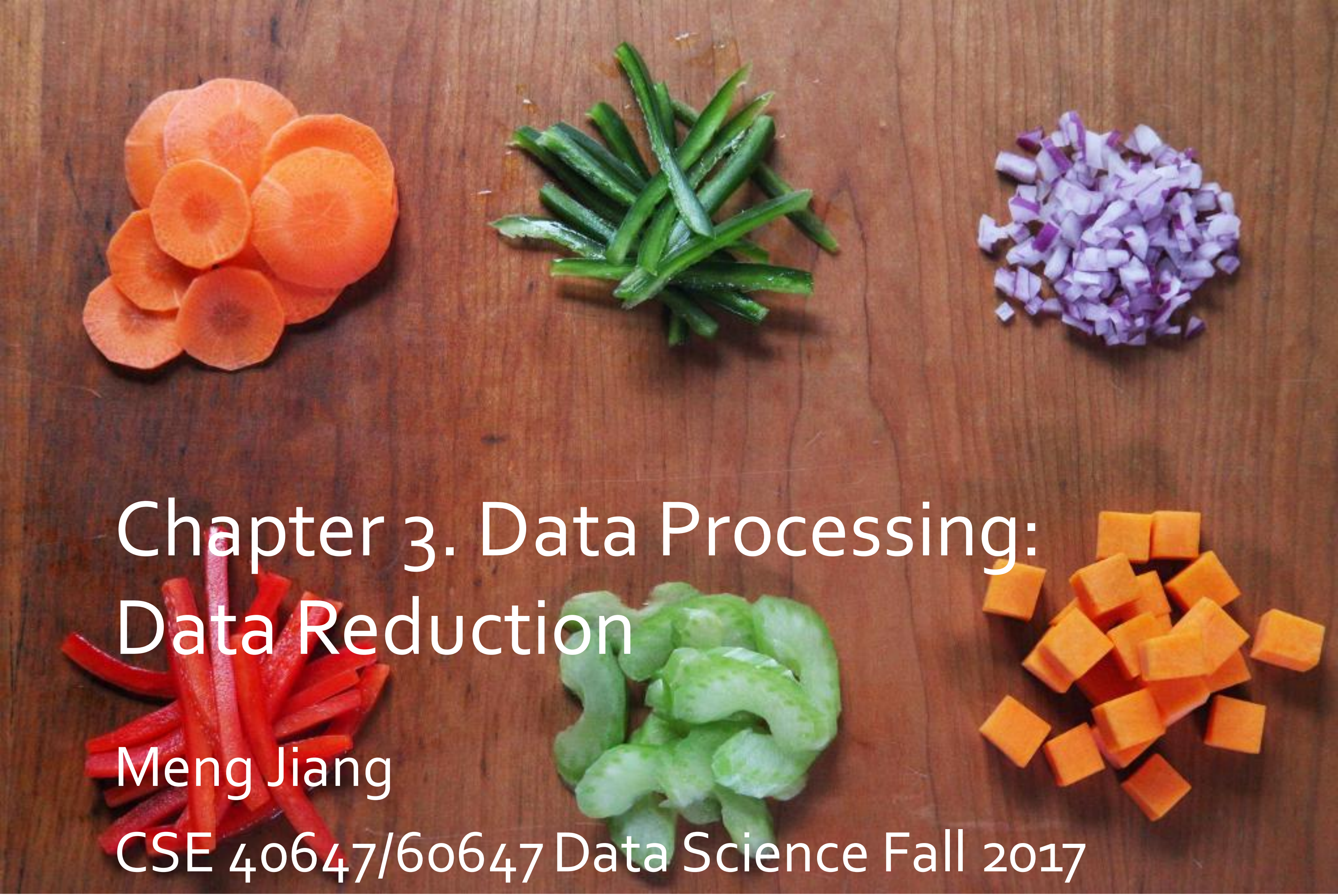


Chapter 3. Data Processing: Data Reduction

Meng Jiang

CSE 40647/60647 Data Science Fall 2017

Introduction to Data Mining

Data Preprocessing

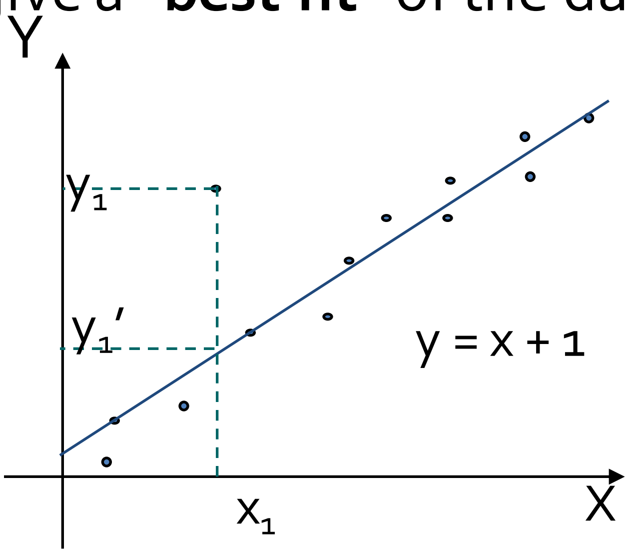
- Data cleaning
- Data integration
- **Data reduction**
  - Reduce data objects
- Dimensionality reduction
  - Reduce dimensions and attributes

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 normalization

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. $y_1' = x_1 + 1$

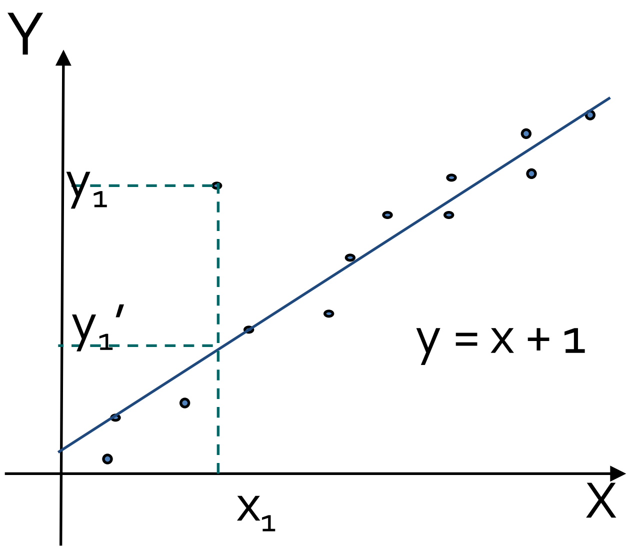


Regression Analysis

- 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 $y'_i = f(x_i, \beta)$

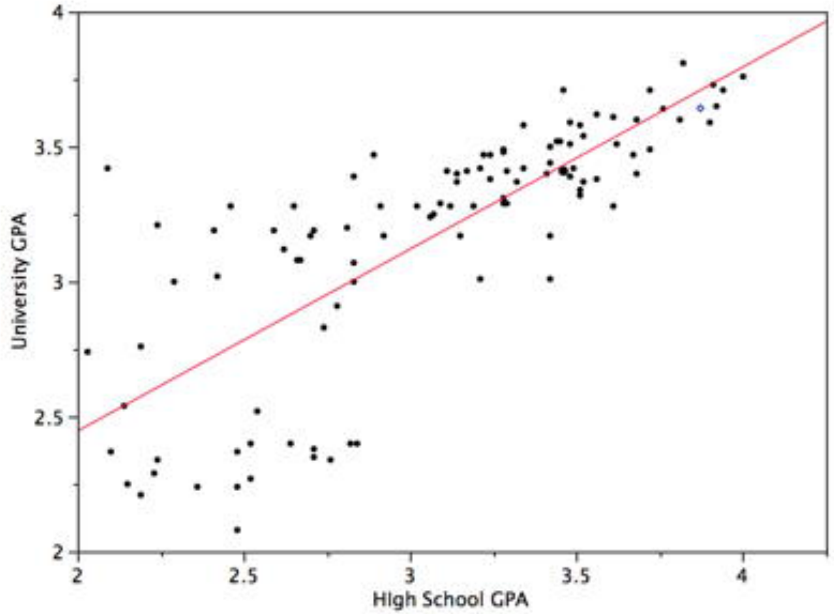
- Used for **prediction** (including forecasting of time-series data), **inference**, **hypothesis testing**, and **modeling of causal relationships**



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

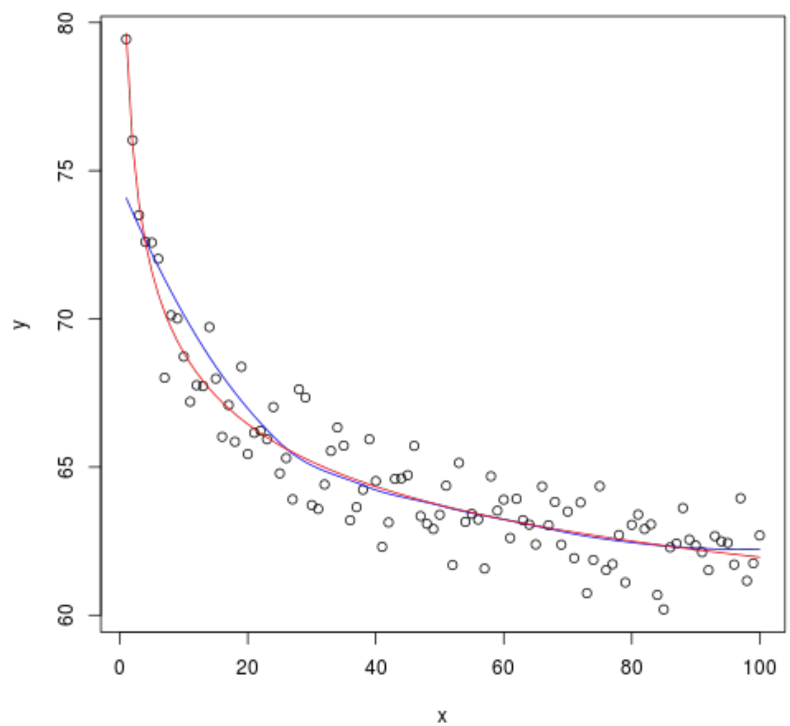
Linear Regression

- 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

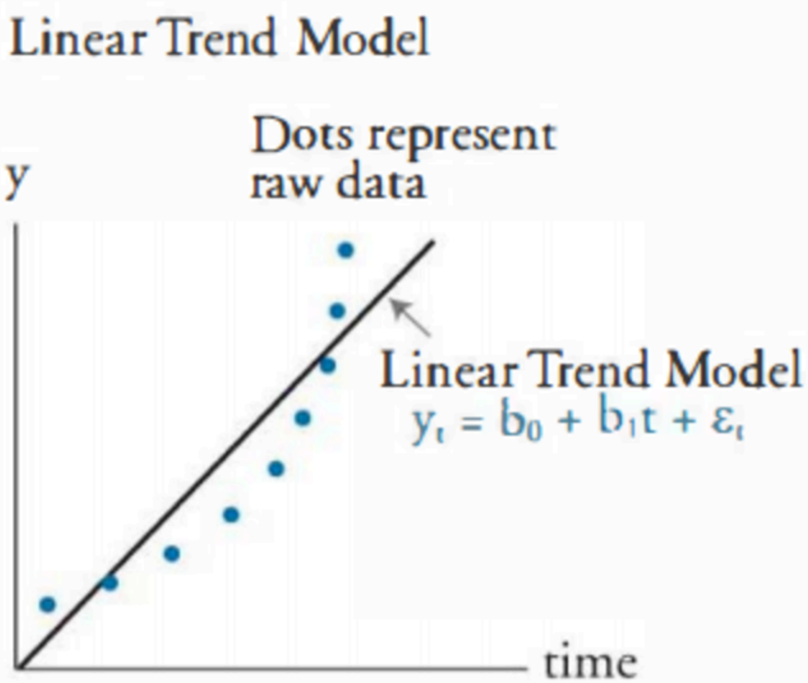
- Nonlinear regression:
  - Data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables

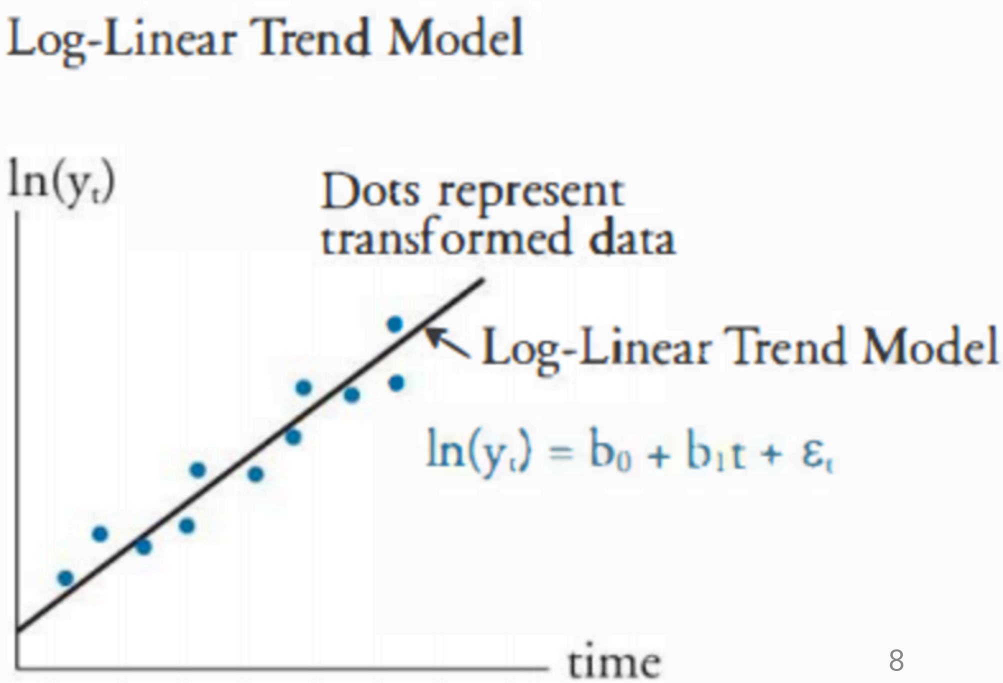


Log-Linear Model

- 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

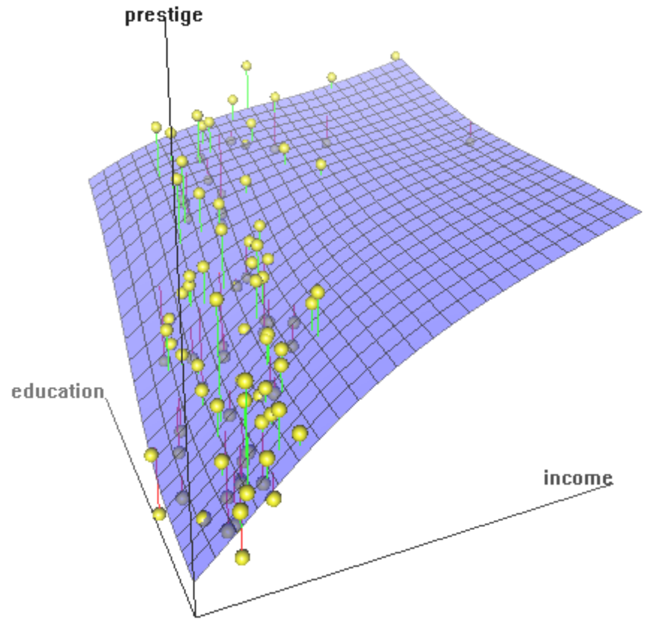
Q: How about Log-Log model?





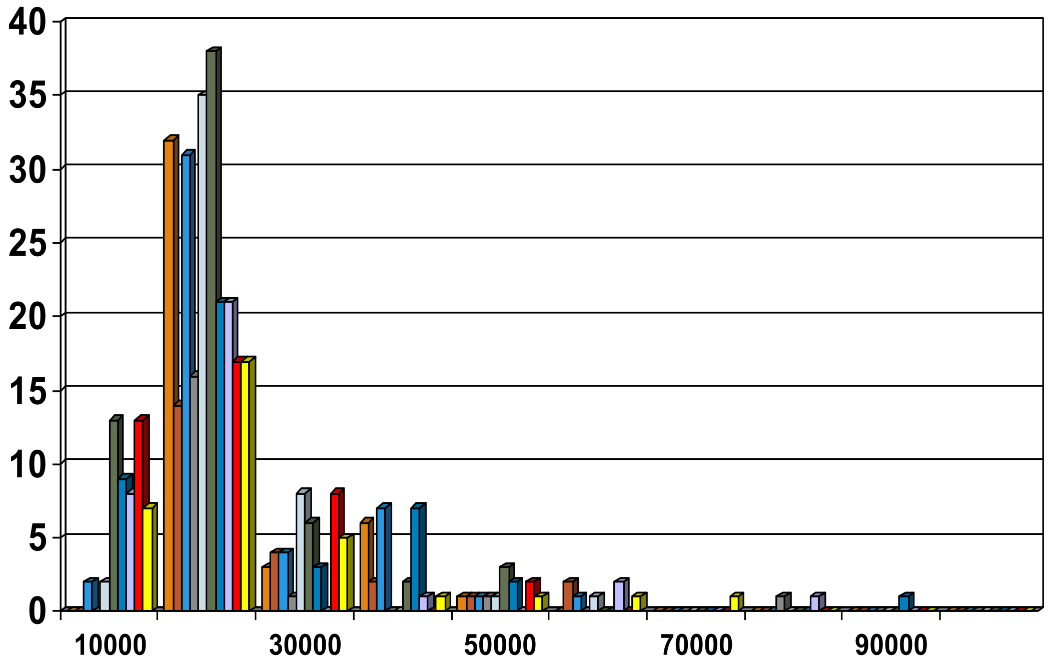
Multiple Regression

- 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

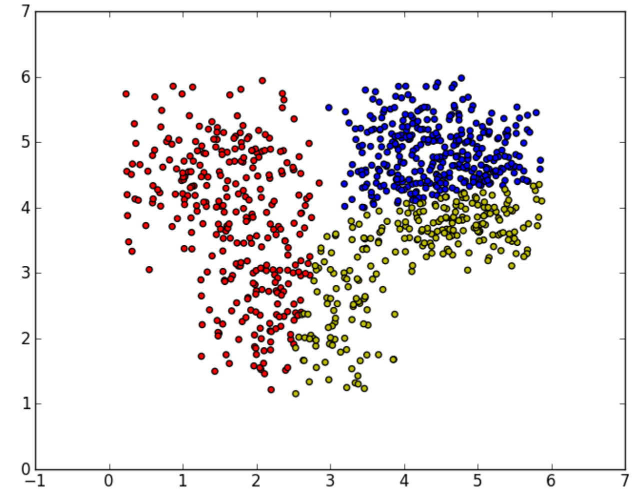
(10,000 , 11,000]

(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”
- There are many choices of clustering definitions and clustering algorithms
- Cluster analysis will be studied in depth in Chapter 10



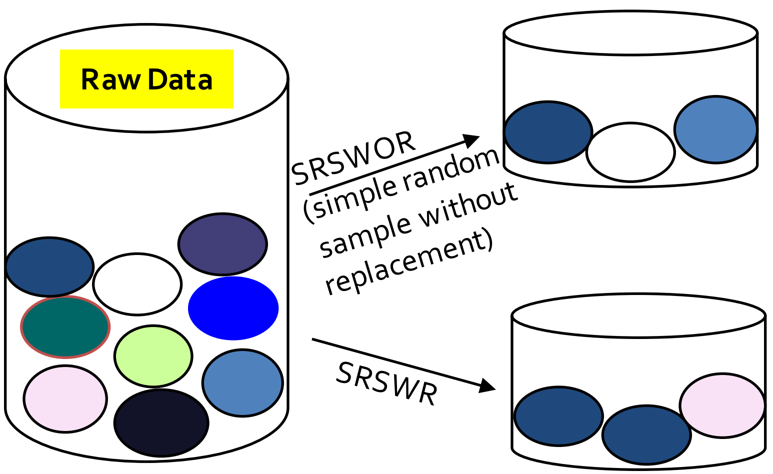
Sampling

- 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

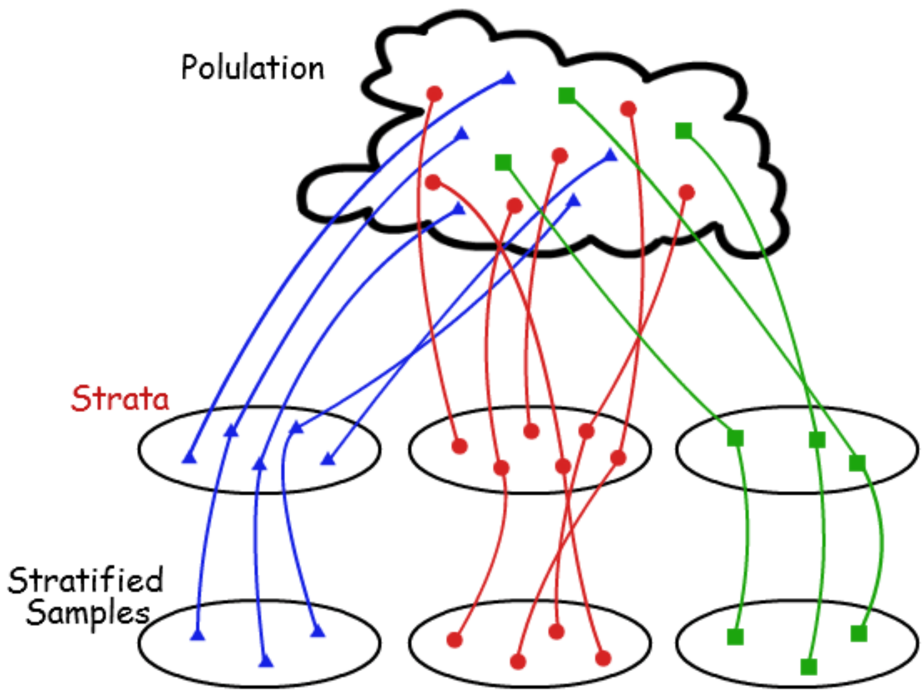
Sampling without replacement:
Once an object is selected, it is removed from the population

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)



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, …

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

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

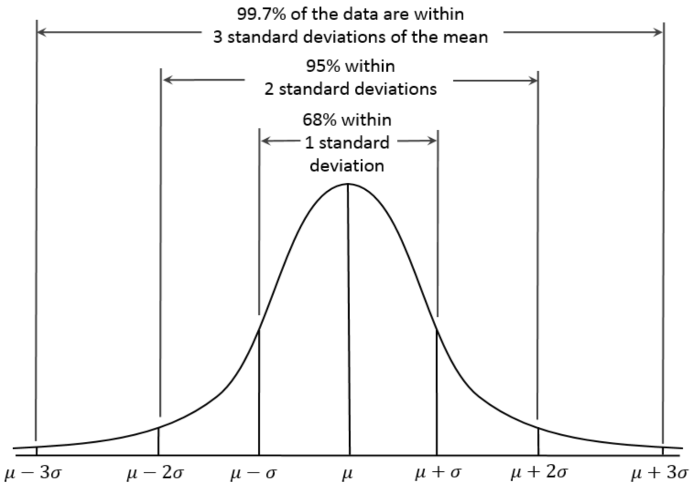
Normalization

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

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

  – Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

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



Normalization

- **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

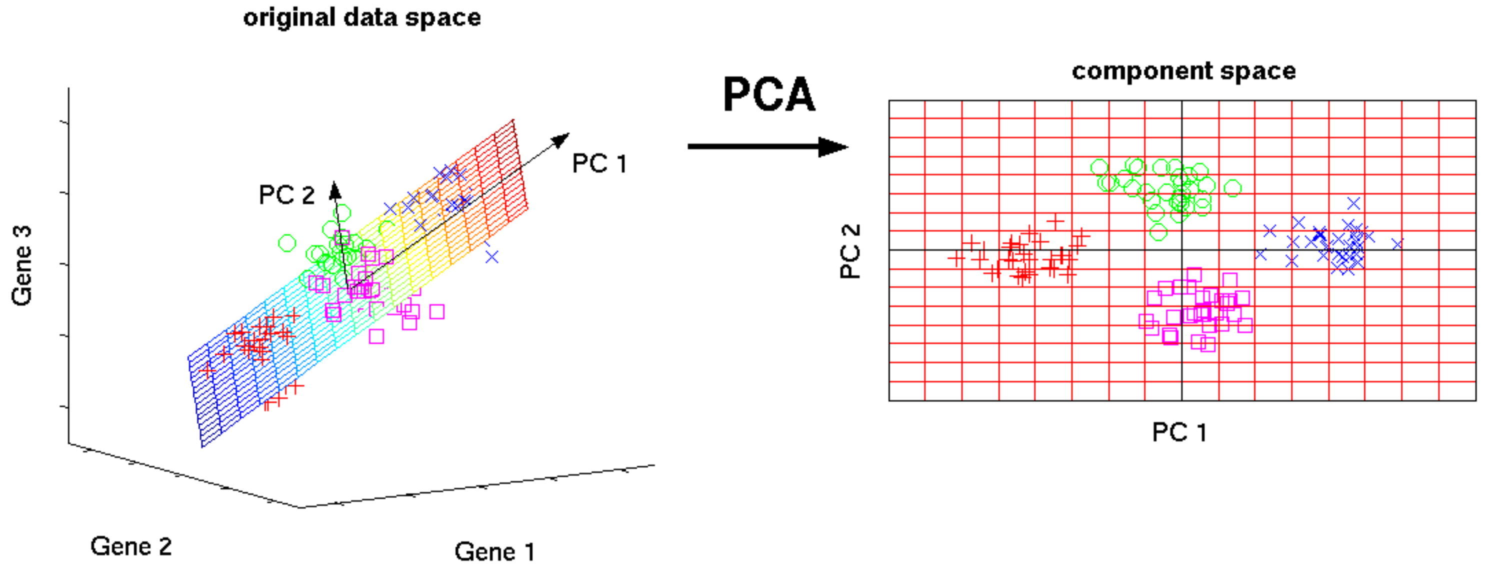
- Curse of dimensionality
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially
- Dimensionality reduction
  - 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

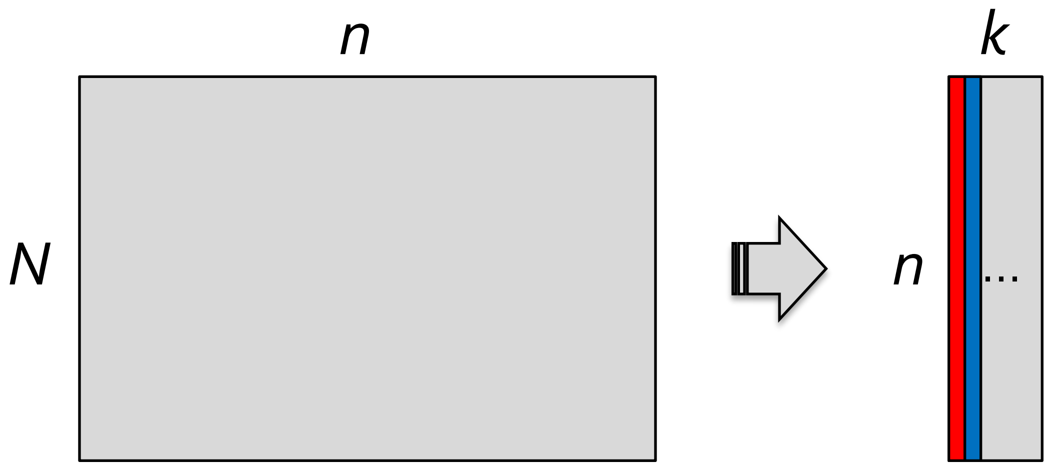
Principal Component Analysis (PCA)

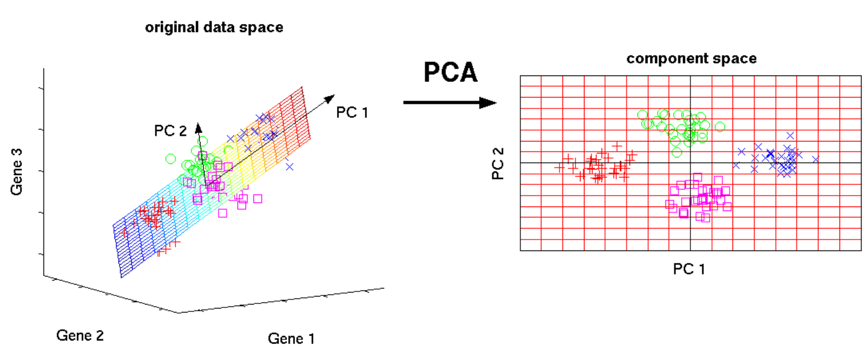
- PCA: A statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called ***principal components***
- The original data are projected onto a **much smaller space**, resulting 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



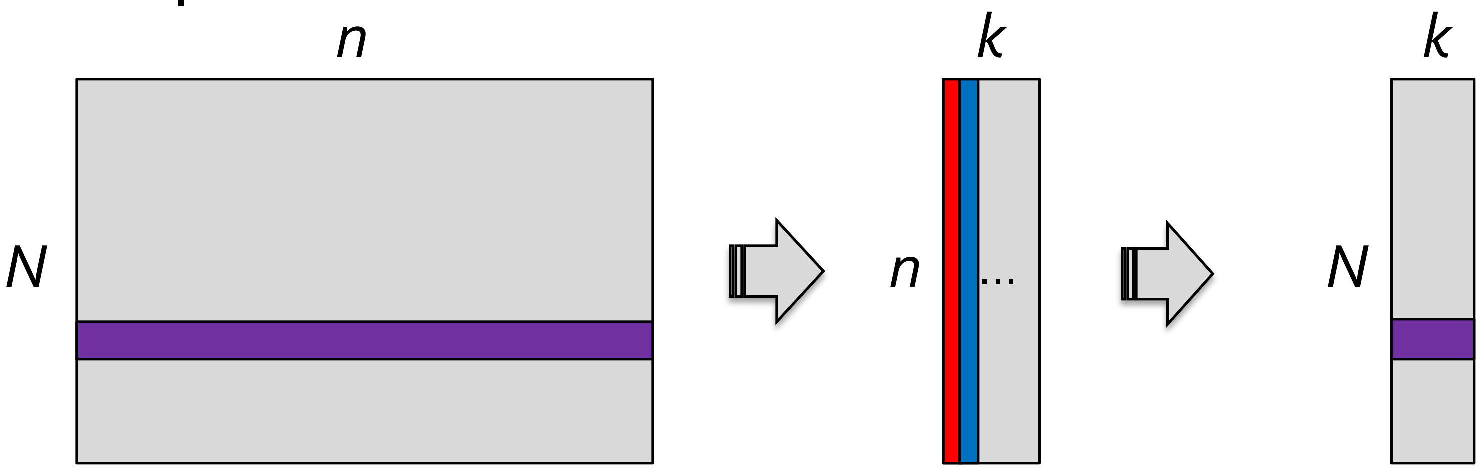


PCA (cont.)

- 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

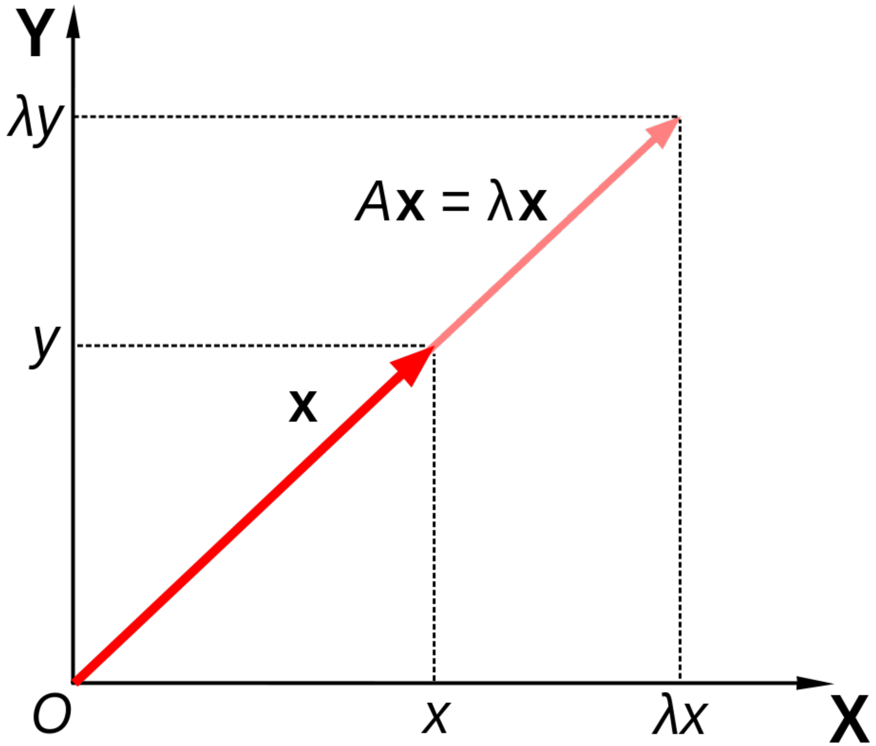
normalized eigenvector

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

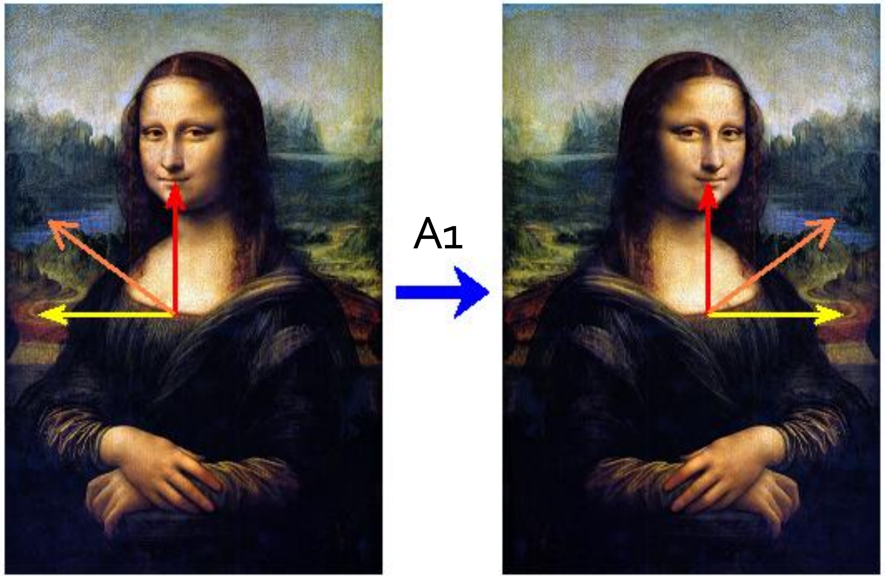


Eigenvectors (cont.)

- 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.)



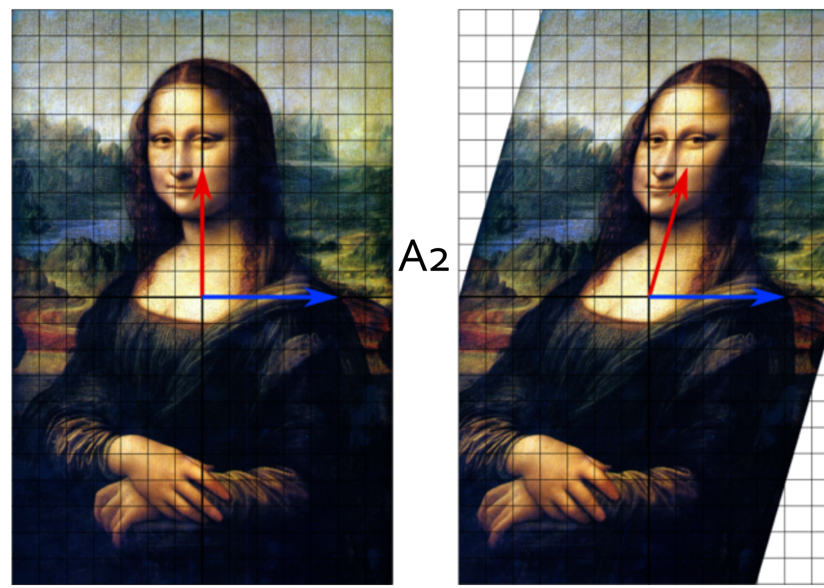
Which vectors are eigenvectors?

- Red
- Orange
- Yellow

What are the eigenvalues?

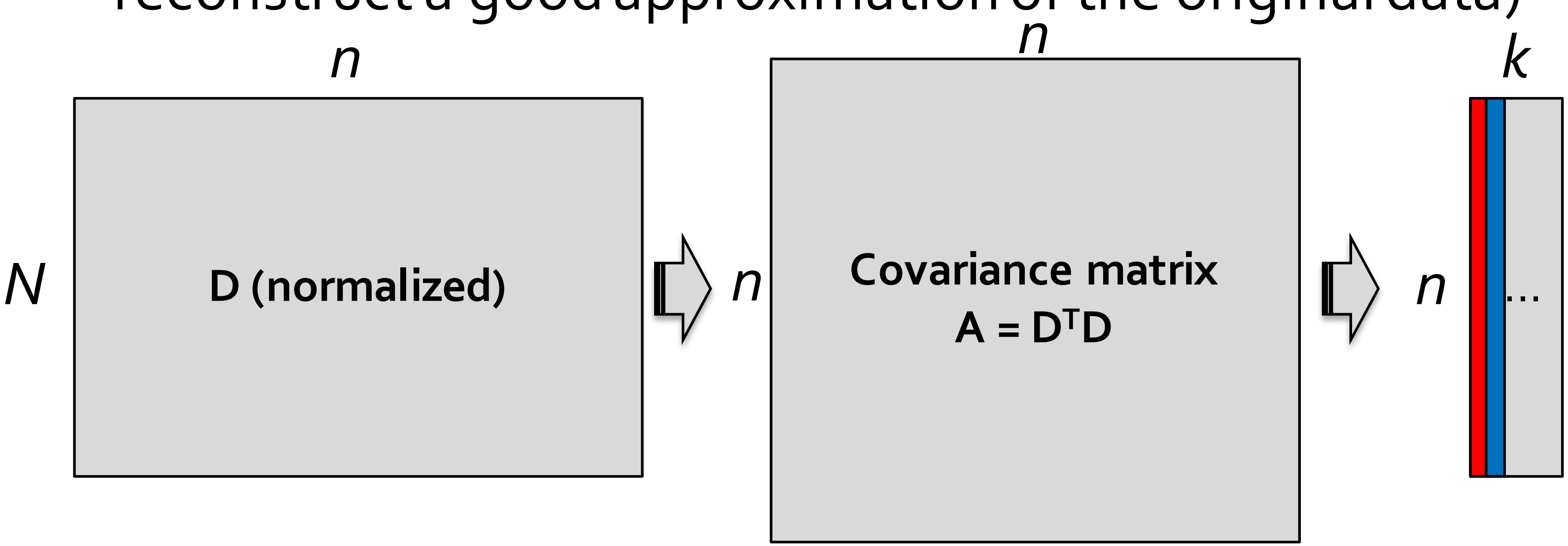
Which vectors are eigenvectors?

- 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

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

The equation $\mathbf{Ax} = \lambda\mathbf{x}$ 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 $\boxed{\mathbf{Ax} = 4\mathbf{x}:}$

$$\begin{bmatrix} 2x_1 + 2x_2 \\ 5x_1 - x_2 \end{bmatrix} = \begin{bmatrix} 4x_1 \\ 4x_2 \end{bmatrix} \;\; \begin{aligned} &\Rightarrow \;\; 2x_1 + 2x_2 = 4x_1 \qquad \text{and} \qquad 5x_1 - x_2 = 4x_2 \\ &\Rightarrow \;\; x_1 = x_2. \end{aligned}$$

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