



# DATA ANALYTICS

## Unit 1:Data Reduction

---

**Dr Mamatha.H.R and Bharathi R**

Department of Computer Science and  
Engineering

# DATA ANALYTICS

---

## Unit 1:Data Reduction

**Mamatha H R**

Department of Computer Science and Engineering

**Data reduction:** Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results

Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.

### Data reduction strategies

- Dimensionality reduction, e.g., remove unimportant attributes
  - Wavelet transforms
  - Principal Components Analysis (PCA)
  - Feature subset selection, feature creation
- Numerosity reduction (some simply call it: Data Reduction)
  - Regression and Log-Linear Models
  - Histograms, clustering, sampling
  - Data cube aggregation
- Data compression- compressed rep. of original data

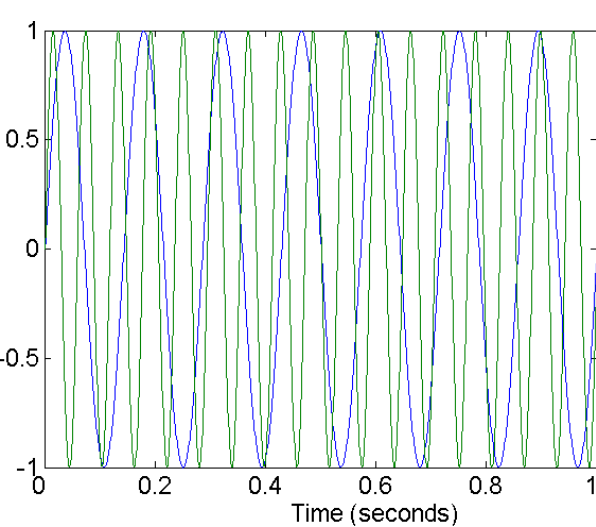
## Data Reduction 1: Dimensionality Reduction

---

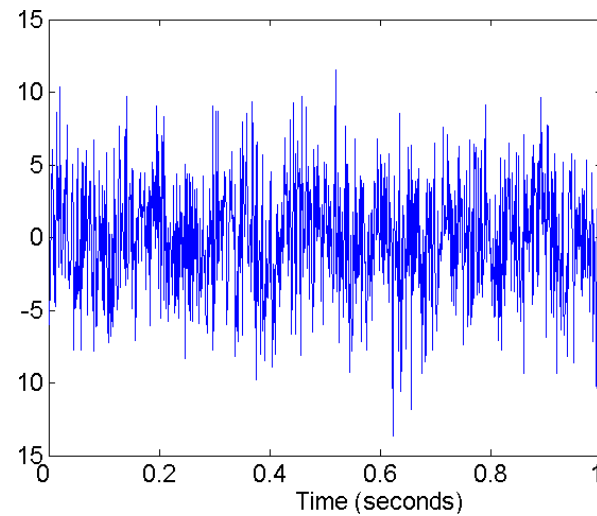
- **Curse of dimensionality**
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially

- **Dimensionality reduction**
  - 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**
  - Wavelet transforms
  - Principal Component Analysis
  - Supervised and nonlinear techniques (e.g., feature selection)

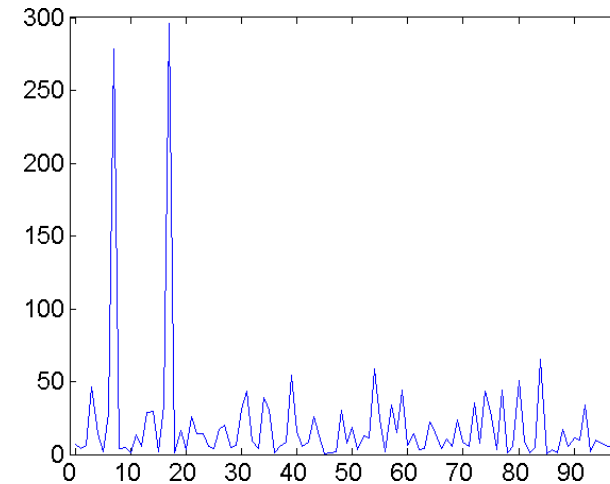
1. Fourier transform
2. Wavelet transform



**Two Sine Waves**



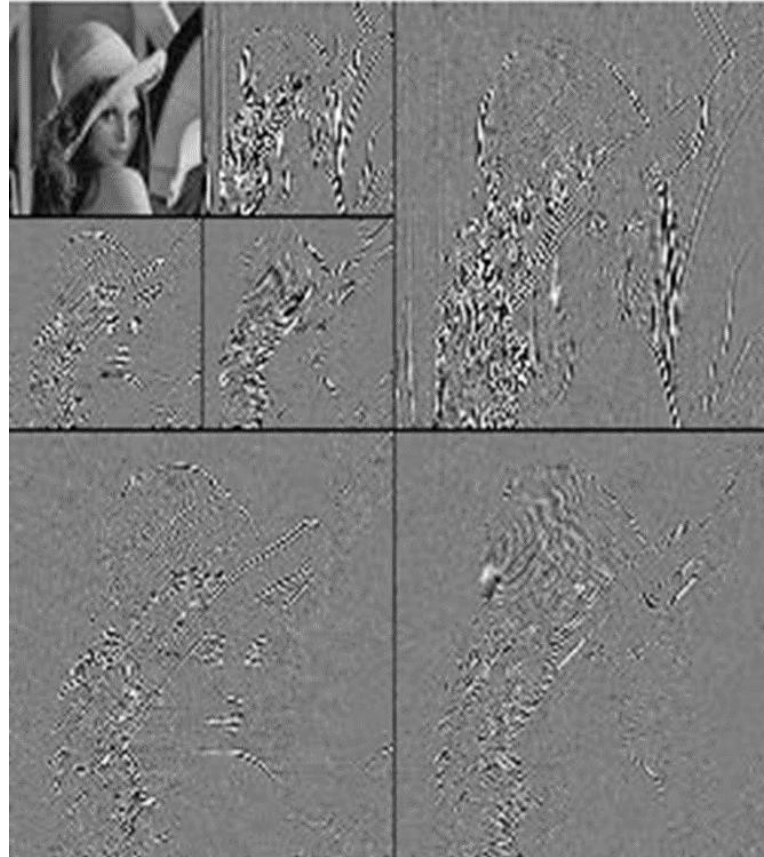
**Two Sine Waves + Noise**



**Frequency**

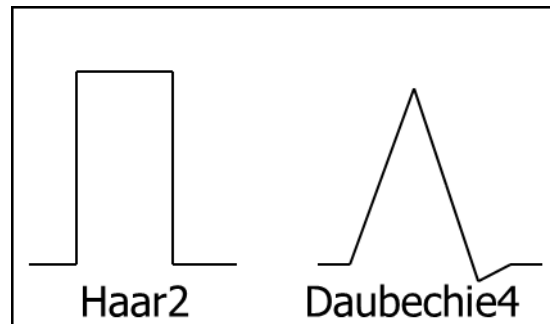
## What Is Wavelet Transform?

- 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





- 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

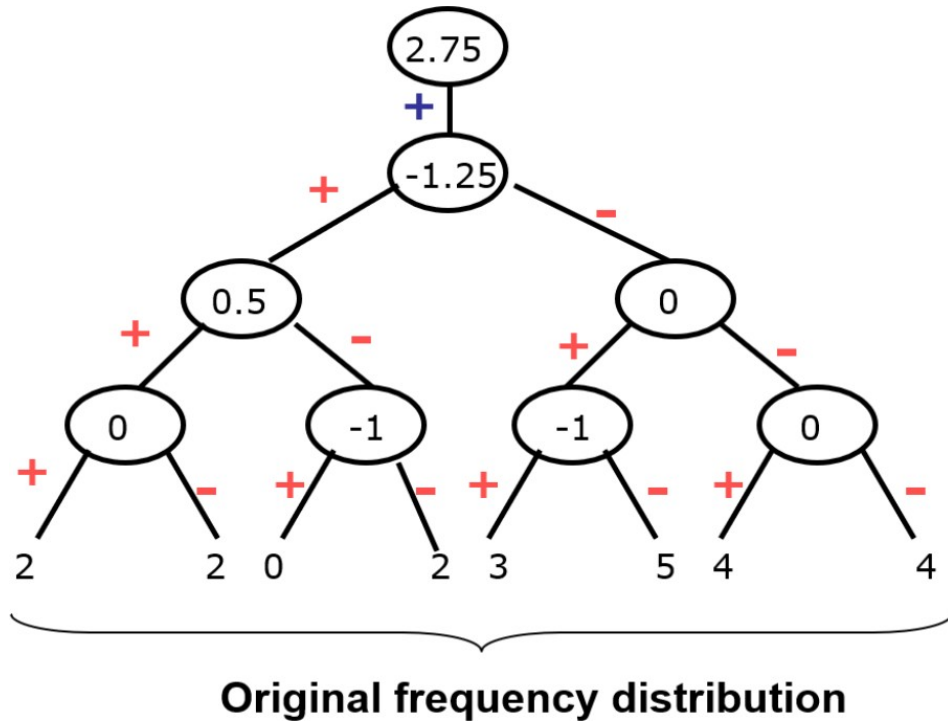


- Method:
  - Length,  $L$ , must be an integer power of 2 (padding with 0's, when necessary)
  - Each transform has 2 functions: smoothing, difference
  - Applies to pairs of data, resulting in two set of data of length  $L/2$
  - Applies two functions recursively, until reaches the desired length

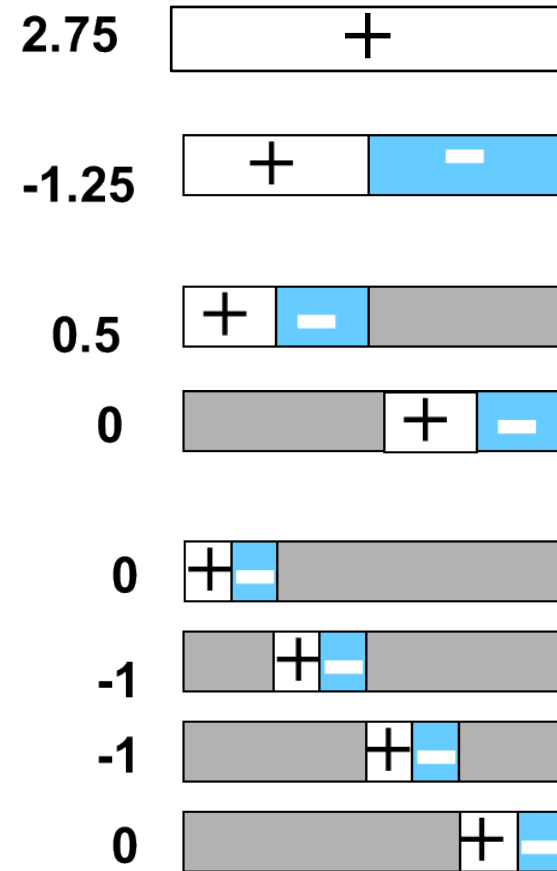
- Wavelets: A math tool for space-efficient hierarchical decomposition of functions
- $S = [2, 2, 0, 2, 3, 5, 4, 4]$  can be transformed to  $S_{\wedge} = [2^3/4, -1^1/4, 1/2, 0, 0, -1, -1, 0]$
- Compression: many small detail coefficients can be replaced by 0's, and only the significant coefficients are retained

Resolution	Averages	Detail Coefficients
8	$[2, 2, 0, 2, 3, 5, 4, 4]$	
4	$[2, 1, 4, 4]$	$[0, -1, -1, 0]$
2	$[1\frac{1}{2}, 4]$	$[\frac{1}{2}, 0]$
1	$[2\frac{3}{4}]$	$[-1\frac{1}{4}]$

Hierarchical decomposition structure (a.k.a. “error tree”)



Coefficient “Supports”



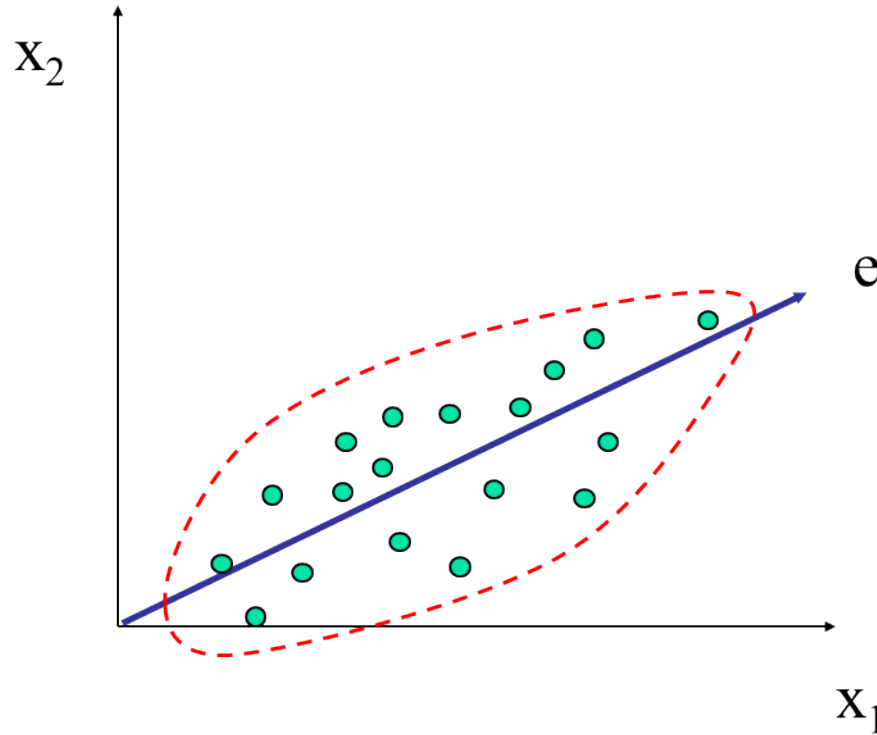
## Why Wavelet Transform?

---

- Use hat-shape filters
  - Emphasize region where points cluster
  - Suppress weaker information in their boundaries
- Effective removal of outliers
  - Insensitive to noise, insensitive to input order
- Multi-resolution
  - Detect arbitrary shaped clusters at different scales
- Efficient
  - Complexity  $O(N)$
- Only applicable to low dimensional data

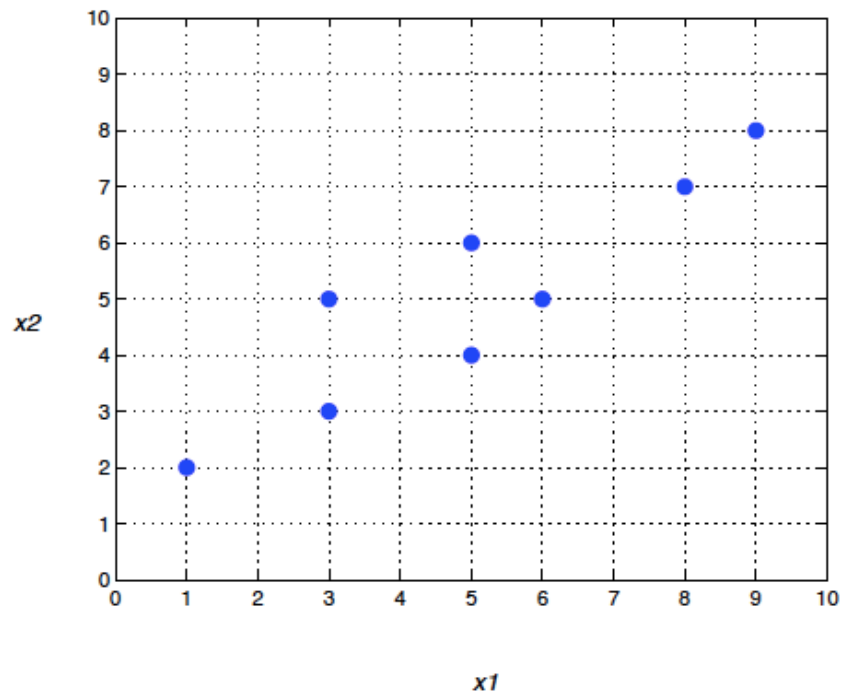
## Principal Component Analysis (PCA)

- Find a projection that captures the largest amount of variation in data
- The original data are projected onto a much smaller space, resulting in dimensionality reduction. We find the eigenvectors of the covariance matrix, and these eigenvectors define the new space



### *PCA – An Example*

- Compute the principal components for the following two-dimensional dataset
  - $X=(x_1, x_2)=\{(1,2), (3,3), (3,5), (5,4), (5,6), (6,5), (8,7), (9,8)\}$

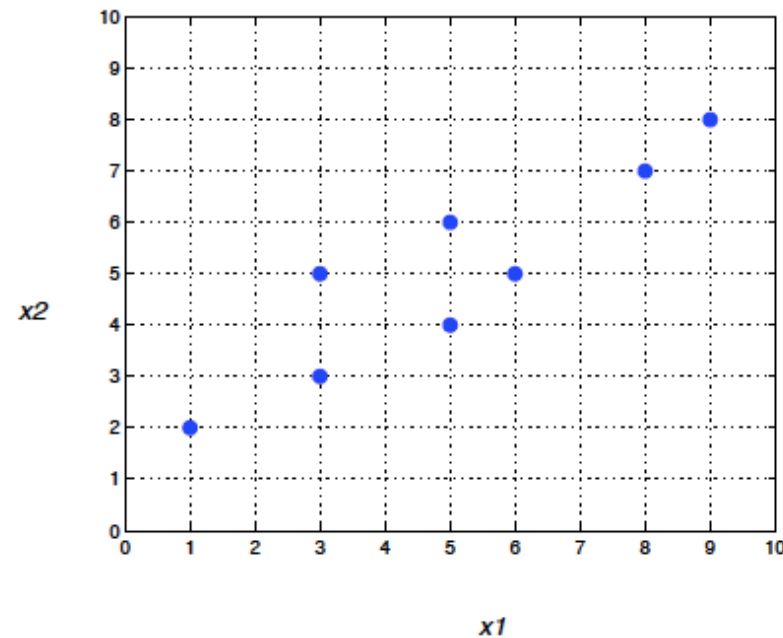


### PCA – An Example

- Compute the principal components for the following two-dimensional dataset

- $X=(x_1,x_2)=\{(1,2),(3,3),(3,5),(5,4),(5,6),(6,5),(8,7),(9,8)\}$

*Step 1: Determine the  
Sample Covariance Matrix*





### PCA – An Example

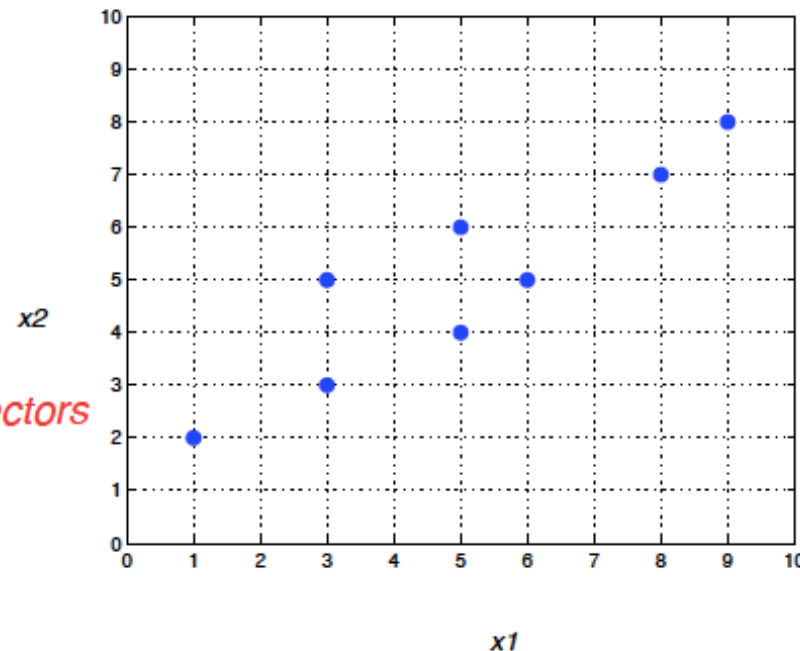
- Compute the principal components for the following two-dimensional dataset

- $X=(x_1,x_2)=\{(1,2),(3,3),(3,5),(5,4),(5,6),(6,5),(8,7),(9,8)\}$

*Step 1: Determine the Sample Covariance Matrix*

$$\Sigma = \begin{pmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{pmatrix}$$

*Step 2: Find eigenvalues and eigenvectors of the covariance matrix*



### PCA – An Example

- Compute the principal components for the following two-dimensional dataset

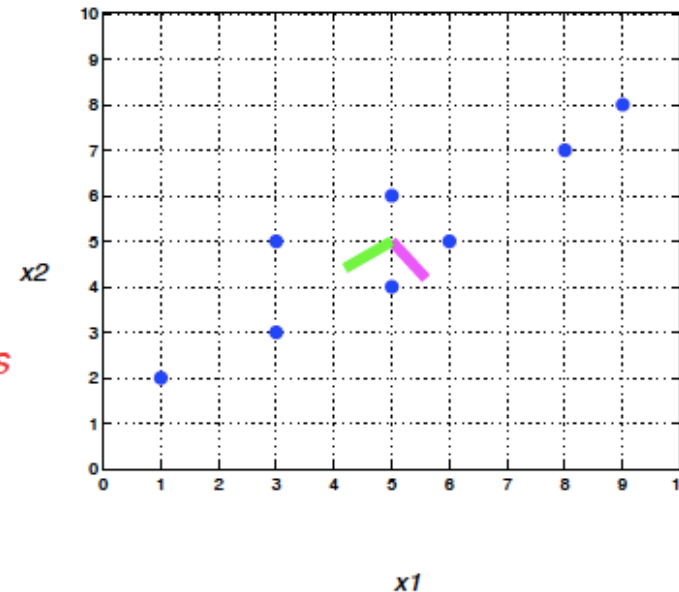
- $X=(x_1,x_2)=\{(1,2),(3,3),(3,5),(5,4),(5,6),(6,5),(8,7),(9,8)\}$

*Step 1: Determine Covariance Matrix*

$$\Sigma = \begin{pmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{pmatrix}$$

*Step 2: Find eigenvalues and eigenvectors of the covariance matrix*

$$\begin{aligned} \lambda &= 0.4081 & \lambda &= 9.3419 \\ \bar{v} &= \begin{pmatrix} 0.5883 \\ -0.8086 \end{pmatrix} & \bar{v} &= \begin{pmatrix} -0.5883 \\ -0.8086 \end{pmatrix} \end{aligned}$$



### PCA – An Example

- Compute the principal components for the following two-dimensional dataset

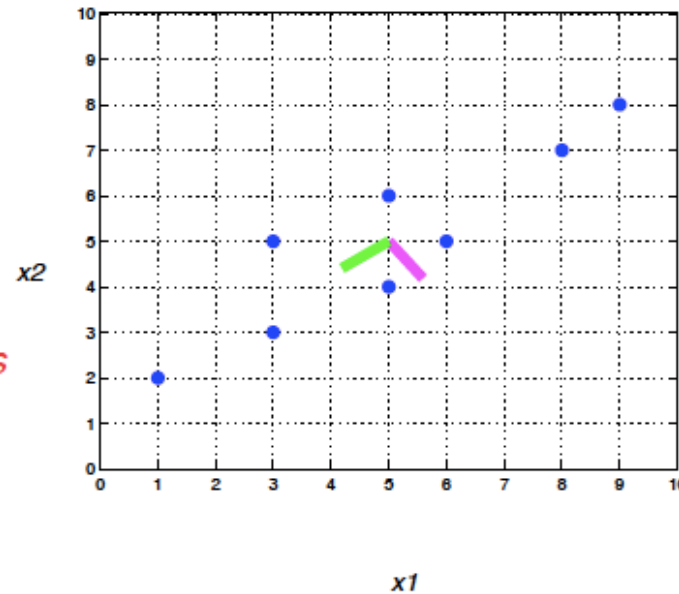
- $X=(x_1,x_2)=\{(1,2),(3,3),(3,5),(5,4),(5,6),(6,5),(8,7),(9,8)\}$

*Step 1: Determine Covariance Matrix*

$$\Sigma = \begin{pmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{pmatrix}$$

*Step 2: Find eigenvalues and eigenvectors of the covariance matrix*

$$\begin{aligned} \lambda &= 0.4081 & \lambda &= 9.3419 \\ \bar{v} &= \begin{pmatrix} 0.5883 \\ -0.8086 \end{pmatrix} & \bar{v} &= \begin{pmatrix} -0.5883 \\ -0.8086 \end{pmatrix} \end{aligned}$$



## Principal Component Analysis (Steps)

---

- Given  $N$  data vectors from  $n$ -dimensions, find  $k \leq n$  orthogonal vectors (*principal components*) that can be best used to represent data
  - Normalize input data: Each attribute falls within the same range
  - Compute  $k$  orthonormal (unit) vectors, i.e., *principal components*
  - Each input data (vector) is a linear combination of the  $k$  principal component vectors
  - The principal components are sorted in order of decreasing “significance” or strength
  - Since the components are sorted, the size of the data can be reduced by eliminating the *weak components*, i.e., those with low variance (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data)
- Works for numeric data only

**Objective:** project the data onto a lower dimensional linear space such that the variance of the projected data is maximized.

Equivalently, it is the linear projection that minimizes the average projection cost (mean squared distance between the data points and their projections).

*Different from the feature subset selection!!!*

Problem to solve: In high dimensional space, we need to learn a large number of parameters. Thus if the dataset is small, this will result in large variance and over-fitting.

We want to represent the vector  $x$  in a different space ( $p$  dimensional) using a set of orthonormal vectors  $U$  (where  $u_i$  is a principle component).

### Method to perform PCA on a data

1. Standardize the data.
2. Obtain the Eigenvectors and Eigenvalues from the covariance matrix or correlation matrix (perform Singular Vector Decomposition )
3. Sort eigenvalues in descending order and choose the  $k$  eigenvectors that correspond to the  $k$  largest eigenvalues where  $k$  is the number of dimensions of the new feature subspace ( $k \leq d$ ) ( $d$  : Number of attributes)
4. Transform the original dataset  $X$  via  $W$  to obtain a  $k$ -dimensional feature subspace  $Y$ .
5. Reconstruct the original dataset.

Method to perform PCA on a data

### Standardize the Data

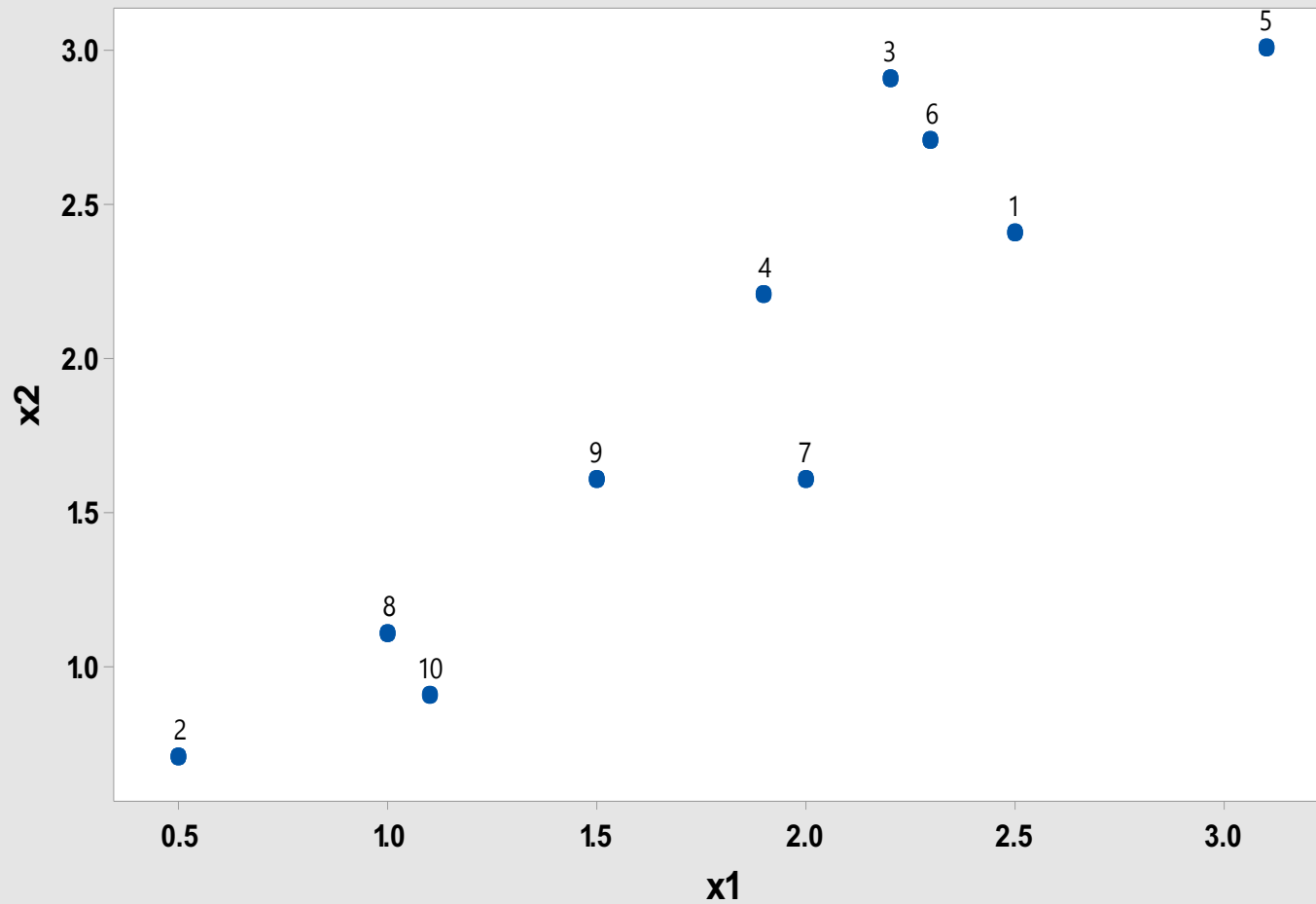
► It involves the following steps :

1. Get data : In this example, a 2 dimensional data set is used

x1	x2
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2.0	1.6
1.0	1.1
1.5	1.6
1.1	0.9

## Method to perform PCA on a data

Scatterplot of x2 vs x1





### Method to perform PCA on a data

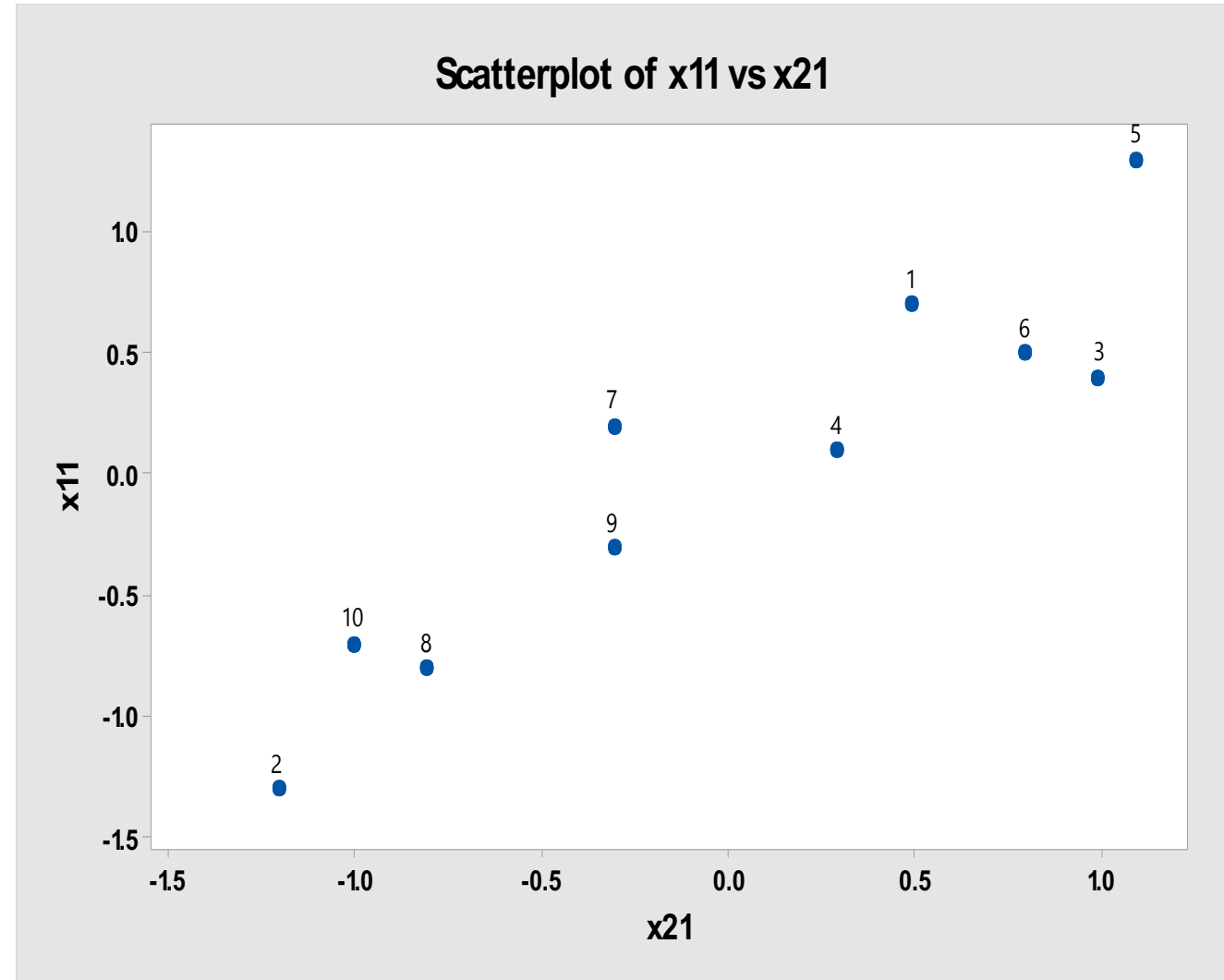
---

## Standardize the Data

- ▶ Mean of  $x_1 = 1.81$
- ▶ Mean of  $x_2 = 1.91$
- 2. Subtract the means from the corresponding data component to re-center the data set.
- 3. Write the "adjusted" data as a matrix. The "adjusted" data set will have means zero.

## Method to perform PCA on a data

$X_1$	$X_2$
0.69	0.49
-1.31	-1.21
0.39	0.99
0.09	0.29
1.29	1.09
0.49	0.79
0.19	-0.31
-0.81	-0.81
-0.31	-0.31
-0.71	-1.01



### Method to perform PCA on a data

---

- . Compute the sample variance-covariance matrix C.

$$C = \begin{bmatrix} 0.616556 & 0.615444 \\ 0.615444 & 0.716556 \end{bmatrix}$$

Since the non-diagonal elements in this covariance matrix are positive, both the X1 and X2 variables increase together.  
Since it is symmetric, the eigenvectors are orthogonal.

### Method to perform PCA on a data

---

2. Compute the eigenvalues  $\lambda$ , and (unit or normalized) eigenvectors  $x$  of  $C$ .

The eigenvalues are

$$\lambda_1 = 1.28403 \text{ and } \lambda_2 = 0.0490834$$

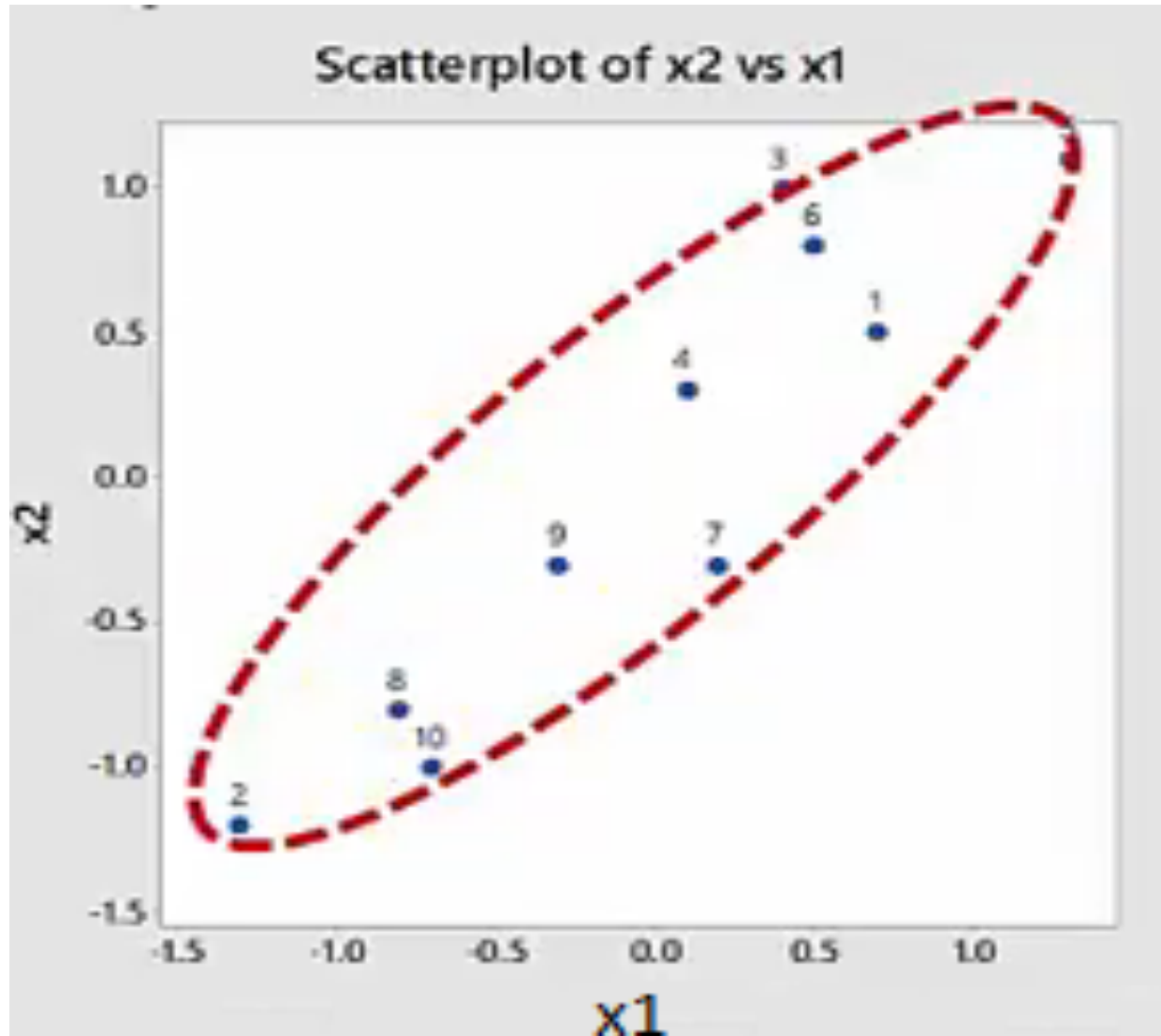
### Method to perform PCA on a data

---

Eigen Vectors are

- ▶  $\begin{bmatrix} 0.677873 \\ 0.735179 \end{bmatrix}$  for  $\lambda_1 = 1.28403$
- ▶  $\begin{bmatrix} 0.735179 \\ -0.677873 \end{bmatrix}$  for  $\lambda_2 = 0.0490834$

## Method to perform PCA on a data



### Method to perform PCA on a data

---

The total sample variance (Trace) = sum of eigen values =  
 $1.2840 + 0.0490 = 1.333$

Variable	Eigenvector 1	Eigenvector 2
$x_1$	0.678	0.735
$x_2$	0.735	-0.678
Eigen Values	1.2840	0.0490
% Of Total Variance	$1.2840/1.333$ = (96.3%)	$0.0490/1.333$ = (3.7%)

### Method to perform PCA on a data

---

## Sort Eigenvalues in Descending Order

- ▶ Once eigenvectors are found, the next step is to order them by eigenvalue, highest to lowest.
- ▶ This gives you the components in order of significance.
- ▶ The eigenvector with the highest eigenvalue is the principle component of the data set.
- ▶ If you like, you can decide to ignore the components of lesser significance.



## DATA ANALYTICS

### Method to perform PCA on a data

---

Form a Feature Vector and construct the projection matrix



### Method to perform PCA on a data

- ▶ This is constructed by taking the eigenvectors that you want to keep from the list of eigenvectors, and forming a matrix with these eigenvectors in the columns.
- ▶ Given our example set of data, and the fact that we have 2 eigenvectors, we have two choices.
- ▶ We can either form a feature vector with both of the eigenvectors:

$$V_1 = \begin{bmatrix} 0.677873 & 0.735179 \\ 0.735179 & -0.677873 \end{bmatrix}$$

- ▶ Or we can choose to leave out the smaller, less significant component and only have a single column:

$$V_2 = \begin{bmatrix} 0.677873 \\ 0.735179 \end{bmatrix}$$

Method to perform PCA on a data

---

# Transform the Original Dataset

► We can derive the new data set by taking

$$Z = XV$$

## Method to perform PCA on a data

▶  $V_2 = \begin{bmatrix} 0.677873 \\ 0.735179 \end{bmatrix}$

▶ Final data =

▶  $\begin{bmatrix} 0.69 & -1.31 & 0.39 & 0.09 & 1.29 & 0.49 & 0.19 & -0.81 & -0.31 & -0.1 \\ 0.49 & -1.21 & 0.99 & 0.29 & 1.09 & 0.79 & -0.31 & -0.81 & -0.31 & -1.1 \end{bmatrix}^T$

▶  $\times \begin{bmatrix} 0.677873 \\ 0.735179 \end{bmatrix}$

### Method to perform PCA on a data

---

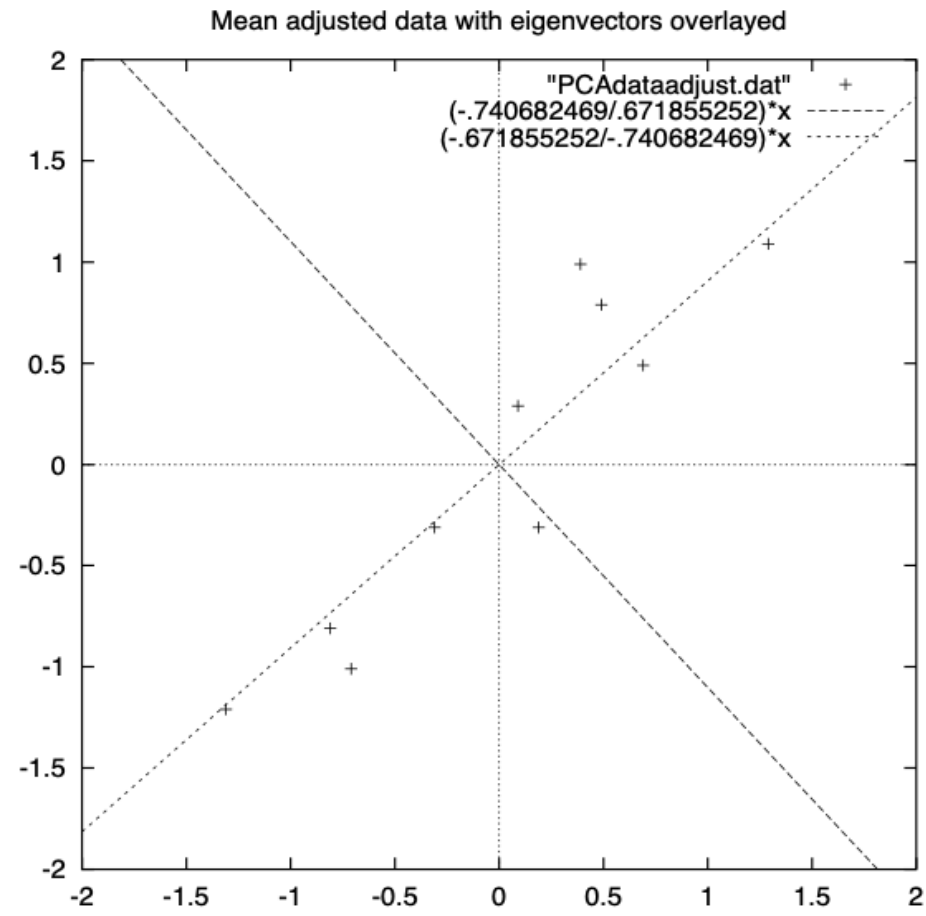
$$\begin{bmatrix} 0.82797 \\ -1.77758 \\ 0.99220 \\ 0.27421 \\ 1.67580 \\ 0.91295 \\ -0.09911 \\ -1.14457 \\ -0.43805 \\ -1.22382 \end{bmatrix}$$

### Method to perform PCA on a data

---

**Getting the old data back : Reconstruction of data( Discussed later)**

## Method to perform PCA on a data



- 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
  - E.g., students' ID is often irrelevant to the task of predicting students' GPA



## Heuristic Search in Attribute Selection

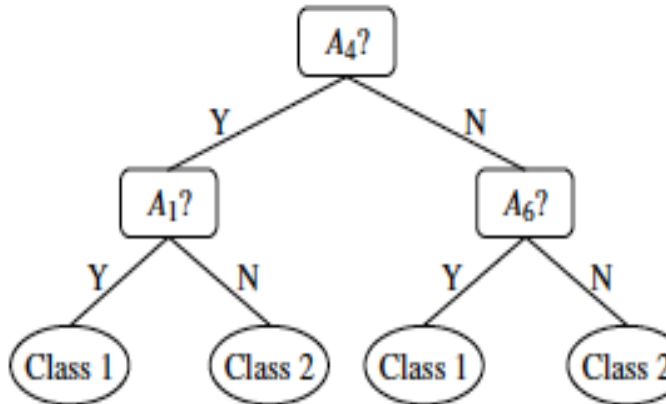
---

- 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 statistical 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
  - Optimal branch and bound:
    - Use attribute elimination and backtracking

## Heuristic Search in Attribute Selection

---

- 1. Stepwise forward selection:** The procedure starts with an empty set of attributes as the reduced set. The best of the original attributes is determined and added to the reduced set. At each subsequent iteration or step, the best of the remaining original attributes is added to the set.
- 2. Stepwise backward elimination:** The procedure starts with the full set of attributes. At each step, it removes the worst attribute remaining in the set.
- 3. Combination of forward selection and backward elimination:** The stepwise forward selection and backward elimination methods can be combined so that, at each step, the procedure selects the best attribute and removes the worst from among the remaining attributes.

Forward selection	Backward elimination	Decision tree induction
<p>Initial attribute set: <math>\{A_1, A_2, A_3, A_4, A_5, A_6\}</math></p> <p>Initial reduced set: <math>\{\}</math>  <math>\Rightarrow \{A_1\}</math>  <math>\Rightarrow \{A_1, A_4\}</math>  <math>\Rightarrow</math> Reduced attribute set:  <math>\{A_1, A_4, A_6\}</math></p>	<p>Initial attribute set: <math>\{A_1, A_2, A_3, A_4, A_5, A_6\}</math></p> <p><math>\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}</math>  <math>\Rightarrow \{A_1, A_4, A_5, A_6\}</math>  <math>\Rightarrow</math> Reduced attribute set:  <math>\{A_1, A_4, A_6\}</math></p>	<p>Initial attribute set: <math>\{A_1, A_2, A_3, A_4, A_5, A_6\}</math></p>  <pre> graph TD     A4["A4?"] -- Y --&gt; A1["A1?"]     A4 -- N --&gt; A6["A6?"]     A1 -- Y --&gt; C1_1((Class 1))     A1 -- N --&gt; C2_1((Class 2))     A6 -- Y --&gt; C1_2((Class 1))     A6 -- N --&gt; C2_2((Class 2))     </pre> <p><math>\Rightarrow</math> Reduced attribute set: <math>\{A_1, A_4, A_6\}</math></p>

**Figure 3.6** Greedy (heuristic) methods for attribute subset selection.

## Heuristic Search in Attribute Selection

---



4. **Decision tree induction**: Decision tree algorithms (e.g., ID3, C4.5, and CART) were originally intended for classification.

Decision tree induction constructs a flowchart like structure where each internal (nonleaf) node denotes a test on an attribute, each branch corresponds to an outcome of the test, and each external (leaf) node denotes a class prediction.

At each node, the algorithm chooses the “best” attribute to partition the data into individual classes.

When decision tree induction is used for attribute subset selection, a tree is constructed from the given data.

All attributes that do not appear in the tree are assumed to be irrelevant. The set of attributes appearing in the tree form the reduced subset of attributes.

## 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 (see: data reduction)
    - E.g., Fourier transformation, wavelet transformation, manifold approaches
  - Attribute construction
    - Combining features
    - Data discretization

- ☐ Mention and explain the different data reduction strategies.
- ☐ Explain how Wavelet transform and Principal Component Analysis are used in the process of data reduction.

### Text Book:

- [Data Mining: Concepts and Techniques](#) by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3rd Edition.
- <https://www.cs.toronto.edu/~mangas/teaching/320/slides/CSC320L11.pdf>
- <https://people.cs.pitt.edu/~milos/courses/cs3750-Fall2007/lectures/PCA.pdf>
- [http://www.cs.otago.ac.nz/cosc453/student\\_tutorials/principal\\_components.pdf](http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf)
- Machine Learning by Anuradha Srinivasaraghavan, Vincy Joseph, ISBN: 9788126578511, Wiley Publications.
- [http://ramanlab.wustl.edu/Lectures/Lecture11\\_PCA.pdf](http://ramanlab.wustl.edu/Lectures/Lecture11_PCA.pdf)



## THANK YOU

---

**Dr.Mamatha H R**

Professor, Department of Computer Science

[mamathahr@pes.edu](mailto:mamathahr@pes.edu)

+91 80 2672 1983 Extn 834