Introduction to Data Mining

Why Data Mining?

- Explosive Growth of Data
  - Data collection and data availability
- Automated data collection tools, Internet, smartphones, …
  - Major sources of abundant data
- Business: Web, e-commerce, transactions, stocks, …
- Science: Remote sensing, biotechnology, scientific simulation, …
- Society and everyone: news, digital cameras, YouTube

- We are drowning in data, but starving for knowledge!

Decision Support

- Typical procedure

– Data -> Knowledge -> Action/Decision -> Goal

- Examples

– Netflix collects user ratings of movies (data) => What types of movies you will like (knowledge) => Recommend new movies to you (action) => Users stay with Netflix (goal)

– Gene sequences of cancer patients (data) => Which genes lead to cancer? (knowledge) => Appropriate treatment (action) => Save life (goal)

– Road traffic (data) => Which road is likely to be congested? (knowledge) => Suggest better routes to drivers (action) => Save time and energy (goal)

What Is Data Mining?

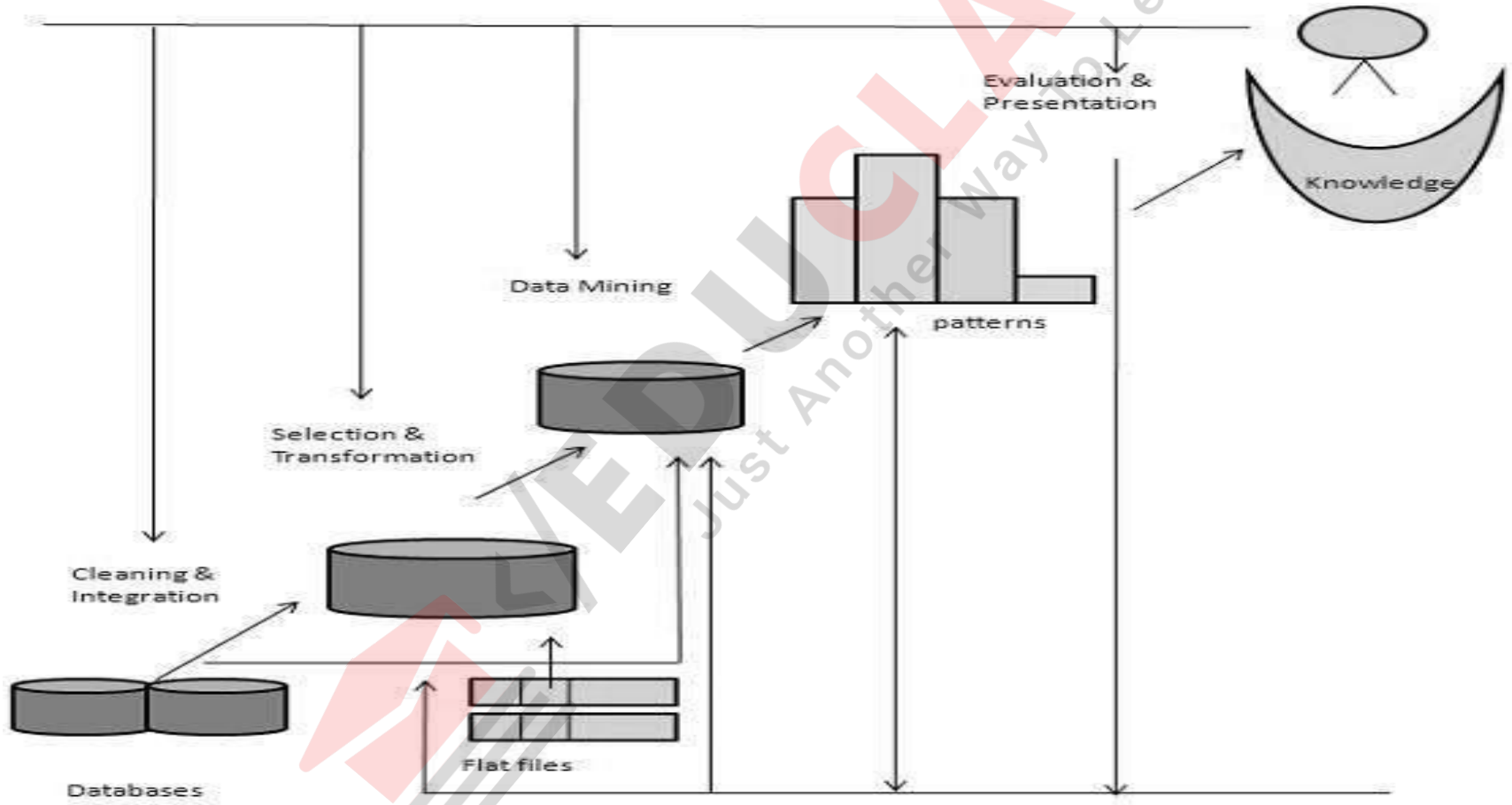
- Data mining

– Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data

- Alternative names

– Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, etc.

KDD Process



Cont…

- **Data Cleaning** − In this step, the noise and inconsistent data is removed.
- **Data Integration** − In this step, multiple data sources are combined.
- **Data Selection** − In this step, data relevant to the analysis task are retrieved from the database.
- **Data Transformation** − In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- **Data Mining** − In this phase, mathematical models are used to determine data patterns.
- Based on the business objectives, suitable modeling techniques should be selected for the prepared dataset.
- Create a scenario to test check the quality and validity of the model.
- Run the model on the prepared dataset.
- Results should be assessed by all stakeholders to make sure that model can meet data mining objectives.

Cont…

- **Pattern Evaluation** − In this phase, patterns identified are evaluated against the business objectives.
- Results generated by the data mining model should be evaluated against the business objectives.
- Gaining business understanding is an iterative process. In fact, while understanding, new business requirements may be raised because of data mining.
- A go or no-go decision is taken to move the model in the deployment phase.
- **Knowledge Presentation** − In the phase, you ship your data mining discoveries to everyday business operations.
- The knowledge or information discovered during data mining process should be made easy to understand for non-technical stakeholders.
- A detailed deployment plan, for shipping, maintenance, and monitoring of data mining discoveries is created.
- A final project report is created with lessons learned and key experiences during the project. This helps to improve the organization's business policy.

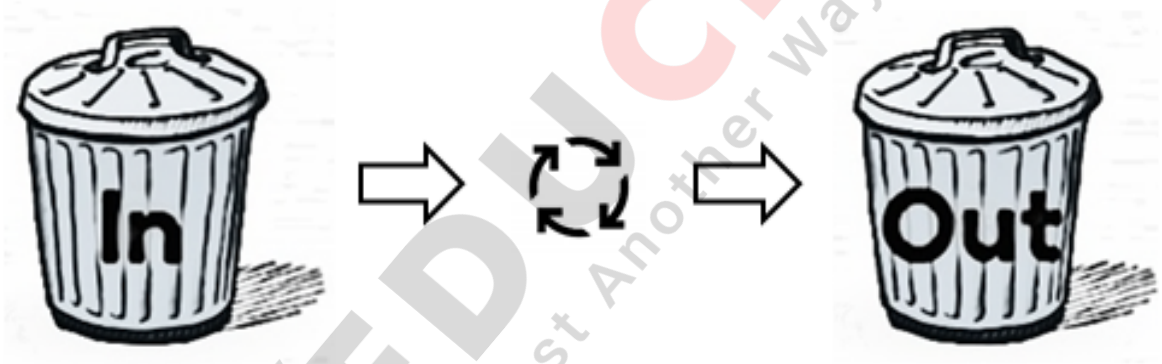
Why to preprocess data?

- Real world data are generally "**dirty**"
  - **Incomplete**: Missing attribute values, lack of certain attributes of interest, or containing only aggregate data.
    - **E.g.** Occupation=" "
  - **Noisy**: Containing errors or outliers.
    - **E.g.** Salary="abcxy"
  - **Inconsistent**: Containing similarity in codes or names.
    - **E.g.** "Gujarat" & "Gujrat" (Common mistakes like **spelling, grammar, articles**)

Why Data processing is important?

"No quality data, No quality results"

- It looks like **Garbage In Garbage Out (GIGO).**



- Quality decisions must be based on **quality data.**
- Duplicate or missing data may cause incorrect or even misleading statistics.
- **Data preparation, cleaning and transformation are the majority task in data mining. (could be as high as 90%).**
- Data preprocessing **prepares** raw data for **further processing.**

Mean

- Mean is the **average** of a dataset.
- To find the mean, calculate the sum of all the data and then divide by the total number of data.
- Example
  - ✔ Find out mean for **12, 15, 11, 11, 7, 13**

First, find the **sum of the data**.

12 + 15 +11 + 11 + 7 + 13 = **69**

Then **divide by the total number of data**.

69 / 6 = **11.5** ⇦ **Mean**

Median

- Median is the middle number in a dataset when the data is arranged in numerical order (Sorted Order).

If count is Odd then middle number is Median

If count is Even then take average of middle two numbers that is Median

Median-Odd

- Example
  - ✓ Find out Median for 12, 15, 11, 11, 7, 13, 15

In above example, count of data is **7**. (Odd)

First, arrange the **data** in **ascending order**.

7, 11, 11, 12, 13, 15, 15

Partitioning data into equal halfs

7, 11, 11, **12**, 13, 15, 15

12 ⇦ Median

Median-Even

- Example
  - ✓ Find out median for 12, 15, 11, 11, 7, 13

In above example, count of data is **6**. (Even)

First, arrange the **data** in **ascending order.**

7, 11, 11, 12, 13, 15

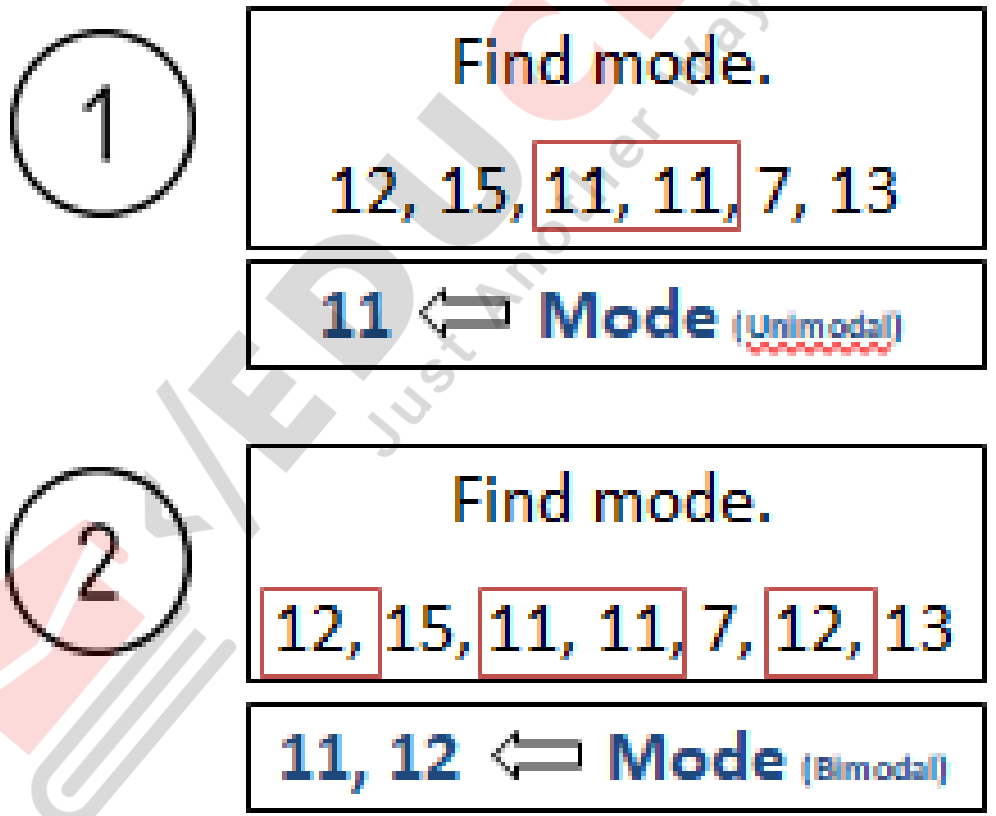
Calculate an **average** of the **two numbers** in the **middle.**

7, 11, 11, 12, 13, 15

(11 + 12)/2 = **11.5** ⇦ **Median**

Mode

- The mode is the number that occurs most often within a set of numbers.
- Example



- Example

3

Find mode.

12, 12, 15, 11, 11, 7, 13, 7

7, 11, 12 ⇦ Mode (Trimodal)

4

Find mode.

12, 15, 11, 10, 7, 14, 13

No Mode

Range

- The range of a set of data is the difference between the largest and the smallest number in the set.
- Example
  - ✓ Find range for given data 40, 30, 43, 48, 26, 50, 55, 40, 34, 42, 47, 50

First, arrange the **data** in **ascending order.**

26, 30, 34, 40, 40, 42, 43, 47, 48, 50, 50, 55

- In our example largest number is 55, and subtract the smallest number is 26.

$55 - 26 = 29 \Leftarrow$ Range

Standard Deviation

- The Standard Deviation is a measure of how spread out any data are.
- Its symbol is σ (the Greek letter sigma).
- *Sample variance* : $(s)^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x - mean)^2$
- Standard Deviation is Square root of sample variance.

Cont..

- The **Variance** is defined as:

> The average of the squared differences from the Mean.

To calculate the variance follow these steps:

1. Calculate the mean, **x**.
2. Write a table that subtracts the mean from each observed value.
3. Square each of the differences, add this column.
4. Divide by **n -1** where **n** is the number of items in the **sample**, this is the **variance** (In actual case take n).
5. To get the **standard deviation** we take the **square root** of the variance.

Cont…

- The owner of the Indian restaurant is interested in how much people spend at the restaurant.
- He examines **10** randomly selected receipts for parties and writes down the following data.

44, 50, 38, 96, 42, 47, 40, 39, 46, 50

1. Find out Mean (1st step)
   - ✓ Mean is **49.2**
2. Write a table that subtracts the mean from each observed value. (2nd step)

Cont…

Step : 3

X	X – Mean	(X – Mean)^2
44	-5.2	27.04
50	0.8	0.64
38	11.2	125.44
96	46.8	2190.24
42	-7.2	51.84
47	-2.2	4.84
40	-9.2	84.64
39	-10.2	104.04
46	-3.2	10.24
50	0.8	0.64
	Total	2600.4

Step : 4

$$= \frac{2600.4}{10 - 1}$$

$$S^2 = 288.7 \sim 289$$

Step : 5

$$S = \sqrt{289}$$

$$S = 17$$

Cont…

- Standard deviation can be thought of measuring how far the data values lie from the mean, we take the mean and move on standard deviation in either direction.
- The mean for this example is 49.2 and the standard deviation is 17.
- Now, 49.2 - 17 = 32.2 and 49.2 + 17 = 66.2
- This means that most of the data probably spend between 32.2 and 66.2.
- If all data are same then variance & standard deviation is 0 (zero).

Attribute Types

- An attribute is a property of the object.
- It also represents different features of the object.
  - E.g. Person → Name, Age, Qualification etc
- Attribute types can be divided into four categories.
  1. Nominal
  2. Ordinal
  3. Interval
  4. Ratio

1. Nominal Attribute

- Nominal attributes are named attributes which can be separated into discrete (individual) categories which do not overlap.
- Nominal attributes values also called as distinct values.
- Example

What is your gender?
Male
Female
Other

What is your hair color?
Black
Brown
Gray
Blonde
Other

2. Ordinal Attribute

- Ordinal attribute is the order of the values, that's important and significant, but the differences between each one is not really known.
- Example
  - Rankings → 1st, 2nd, 3rd
  - Ratings → ★ ★ ★ , ★ ★ ★ ★ ★
- We know that a 5 star is better than a 2 star or 3 star, but we don't know and cannot quantify–how much better it is?

3. Interval Attribute

- Interval attribute comes in the form of a numerical value where the difference between points is meaningful.
- Example
  - Temperature → 10°-20°, 30°-50°, 35°-45°
  - Calendar Dates → $15^{th} - 22^{nd}$, $10^{th} - 30^{th}$
- We can not find true zero (absolute) value with interval attributes.

4. Ratio Attribute

- Ratio attribute is looks like interval attribute, but it must have a true zero (absolute) value.
- It tells us about the order and the exact value between units or data.
- Example
  - **Age Group** → 10-20, 30-50, 35-45 (In years)
  - **Mass** → 20-30 kg, 10-15 kg
- It does have a true zero (absolute) so, it is possible to compute ratios.

5. Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes
- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variables.

Data Cleaning

- Importance
  - “Data cleaning is one of the three biggest problems in data warehousing”—Ralph Kimball
  - “Data cleaning is the number one problem in data warehousing”—DCI survey
- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
- Missing data may need to be inferred.

Customer Data

Name	Age	Sex	Income	Class
Mike	40	Male	150k	Big spender
Jenny	20	Female	?	Regular
…				

How to Handle Missing Data?

- **Ignore the tuple**: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably.
- **Fill in the missing value manually**: tedious + infeasible?
- **Fill in it automatically** with
  - a global constant : e.g., "unknown", a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree (e.g., predict my age based on the info at my web site?)

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention

How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Simple Discretization Methods: Binning

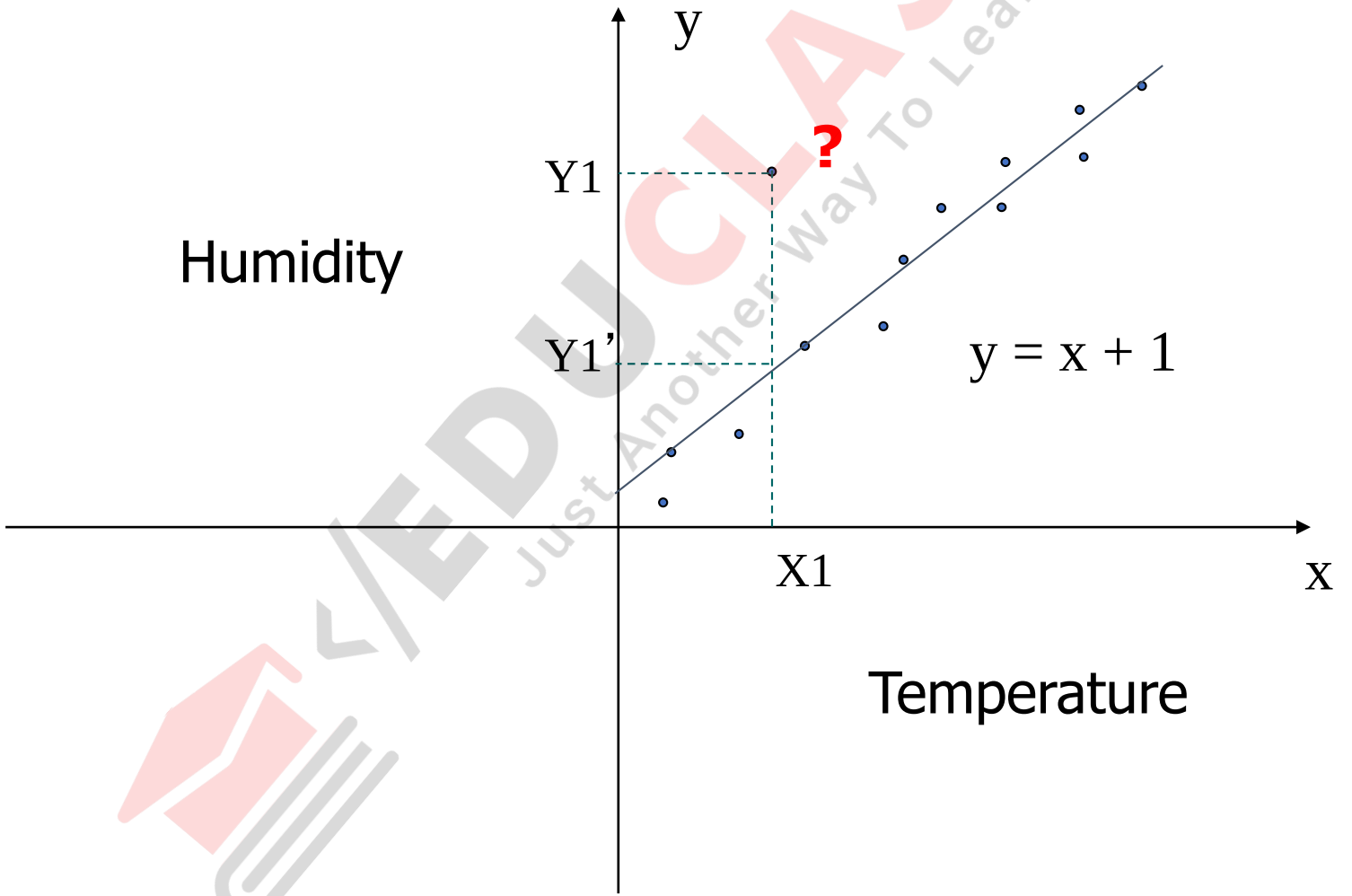
- Equal-width (distance) partitioning
  - Divides the range into *N* intervals of equal size: uniform grid
  - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well
- Equal-depth (frequency) partitioning
  - Divides the range into *N* intervals, each containing approximately same number of samples
  - Good data scaling

Binning Methods for Data Smoothing

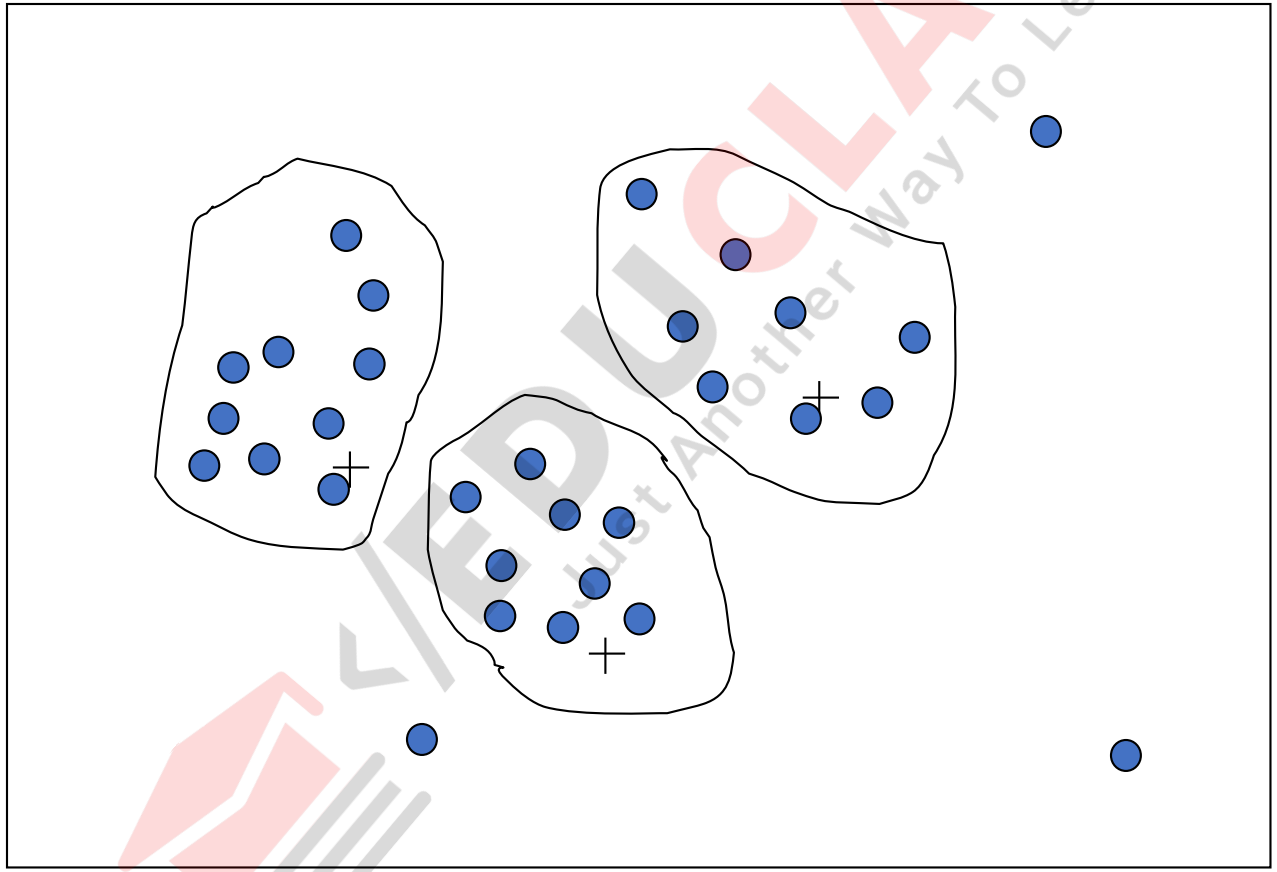
- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Regression



Cluster Analysis



Correcting Inconsistent Data

- Inconsistent: containing discrepancies in codes or names
- Examples: – the data codes for pay_type in one database may be “H” and “S”, and 1 and 2 in another.
- – a weight attribute may be stored in metric units in one Data Cleaning system and British imperial units in another.
- – For a hotel chain, the price of rooms in different cities may involve not only different currencies but also different services (such as free breakfast) and taxes.

Data Integration

- Data integration:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units (e.g., GPA in US and China)

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by *correlation analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Correlation Analysis (Numerical Data)

- Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum (A - \bar{A})(B - \bar{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum (AB) - n\bar{A}\bar{B}}{(n-1)\sigma_A \sigma_B}$$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and Σ(AB) is the sum of the AB cross-product.

- If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.
- $r_{A,B} = 0$: independent; $r_{A,B} < 0$: negatively correlated

Correlation Analysis (Categorical Data)

- X^2 (chi-square) test (Example: Grade and Sex)

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The larger the X^2 value, the more likely the variables are related
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count
- 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

Chi-Square Calculation: An Example

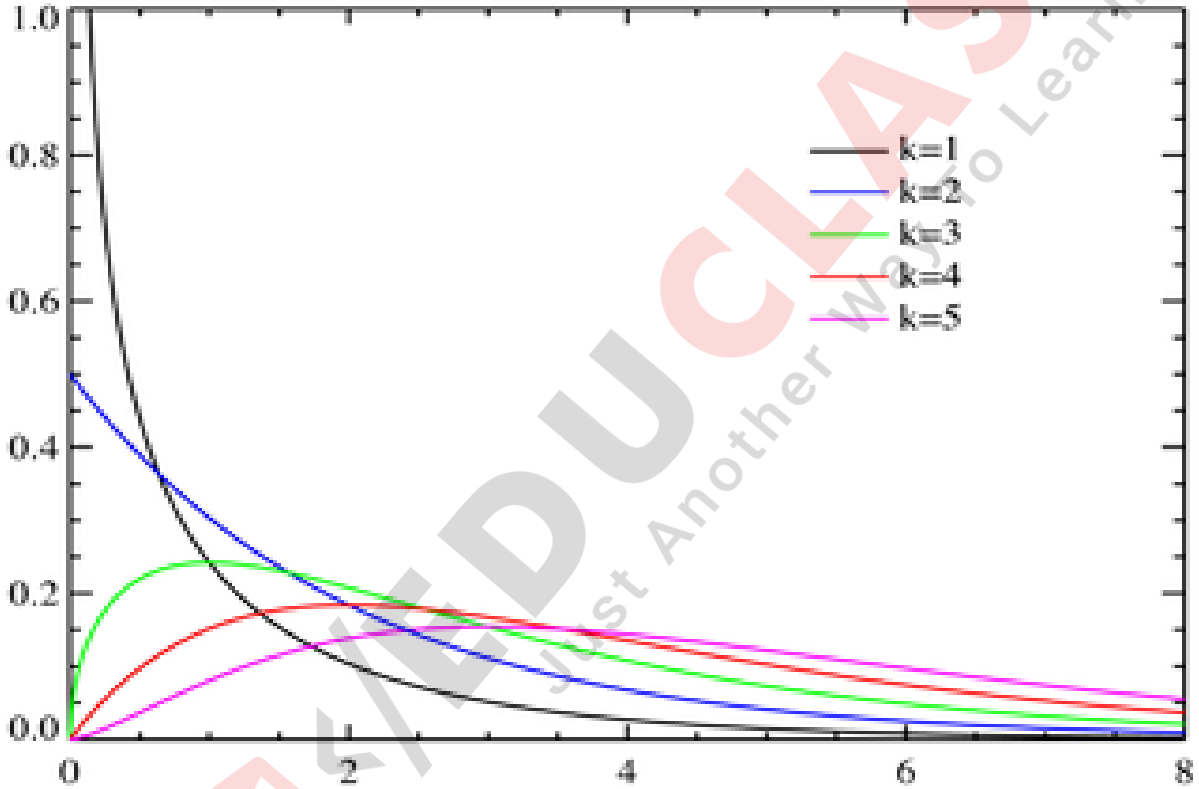
	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

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

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

- It shows that like_science_fiction and play_chess are correlated in the group

Chi Square Distribution



Degree of freedom = 1 (e.g., $(r-1)(c-1)$)

For 0.001 significance, threshold = 10.828

http://en.wikipedia.org/wiki/Pearson's_chi-squared_test

Correlation (viewed as linear relationship)

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, A and B, and then take their dot product

$$a'_k = (a_k - mean(A)) / std(A)$$

$$b'_k = (b_k - mean(B)) / std(B)$$

$$correlation(A, B) = A' \bullet B'$$

Covariance (Numeric Data)

- Covariance is similar to correlation

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

Correlation coefficient: $$r_{A,B} = \frac{Cov(A, B)}{\sigma_A \sigma_B}$$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective mean or **expected values** of A and B, σ_A and σ_B are the respective standard deviation of A and B.

- **Positive covariance**: If $Cov_{A,B} > 0$, then A and B both tend to be larger than their expected values.
- **Negative covariance**: If $Cov_{A,B} < 0$ then if A is larger than its expected value, B is likely to be smaller than its expected value.
- **Independence**: $Cov_{A,B} = 0$ but the converse is not true:
  - Some pairs of random variables may have a covariance of 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

Co-Variance: An Example

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

- It can be simplified in computation as

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$$

- Suppose two stocks A and B 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?
  - E(A) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4
  - E(B) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6
  - Cov(A,B) = (2×5+3×8+5×10+4×11+6×14)/5 − 4 × 9.6 = 4
- Thus, A and B rise together since Cov(A, B) > 0.

Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization
- Discretization
- Attribute Construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Attribute/feature construction
  - New attributes constructed from the given ones

Cont..

- **1. Smoothing:**
  It is a process that is used to remove noise from the dataset using some algorithms.
- It allows for highlighting important features present in the dataset.
- It helps in predicting the patterns.
- When collecting data, it can be manipulated to eliminate or reduce any variance or any other noise form.
- The concept behind data smoothing is that it will be able to identify simple changes to help predict different trends and patterns.
- This serves as a help to analysts or traders who need to look at a lot of data which can often be difficult to digest for finding patterns that they wouldn't see otherwise.

- **2. Aggregation:**
  Data collection or aggregation is the method of storing and presenting data in a summary format.
- The data may be obtained from multiple data sources to integrate these data sources into a data analysis description.
- This is a crucial step since the accuracy of data analysis insights is highly dependent on the quantity and quality of the data used.
- Gathering accurate data of high quality and a large enough quantity is necessary to produce relevant results.
- The collection of data is useful for everything from decisions concerning financing or business strategy of the product, pricing, operations, and marketing strategies.

  For **example**, Sales, data may be aggregated to compute monthly& annual total amounts.

- **3. Discretization:**
  It is a process of transforming continuous data into set of small intervals. Most Data Mining activities in the real world require continuous attributes. Yet many of the existing data mining frameworks are unable to handle these attributes.
- Also, even if a data mining task can manage a continuous attribute, it can significantly improve its efficiency by replacing a constant quality attribute with its discrete values.
- For **example**, (1-10, 11-20) (age:- young, middle age, senior).

- **4. Attribute Construction:**
  Where new attributes are created & applied to assist the mining process from the given set of attributes. This simplifies the original data & makes the mining more efficient.
- **5. Generalization:**
  It converts low-level data attributes to high-level data attributes using concept hierarchy.
- For Example Age initially in Numerical form (22, 25) is converted into categorical value (young, old).

For **example**, Categorical attributes, such as house addresses, may be generalized to higher-level definitions, such as town or country.

- **6. Normalization:** Data normalization involves converting all data variable into a given range.
  Techniques that are used for normalization are:
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- **Min-Max Normalization:**This transforms the original data linearly.
- Suppose that: min_A is the minima and max_A is the maxima of an attribute, P
- We Have the Formula:

Data Transformation: Normalization

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

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

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

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

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

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

- Normalization by decimal scaling

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

Data Reduction Strategies

- 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

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

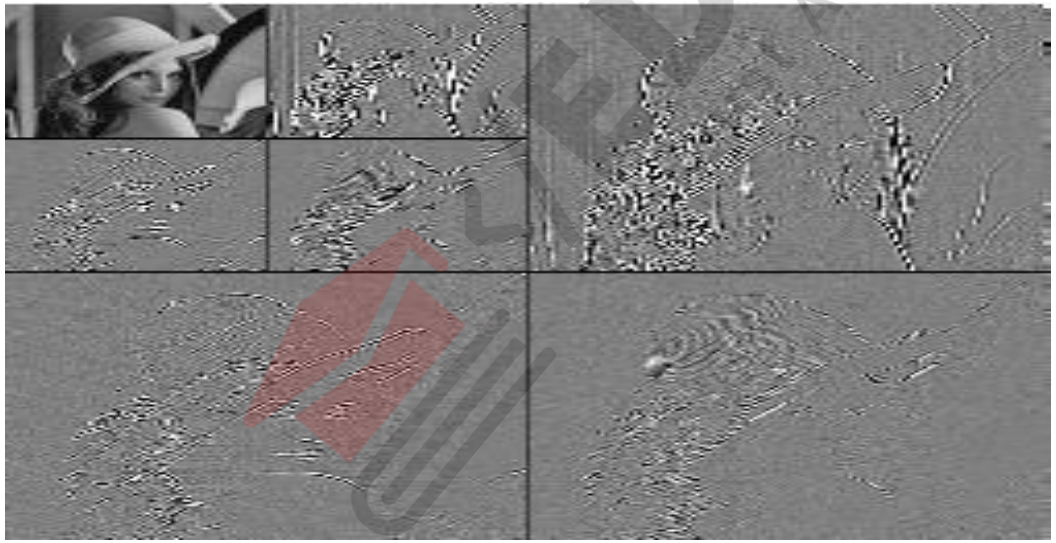
- Avoid the curse of dimensionality
- Help eliminate irrelevant features and reduce noise
- Reduce time and space required in data mining
- Allow easier visualization

Dimensionality reduction techniques

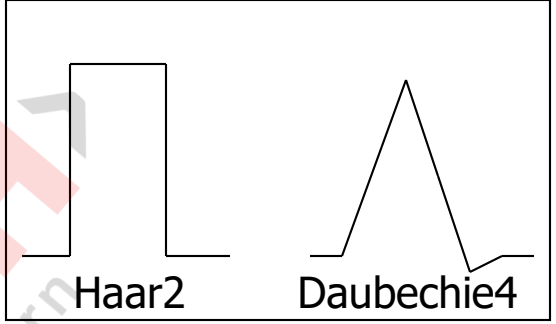
- Wavelet transforms
- Principal Component Analysis
- Supervised and nonlinear techniques (e.g., feature selection)

What Is Wavelet Transform?

- Decomposes a signal into different frequency subbands
  - Applicable to ndimensional 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



Dimensionality Reduction: Wavelet Transformation



- Discrete wavelet transform (DWT): linear signal processing, multi-resolutional analysis
- 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: sum, 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
- Compressed approximation: store only a small fraction of the strongest of the wavelet coefficients

An example

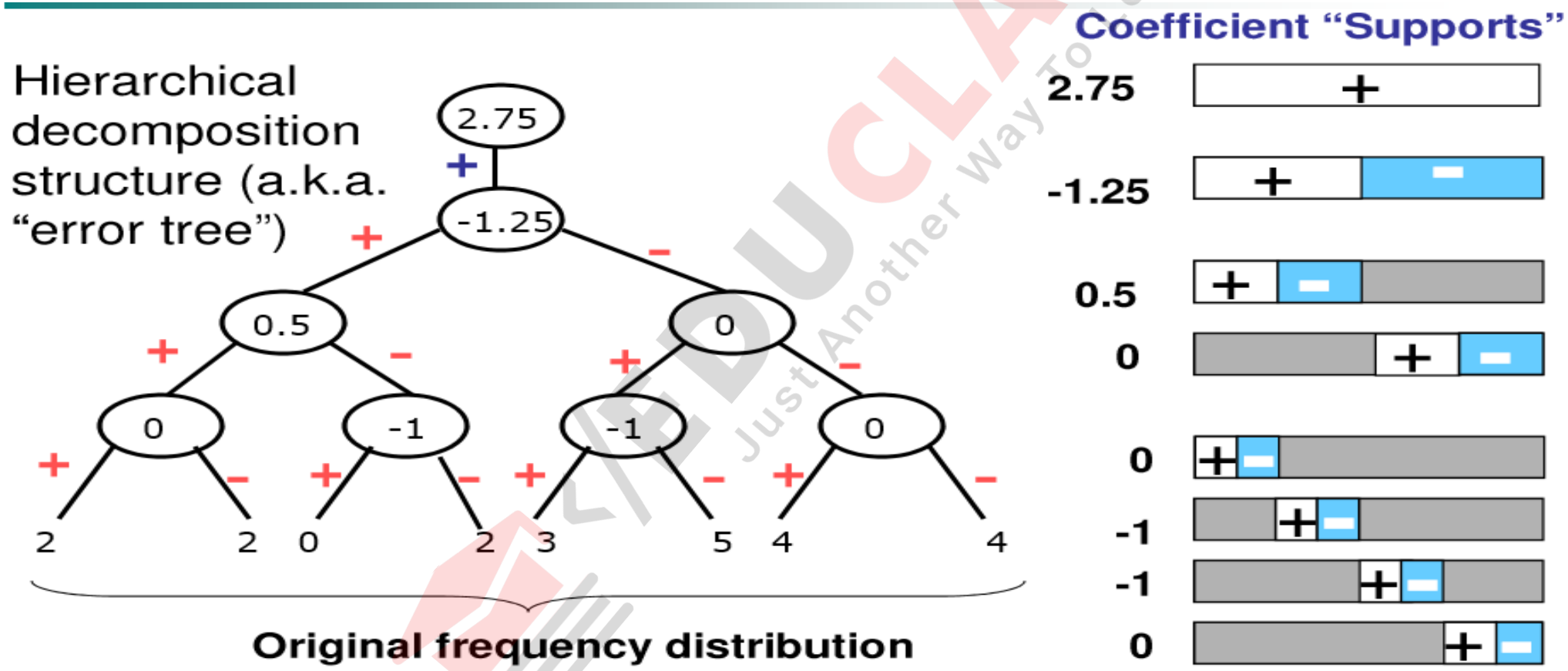
0	1	2	3	4	3	2	1
1	1	-1	-1				
0	0	2	-2	4	-4	6	10
0	**0**	**-4**	**0**	**-8**	**0**	**4**	**16**

Wavelet Decomposition

- 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	$[2\frac{3}{4}]$	$[-1\frac{1}{4}]$

Haar Wavelet Coefficients



Why Wavelet Transform?

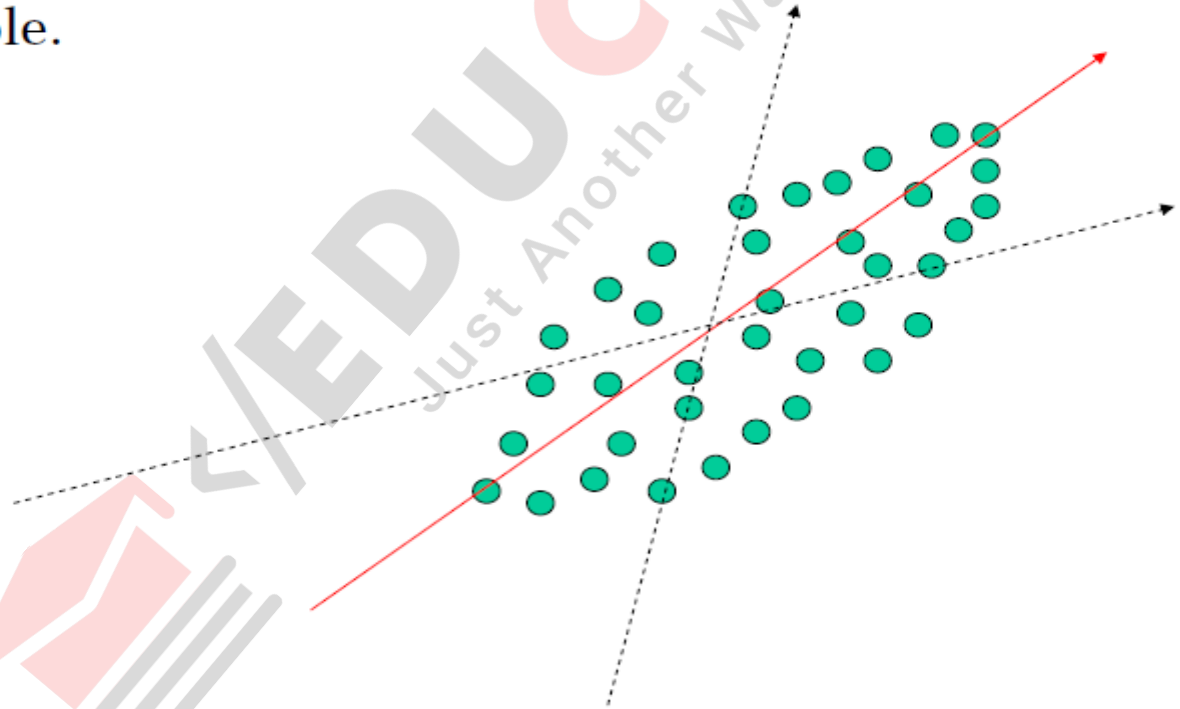
- Use hat-shape filters
  - Emphasize region where points cluster
  - Suppress weaker information in their boundaries
- Effective removal of outliers
  - Insensitive to noise, insensitive to input order
- Multi-resolution
  - Detect arbitrary shaped clusters at different scales
- Efficient
  - Complexity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

Dimensionality Reduction: Principal Component Analysis (PCA)

- Given *N* data vectors from *d*-dimensions, find $k \leq d$ orthogonal vectors (*principal components*) that can be best used to represent data
- Steps
  - Normalize input data
  - Compute *k* orthonormal (unit) vectors, i.e., *principal components*
  - 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)
- Used when the number of dimensions is large

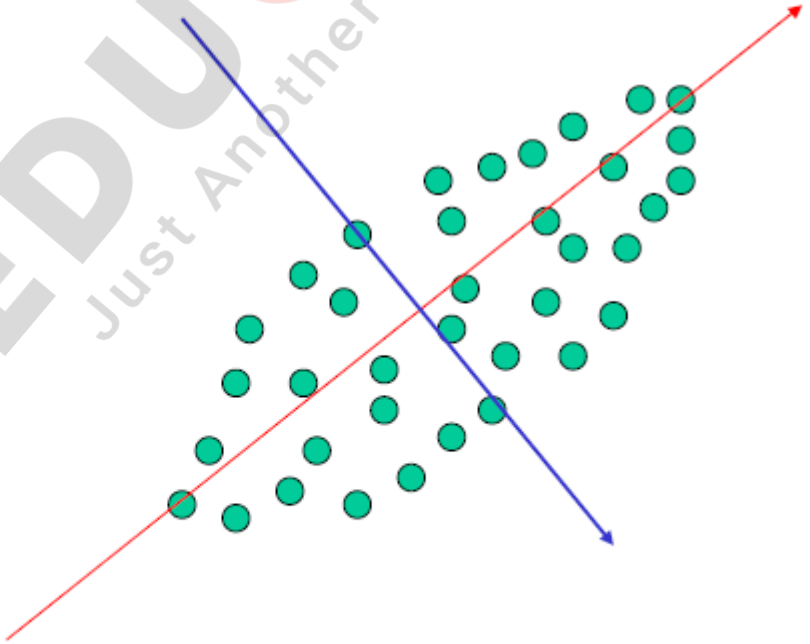
Basic Idea of PCA

Goal: Map data points into a few dimension while trying to preserve the variance of data as much as possible.



Basic Idea of PCA

Goal: Map data points into a few dimension while trying to preserve the variance of data as much as possible.



Numerosity Reduction

- Numerosity reduction techniques replace the original data volume by alternative, smaller forms of data representation.

These techniques are of two types:

- For parametric methods, a model is used to estimate the data, so that typically only the data parameters need to be stored, instead of the actual data.
  - Regression
  - log-linear models
- Non-parametric methods for storing reduced representations of the data include
  - histograms
  - clustering
  - sampling
  - data cube aggregation

Regression

- It's a data smoothing technique that conforms data values to a function.
- Linear regression involves finding the "best" line to fit two attributes (or variables) so that one attribute can be used to predict the other.
- Multiple linear regression is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.

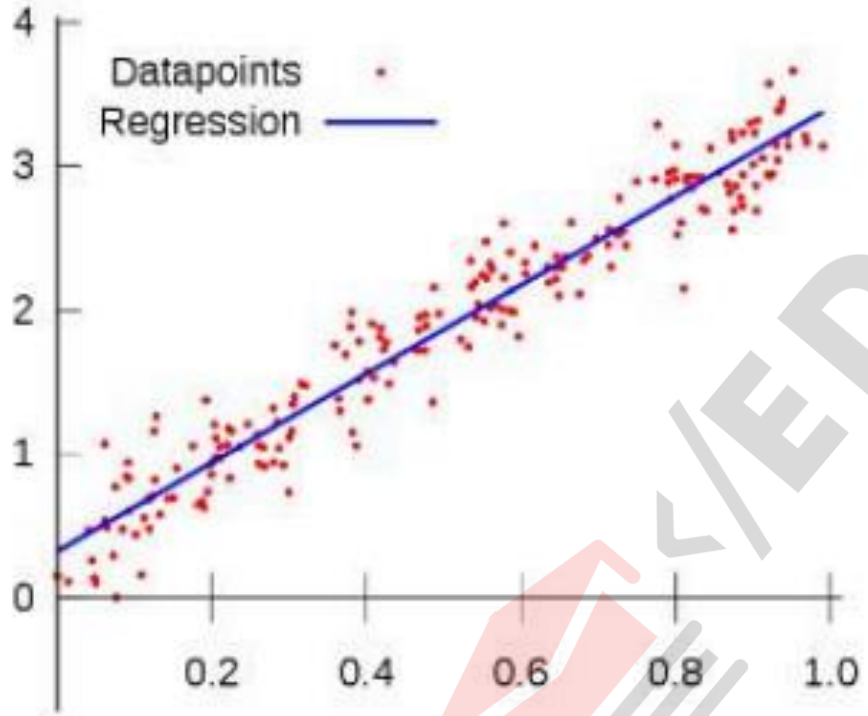
Cont..

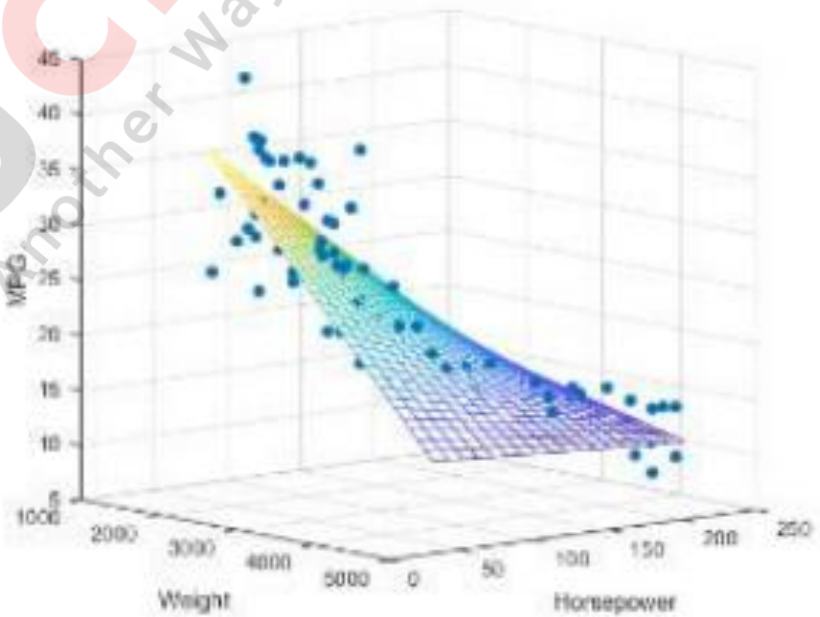
- A random variable Y (response variable), can be modeled as a linear function of another random variable, X (predictor variable), with the equation

$$Y = \alpha + \beta X$$

- Y is assumed to be constant
- α and β (regression coefficients) - Y-intercept and the slope line.
  - Can be solved by the method of least squares. (minimizes the error between actual line separating data and the estimate of the line)

Cont..





Multiple Regression

- Extension of linear regression
- Involve more than one predictor variable
- Response variable Y can be modeled as a linear function of a multidimensional feature vector.
- Eg: multiple regression model based on 2 predictor variables X_1 and X_2

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2$$

Histograms

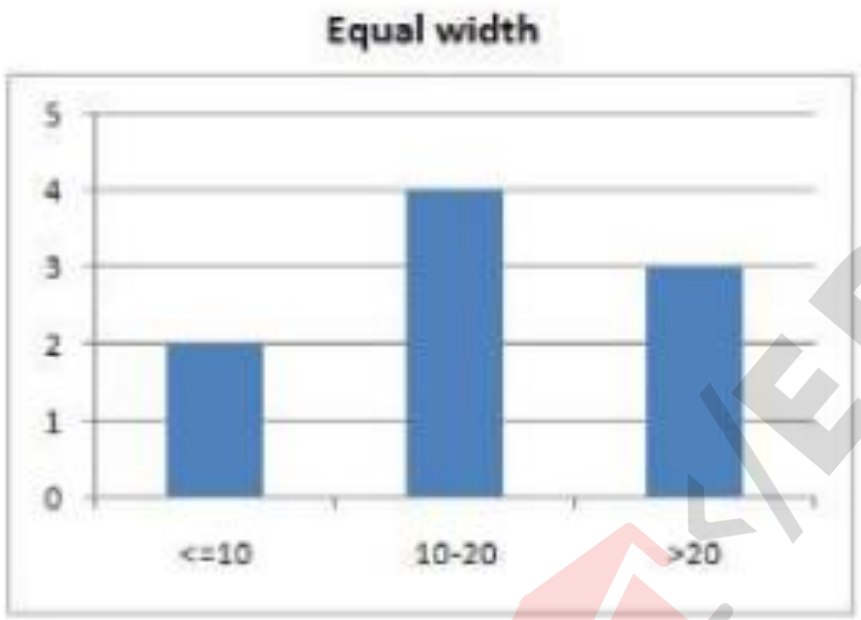
- Divide data into buckets and store average (sum) for each bucket
- Use binning to approximate data distributions
- Bucket - horizontal axis
- Height (area) of bucket - the average frequency of the values represented by the bucket
- Bucket for single attribute-value/frequency pair - singleton buckets
- Continuous ranges for the given attribute

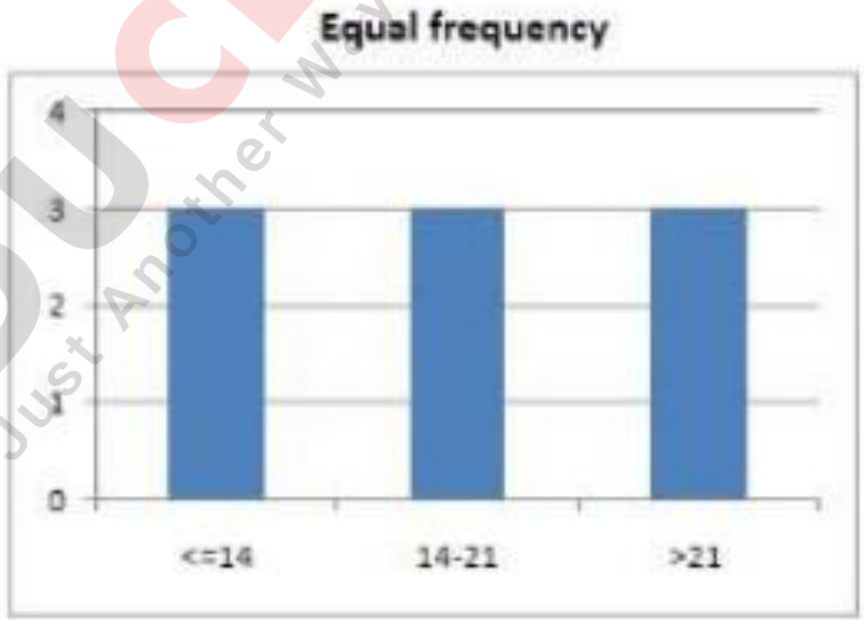
Cont..

There are several partitioning rules, including the following:

- **Equal-width:**
  In an equal-width histogram, the width of each bucket range is uniform
- **Equal-frequency (or equal-depth):**
  In an equal-frequency histogram, the buckets are created so that, roughly, the frequency of each bucket is constant (i.e., each bucket contains roughly the same number of contiguous data samples).

Cont..



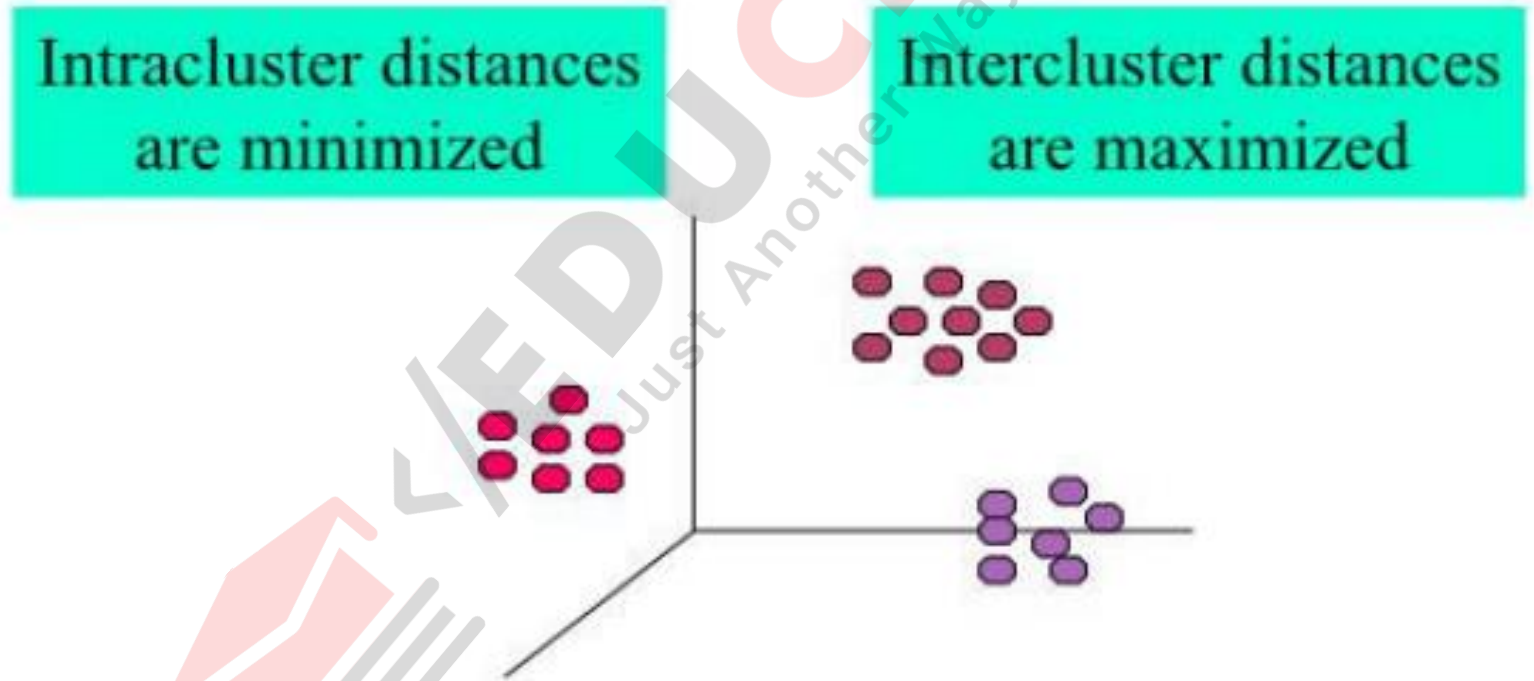


Cluster Analysis

- Cluster: A collection of data objects
  - similar (or related) to one another within the same group
  - dissimilar (or unrelated) to the objects in other groups
- Cluster analysis (or *clustering*, *data segmentation*)
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters
- Unsupervised learning: no predefined classes (i.e., *learning by observations*)

Illustrating Clusters

- Euclidean Distance Based Clustering in 3-D space.



Difference between Dimensionality and Numerosity Reduction

Dimensionality Reduction	Numerosity Reduction
In dimensionality reduction, data encoding or data transformations are applied to obtain a reduced or compressed for of original data.	In Numerosity reduction, data volume is reduced by choosing suitable alternating forms of data representation.
It can be used to remove irrelevant or redundant attributes.	It is merely a representation technique of original data into smaller form.
In this method, some data can be lost which is irrelevant.	In this method, there is no loss of data.
1.Methods for dimensionality reduction are:Wavelet transformations. 2.Principal Component Analysis.	1.Methods for Numerosity reduction are:Regression or log-linear model (parametric). 2.Histograms, clusturing, sampling (non-parametric).
The components of dimensionality reduction are feature selection and feature extraction.	It has no components but methods that ensure reduction of data volume.
It leads to less misleading data and more model accuracy.	It preserves the integrity of data and the data volume is also reduced.

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:

Note: Sampling may not reduce database I/Os (page at a time)

Types of Sampling

- Simple random sampling

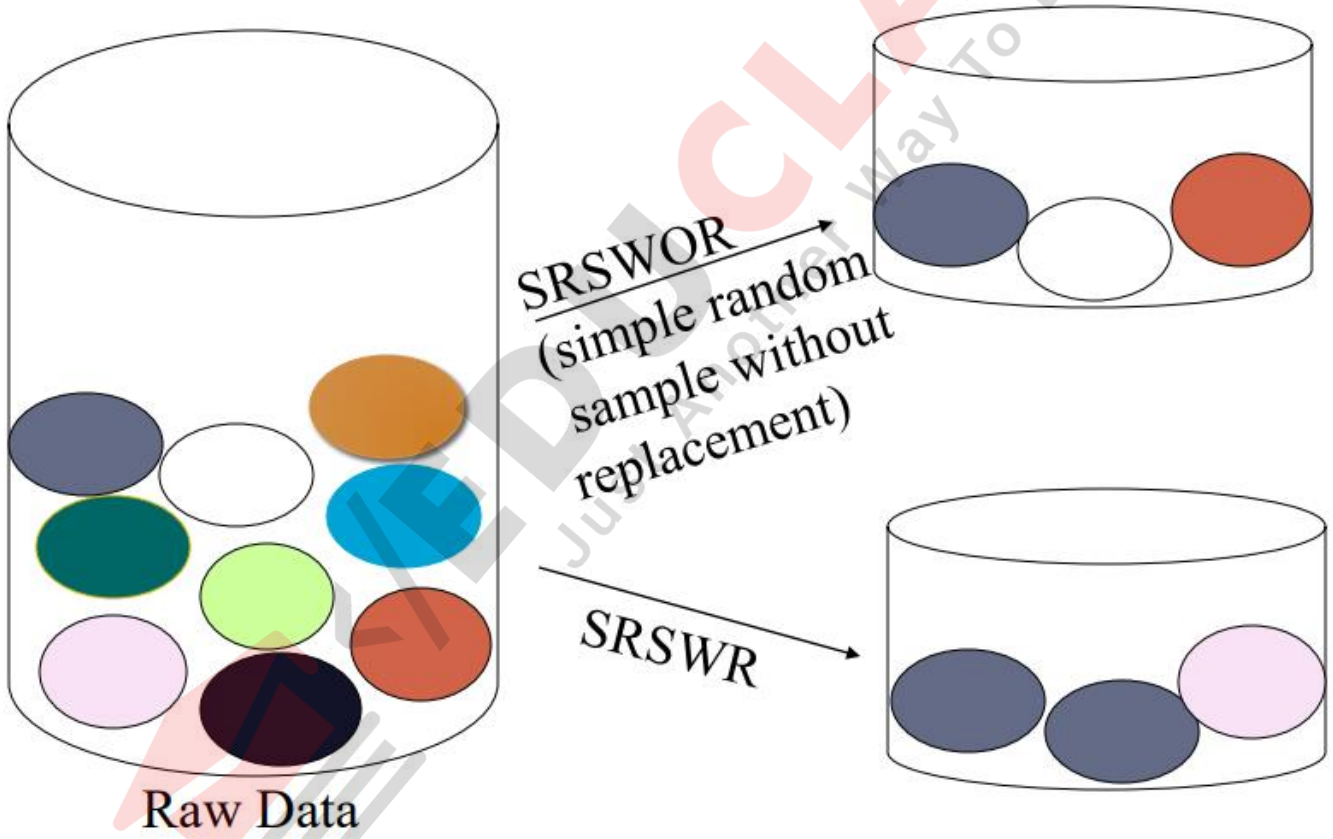
  There is an equal probability of selecting any particular item
- Sampling without replacement

  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:

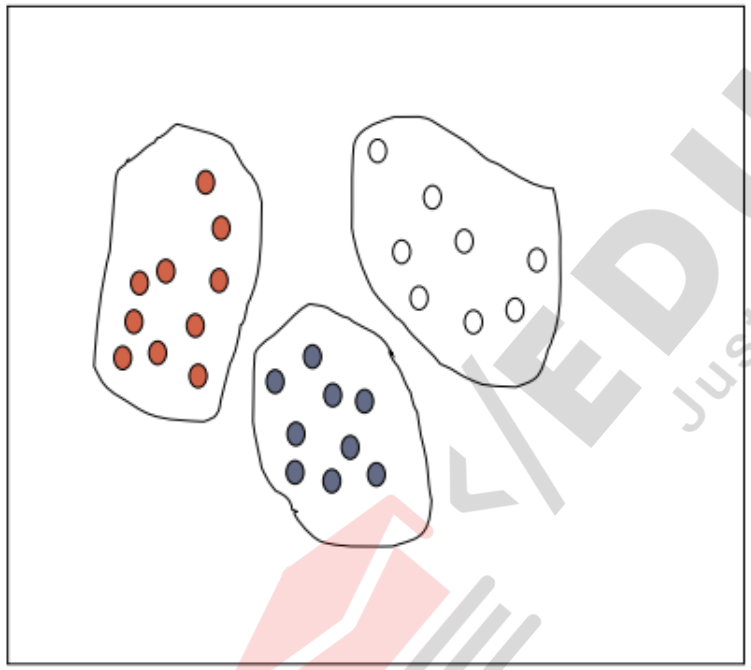
  Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data) • Used in conjunction with skewed data

Sampling: With or without Replacement

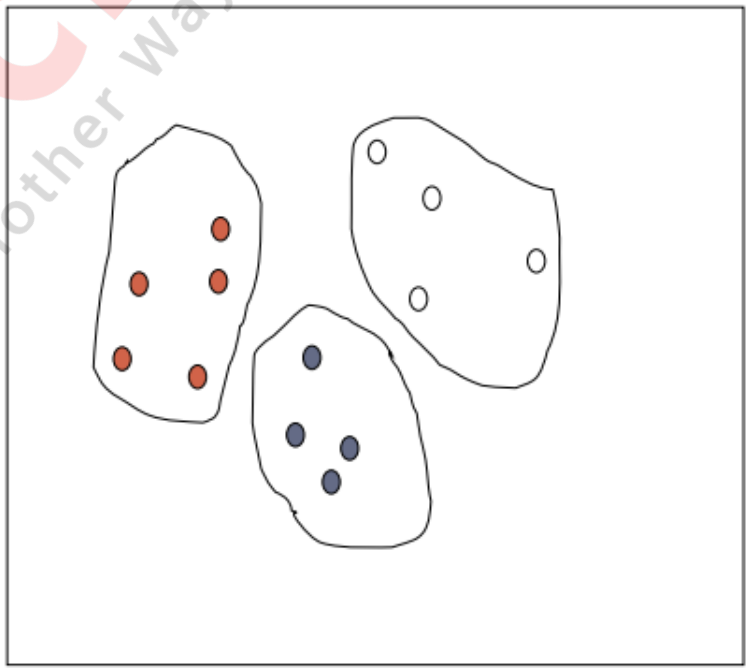


Sampling: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample



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
- Time sequence
  - Typically short and vary slowly with time
- Dimensionality and numerosity reduction may also be considered as forms of data compression