Chapter 3. Data Processing

Meng Jiang

CS412 Summer 2017:

Introduction to Data Mining

Why? Data Quality Issues

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

Data Preprocessing

- **Data cleaning**
- Data integration
- Data reduction
- Dimensionality reduction

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, and transmission error
  - 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* = "03/07/2010"
    - Was rating "1, 2, 3", now rating "A, B, C"
  - Intentional (e.g., *disguised missing* data)
    - Jan. 1 as everyone's birthday?

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
- Missing data may be due to
  - 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
- 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

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
  - Faulty data collection instruments
  - Data transmission problems
  - Technology limitation
  - Inconsistency in naming convention
- Other data problems
  - 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
- Semi-supervised: Combined computer and human inspection
  - Detect suspicious values and check by human (e.g., deal with possible outliers)

Data Preprocessing

- Data cleaning
- **Data integration**
- Data reduction
- Dimensionality reduction

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
- 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* 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

	Play chess	Not play chess	Sum (row)
Like science fiction			450
Not like science fiction			1050
Sum(col.)	300	1200	1500

Correlation Analysis

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

How to derive 90?

450/1500 * 300 = 90

Correlation Analysis

	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

Correlation Analysis (for Categorical Data)

- **X^2 (chi-square) test:**

observed

$$\chi^2 = \sum_{i}^{n} \frac{(O_i - E_i)^2}{E_i}$$

expected

- **Null hypothesis:** The two distributions are independent
- The cells that contribute the most to the X^2 value are those whose actual count is different from the expected count
  - The larger the X^2 value, the more the null hypothesis of independence is rejected, and the more likely the variables are related

	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

Example: Chi-Square Calculation

	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)

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

We can reject the null hypothesis of independence at a confidence level of 0.001.

- It shows that like_science_fiction and play_chess are correlated.

Example: Chi-Square Calculation

Degrees of freedom (df)	χ^2 value[19]										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001

Correlation Analysis (for Categorical Data)

- **X^2 (chi-square) test:**

observed

$$\chi^2 = \sum_{i}^{n} \frac{(O_i - E_i)^2}{E_i}$$

expected

- **Null hypothesis:** The two distributions are independent
- The cells that contribute the most to the X^2 value are those whose actual count is different from the expected count
  - The larger the X^2 value, the more the null hypothesis of independence is rejected, and the more likely the variables are related
- Note: 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

Variance for Single Variable (for 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:*

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

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

    μ is the mean, and **μ = E[X]** is the expected value of X
  - That is, variance is the expected value of the square deviation from the mean
  - It can also be written as:

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

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 respective 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:
  – (2, 5), (3, 8), (5, 10), (4, 11), (6, 14)
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
- Covariance formula

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

- Its computation can be simplified as: $\sigma_{12} = E[X_1 X_2] - E[X_1]E[X_2]$
  – $E(X_1) = (2 + 3 + 5 + 4 + 6) / 5 = 20/5 = 4$
  – $E(X_2) = (5 + 8 + 10 + 11 + 14) / 5 = 48/5 = 9.6$
  – $\sigma_{12} = (2\times5 + 3\times8 + 5\times10 + 4\times11 + 6\times14) / 5 - 4 \times 9.6 = 4$
- Thus, X_1 and X_2 rise together since $\sigma_{12} > 0$

Correlation between Two Numerical Variables

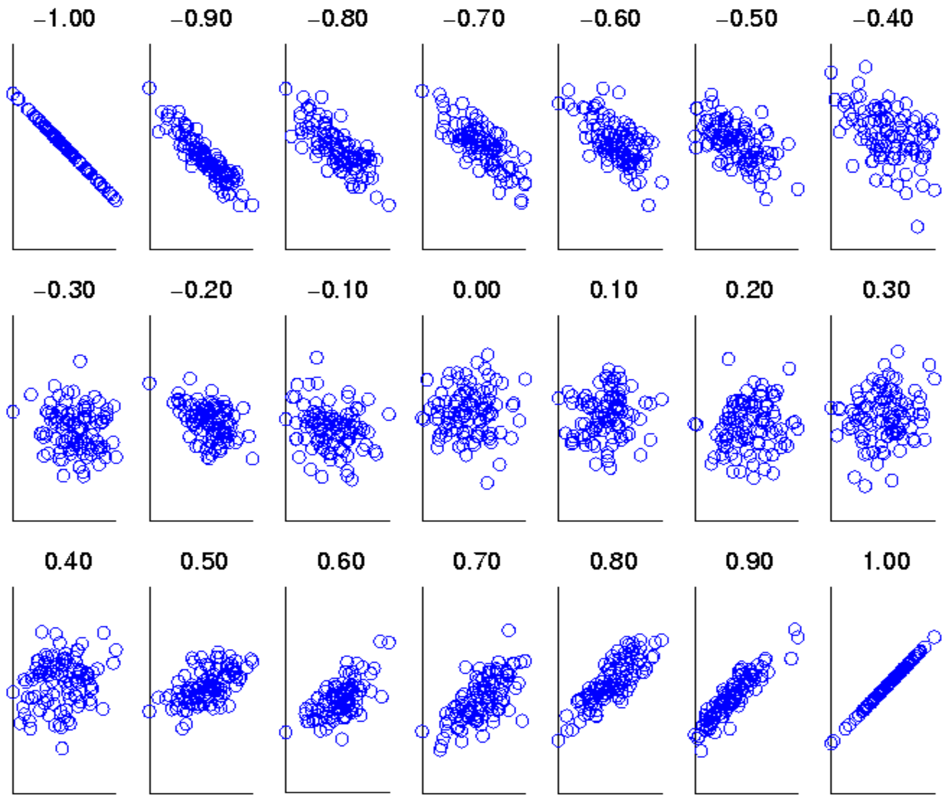
- **Correlation** between two variables X1 and X2 is the standard covariance, obtained by normalizing the covariance with the standard deviation of each variable

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^{\ 2} \sigma_2^{\ 2}}}$$

- If $\rho_{12} > 0$: A and B are positively correlated (X1's values increase as X2's)
  - The higher, the stronger correlation
- If $\rho_{12} = 0$: independent (under the same assumption as discussed in co-variance)
- If $\rho_{12} < 0$: negatively correlated

Visualizing Changes of Correlation Coefficient

- Correlation coefficient value range: [–1, 1]
- A set of scatter plots shows sets of points and their correlation coefficients changing from –1 to 1



Covariance Matrix

- The variance and covariance information for the two variables X_1 and X_2 can be summarized as 2 * 2 covariance matrix as

$$\Sigma = E[(\mathbf{X}-\mu)(\mathbf{X}-\mu)^T] = E[\begin{pmatrix} X_1-\mu_1 \\ X_2-\mu_2 \end{pmatrix}(X_1-\mu_1 \quad X_2-\mu_2)]$$

$$= \begin{pmatrix} E[(X_1-\mu_1)(X_1-\mu_1)] & E[(X_1-\mu_1)(X_2-\mu_2)] \\ E[(X_2-\mu_2)(X_1-\mu_1)] & E[(X_2-\mu_2)(X_2-\mu_2)] \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

- Generalizing it to *d* dimensions, we have,

$$D = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & \cdots & x_{dd} \end{pmatrix} \quad \mathbf{\Sigma} = E[(\mathbf{X}-\mu)(\mathbf{X}-\mu)^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

Announcement

- Assignment 1 is out!
- Due date: June 15th.
- Compass
- TAs: Xuan Wang (xwang174@illinois.edu) and Sheng Wang (swang141@illinois.edu)

Chapter 3. Data Processing

Meng Jiang

CS412 Summer 2017:

Introduction to Data Mining

Data Preprocessing

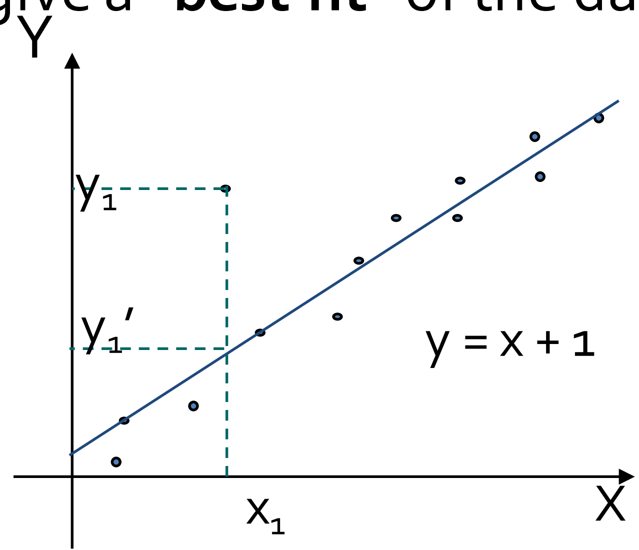
- Data cleaning
- Data integration
- **Data reduction**
  - Reduce by data objects
- Dimensionality reduction
  - Reduce by dimensions and attributes

Data Reduction

- Data reduction
  - Obtain a reduced representation of the data set
  - Why? Complex analysis may take a very long time to run on the complete data set
- Methods for data reduction
  - Regression and Log-Linear Models
  - Histograms, Clustering, Sampling
  - Data compression

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***): Y
  - and of one or more *independent variables* (also known as ***explanatory variables*** or ***predictors***): X, or X_1, X_2, …X_n
- Parameters are estimated to give a "**best fit**" of the data
  - Data: (x_1, y_1)
  - Fit of the data: (x_1, y_1')
    - Ex. $y_1' = x_1 + 1$

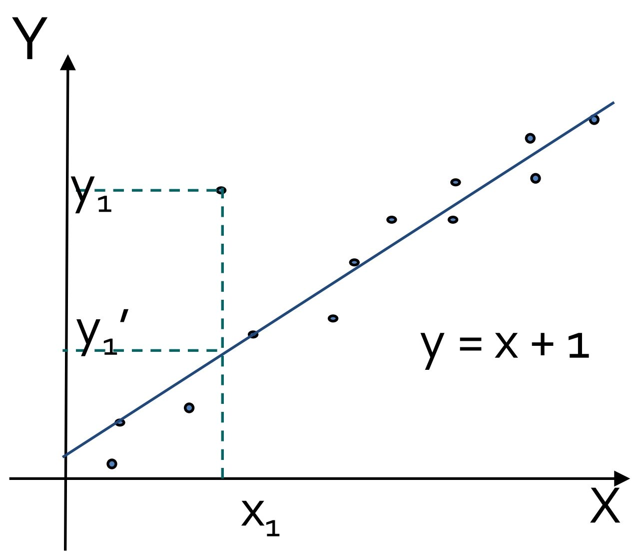


Regression Analysis (cont.)

- Most commonly the best fit is evaluated by using the ***least square method***, but other criteria have also been used

  min $g = \sum_{i=1}^{n}(y_i - y'_i)^2$, where $y'_i = f(x_i, \beta)$

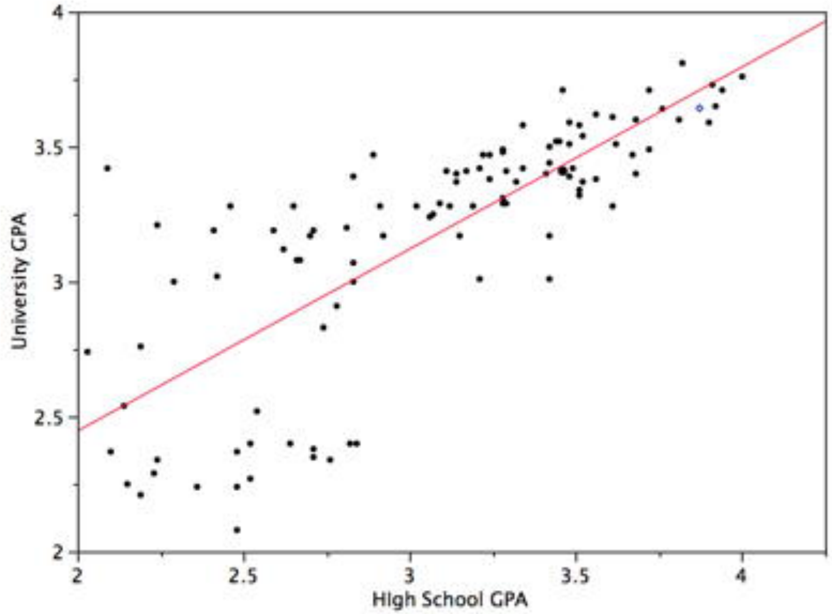
- Used for **prediction** (including forecasting of time-series data), **inference**, **hypothesis testing**, and **modeling of causal relationships**



Set up y = f(x) = β_1 x + β_2
Learn β by minimizing the least square error

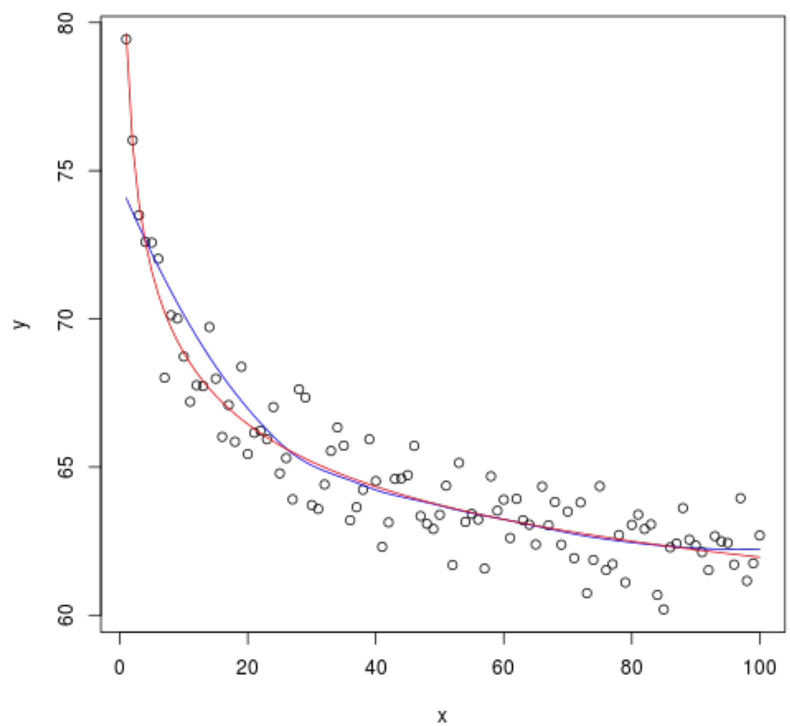
Linear Regression

- 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



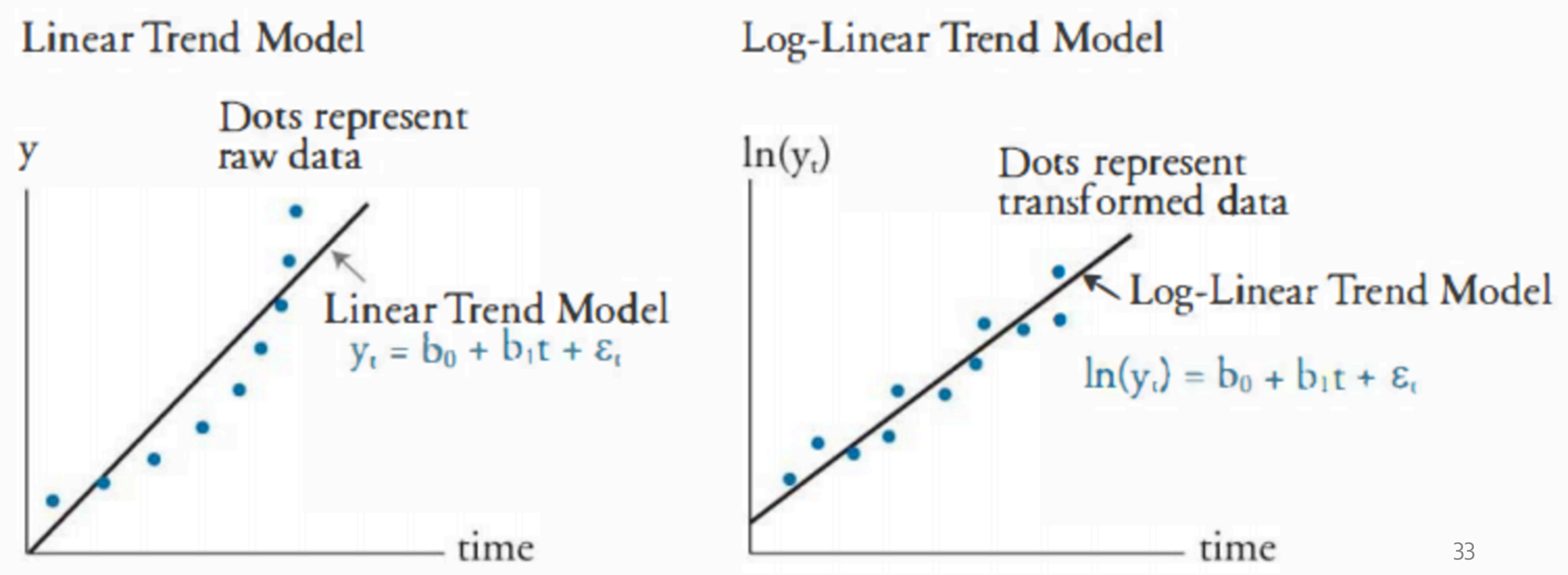
Nonlinear Regression

- Nonlinear regression:
  - Data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables



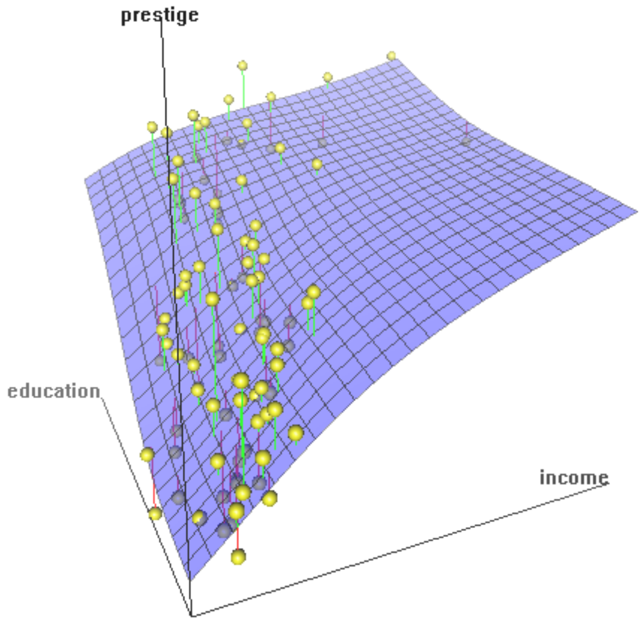
Log-Linear Model

- Log-linear model
  - A math model that takes the form of a **function whose logarithm** is a linear combination of the parameters of the model



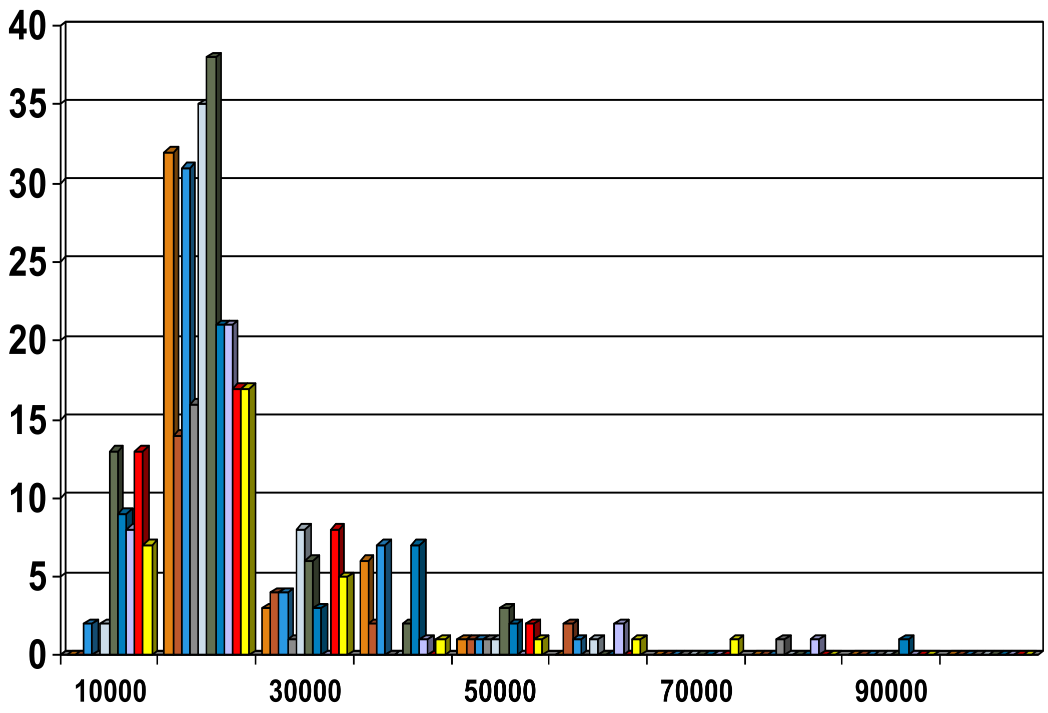
Multiple Regression

- Multiple regression: $Y = b_0 + b_1 X_1 + b_2 X_2$
  - Allows a response variable Y to be modeled as a linear function of multidimensional feature vector
  - Many nonlinear functions can be **transformed** into the above



Histogram Analysis

- Divide data into buckets and store average (sum) for each bucket
- One popular partitioning rules - Equal-width: equal bucket range



(10,000 , 10,001] = 10,001

...

to

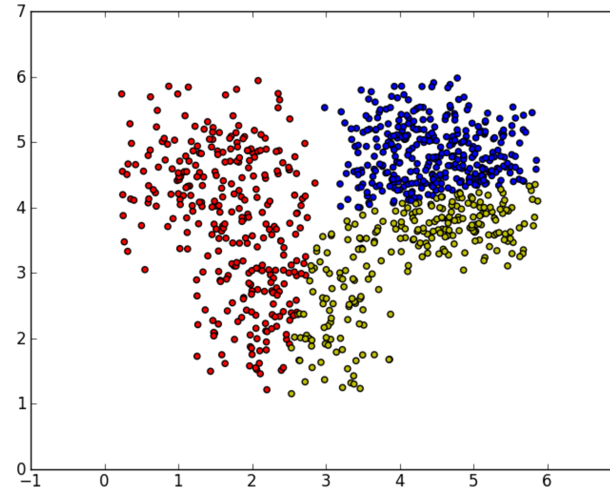
(10,000 , 11,000]

(11,000 , 12,000]

...

Clustering

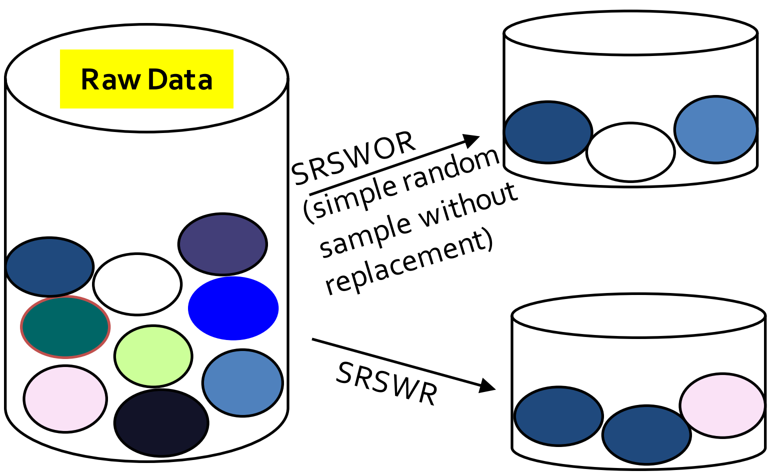
- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only
- Can be very effective if data is clustered but not if data is "smeared"
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms
- Cluster analysis will be studied in depth in Chapter 10



Sampling

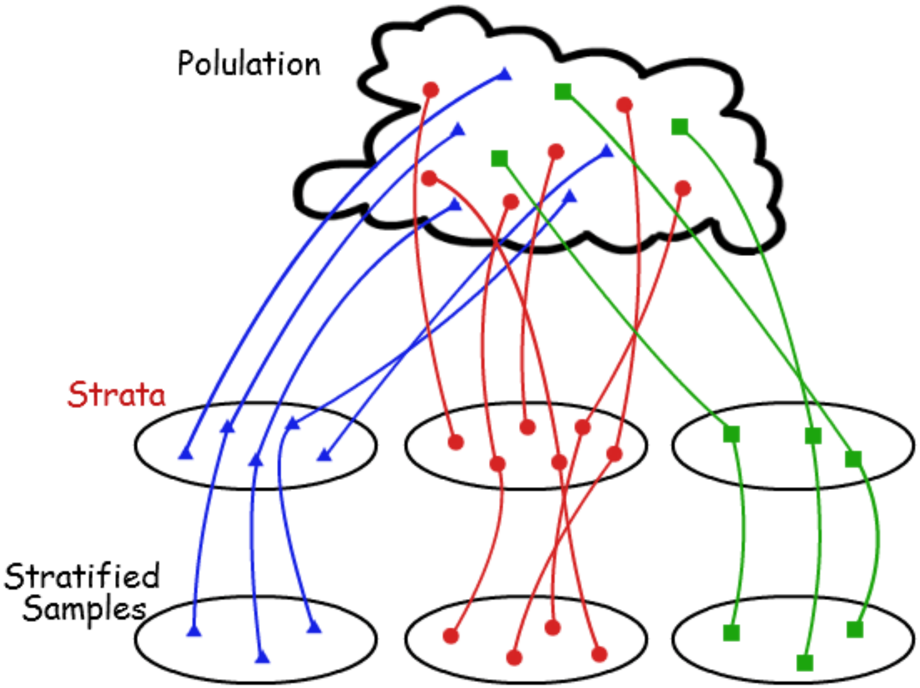
- Sampling: obtaining a small sample *s* to represent the whole data set *N*
- Key principle: Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew

Simple random sampling:
Equal probability of selecting any particular item
Sampling without replacement:
Once an object is selected, it is removed from the population
Sampling with replacement:
A selected object is not removed from the population



Stratified Sampling

- **Stratified sampling**
  - Partition (or cluster) the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)



Recall: Data Reduction

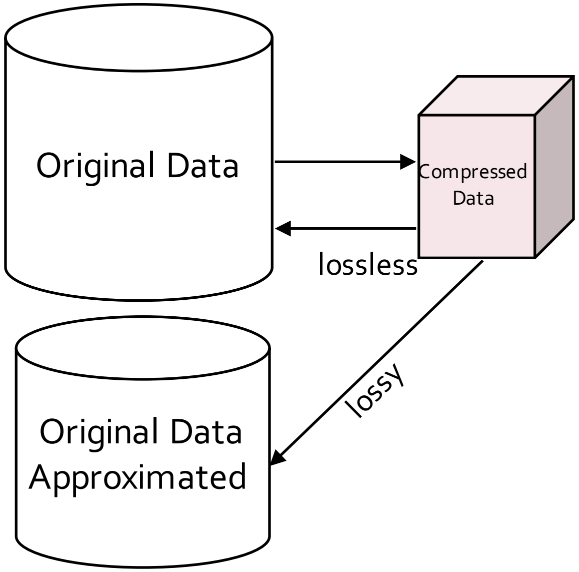
- Data reduction
  - Obtain a reduced representation of the data set
  - Why? Complex analysis may take a very long time to run on the complete data set
- Methods for data reduction
  - **Regression and Log-Linear Models**
  - **Histograms, Clustering, Sampling**
  - Data compression

Parametric vs. Non-Parametric Data Reduction Methods

- **Parametric methods** (e.g., regression)
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
- **Non-parametric** methods
  - Do not assume models
  - Major families: histograms, clustering, sampling, …

Data Compression

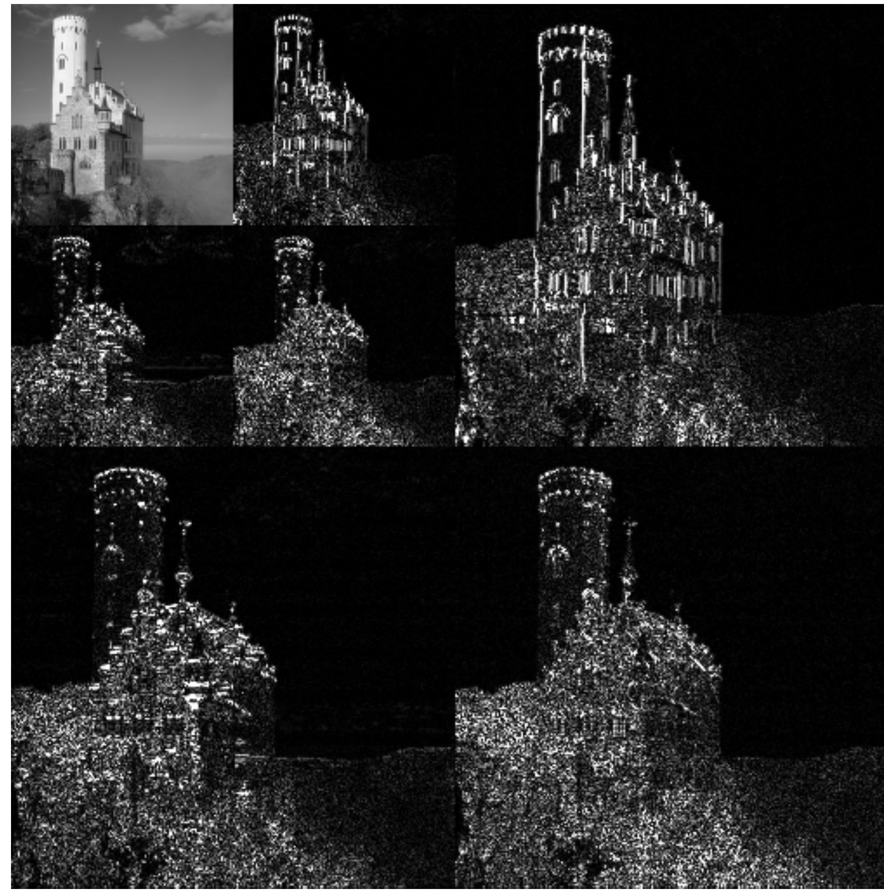
- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless, but only limited manipulation is possible without expansion
- Audio/video compression
  - Typically lossy compression, with progressive refinement
  - Sometimes small fragments of signal can be reconstructed without reconstructing the whole
- Data reduction and dimensionality reduction may also be considered as forms of data compression



Lossy vs. lossless compression

Wavelet Transform: A Data Compression Technique

- Wavelet Transform
  - Decomposes a signal into different frequency subbands
  - Applicable to n-dimensional signals
- Data are transformed to preserve relative distance between objects at different levels of resolution
- Allow natural clusters to become more distinguishable
- Used for image compression



Wavelet Transformation

- Discrete wavelet transform (DWT) for linear signal processing, multi-resolution analysis
- Compressed approximation: Store only a small fraction of the strongest of the wavelet coefficients
- Similar to discrete Fourier transform (DFT), but better lossy compression, localized in space

Transform	Representation	Input
Fourier transform	$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \xi}\, dx$	ξ, frequency
Time-frequency analysis	$X(t, f)$	t, time; f, frequency
Wavelet transform	$X(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} \overline{\Psi\left(\frac{t-b}{a}\right)}\, x(t)\, dt$	a, scaling; b, time

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 \$73,600 is mapped to

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

Normalization (cont.)

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

- **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 by decimal scaling**

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

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

Data Preprocessing

- Data cleaning
- Data integration
- Data reduction
- **Dimensionality reduction**

Dimensionality Reduction

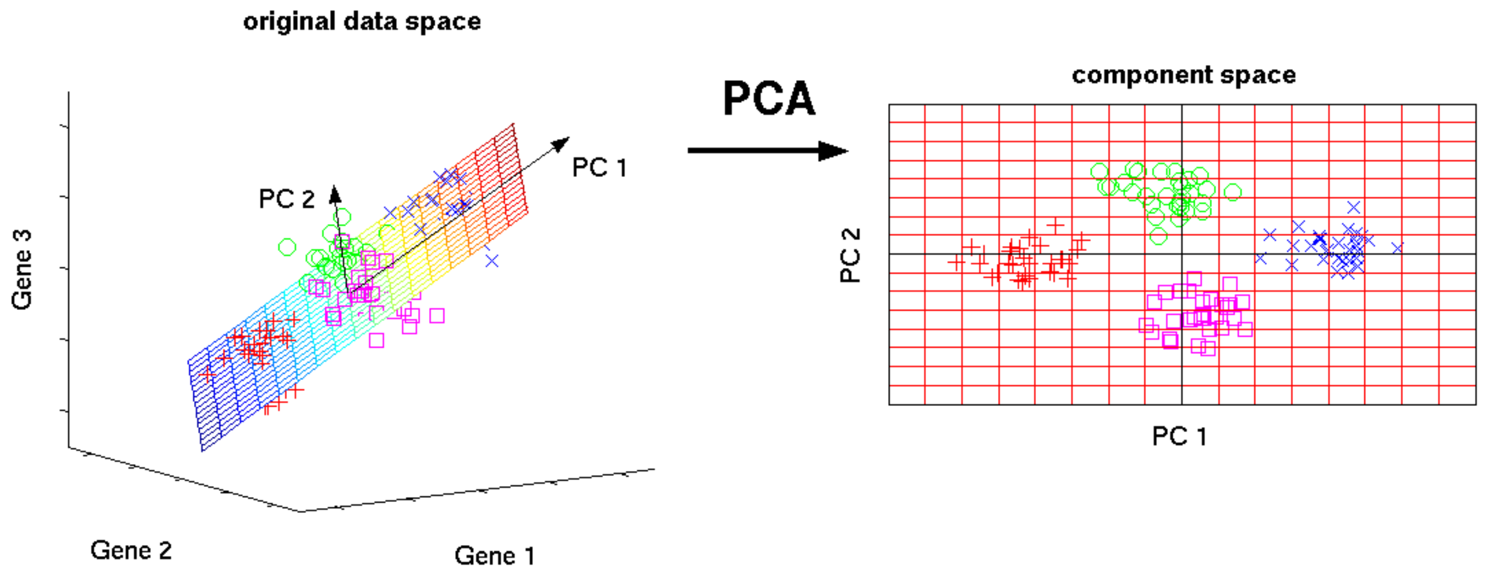
- Curse of dimensionality
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially
- Dimensionality reduction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

- Dimensionality reduction methodologies
  - **Feature selection (FS)**: Find a subset of the original variables (or features, attributes)
  - **Feature extraction (FE)**: Transform the data in the **high-dimensional** space to a space of **fewer** dimensions
- Some typical dimensionality methods
  - FE: Principal Component Analysis
  - FS: Attribute Subset Selection = Attribute Selection

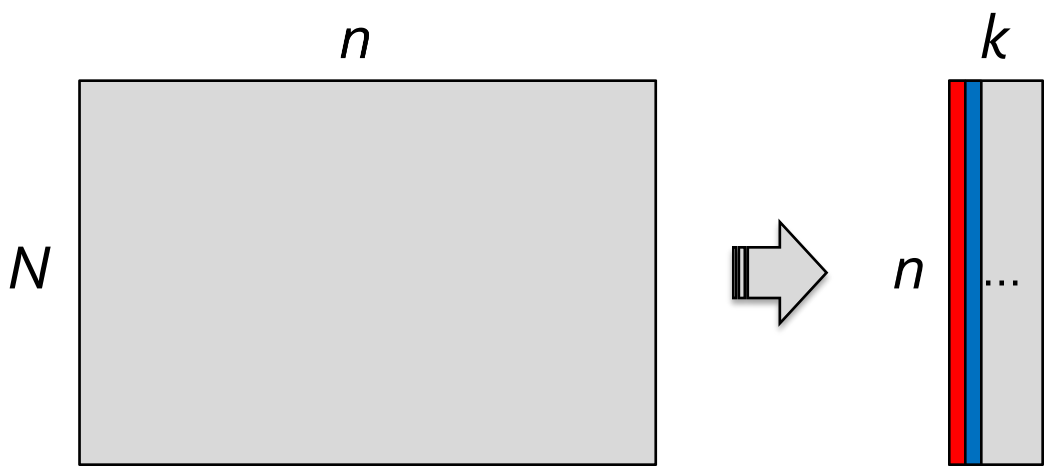
Principal Component Analysis (PCA)

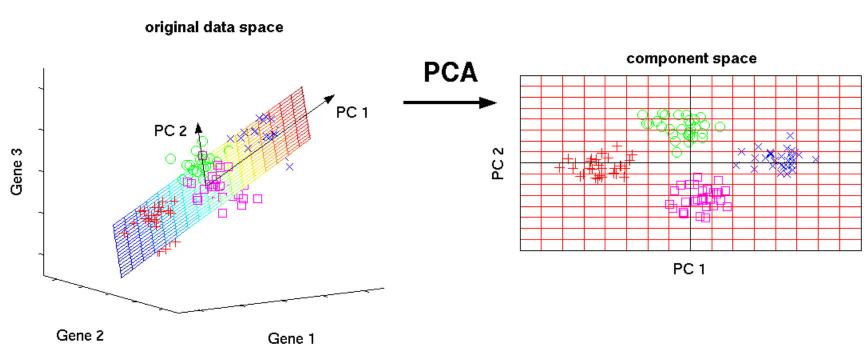
- PCA: A statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called ***principal components***
- The original data are projected onto a **much smaller space**, resulting in dimensionality reduction (e.g., n=3 to k=2)



PCA (cont.)

- Given N data vectors from n-dimensions, find $\boldsymbol{k \leq n}$ **orthogonal vectors** (*principal components*) best used to represent data



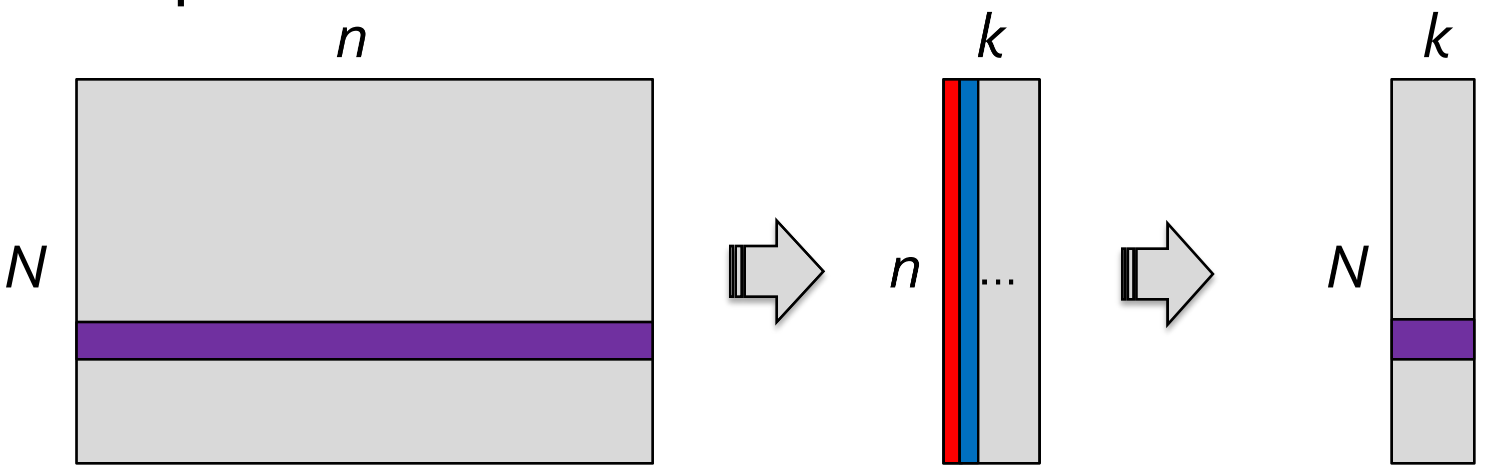


PCA (cont.)

- Given N data vectors from n-dimensions, find k ≤ n orthogonal vectors (principal components) best used to represent data
  - Normalize input data: Each attribute falls within the same range
  - Compute k **orthonormal (unit) vectors**, i.e., principal components

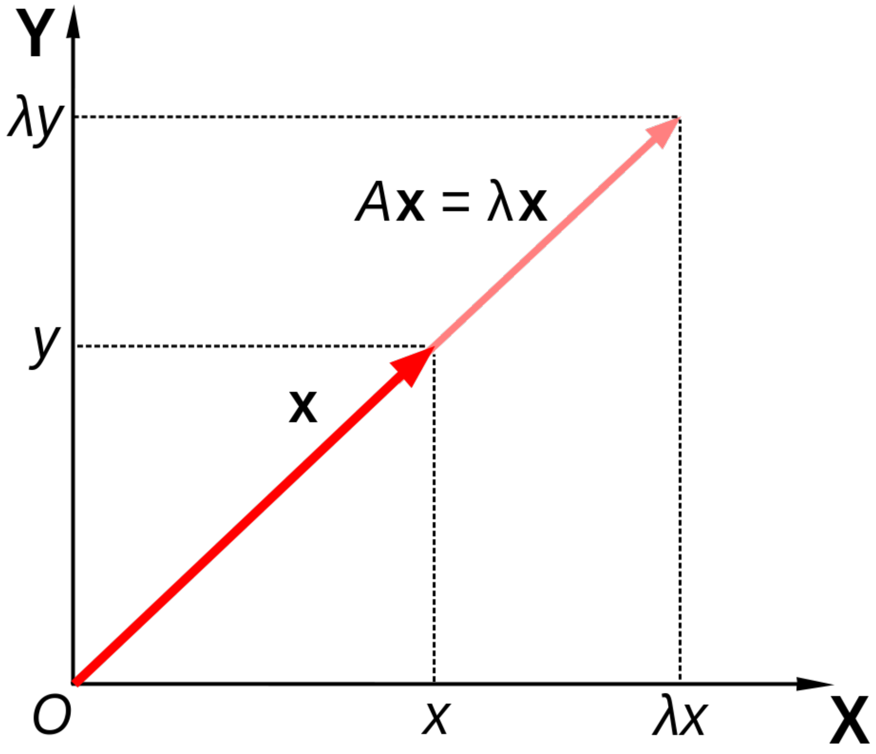
normalized eigenvector

- Each input data (vector) is a linear combination of the k **principal component vectors**

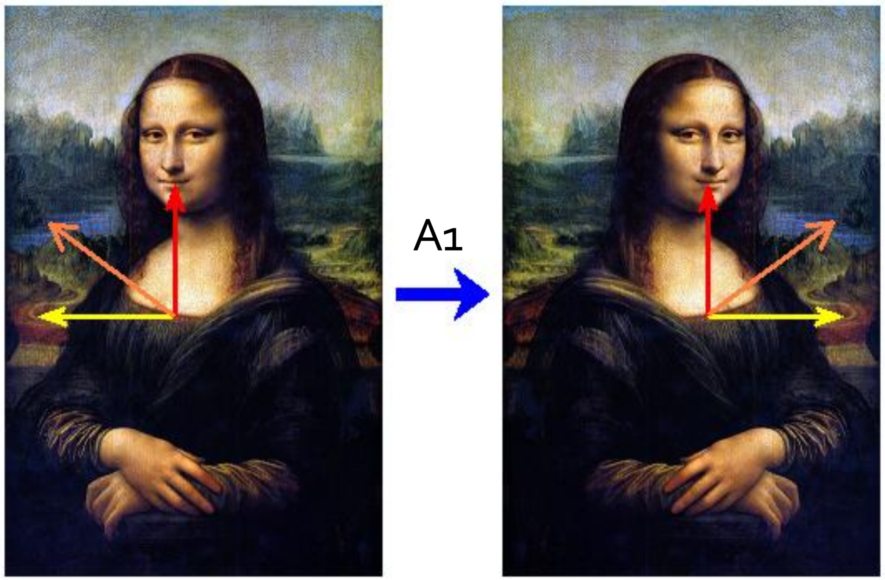


Eigenvectors (cont.)

- For a square matrix **A** (n*n), find the eigenvector **x** (n*1).
  - **A** represents the linear transformation (from n to n)
- Matrix **A** acts by stretching the vector **x**, not changing its direction, so **x** is an eigenvector of **A**.



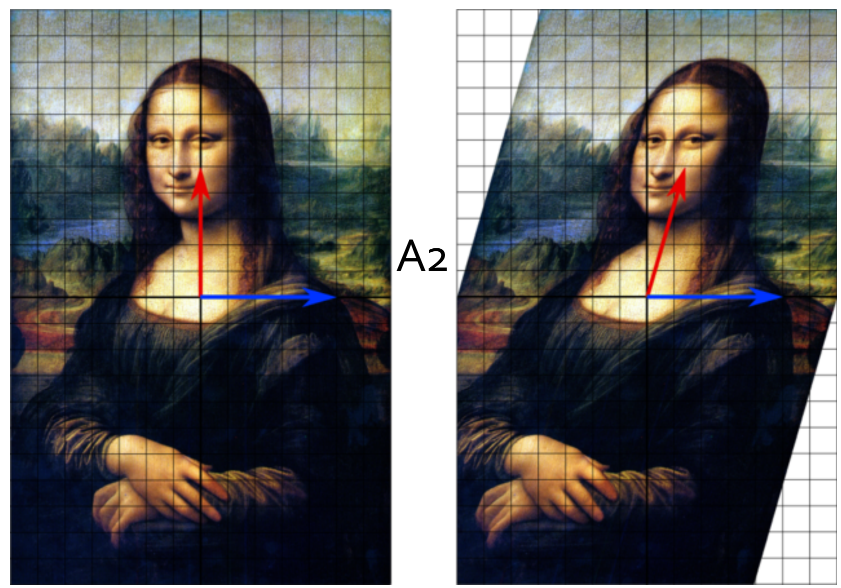
Eigenvectors (cont.)



Which vectors are eigenvectors?
- Red
- Orange
- Yellow

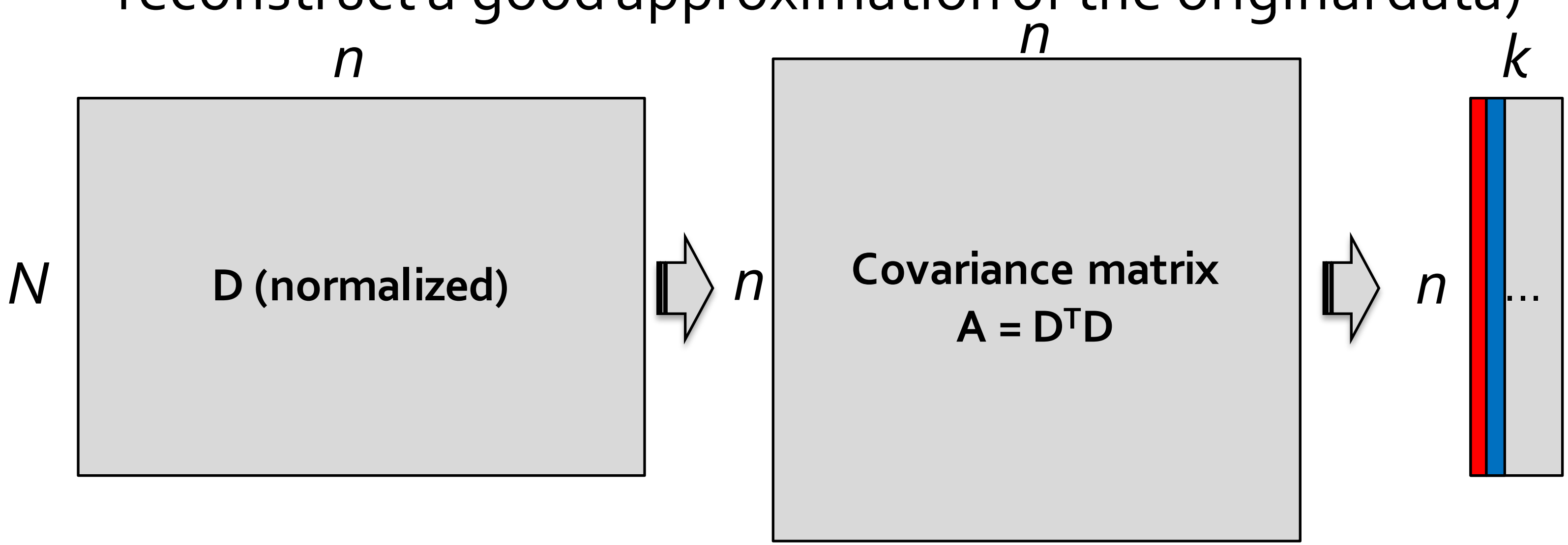
What are the eigenvalues?

Which vectors are eigenvectors?
- Red
- Blue



PCA and Eigenvectors

- For ***Square Matrix***: Data matrix to Covariance matrix
- The principal components are sorted in order of **decreasing "significance" or strength**
- **From n to k:** Since the components are sorted, the size of the data can be reduced by eliminating the weak components (i.e., using the strongest principal components, to reconstruct a good approximation of the original data)



PCA and Eigenvectors (cont.)

- Method: Find the **eigenvectors of covariance (square) matrix**, and these eigenvectors define the new space

$$\begin{aligned} \mathbf{A}\mathbf{x} = \lambda\mathbf{x} \quad &\Leftrightarrow \quad \mathbf{A}\mathbf{x} - \lambda\mathbf{x} = \mathbf{0} \\ &\Leftrightarrow \quad \mathbf{A}\mathbf{x} - \lambda\mathbf{I}\mathbf{x} = \mathbf{0} \\ &\Leftrightarrow \quad (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}. \end{aligned}$$

The equation $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ has nonzero solutions for the vector x if and only if the matrix $\mathbf{A} - \lambda\mathbf{I}$ has zero determinant.

Example: Find the eigenvalues of the matrix $\mathbf{A} = \begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix}$.

Ex. Eigenvalues

Example: Find the eigenvalues of the matrix $\mathbf{A} = \begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix}$.

The eigenvalues are those λ for which $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$. Now

$$
\begin{aligned}
\det(\mathbf{A} - \lambda\mathbf{I}) &= \det\left(\begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix} - \lambda\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \\
&= \det\left(\begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right) \\
&= \begin{vmatrix} 2-\lambda & 2 \\ 5 & -1-\lambda \end{vmatrix} \\
&= (2-\lambda)(-1-\lambda) - 10 \\
&= \lambda^2 - \lambda - 12.
\end{aligned}
$$

The eigenvalues of $\mathbf{A}$ are the solutions of the quadratic equation $\lambda^2 - \lambda - 12 = 0$, namely $\lambda_1 = -3$ and $\lambda_2 = 4$.

Ex. Eigenvectors

First, we work with $\lambda = -3$. The equation $\mathbf{Ax} = \lambda\mathbf{x}$ becomes $\boxed{\mathbf{Ax} = -3\mathbf{x}.}$ Writing

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

and using the matrix $\mathbf{A}$ from above, we have

$$\mathbf{Ax} = \boxed{\begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}} = \begin{bmatrix} 2x_1 + 2x_2 \\ 5x_1 - x_2 \end{bmatrix},$$

while

$$-3\mathbf{x} = \begin{bmatrix} -3x_1 \\ -3x_2 \end{bmatrix}.$$

Setting these equal, we get

$$\boxed{\begin{bmatrix} 2x_1 + 2x_2 \\ 5x_1 - x_2 \end{bmatrix} = \begin{bmatrix} -3x_1 \\ -3x_2 \end{bmatrix}} \Rightarrow \quad 2x_1 + 2x_2 = -3x_1 \quad \text{and} \quad 5x_1 - x_2 = -3x_2$$

$$\Rightarrow \quad 5x_1 = -2x_2$$

$$\Rightarrow \quad \boxed{x_1 = -\frac{2}{5}x_2.} \qquad \boxed{\mathbf{u_1} = \begin{bmatrix} 2 \\ -5 \end{bmatrix}}$$

Ex. Eigenvectors (cont.)

Similarly, we can find eigenvectors associated with the eigenvalue $\lambda = 4$ by solving $\boxed{\mathbf{Ax} = 4\mathbf{x}:}$

$$\begin{bmatrix} 2x_1 + 2x_2 \\ 5x_1 - x_2 \end{bmatrix} = \begin{bmatrix} 4x_1 \\ 4x_2 \end{bmatrix} \;\Rightarrow\; 2x_1 + 2x_2 = 4x_1 \quad \text{and} \quad 5x_1 - x_2 = 4x_2$$

$$\Rightarrow\; x_1 = x_2.$$

Hence the set of eigenvectors associated with $\lambda = 4$ is spanned by

$$\boxed{\mathbf{u_2} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}}.$$

Ex. Eigenvalues (cont.)

Example: Find the eigenvalues and associated eigenvectors of the matrix

$$\mathbf{A} = \begin{bmatrix} 7 & 0 & -3 \\ -9 & -2 & 3 \\ 18 & 0 & -8 \end{bmatrix}.$$

First we compute $\det(\mathbf{A} - \lambda\mathbf{I})$ via a cofactor expansion along the second column:

$$\begin{aligned} \begin{vmatrix} 7-\lambda & 0 & -3 \\ -9 & -2-\lambda & 3 \\ 18 & 0 & -8-\lambda \end{vmatrix} &= (-2-\lambda)(-1)^4 \begin{vmatrix} 7-\lambda & -3 \\ 18 & -8-\lambda \end{vmatrix} \\ &= -(2+\lambda)[(7-\lambda)(-8-\lambda)+54] \\ &= -(\lambda+2)(\lambda^2+\lambda-2) \\ &= -(\lambda+2)^2(\lambda-1). \end{aligned}$$

Thus $\mathbf{A}$ has two distinct eigenvalues, $\lambda_1 = -2$ and $\lambda_3 = 1$. (Note that we might say $\lambda_2 = -2$, since, as a root, -2 has multiplicity two. This is why we labelled the eigenvalue 1 as λ_3.)

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 condition to the first, ...
  - Step-wise attribute elimination:
    - Repeatedly eliminate the worst attribute
  - Best combined attribute selection and elimination

Summary

- **Data quality**: accuracy, completeness, consistency, timeliness, believability, interpretability
- **Data cleaning**: e.g. missing/noisy values, outliers
- **Data integration** from multiple sources:
  - Correlation analysis: Chi-Square test, Covariance
- **Data reduction and data transformation**
  - Normalization: Z-score normalization
- **Dimensionality reduction**
  - PCA, Heuristic Search in Attribute Selection

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