Data Preprocessing

Vandana Jha

Data Preprocessing

- Data Preprocessing: An Overview
- Data Cleaning
- Data Integration
- Data Reduction and Transformation
- Dimensionality Reduction
- Summary

Major tasks

Data Preprocessing

- Data Preprocessing: An Overview
- Data Cleaning
- Data Integration
- Data Reduction and Transformation
- Dimensionality Reduction
- Summary

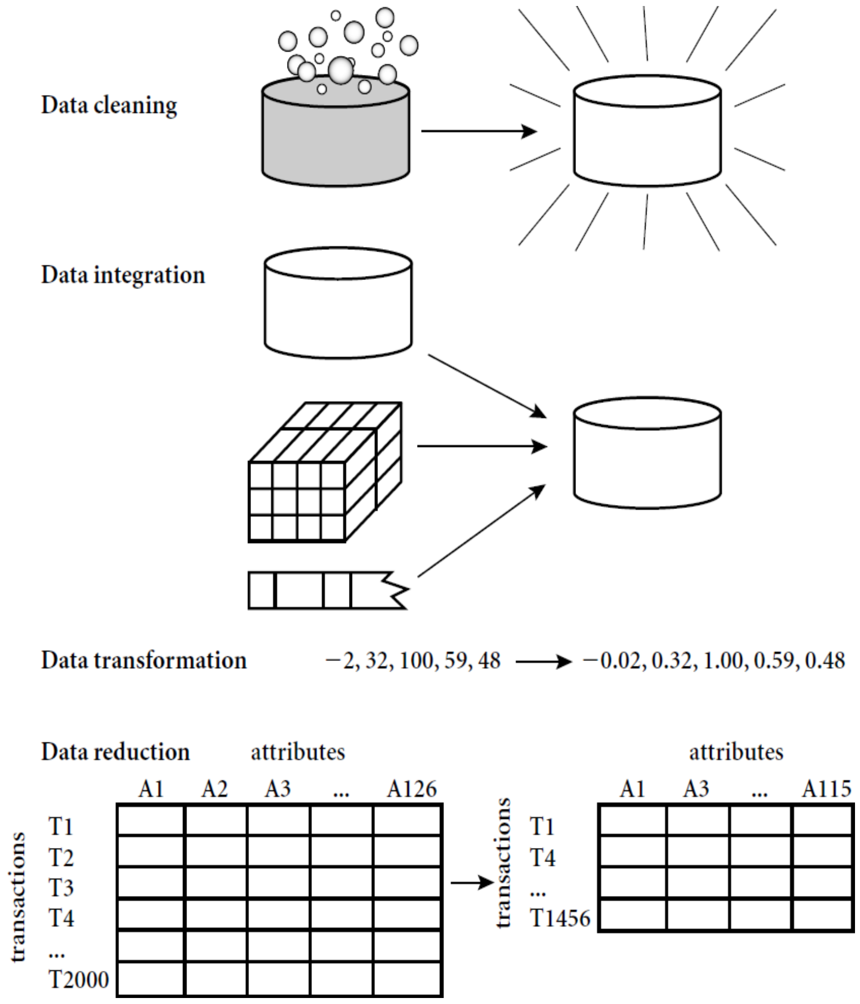
Major tasks

Why Preprocess the Data? — Data Quality Issues

- Measures for data quality: A multidimensional view
  - **Accuracy**: correct or wrong, accurate or not
  - **Completeness**: not recorded, unavailable, …
  - **Consistency**: some modified but some not, dangling, …
  - **Timeliness**: timely update?
  - **Interpretability**: how easy the data can be understood?
  - **Trustworthiness**: how trustable the data are correct?

What is Data Preprocessing? — Major Tasks

- **Data cleaning**
  - Handle missing data, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation



Data Preprocessing

- Data Preprocessing: An Overview
- Data Cleaning
- Data Integration
- Data Reduction and Transformation
- Dimensionality Reduction
- Summary

Incomplete (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
- Various reasons for missing data:
  - Equipment malfunction
  - Inconsistent with other recorded data and thus deleted
  - Data were not entered due to misunderstanding
  - Certain data may not be considered important at the time of entry
  - Did not register history or changes of the data
- Missing data may need to be inferred

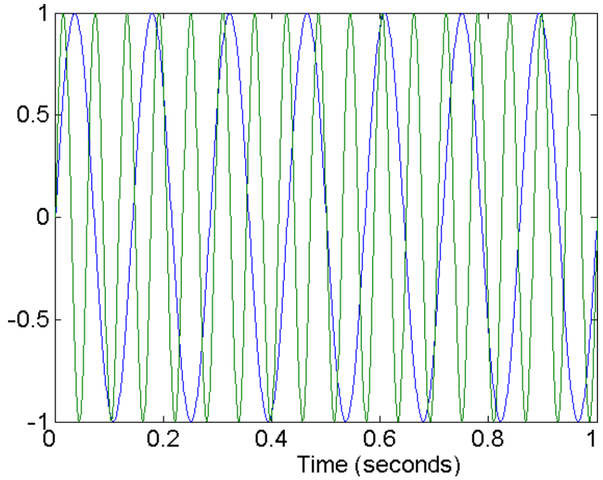
How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)— not effective when the percentage of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Automatically fill it in 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

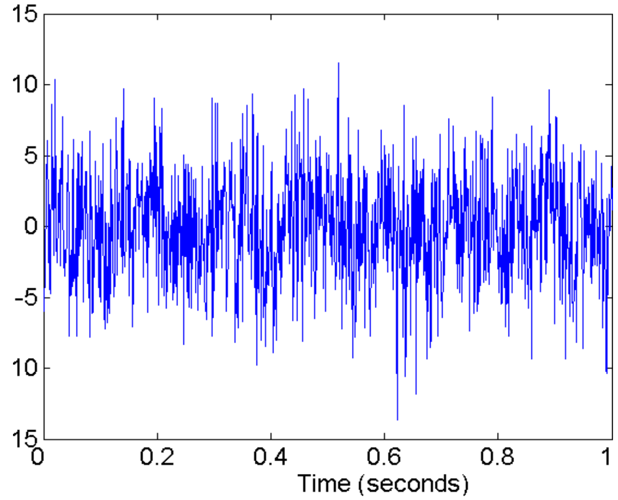
Noisy Data

- Noise: random error(s) or variance in a measured variable
- Incorrect attribute values may be due to
  - Faulty data collection instruments
  - Data entry problems
  - Data transmission problems
  - Technology limitations
  - Inconsistency in naming conventions
- Other data problems
  - Duplicate records
  - Inconsistent data

Two Sine Waves



Two Sine Waves + Noise



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.

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

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
- Semi-supervised: Combined computer and human inspection
  - Detect suspicious values and check by human (e.g., deal with possible outliers)

Data Preprocessing

- Data Preprocessing: An Overview
- Data Cleaning
- Data Integration
- Data Reduction and Transformation
- Dimensionality Reduction
- Summary

Data Integration

- Data integration
  - Combining data from multiple sources into a coherent store
    - Schema integration: e.g., A.cust-id = B.cust-#
      - Integrate metadata from different sources
- Entity identification:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
  - Often need domain knowledge or machine learning or both
- 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
  - Need case-by-case analysis

Handling Redundancy in Data Integration

- Redundant data occur often when integrating 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

Handling Redundancy in Data Integration

- Redundant data occur often when integrating 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 detected by correlation analysis and covariance analysis
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Correlation Analysis (for Categorical Data)

- **$\chi 2$ (chi-square) test:**
  - To discover the correlation relationship between two nominal attributes, A and B.

Correlation Analysis (for Categorical Data)

- χ^2 **(chi-square) test:**
  - To discover the correlation relationship between two nominal attributes, A and B.
  - Suppose A has c distinct values $\{a_1, a_2, \ldots, a_c\}$, B has r distinct values $\{b_1, b_2, \ldots, b_r\}$.
  - Contingency table: How many times the joint event (A_i, B_j), "attribute A takes on values a_i and attribute B takes on value b_j", happens based on the observed data tuples.

Correlation Analysis (for Categorical Data)

- χ^2 **(chi-square) test:**
  - To discover the correlation relationship between two nominal attributes, A and B.
  - Suppose A has c distinct values $\{a_1, a_2, \ldots, a_c\}$, B has r distinct values $\{b_1, b_2, \ldots, b_r\}$.
  - Contingency table: How many times the joint event (A_i, B_j), "attribute A takes on values a_i and attribute B takes on value b_j", happens based on the observed data tuples.

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Where o_{ij} is the observed frequency (or, actual count) of the joint event ($\underline{A_i}$, $\underline{B_j}$), and e_{ij} is the expected frequency:

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n},$$

Correlation Analysis (for Categorical Data)

- **χ2 (chi-square) test:**

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

  - Null hypothesis: The two variables are independent
  - The cells that contribute the most to the χ^2 value are those whose actual count is very different from the expected count
    - The larger the χ2 value, the more likely the variables are related

Chi-Square Calculation: An Example

Karlsruher Institut für Technologie

	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

Contingency Table

Numbers outside bracket mean the observed frequencies of a joint event, and numbers inside bracket mean the expected frequencies.

Chi-Square Calculation: An Example

Karlsruher Institut für Technologie

	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

Contingency Table

Numbers outside bracket mean the observed frequencies of a joint event, and numbers inside bracket mean the expected frequencies.

How to derive expected frequency (e_{ij})?

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

Contingency Table

Numbers outside bracket mean the observed frequencies of a joint event, and numbers inside bracket mean the expected frequencies.

How to derive expected frequency (e_{ij})?
(450* 300)/1500 = 90

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n},$$

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

Contingency Table

χ^2 (chi-square) calculation

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

$$\boxed{\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}}$$

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

Contingency Table

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

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

$$\boxed{\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}}$$

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

Contingency Table

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

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

Given a threshold 10.828

It shows that like_science_fiction and play_chess are correlated in the group

Review: Variance for Single Variable (Numerical Data)

- The variance of a random variable *X* provides a measure of how much the value of *X* deviates from the mean or expected value of *X:*

$$\sigma^2 = \mathrm{var}(X) = E[(X-\mu)^2] = \begin{cases} \sum_x (x-\mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx & \text{if } X \text{ is continuous} \end{cases}$$

  - where σ^2 is the variance of X, σ is called *standard deviation*

    μ = E[X] is the mean (or expected value) of X

Review: Variance for Single Variable (Numerical Data)

- The variance of a random variable *X* provides a measure of how much the value of *X* deviates from the mean or expected value of *X:*

$$\sigma^2 = \operatorname{var}(X) = E[(X-\mu)^2] = \begin{cases} \sum_x (x-\mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx & \text{if } X \text{ is continuous} \end{cases}$$

  - where σ^2 is the variance of X, σ is called *standard deviation*

    $\mu = E[X]$ is the mean (or expected value) of X

  - It can also be written as:

$$\boxed{\sigma^2 = \operatorname{var}(X) = E[(X-\mu)^2] = E[X^2] - \mu^2 = E[X^2] - [E(X)]^2}$$

Covariance for Two Variables

- Covariance between two variables X_1 and X_2

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1\mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

where $\mu_1 = E[X_1]$ is the mean (or expected value) of X_1; similarly for μ_2

Single variable

$$\boxed{\sigma^2 = \mathrm{var}(X) = E[(X - \mu)^2] = E[X^2] - \mu^2 = E[X^2] - [E(X)]^2}$$

Covariance for Two Variables

- Covariance between two variables X_1 and X_2

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1\mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

where $\mu_1 = E[X_1]$ is the mean (or expected value) of X_1; similarly for μ_2

- **Positive covariance:** If $\sigma_{12} > 0$

- **Negative covariance**: If $\sigma_{12} < 0$

- **Independence**: If X_1 and X_2 are independent, $\sigma_{12} = 0$, but the reverse is not true
  - Some pairs of random variables may have a covariance 0 but are not independent
  - Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence

Example: Calculation of Covariance

- Suppose two stocks X_1 and X_2 have the following values in one week:
  - Day 1: $(X_1, X_2) = (2, 5)$,
  - Day 2: $(X_1, X_2) = (3, 8)$,
  - Day 3: $(X_1, X_2) = (5, 10)$,
  - Day 4: $(X_1, X_2) = (4, 11)$,
  - Day 5: $(X_1, X_2) = (6, 14)$.

Example: Calculation of Covariance

- Suppose two stocks X_1 and X_2 have the following values in one week:

  (2, 5), (3, 8), (5, 10), (4, 11), (6, 14)

- Covariance formula:

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

$$\sigma_{12} = E[X_1 X_2] - E[X_1]E[X_2]$$

Example: Calculation of Covariance

- Suppose two stocks X_1 and X_2 have the following values in one week:

  (2, 5), (3, 8), (5, 10), (4, 11), (6, 14)

- Covariance formula:

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

$$\sigma_{12} = E[X_1 X_2] - E[X_1]E[X_2]$$

- Its computation can be simplified as:
  - $E(X_1) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4$
  - $E(X_2) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6$

Example: Calculation of Covariance

- Suppose two stocks X_1 and X_2 have the following values in one week:

  (2, 5), (3, 8), (5, 10), (4, 11), (6, 14)

- Covariance formula:

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

$$\sigma_{12} = E[X_1 X_2] - E[X_1]E[X_2]$$

- Its computation can be simplified as:
  - $E(X_1) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4$
  - $E(X_2) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6$
  - $\sigma_{12} = \underbrace{(2\times5 + 3\times8 + 5\times10 + 4\times11 + 6\times14)/5}_{E[X1X2]} - 4 \times 9.6 = 4$

Example: Calculation of Covariance

- Suppose two stocks X_1 and X_2 have the following values in one week:

  (2, 5), (3, 8), (5, 10), (4, 11), (6, 14)

- Covariance formula:

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1\mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

$$\sigma_{12} = E[X_1 X_2] - E[X_1]E[X_2]$$

- Its computation can be simplified as:
  - $E(X_1) = (2 + 3 + 5 + 4 + 6)/5 = 20/5 = 4$
  - $E(X_2) = (5 + 8 + 10 + 11 + 14)/5 = 48/5 = 9.6$
  - $\sigma_{12} = (2\times5 + 3\times8 + 5\times10 + 4\times11 + 6\times14)/5 - 4 \times 9.6 = 4$

  Therefore, X1 and X2 rise together since σ12 > 0

Data Preprocessing

- Data Preprocessing: An Overview
- Data Cleaning
- Data Integration
- Data Reduction and Transformation
- Dimensionality Reduction
- Summary

Data Reduction

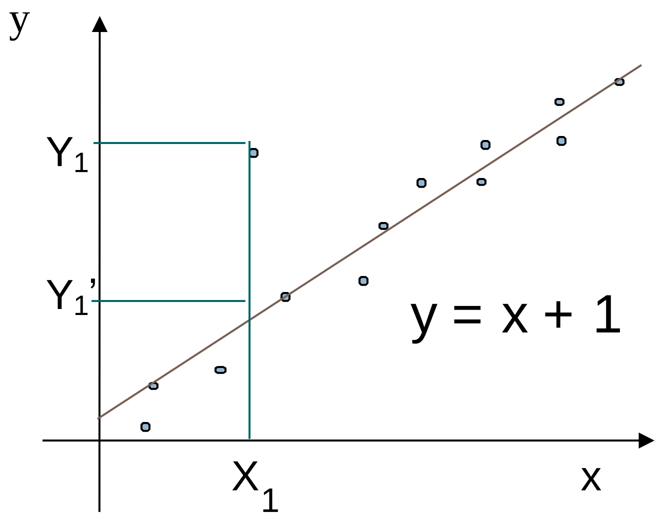
- Data reduction:
  - Obtain a reduced representation of the data set
    - much smaller in volume but yet produces almost the same analytical results
- Why data reduction?—A database/data warehouse may store terabytes of data
  - Complex analysis may take a very long time to run on the complete data set

Data Reduction

- Data reduction:
  - Obtain a reduced representation of the data set
    - much smaller in volume but yet produces almost the same analytical results
- Why data reduction?—A database/data warehouse may store terabytes of data
  - Complex analysis may take a very long time to run on the complete data set
- Methods for data reduction (also *data size reduction* or *numerosity reduction*)
  - Regression and Log-Linear Models
  - Histograms, clustering, sampling
  - Data cube aggregation
  - Data compression

Data Reduction: Regression Analysis

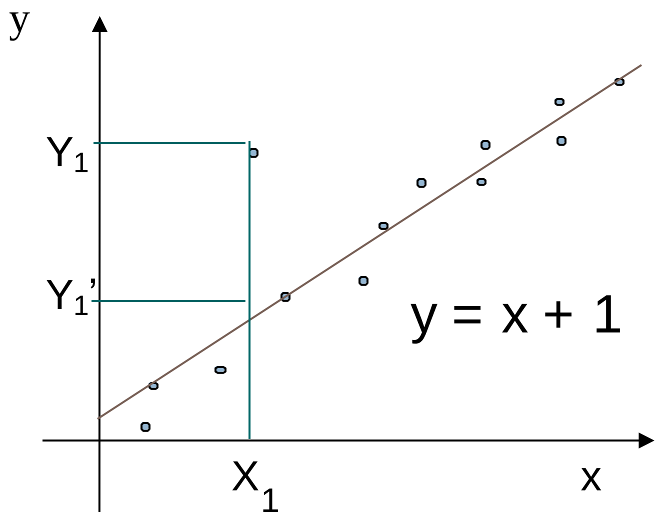
- Regression analysis: A collective name for techniques for the modeling and analysis of numerical data consisting of values of a dependent variable (also called response variable or *measurement*) and of one or more *independent variables* (also known as explanatory variables or predictors)



- Used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships

Data Reduction: Regression Analysis

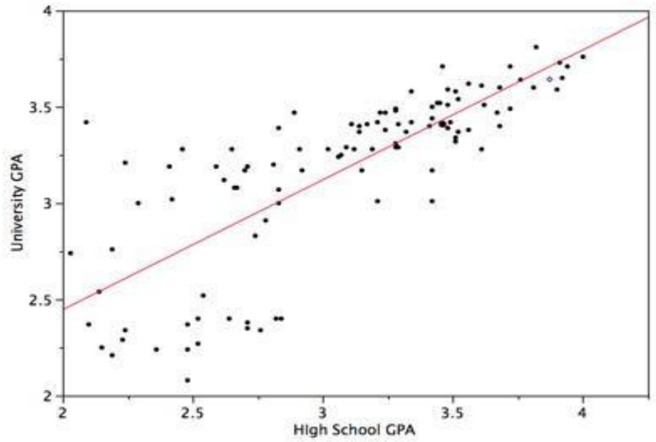
- Regression analysis: A collective name for techniques for the modeling and analysis of numerical data consisting of values of a dependent variable (also called response variable or *measurement*) and of one or more *independent variables* (also known as explanatory variables or predictors)
- The parameters are estimated so as to give a "best fit" of the data
- Mostly the best fit is evaluated by using the least squares method, but other criteria have also been used



- Used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships

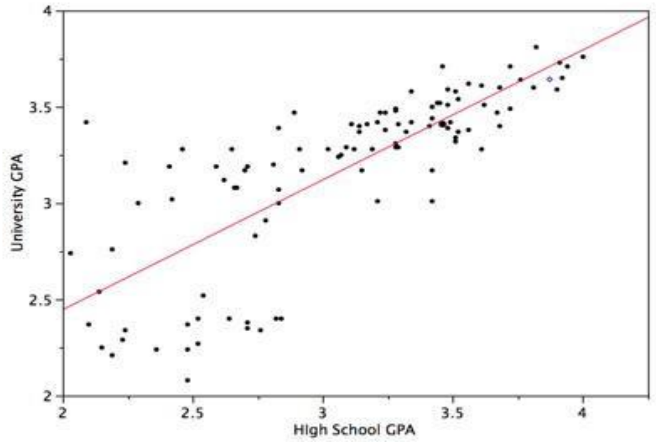
Linear and Multiple Regression

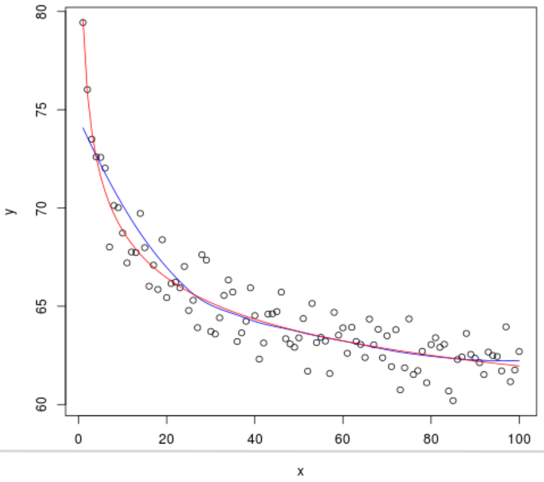
- Linear regression: $Y = w X + b$
  - Data modeled to fit a straight line
  - Often uses the least-square method to fit the line
  - Two regression coefficients, *w* and *b,* specify the line and are to be estimated by using the data at hand
  - Using the least squares criterion to the known values of $(X_{1,} Y_{1}), (X_{2,} Y_{2}), \ldots, (X_{n\,,} Y_{n})$



Linear and Multiple Regression

- <u>Linear regression</u>: $Y = w X + b$
  - Data modeled to fit a straight line
  - Often uses the least-square method to fit the line
  - Two regression coefficients, *w* and *b*, specify the line and are to be estimated by using the data at hand
  - Using the least squares criterion to the known values of $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$
- <u>Nonlinear regression</u>:
  - Data modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables
  - Data are fitted by a method of successive approximations

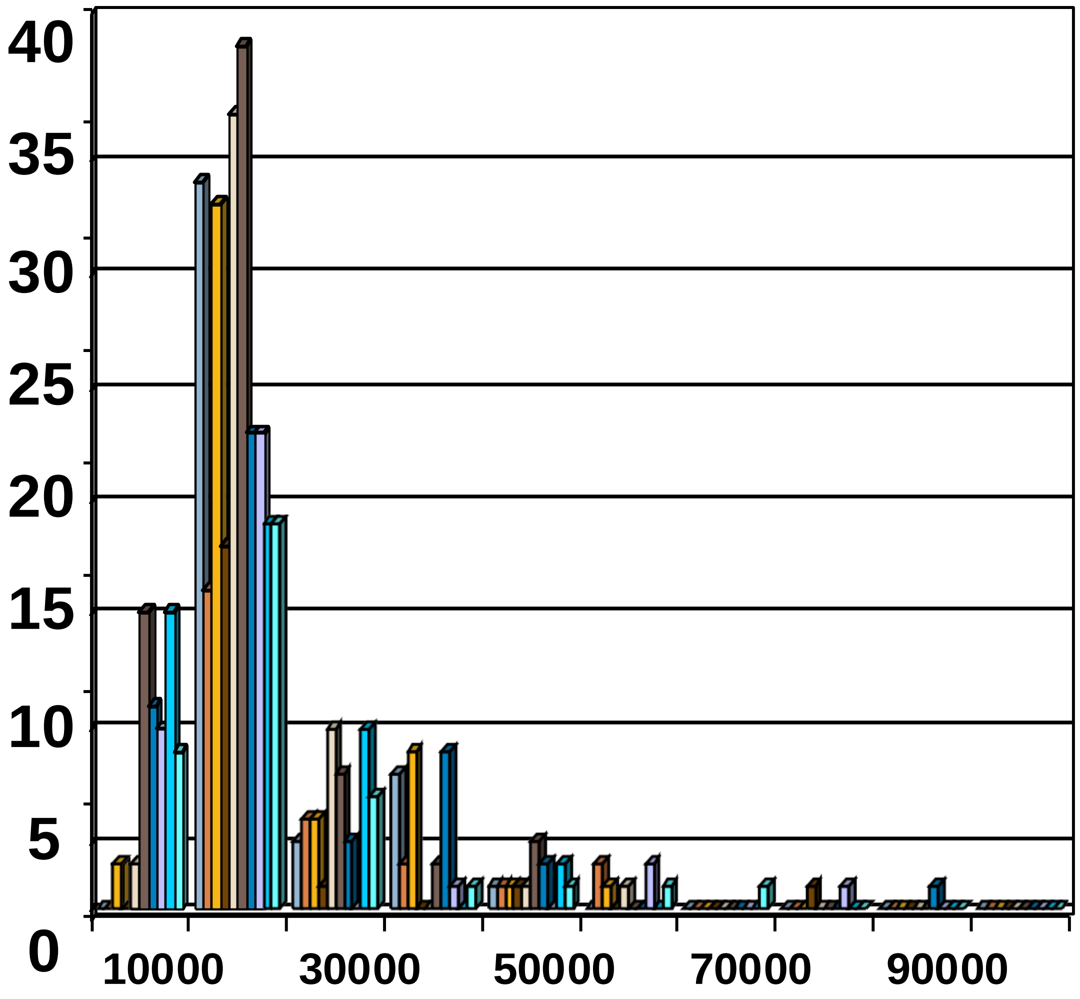




Scientific Computing Center SCC

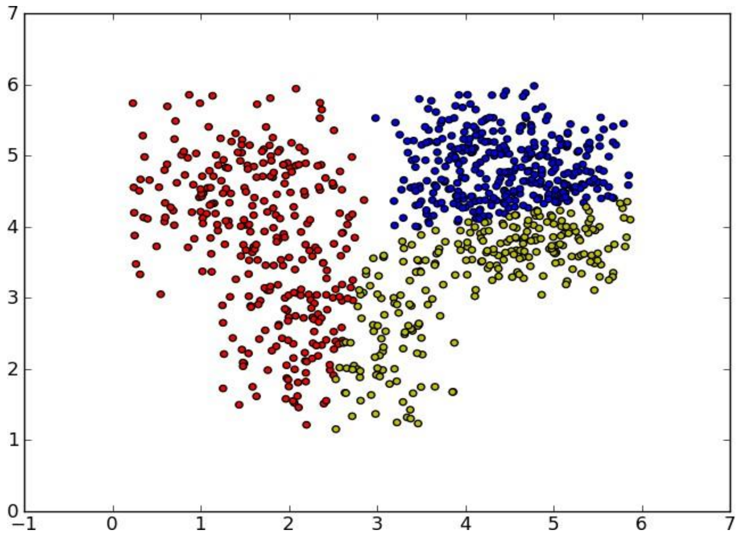
Histogram Analysis

- Divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
  - Equal-width: equal bucket range
  - Equal-frequency (or equal-depth)



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 clustered 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



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

Data Preprocessing

- Data Preprocessing: An Overview
- Data Cleaning
- Data Integration
- Data Reduction and Transformation
- Dimensionality Reduction
- Summary

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

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
  - Attribute/feature construction
    - New attributes constructed from the given ones
  - Aggregation: Summarization, data cube construction
  - Normalization: Scaled to fall within a smaller, specified range
    - min-max normalization; z-score normalization; normalization by decimal scaling
  - Discretization: Concept hierarchy climbing

Normalization

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

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

Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]

- Then \$73,600 is mapped to

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

Normalization

Karlsruher Institut für Technologie

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

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

- Z-score normalization (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

Z-score: The distance between the raw score and the population mean in the unit of the standard deviation

  - Ex. Let μ = 54,000, σ = 16,000. Then,

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

Normalization

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

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

- Z-score normalization (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

Z-score: The distance between the raw score and the population mean in the unit of the standard deviation

- Normalization by decimal scaling

$$v' = v / 10^j$$, Where j is the smallest integer such that Max(|v'|) < 1

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
  - Numeric—real numbers, e.g., integer or real numbers

- Discretization: Divide the range of a continuous attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Reduce data size by discretization
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute
  - Prepare for further analysis, e.g., classification

Data Discretization Methods

- Binning
  - Top-down split, unsupervised
- Histogram analysis
  - Top-down split, unsupervised
- Clustering analysis
  - Unsupervised, top-down split or bottom-up merge
- Decision-tree analysis
  - Supervised, top-down split
- Correlation (e.g., χ^2) analysis
  - Unsupervised, bottom-up merge
- Note: All the methods can be applied recursively

Example: Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 22, 24, 25, 26, 28, 29, 33
- Partition into equal-frequency (equi-width) bins:

  - Bin 1: 4, 8, 9, 15

  - Bin 2: 21, 22, 24, 25

  - Bin 3: 26, 28, 29, 33
- 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, 33

Data Preprocessing

- Data Preprocessing: An Overview
- Data Cleaning
- Data Integration
- Data Reduction and Transformation
- Dimensionality Reduction
- Summary

Dimensionality Reduction

- Curse of dimensionality
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful

Dimensionality Reduction

- Curse of dimensionality
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
- Dimensionality reduction
  - Reducing the number of random variables under consideration, via obtaining a set of principal variables

Dimensionality Reduction

- Curse of dimensionality
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
- Dimensionality reduction
  - Reducing the number of random variables under consideration, via obtaining a set of principal variables

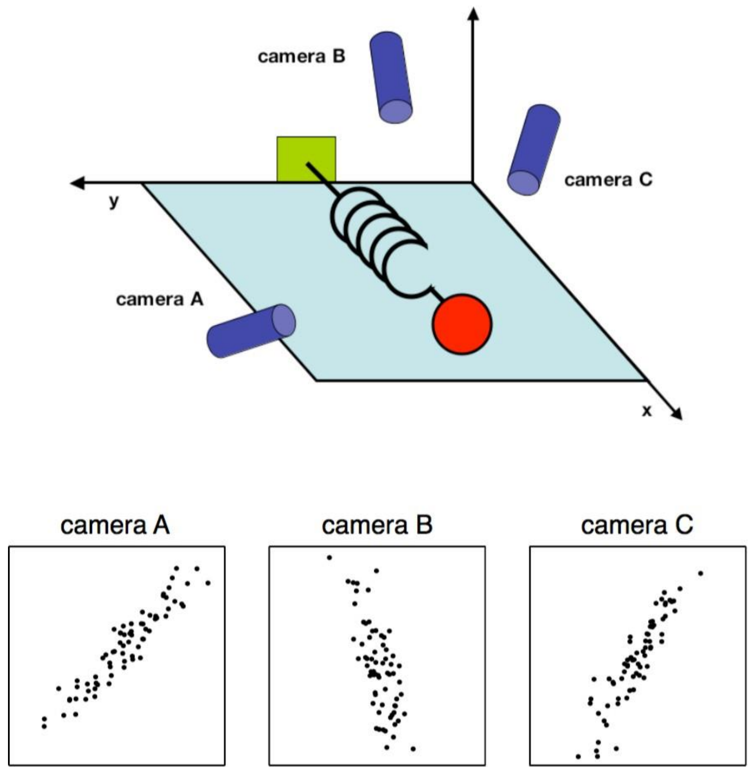
- Advantages of dimensionality reduction
  - Mitigate the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualization

Dimensionality Reduction Techniques

- Dimensionality reduction methodologies
  - **Feature selection**: Find a subset of the original variables (or features, attributes)
  - **Feature extraction**: Transform the data in the high-dimensional space to a space of fewer dimensions
- Some typical dimensionality reduction methods
  - Principal Component Analysis
  - Supervised and nonlinear techniques
    - Feature subset selection
    - Feature creation

Principal Component Analysis (PCA)

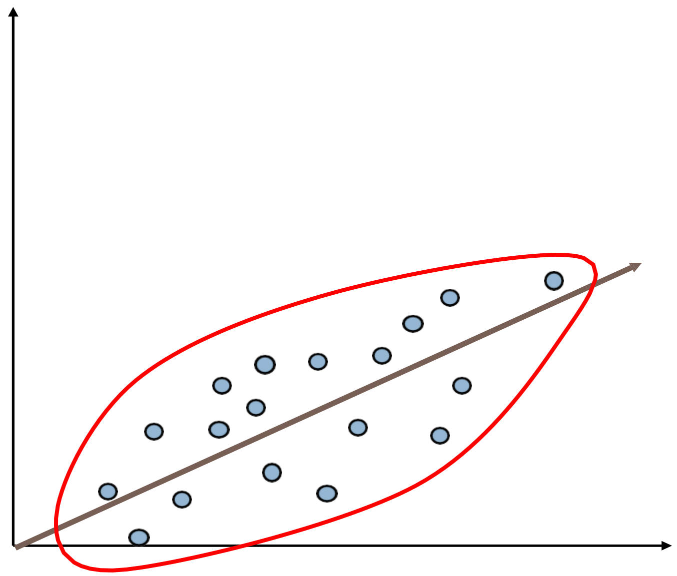
- PCA: A statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called ***principal components***
- The original data are projected onto a much smaller space, resulting in dimensionality reduction
- Method: Find the eigenvectors of the covariance matrix, and these eigenvectors define the new space



Ball travels in a straight line. Data from three cameras contain much redundancy

Principal Components Analysis: Intuition

- Goal is to find a projection that captures the largest amount of variation in data
- Find the eigenvectors of the covariance matrix
- The eigenvectors define the new space



Principal Component Analysis: Details

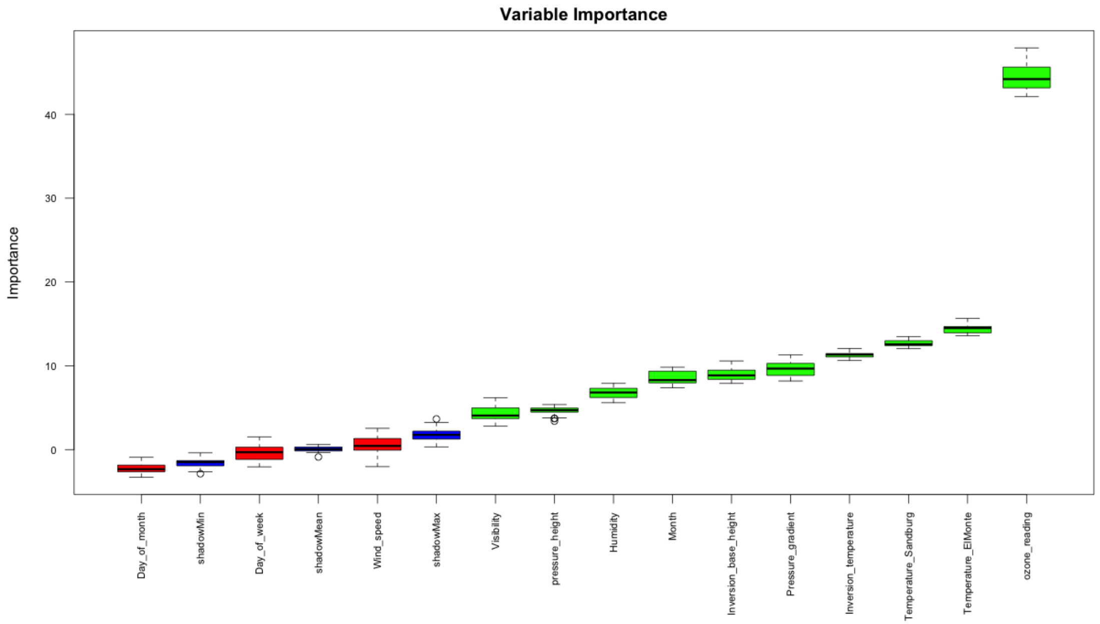
- Let A be an $n \times n$ matrix representing the covariance of the data.
  - λ is an **eigenvalue** of A if there exists a non-zero vector $\boldsymbol{v}$ such that:

$$A\boldsymbol{v} = \lambda\boldsymbol{v}$$

- In this case, vector $\boldsymbol{v}$ is called an **eigenvector** of A corresponding to λ. For each eigenvalue λ, the set of all vectors $\boldsymbol{v}$ satisfying $A\boldsymbol{v} = \lambda\boldsymbol{v}$ is called the **eigenspace** of A corresponding to λ.

Attribute Subset Selection

- Another way to reduce dimensionality of data
- Redundant attributes
  - 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 attributes
  - Contain no information that is useful for the data mining task at hand
    - Ex. A student’s ID is often irrelevant to the task of predicting his/her GPA



Heuristic Search in Attribute Selection

- There are 2^d possible attribute combinations of *d* attributes
- Typical heuristic attribute selection methods:
  - Best single attribute under the attribute independence assumption: choose by significance tests
  - Best step-wise feature selection:
    - The best single-attribute is picked first
    - Then next best attribute conditioned to the first, ...
  - Step-wise attribute elimination:
    - Repeatedly eliminate the worst attribute
  - Best combined attribute selection and elimination
  - Optimal branch and bound:
    - Use attribute elimination and backtracking

Attribute Creation (Feature Generation)

- Create new attributes (features) that can capture the important information in a data set more effectively than the original ones
- Three general methodologies
  - Attribute extraction
    - Domain-specific
  - Mapping data to new space
    - E.g., Fourier transformation, wavelet transformation, manifold approaches
  - Attribute construction
    - Combining features
    - Data discretization

Summary

- **Data quality**: accuracy, completeness, consistency, timeliness, interpretability, trustworthiness
- **Data cleaning**: e.g. missing/noisy values, outliers
- **Data integration** from multiple sources:
  - Entity identification problem; Remove redundancies; Detect inconsistencies
- **Data reduction**
  - Dimensionality reduction; Numerosity reduction; Data compression
- **Data transformation and data discretization**
  - Normalization; Concept hierarchy generation

References

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. Comm. of ACM, 42:73-78, 1999
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- T. Dasu, T. Johnson, S. Muthukrishnan, V. Shkapenyuk. Mining Database Structure; Or, How to Build a Data Quality Browser. SIGMOD'02
- H. V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data Engineering, 20(4), Dec. 1997
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- E. Rahm and H. H. Do. Data Cleaning: Problems and Current Approaches. IEEE Bulletin of the Technical Committee on Data Engineering. Vol.23, No.4
- V. Raman and J. Hellerstein. Potters Wheel: An Interactive Framework for Data Cleaning and Transformation, VLDB'2001
- T. Redman. Data Quality: Management and Technology. Bantam Books, 1992
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. IEEE Trans. Knowledge and Data Engineering, 7:623-640, 1995