

Data Preprocessing

Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization
- Summary

Why Data Preprocessing?

- Data in the real world is dirty
 - **incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
 - e.g., occupation=" "
 - **noisy**: containing errors or outliers
 - e.g., Salary="-10"
 - **inconsistent**: containing discrepancies in codes or names
 - e.g., Age="42" Birthday="03/07/1997"
 - e.g., Was rating "1,2,3", now rating "A, B, C"
 - e.g., discrepancy between duplicate records

Why Is Data Dirty?

- Incomplete data may come from
 - “Not applicable” data value when collected
 - Different considerations between the time when the data was collected and when it is analyzed.
 - Human/hardware/software problems
- Noisy data (incorrect values) may come from
 - Faulty data collection instruments
 - Human or computer error at data entry
 - Errors in data transmission
- Inconsistent data may come from
 - Different data sources
 - Functional dependency violation (e.g., modify some linked data)
- Duplicate records also need data cleaning

Why Is Data Preprocessing Important?

- 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 warehouse needs consistent integration of quality data
- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse

Multi-Dimensional Measure of Data Quality

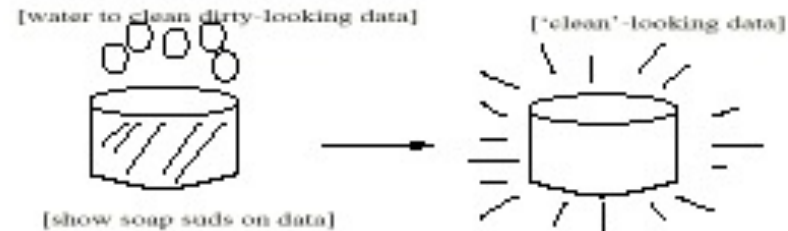
- A well-accepted multidimensional view:
 - Accuracy
 - Completeness
 - Consistency
 - Timeliness
 - Believability
 - Value added
 - Interpretability
 - Natural,
 - Representational
 - Accessibility

Major Tasks in Data Preprocessing

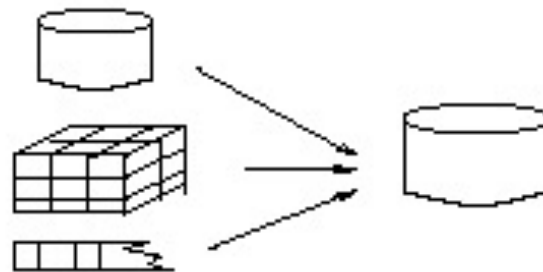
- **Data cleaning**
 - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
 - Integration of multiple databases, data cubes, or files
- **Data transformation**
 - Normalization and aggregation
- **Data reduction**
 - Obtains reduced representation in volume but produces the same or similar analytical results
- **Data discretization**
 - Part of data reduction but with particular importance, especially for numerical data

Forms of Data Preprocessing

Data Cleaning



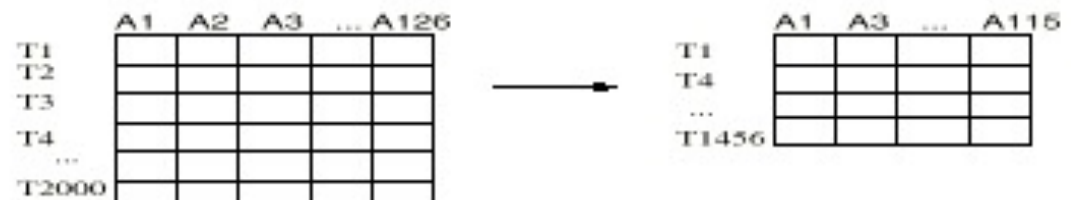
Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization
- Summary

Mining Data Descriptive Characteristics

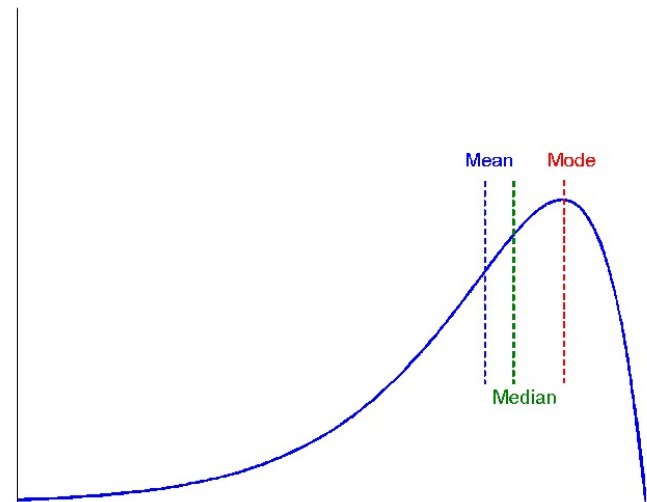
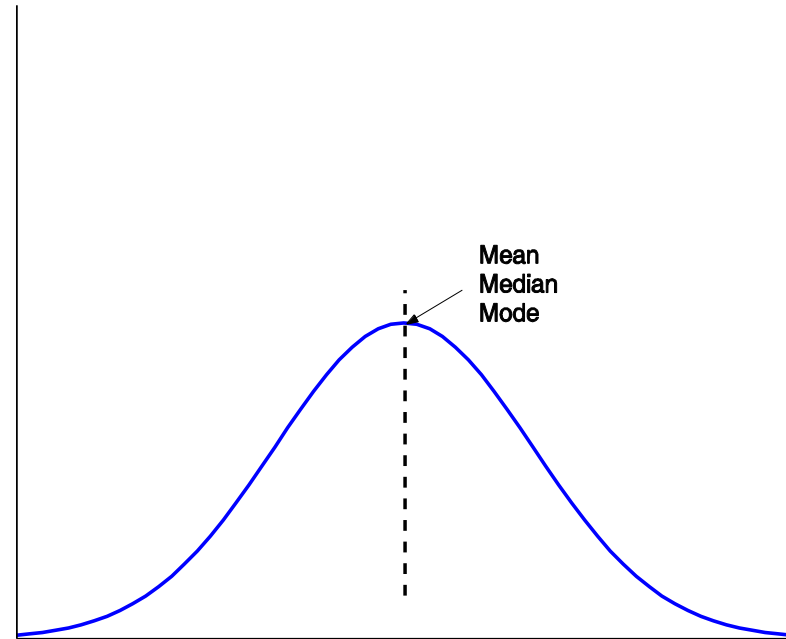
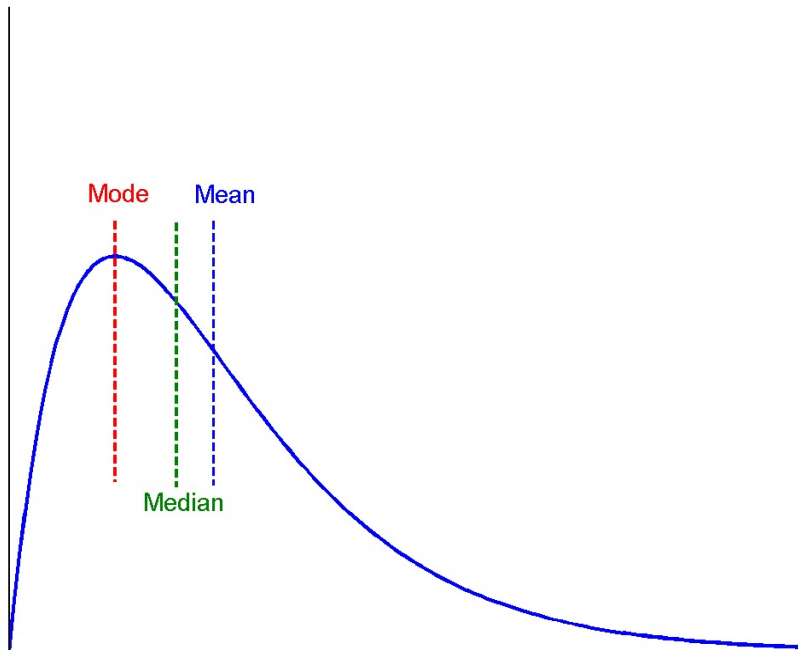
- Motivation
 - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
 - median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
 - Data dispersion: analyzed with multiple granularities of precision
 - 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

- Mean (algebraic measure) (sample vs. population): $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ $\mu = \frac{\sum x}{N}$
 - Weighted arithmetic mean:
$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$
 - Trimmed mean: chopping extreme values
- Median: A holistic measure
 - Middle value if odd number of values, or average of the middle two values otherwise
- Mode
 - Value that occurs most frequently in the data
 - Unimodal, bimodal, trimodal

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, right and left skewed data(tail's direction)

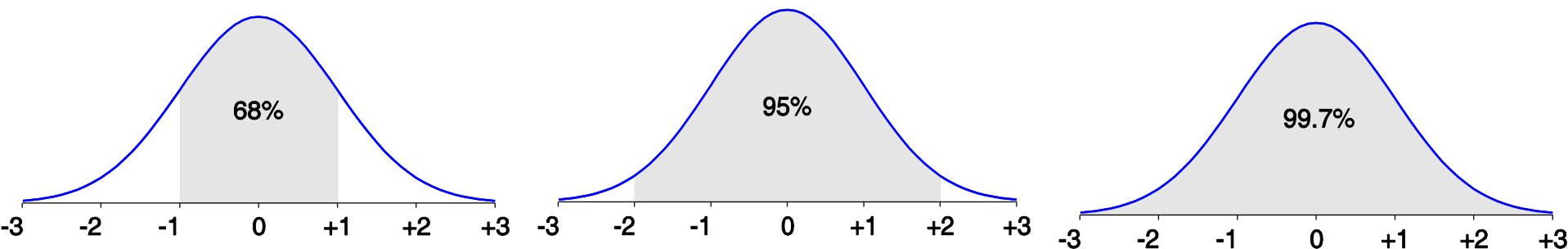


Measuring the Dispersion of Data

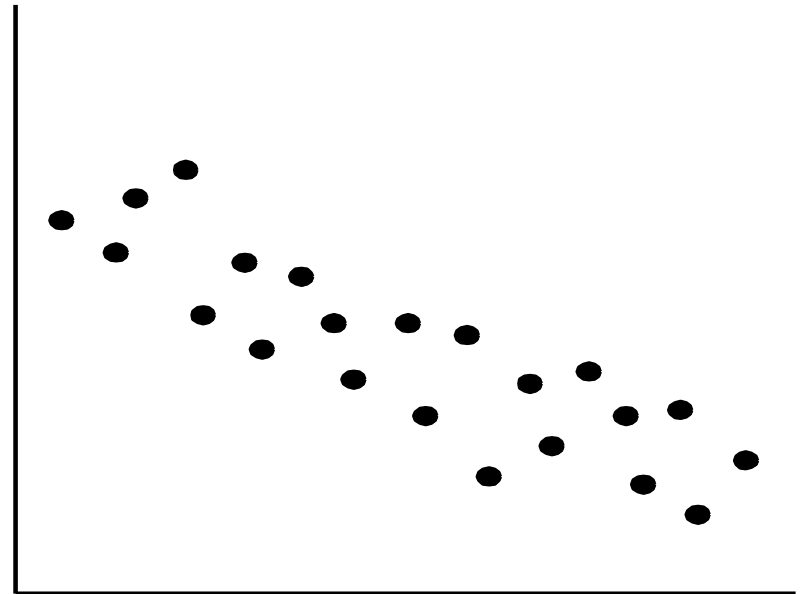
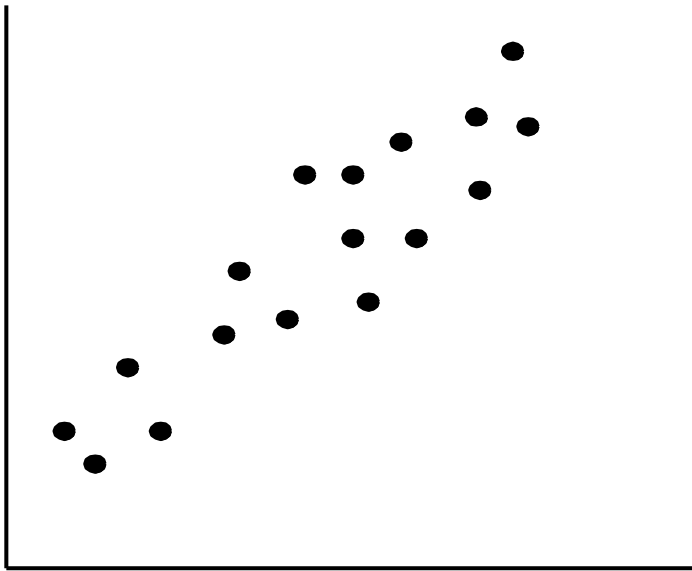
- Quartiles, outliers and boxplots
 - **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
 - **Inter-quartile range**: $IQR = Q_3 - Q_1$
 - **Five number summary**: min, Q_1 , M, Q_3 , max
 - **Boxplot**: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually
 - **Outlier**: usually, a value higher/lower than $1.5 \times IQR$
- Variance and standard deviation (*sample: s , population: σ*)
 - **Variance**: (algebraic, scalable computation)
$$\sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^n x_i^2 - \mu^2$$
$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right]$$
 - **Standard deviation** s (*or* σ) is the square root of variance s^2 (*or* σ^2)

Properties of Normal Distribution Curve

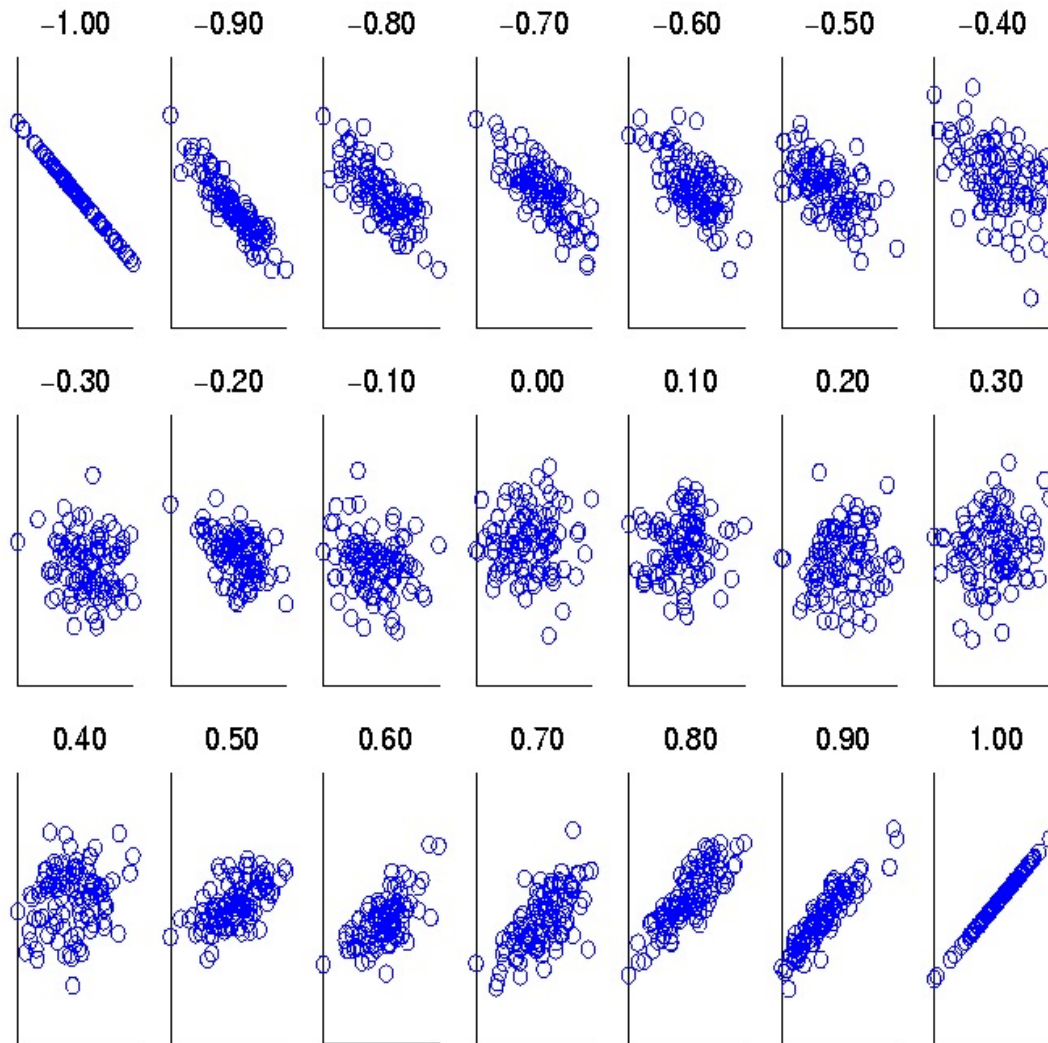
- The normal (distribution) curve
 - From $\mu - \sigma$ to $\mu + \sigma$: contains about 68% of the measurements (μ : mean, σ : standard deviation)
 - From $\mu - 2\sigma$ to $\mu + 2\sigma$: contains about 95% of it
 - From $\mu - 3\sigma$ to $\mu + 3\sigma$: contains about 99.7% of it



Positively and Negatively Correlated Data



Korelasyonu görsel değerlendirme



Scatter plots
showing the
similarity from
-1 to 1.

Korelasyon(Correlation)

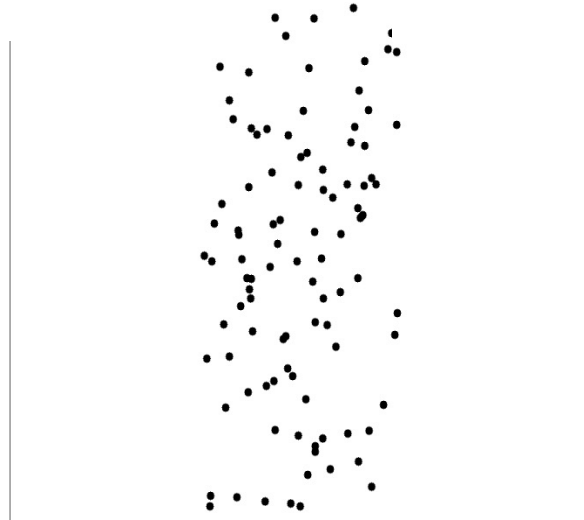
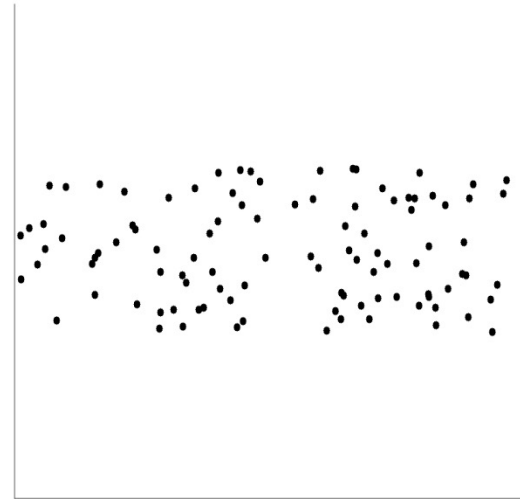
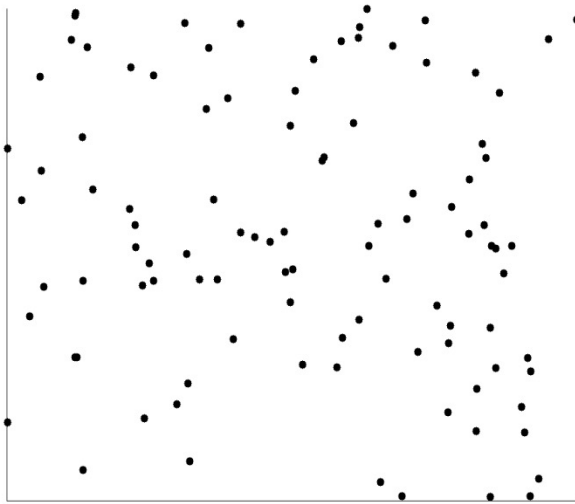
- Korelasyon, objeler arasındaki lineer ilişki ölçütlerini ifade eder.
- Korelasyonu hesaplamak için, p ve q data objelerini standardize edip dot product alırız.

$$p'_k = (p_k - \text{mean}(p)) / \text{std}(p)$$

$$q'_k = (q_k - \text{mean}(q)) / \text{std}(q)$$

$$\text{correlation}(p, q) = p' \bullet q'$$

Not Correlated Data



Benzerlik(Similarity)

- Benzerlik(Similarity)
 - İki objenin benzerliğinin sayısal değeri
 - Yüksek değer daha çok benzerlik ifade eder
 - Genellikle $[0,1]$ aralık değerlerindedir
 - Farklılık(Dissimilarity) tam tersini ifade eder

Basit Attributeler için Similarity/Dissimilarity

p ve q veri iki objesi için attribute değerlerdir.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{ p-q }{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{ p-q }{n-1}$
Interval or Ratio	$d = p - q $	$s = -d, s = \frac{1}{1+d} \text{ or } s = 1 - \frac{d - \min_d}{\max_d - \min_d}$

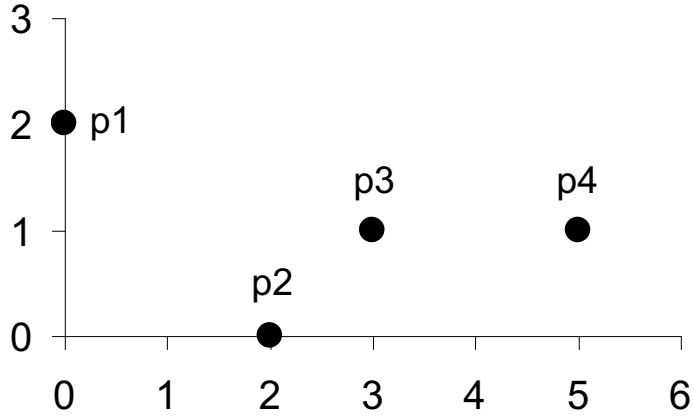
Benzerlik Ölçüleri

- Öklit uzaklığı(Euclidean distance)

$$\mathit{dist} = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

- n boyut(attribute) sayısını, p_k ve q_k sırasıyla, p ve q objelerinin k'ninci değerlerini ifade eder.

Öklit uzaklığı (Euclidean Distance)



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Uzaklık Matrisi

Minkovski uzaklığı

- Minkovski uzaklığı, Öklit uzaklığının genelleştirilmiş versiyonudur.

$$\mathit{dist} = \left(\sum_{k=1}^n |p_k - q_k|^r \right)^{\frac{1}{r}}$$

- r bir parametre olsun, n boyut(attribute) sayısını, p_k ve q_k sırasıyla, p ve q objelerinin k 'ninci değerlerini ifade eder.

Minkovski uzaklığı

- $r = 1$. City block(Manhattan, taxicab, L_1 norm) distance.
 - Hamming distance bunun genel kullanım örneklerindendir, iki binary vektör arası uzaklığı bulur.
- $r = 2$. Euclidean distance
- $r \rightarrow \infty$. “supremum” (L_{\max} norm, L_{∞} norm) distance.
 - Bu vektörlerin değerleri arasındaki maksimum farkı ifade eder
- r ile n karıştırılmamalıdır, bütün bu uzaklıklar tüm boyutlar için tanımlanmıştır.

Minkovski uzaklığı

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_{∞}	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Uzaklık Matrisi

Uzaklıkların Ortak özellikleri

Öklit uzaklığı gibi uzaklıkların bazı temel özellikleri vardır:

1. $d(p, q) \geq 0$ tüm p ve q için ve $d(p, q) = 0$ sadece $p = q$. (Positive definiteness)
2. $d(p, q) = d(q, p)$ tüm p ve q için. (Symmetry)
3. $d(p, r) \leq d(p, q) + d(q, r)$ tüm p, q , ve r noktaları için. (Triangle Inequality)

$d(p, q)$, p ve q noktaları(veri objeleri) için uzaklık değeridir.

Bu özellikleri sağlayan tüm uzaklık değerlerine metrik(metric) denir.

Binary Vektörler arası benzerlik

- p ve q objelerinin binary attributeler içeren vektörler olarak ifade edilmesi yaygındır.

- Bu değerlerin benzerlikleri hesaplınsın

M_{01} = p 'nin 0 ve q 'nın 1 olduğu değerlerin sayısı olsun

M_{10} = p 'nin 1 ve q 'nin 0 olduğu değerlerin sayısı olsun

M_{00} = p 'nin 0 ve q 'nin 0 olduğu değerlerin sayısı olsun

M_{11} = p 'nin 1 ve q 'nin 1 olduğu değerlerin sayısı olsun

- Simple Matching ve Jaccard Coefficients

SMC = number of matches / number of attributes

$$= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$$

Jaccard = number of 11 matches / number of not-both-zero attributes values

$$= (M_{11}) / (M_{01} + M_{10} + M_{11})$$

SMC ve Jaccard Örneği

$$p = 1000000000$$

$$q = 0000001001$$

$$M_{01} = 2 \quad (\text{p'nin 0 ve q'nun 1 olduğu değerlerin sayısı})$$

$$M_{10} = 1 \quad (\text{p'nin 1 ve q'nun 0 olduğu değerlerin sayısı})$$

$$M_{00} = 7 \quad (\text{p'nin 0 ve q'nun 0 olduğu değerlerin sayısı})$$

$$M_{11} = 0 \quad (\text{p'nin 1 ve q'nun 1 olduğu değerlerin sayısı})$$

$$SMC = (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$$

$$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$$

Cosine Benzerliği

d_1 ve d_2 iki doküman vektörü olsun

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| ||d_2|| ,$$

Vektörler arası dot product demektir. $||d||$ d vektörü uzunluğunu ifade eder.

örnek:

$$d_1 = \mathbf{3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0}$$

$$d_2 = \mathbf{1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2}$$

$$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$||d_1|| = (3*3 + 2*2 + 0*0 + 5*5 + 0*0 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{0.5} = (42)^{0.5} = 6.481$$

$$||d_2|| = (1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 2*2)^{0.5} = (6)^{0.5} = 2.245$$

$$\cos(d_1, d_2) = 0.3150$$

Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization
- Summary

Data Cleaning

- Importance
 - Data cleaning is one of the biggest problems in data analysis
- Data cleaning tasks
 - Fill in missing values
 - Identify outliers and smooth out noisy data
 - Correct inconsistent data
 - Resolve redundancy caused by data integration

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 (assuming the tasks in classification—not effective when the percentage 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

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
 - technology limitation
 - inconsistency in naming convention
- Other data problems which requires data cleaning
 - duplicate records
 - incomplete data
 - inconsistent data

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

Then, 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

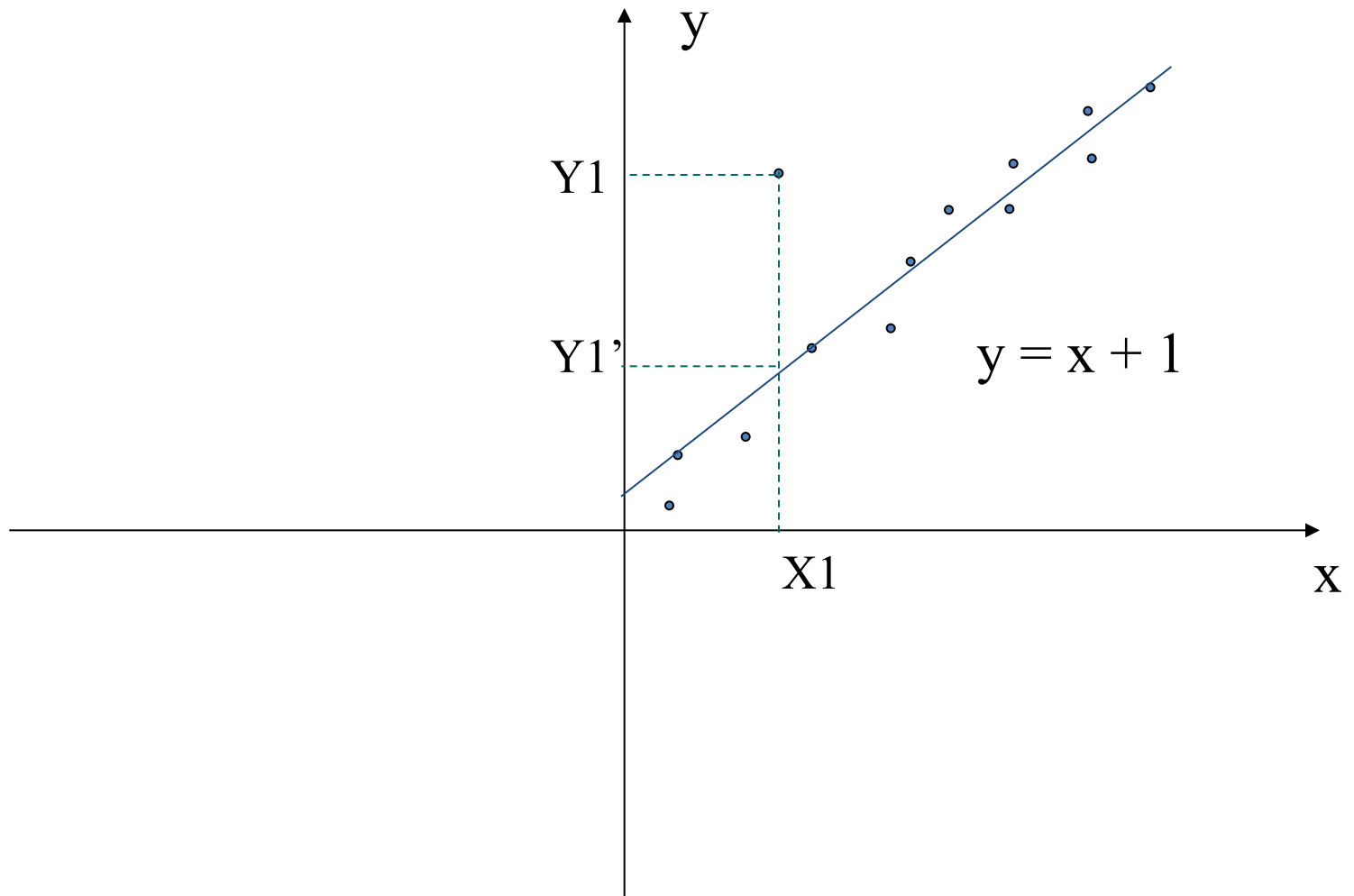
* Smoothing by bin means:

- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

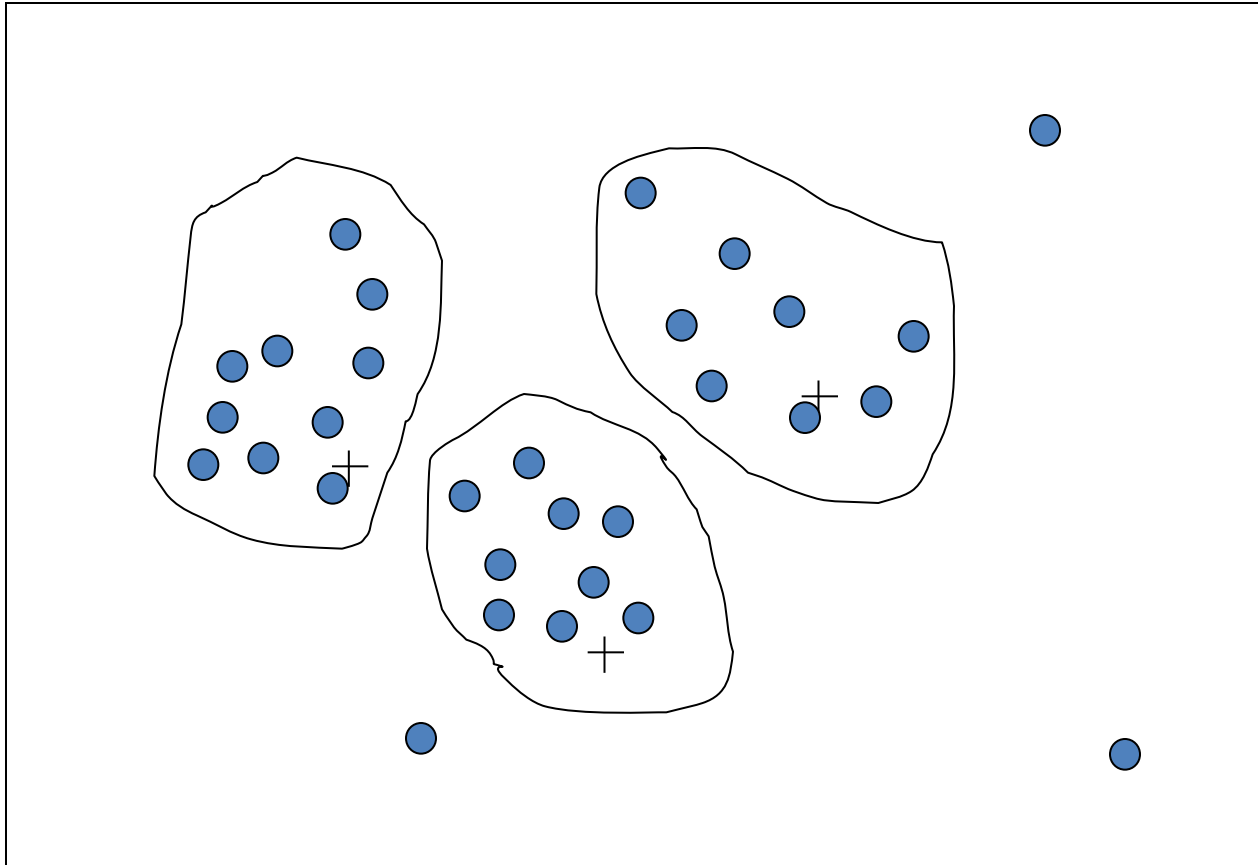
* Smoothing by bin boundaries:

- Bin 1: 4, 4, 4, 15
- Bin 2: 21, 21, 25, 25
- Bin 3: 26, 26, 26, 34

Regression



Cluster Analysis



Data Preprocessing

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

Data Integration

- Data integration:
 - Combines data from multiple sources into a coherent store
- Schema integration: e.g., $A.cust-id \equiv B.cust-\#$
 - 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
- Redundant attributes may be able to be detected by *correlation analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization
- Generalization: concept hierarchy climbing
- 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

Data Transformation: 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 $\frac{73,000 - 12,000}{98,000 - 12,000} (1.0 - 0) + 0 = 0.709$

- Z-score normalization (μ : mean, σ : standard deviation): $v' = \frac{v - \mu_A}{\sigma_A}$

- Ex. Let $\mu = 54,000$, $\sigma = 16,000$. Then $\frac{73,000 - 54,000}{16,000} = 1.188$

- Normalization by decimal scaling

$$v' = \frac{v}{10^j} \quad \text{Where } j \text{ is the smallest integer such that } \text{Max}(|v'|) < 1$$

Data Preprocessing

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

Data Reduction Strategies

- Why data reduction?
 - You may need to process 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
- **Data reduction strategies**
 - Data cube aggregation:
 - Dimensionality reduction — e.g., remove unimportant attributes
 - Data Compression
 - Numerosity reduction — e.g., fit data into models
 - Discretization and concept hierarchy generation

Data Preprocessing

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

Discretization

- Discretization
 - Reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals
 - Interval labels can then be used to replace actual data values
 - Reduce data size by discretization
 - Also, some classification algorithms only accept categorical attributes.

Discretization and Concept Hierarchy

- 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)