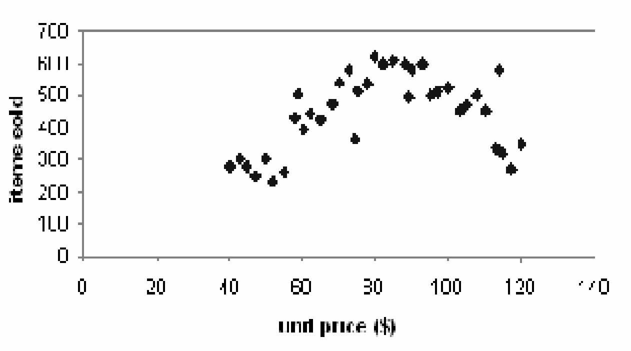
Quantile-Quantile (Q-Q) Plot

- 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



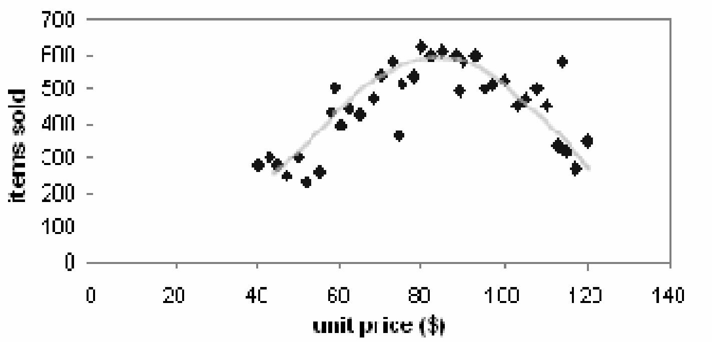
Scatter plot

- 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



Loess Curve (local regression)

- 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



Chapter 5: 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
- Descriptive statistical measures in large databases
- Summary

Summary

- Concept description: characterization and discrimination
- OLAP-based vs. attribute-oriented induction (AOI)
- Efficient implementation of AOI
- Analytical characterization and comparison
- Descriptive statistical measures in large databases

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.
- M. Ester, R. Wittmann. Incremental Generalization for Mining in a Data Warehousing Environment. Proc. Int. Conf. on Extending Database Technology, pp.135-149, 1998.
- 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.

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.