Chapter 2
Data Preprocessing

Outline

- General data characteristics
- Data cleaning
- Data integration and transformation
- Data reduction
- Summary

Types of Data Sets

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Spatial data: maps
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data

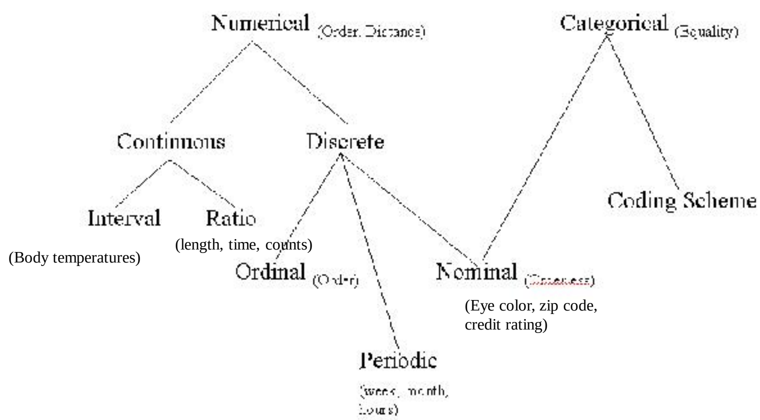
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Discrete vs. Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

Data Hierarchy



Important Characteristics of Structured Data

- Dimensionality
  - Curse of dimensionality
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Similarity
  - Distance measure

Mining Data Descriptive Characteristics

- <u>Motivation</u>
  - To better understand the data: central tendency, variation and spread
- <u>Data dispersion characteristics</u>
  - median, max, min, quantiles, outliers, variance, etc.

Measuring the Central Tendency

- <u>Mean (algebraic measure) (sample vs. population):</u>
  - Weighted arithmetic mean:
  - Trimmed mean: chopping extreme values

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad\qquad \text{«} = \frac{\sum x}{N}$$

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

Measuring the Central Tendency

- Median: A holistic measure
  - Middle value if odd number of values, or average of the middle two values otherwise
- Estimated by interpolation (for *grouped data*):
  - Group data in intervals according to x_i and record data frequency (number of unique data values).
  - Pick median interval, containing the median frequency.
  - L_1 is the lower boundary of the median interval, N is the number of values in data set. $(\Sigma freq)_l$ is the sum of frequency of all the intervals that are lower than the median interval, $freq_{median}$ is the frequency of the median interval, and width is the width of the median interval

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}}\right) width$$

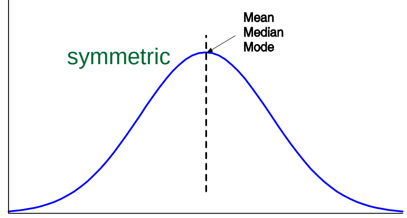
Measuring the Central Tendency

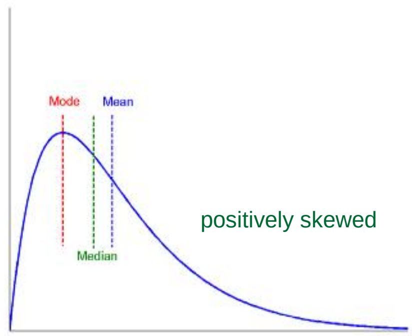
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula for unimodal frequency curves that are moderately skewed (asymmetrical) :

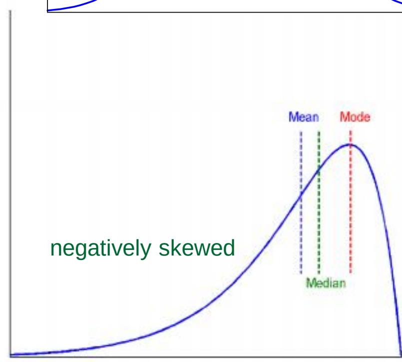
$$mean - mode = 3 \times (mean - median)$$

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data





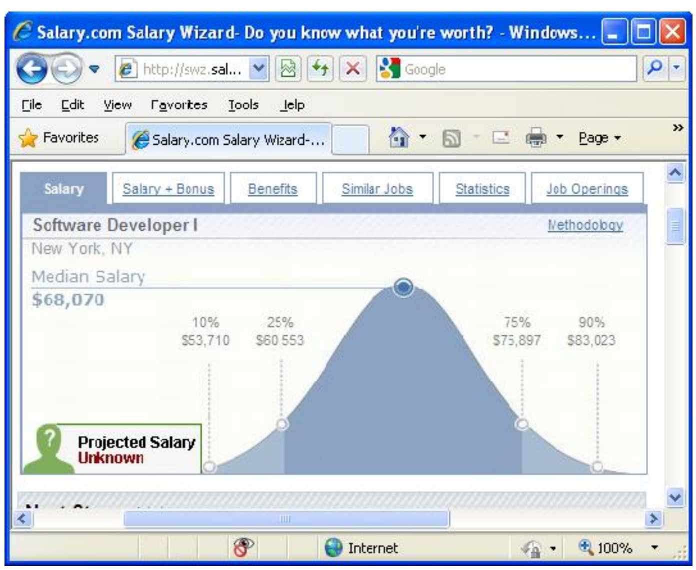


Measuring the Dispersion of Data

- The degree to which numerical data tend to spread is called the dispersion or variance of the data.
- Common measures:
  - Range
  - Interquartile range
  - Five-number summary: min, Q_1, M, Q_3, max
  - Standard deviation.

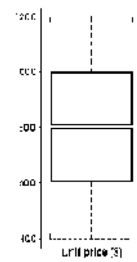
Range, Quartiles, Outliers

- Range: difference between min. and max.
- Quartiles: Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
  - The *k*th percentile of a set of data in numerical order is the value x_i having the property that k percent of the data entries lie at or below x_i.
  - Median is Q_2.
- Inter-quartile range: IQR = $Q_3 - Q_1$
- Five number summary: min, Q_1, M(Median), Q_3, max
- Outlier: usually, a value falling at least 1.5 x IQR above the Q3 or below Q1.

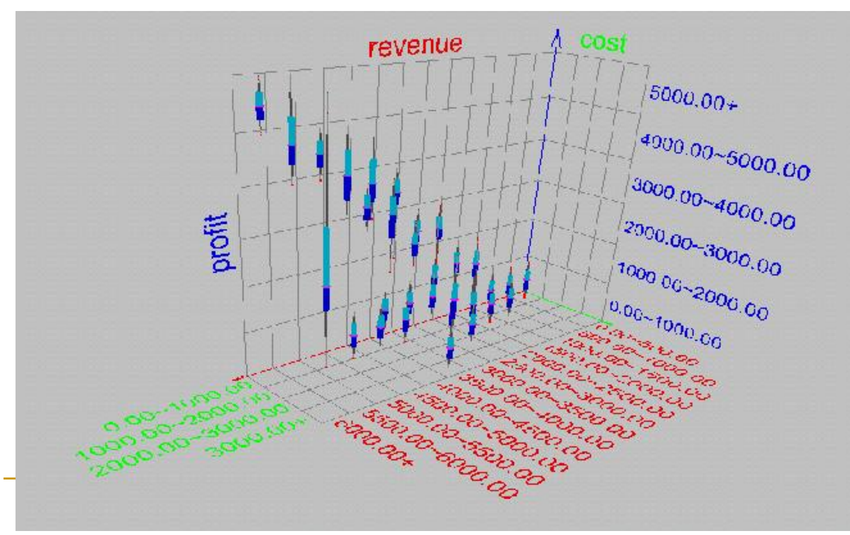


Boxplot Analysis

- Incorporates the Five-number summary of a distribution:
  - Data is represented with a box
  - The ends of the box are at the Q_1 and Q_3, 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
  - plot **outlier** individually



Visualization of Data Dispersion: 3-D Boxplots



Variance and Standard Deviation

- Variance and standard deviation (*sample: s, population: σ*)

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

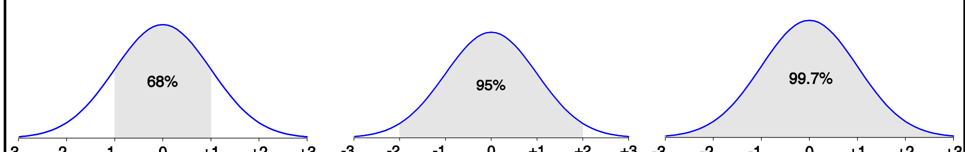
$$\pm^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \text{«})^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \text{«}^2$$

  - Variance: (algebraic, scalable computation)
  - Standard deviation *s (or σ)* is the square root of variance s^2 (or σ^2)

CISC 4631 17

Properties of Normal Distribution Curve

- The normal (distribution) curve
  - From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: contains about 99.7% of it



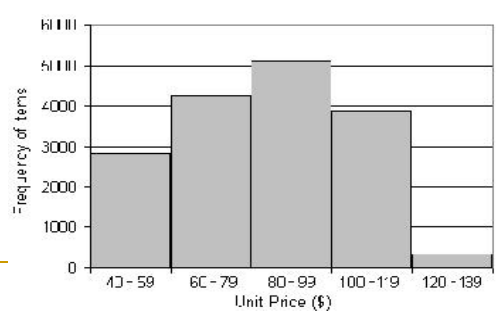
CISC 4631 18

Graphic Displays of Basic Statistical Descriptions

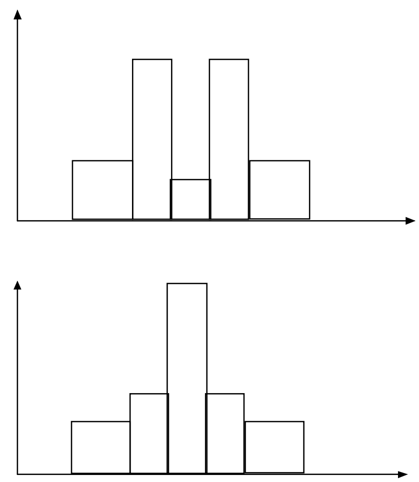
- Boxplot: graphic display of five-number summary
- Histogram: x-axis are values, y-axis repres. frequencies
- Quantile plot: each value x_i is paired with f_i indicating that approximately 100 f_i% of data are $\leq x_i$
- Quantile-quantile (q-q) plot: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane

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



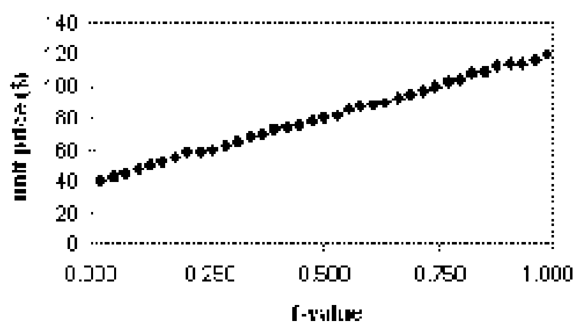
Histograms Often Tells More than Boxplots



- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

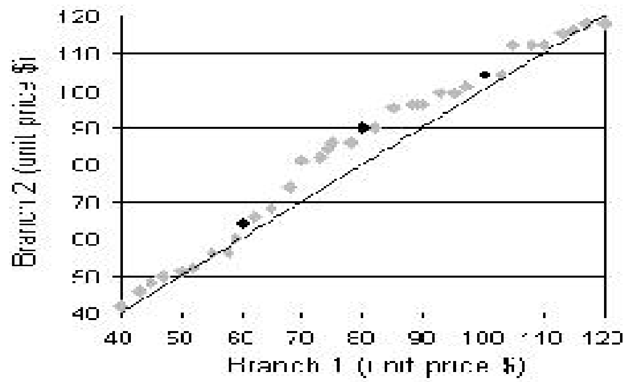
Quantile Plot

- 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

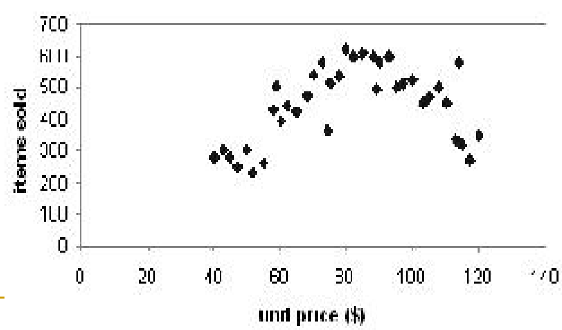
- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- Allows the user to view whether there is a shift in going from one distribution to another



23

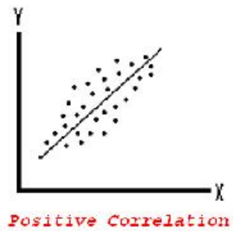
Scatter plot

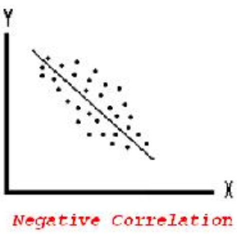
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



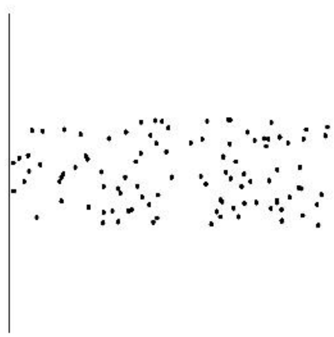
24

Positively and Negatively Correlated Data





Not Correlated Data



Outline

- General data characteristics
- Data cleaning
- Data integration and transformation
- Data reduction
- Summary

Data Cleaning

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics
  - "Data cleaning is the number one problem in data warehousing"—DCI survey
  - Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse

Major Tasks of Data Cleaning

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

Data in the Real World Is Dirty

- incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - e.g., occupation=" " (missing data)
- noisy: containing noise, errors, or outliers
  - e.g., Salary="−10" (an error)
- inconsistent: containing discrepancies in codes or names, e.g.,
  - Age="42" Birthday="03/07/1997"
  - Was rating "1,2,3", now rating "A, B, C"
  - discrepancy between duplicate records

Multi-Dimensional Measure of Data Quality

- A well-accepted multidimensional view:
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Value added
  - Interpretability
  - Accessibility
- Broad categories:
  - Intrinsic, contextual, and representational.

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
  - not register history or changes of the data
- Missing data may need to be inferred

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree (**Prediction**)

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

How to Handle Noisy Data?

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

Simple Discretization Methods: Binning

- Equal-width (distance) partitioning
  - 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
  - Divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

Binning Methods for Data Smoothing

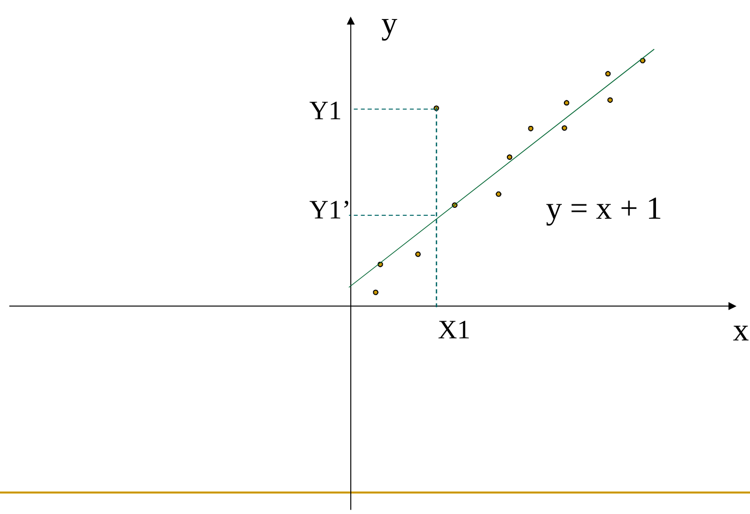
- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34

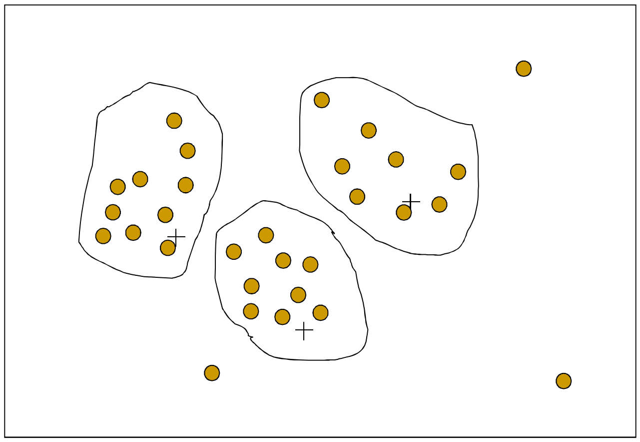
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29

* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Regression



Cluster Analysis



Outline

- General data characteristics
- Data cleaning
- Data integration and transformation
- Data reduction
- Summary

Data Integration

- Data integration:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- 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. British units

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality
- Redundant attributes may be able to be detected by *correlation analysis*

Correlation Analysis (Numerical Data)

- Correlation coefficient (also called *Pearson's product moment coefficient*)

$$r_{p,q} = \frac{\sum (p - \bar{p})(q - \bar{q})}{(n-1)\sigma_p \sigma_q} = \frac{\sum (pq) - n\bar{p}\bar{q}}{(n-1)\sigma_p \sigma_q}$$

where n is the number of tuples, $\bar{p}$ and $\bar{q}$ are the respective means of p and q, σ_p and σ_q are the respective standard deviation of p and q, and Σ(pq) is the sum of the pq cross-product.

- If $r_{p,q} > 0$, p and q are positively correlated (p's values increase as q's). The higher, the stronger correlation.
- $r_{p,q} = 0$: independent; $r_{pq} < 0$: negatively correlated

Correlation Analysis (Categorical Data)

- X^2 (chi-square) test

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The larger the X^2 value, the more likely the variables are related
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count
- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

X^2 (chi-square) Test

- χ^2 term-category independency test:

$$E(i,j) = \frac{\sum_{a \in \{w, \neg w)} O(a,j) \times \sum_{b \in \{c, \neg c)} O(i,b)}{N}$$

	c	$\neg c$	Σ
w	40	80	120
$\neg w$	60	320	380
Σ	100	400	500 (N)

A 2-way contingency table

$$\chi^2_{w,c} = \sum_{i \in \{w, \neg w\}} \sum_{j \in \{c, \neg c\}} \frac{(O(i,j) - E(i,j))^2}{E(i,j)}$$

$$\chi^2_{w,c} = 17.61$$

Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- It shows that like_science_fiction and play_chess are correlated in the group

Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values
- Methods
  - Smoothing: Remove noise from data
  - Aggregation: Summarization (*annual total*), data cube construction
  - Generalization: Concept hierarchy climbing (*street* -> *city*)
  - Normalization: Scaled to fall within a small, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling
  - Attribute/feature construction
    - New attributes constructed from the given ones

Min-max Normalization

- Min-max normalization: to [new_min$_A$, new_max$_A$]

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

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to

$$\frac{73,600 - 12,000}{98,000 - 12,000}(1.0 - 0) + 0 = 0.716$$

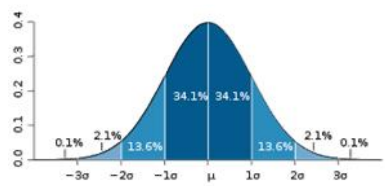
- If a future input falls outside of the original data range?

Z-score Normalization

- Z-score normalization (μ: mean, σ: standard deviation): $v' = \frac{v - \ll_A}{\pm_A}$
  - Ex. Let μ = 54,000, σ = 16,000. Then

$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

- When actual min. and max. values are unknown or there is outliers.



Decimal Normalization

- Normalization by decimal scaling

$$v' = \frac{v}{10^j}$$

Where j is the smallest integer such that Max(|v'|) < 1

- -986 to 917 => -0.986 to 0.917

Outline

- General data characteristics
- Data cleaning
- Data integration and transformation
- Data reduction
- Summary

Data Reduction

- 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

Dimensionality Reduction

- Curve of dimensionality
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially
- Dimensionality reduction
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualization

Dimensionality Reduction Techniques

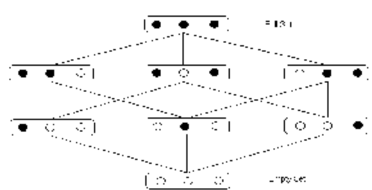
- Feature selection
  - Select *m* from *n* features, $m \leq n$
  - Remove *irrelevant*, *redundant* features
  - Saving in search space
- Feature transformation
  - Form new features (**a**) in a new domain from original features (**f**)

Feature Selection

- Redundant features
  - duplicate much or all of the information contained in one or more other attributes
  - E.g., purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - contain no information that is useful for the data mining task at hand
  - E.g., students' ID is often irrelevant to the task of predicting students' GPA

Feature Selection

- Problem illustration
  - Full set
  - Empty set
  - Enumeration



Feature Selection (2)

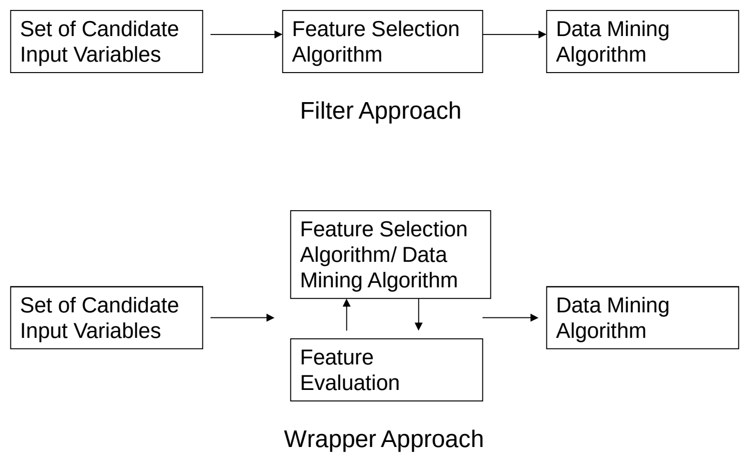
- Goodness metrics
  - Dependency: dependence on classes
  - Distance: separating classes
  - Information: entropy
  - Consistency:
    - Inconsistency Rate - #inconsistencies/N
    - Example: (F1, F2, F3) and (F1,F3)
    - Both sets have 2/6 inconsistency rate
  - Accuracy (classifier based): 1 - errorRate

F 1	F 2	F 3	C
0	0	1	1
0	0	1	0
0	0	1	1
1	0	0	1
1	0	0	0
1	0	0	0

Feature Subset Selection Techniques

- Brute-force approach:
  - Try all possible feature subsets as input to data mining algorithm
- Embedded approaches:
  - Feature selection occurs naturally as part of the data mining algorithm
- Filter approaches:
  - Features are selected before data mining algorithm is run
- Wrapper approaches:
  - Use the data mining algorithm as a black box to find best subset of attributes

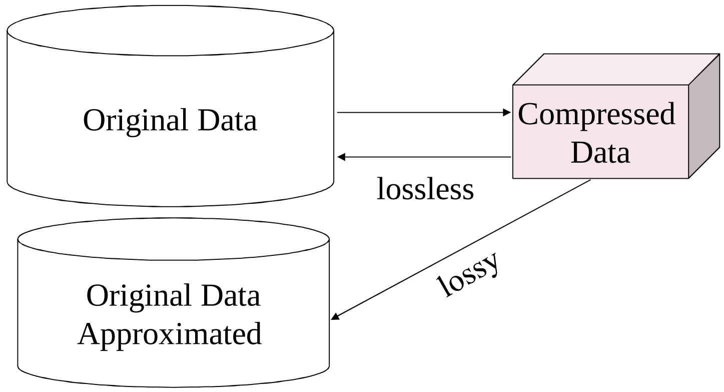
Filter vs. Wrapper Model



Data Compression

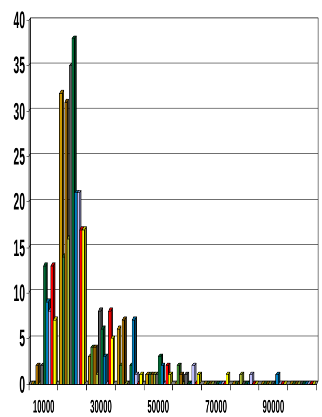
- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless
  - But only limited manipulation is possible without expansion
- Audio/video compression
  - Typically lossy compression, with progressive refinement
  - Sometimes small fragments of signal can be reconstructed without reconstructing the whole

Data Compression



Data Reduction: 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 β–1 largest differences



Data Reduction Method: 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 organized into distinct clusters 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 later

Data Reduction Method: 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
- Key principle: Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
  - Develop adaptive sampling methods, e.g., stratified sampling.
- Note: Sampling may not reduce database I/Os (page at a time)

Types of Sampling

- Simple random sampling (SRS)
  - There is an equal probability of selecting any particular item
- Sampling without replacement (SRSWOR)
  - Once an object is selected, it is removed from the population
- Sampling with replacement (SRSWR)
  - A selected object is not removed from the population
- Cluster sample
  - First get **M** clusters, then SRS **s** clusters, **s < M**.
- Stratified sampling:
  - Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
  - Used in conjunction with skewed data

Data Reduction: Discretization

- 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
  - Numerical — real numbers, e.g., integer or real numbers
- Discretization:
  - Divide the range of a numerical attribute into intervals
  - Some classification algorithms only accept categorical attributes.
  - Reduce data size by discretization
  - Prepare for further analysis

Discretization and Concept Hierarchy

- Discretization
  - Reduce the number of values for a given numerical 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)

Discretization and Concept Hierarchy Generation for Numerical Data

- Typical methods: All the methods can be applied recursively
  - Binning (covered above)
    - Top-down split, unsupervised,
  - Histogram analysis (covered above)
    - Top-down split, unsupervised
  - Clustering analysis (covered above)
    - Either top-down split or bottom-up merge, unsupervised
  - Entropy-based discretization: supervised, top-down split
  - Interval merging by χ^2 Analysis: unsupervised, bottom-up merge
  - Segmentation by natural partitioning: top-down split, unsupervised

Entropy-Based Discretization

- In information theory, entropy quantifies, in the sense of an expected value, the information in a message.
- Equivalently, entropy is a measure of the average information content which is missing when the value of the random variable in unknown.
- To discretize a numerical attribute A, the method selects the value of A that has the min. entropy as a split-point, and do this recursively.

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

$$I(S,T) = \frac{|S_1|}{|S|} Entropy(S_1) + \frac{|S_2|}{|S|} Entropy(S_2)$$

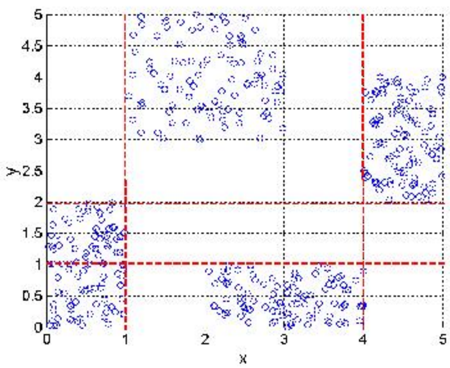
- Entropy is calculated based on class distribution of the samples in the set. Given *m* classes, the entropy of S_1 is

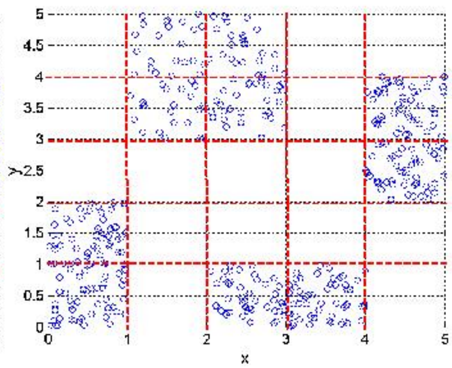
$$Entropy(S_1) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

  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.

Discretization Using Class Labels

- **Entropy based approach**

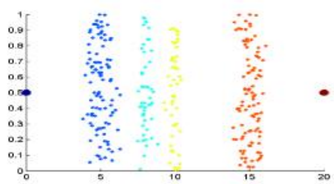




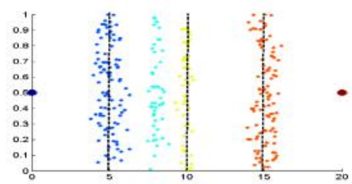
3 categories for both x and y

5 categories for both x and y

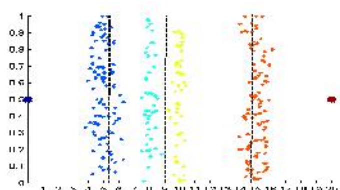
Discretization Without Using Class Labels



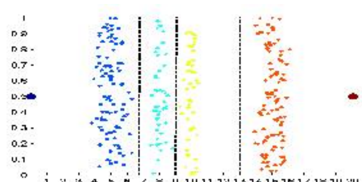
Data



Equal interval width



Equal frequency



K-means

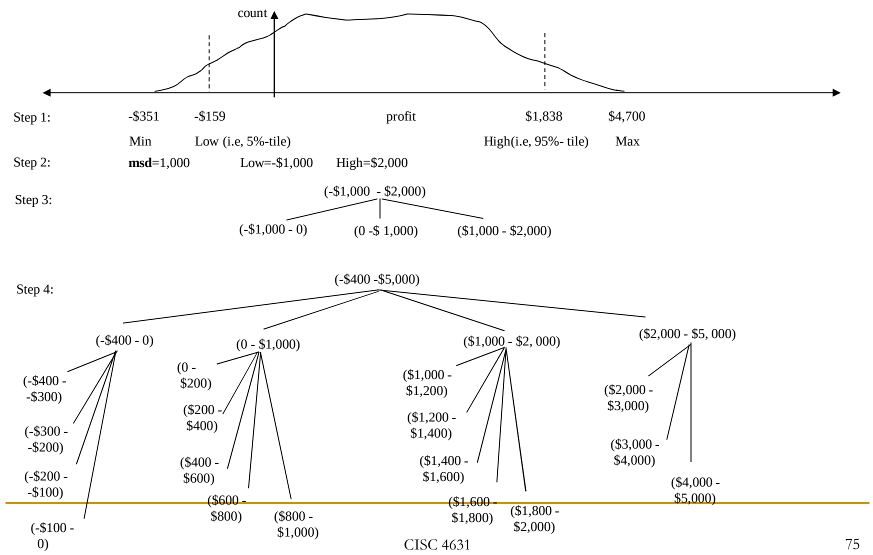
Interval Merge by χ^2 Analysis

- 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
  - χ^2 tests are performed for every pair of adjacent intervals
  - Adjacent intervals with the least χ^2 values are merged together, since low χ^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.)

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 (msd), 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

Example of 3-4-5 Rule

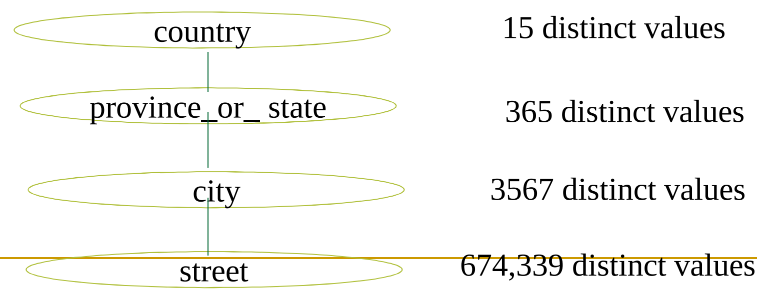


Concept Hierarchy Generation for Categorical Data

- 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 a set of attributes, but not their orders.
  - Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values
- Specification of only a partial set of attributes
  - E.g., only street < city, not others

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



Outline

- General data characteristics
- Data cleaning
- Data integration and transformation
- Data reduction
- Summary

Summary

- Data preparation/preprocessing: A big issue for data mining
- Data description, data exploration, and measure data similarity set the base for quality data preprocessing
- Data preparation includes
  - Data cleaning
  - Data integration and data transformation
  - Data reduction (dimensionality and numerosity reduction)
- A lot a methods have been developed but data preprocessing still an active area of research