

UNIT-II

Data Preprocessing

Major tasks in data pre-processing, Data cleaning: Missing values, noisy Data; Data reduction: Overview of data reduction strategies, principal components analysis, attribute subset selection, histograms, sampling; Data transformation: Data transformation strategies overview, data transformation by normalization.

Learning Material

Why Data Preprocessing?

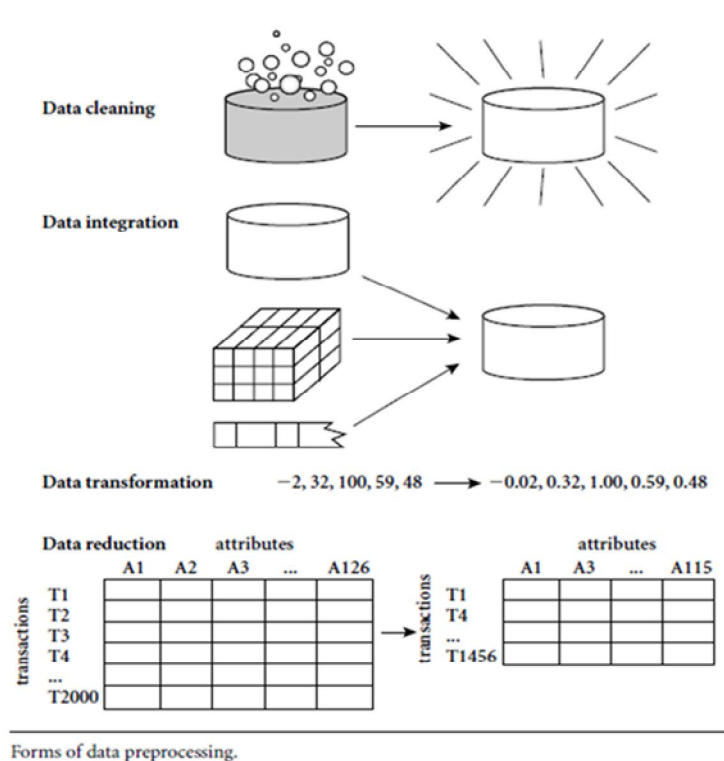
- ❖ Data in the real world is huge size, may contain
 - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
 - Noisy: containing errors or outliers
 - Inconsistent: containing discrepancies in codes or names
- ❖ No quality data, no quality mining results!
 - Quality decisions must be based on quality data
 - Data warehouse needs consistent integration of quality data
- ❖ Real world data tend to dirty, incomplete and inconsistent. Data preprocessing techniques can improve the quality of the data there by helping to improve the accuracy and efficiency of the subsequent mining process.

- ❖ Detecting anomalies, rectifying them early and reducing the data to be analyzed can lead to huge profit for decision making. So need preprocessing of data.

2.1 Major tasks in Data preprocessing

- **Data cleaning** : to remove noise and inconsistencies in the data.
- **Data Integration** : merge data from multiple sources.
- **Data transformation** : normalization may be applied to improve the accuracy and efficiency of the algorithms
- **Data reduction**: can reduce the data size by aggregating, eliminating redundant features or clustering.

Forms of data preprocessing



2.2 Data Cleaning

Real world data tend to be incomplete noisy and inconsistent. Data cleaning routines attempt to

1. Fill in missing values
2. Smooth out noisy data while identifying outliers

2.2.1 Missing values

- 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
 - not register history or changes of the data
- Missing data may need to be inferred.

Handling of Missing Data

1. Ignore the tuple: usually done when class label is missing (assuming the task is classification)—not effective unless the tuple contains several attributes with missing values

2. Fill in the missing value manually: Time consuming and infeasible for a large data set with many missing values.
3. Use a global constant to fill in the missing value: replace all missing attribute values by the same constant e.g., “unknown”, or $-\infty$.
4. Use the attribute mean to fill in the missing value: Average income of AE customers is \$28,000 use this value to replace the missing value for income.
5. Use the attribute mean for all samples belonging to the same class to fill in the missing value: smarter ex: if classify customers according to credit-risk, replace the missing value with a avg income value for customers in the same credit risk category as that of the given tuple.
6. Use the most probable value to fill in the missing value: This may be determined with regression, Bayesian formula or decision tree

2.2.2 Noisy Data

- Noise: random error or variance in a measured variable.
- Smooth out the data to remove the noise
- Incorrect attribute values may due to
 1. Faulty data collection instruments
 2. Data entry problems i.e May have been human or computer errors occurring at data entry
 3. Errors in data transmission
 4. Technology limitation ex: Limited buffer size
 5. inconsistency in naming convention

- Other data problems which requires data cleaning
 1. duplicate records
 2. incomplete data
 3. inconsistent data

Handling Noisy Data

A) Binning method: first sort data and partition into (equi-depth) bins then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

B) Clustering : detect and remove outliers

C) Combined computer and human inspection : detect suspicious values and check by human

D) Regression : smooth by fitting the data into regression functions

A) Simple Discretization Methods: Binning

Binning methods smooth a sorted data value by consulting its neighborhood (value around it). The sorted values are distributed in to a number of buckets or bins.

Binning Methods for Data Smoothing

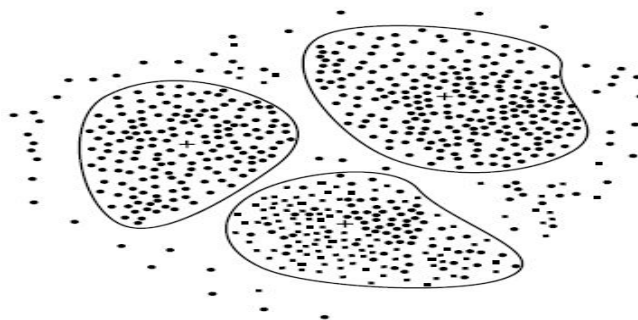
Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34
Smoothing by bin means:
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29
Smoothing by bin boundaries:
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Smoothing by bin medians – In which each bin values is replaced by closest boundary values

B) Cluster Analysis

- Outliers may be detected by clustering; where similar values are organized in to groups or clusters.
- Values that fall outside of the set of clusters may be considered outliers



A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters. Each cluster centroid is marked with a "+", representing the average point in space for that cluster. Outliers may be detected as values that fall outside of the sets of clusters.

C) Combined computer and human inspection

- Outliers may be identified through a combination of computer and human inspection

Ex : in one application an information-theoretic measure was used to help identify outlier patterns in a handwritten character database for classification.

- The measure's value reflected the "surprise" content of the predicted character label with respect to the known label.
- Patterns whose surprise content is above a threshold are output to a list.
- A human can then sort through the patterns in the list to identify the actual garbage ones.
- This is much faster than having to manually search through the entire database.
- Outlier patterns may be informative (ex: identifying useful data exceptions, such as different versions of the characters "0" or "7" or garbage)
- The garbage values can then be excluded from use in subsequent data mining.

2.3 Data reduction

- Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set

- Data reduction technique can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data.

2.3.1 Overview of Data reduction strategies

- 1) **Dimensionality reduction** : irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed.
- 2) **Data compression** : encoding mechanisms are used to reduce the data set size.
- 3) **Numerosity reduction** : the data are replaced or estimated alternative, smaller data representations such as parametric models or nonparametric methods such as clustering, sampling and the use of histograms.

2.3.1.1 Dimensionality reduction

A) Attribute subset selection

- Data sets for analysis may contain hundreds of attributes, many of which may be irrelevant to the mining task, or redundant can slow down the mining process.
- Dimensionality reduction reduces the data set size by removing such attributes(or dimensions) from it. For this methods of attribute subset selection are applied.
- The goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes.

Heuristic methods of attribute subset selection include the following techniques.

1. step-wise forward selection
2. step-wise backward elimination
3. combining forward selection and backward elimination
4. decision-tree induction

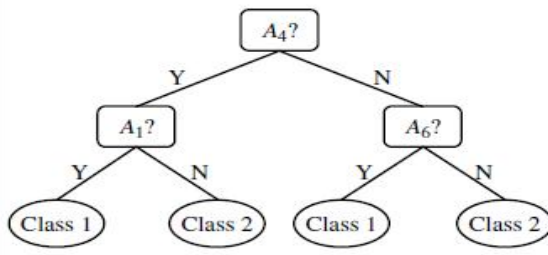
1. Stepwise forward selection: The procedure starts with an empty set of attributes as the reduced set. The best of the original attributes is determined and added to the reduced set. At each subsequent iteration or step, the best of the remaining original attributes is added to the set.

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.

4. Decision tree induction: Decision tree induction constructs a flowchart like structure where each internal (non leaf) 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.

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$

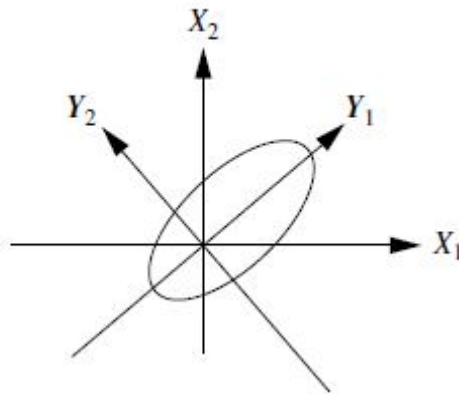
Greedy (heuristic) methods for attribute subset selection.

B) Principal Component Analysis

The principal components analysis is a method of dimensionality reduction.

The basic procedure is as follows:

1. The input data are normalized, so that each attribute falls within the same range. This step helps ensure that attributes with large domains will not dominate attributes with smaller domains.
2. PCA computes k orthonormal vectors that provide a basis for the normalized input data. These are unit vectors that each point in a direction perpendicular to the others. These vectors are referred to as the *principal components*. The input data are a linear combination of the principal components.
3. The principal components are sorted in order of decreasing “significance” or strength. The principal components essentially serve as a new set of axes for the data, providing important information about variance.



That is, the sorted axes are such that the first axis shows the most variance among the data, the second axis shows the next highest variance, and so on. For example, Figure 3.5 shows the first two principal components, Y_1 and Y_2 , for the given set of data originally mapped to the axes X_1 and X_2 . This information helps identify groups or patterns within the data.

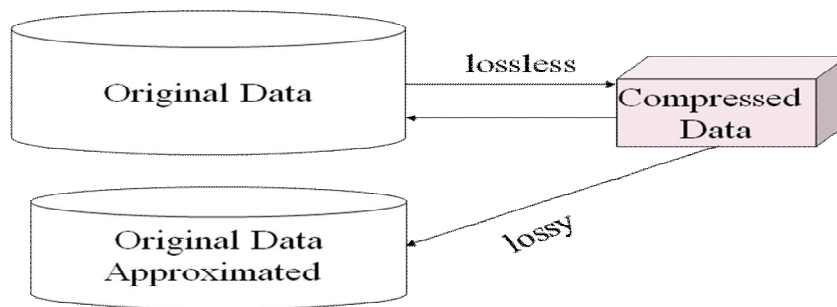
4. Because the components are sorted in decreasing order of “significance,” the data size can be reduced by eliminating the weaker components, that is, those with low variance.

Using the strongest principal components, it should be possible to reconstruct a good approximation of the original data.

PCA can be applied to ordered and unordered attributes, and can handle sparse data and skewed data. Multidimensional data of more than two dimensions can be handled by reducing the problem to two dimensions. Principal components may be used as inputs to multiple regression and cluster analysis. In comparison with wavelet transforms, PCA tends to be better at handling sparse data, whereas wavelet transforms are more suitable for data of high dimensionality.

2.3.1.2 Data Compression

- Data data encoding or transformations are applied so as to obtain a reduced or “compressed” representation of the original data.
- If the original data can be reconstructed from the compressed data without any loss of information, the data reduction is called lossless.
- If, instead, we can reconstruct only an approximation of the original data, then the data reduction is called lossy.



Two popular and effective methods of lossy data compression are wavelet transforms and principal component analysis

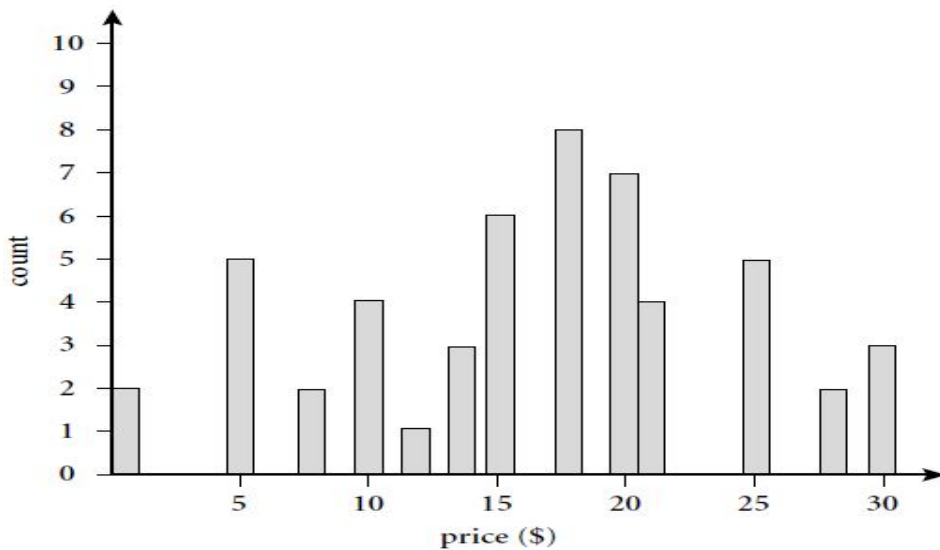
2.3.1.3 Numerosity reduction

- Can reduce the data volume by choosing alternative smaller forms of representations. Techniques may be parametric or non parametric.
- Parametric methods
 - A model is used to estimate the data, so that only the data parameters need to be stored, instead of the actual data (outliers also stored)
 - Regression and log-linear models can be used to approximate the given data. both methods used for data compression, both handle sparse & skewed data.

- Non-parametric methods
 - Do not assume models
 - histograms, clustering, sampling

A) Histograms (a popular data reduction technique)

- Histograms use binning to approximate data distributions. A histogram for an attribute, A, partitions the data distribution of A into disjoint subsets, or buckets.
- Histograms. If each bucket represents only a single attribute-value/frequency pair, the buckets are called *singleton buckets*.
- The following data are a list of prices of commonly sold items at AllElectronics
- (rounded to the nearest dollar). The numbers have been sorted: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.



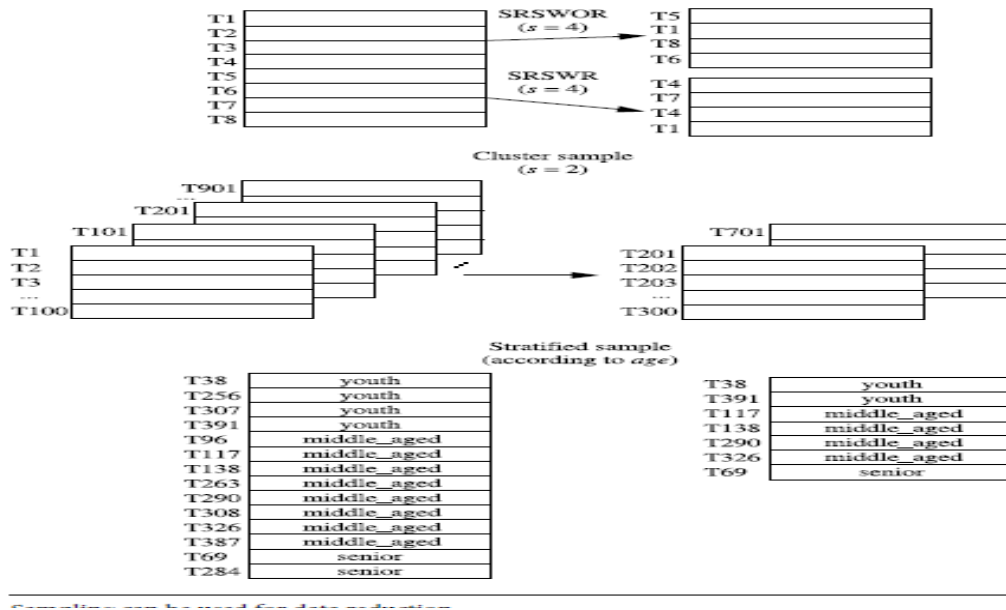
A histogram for *price* using singleton buckets—each bucket represents one price-value/frequency pair.

- Partitioning rules for the bucket:

B) Sampling

- Sampling can be used as a data reduction technique because it allows a large data set to be represented by a much smaller random sample (or subset) of the data.
- Suppose that a large data set, D , contains N tuples. Let's look at the most common ways that we could sample D for data reduction are
- Simple random sample without replacement (SRSWOR) of size s : This is created by drawing s of the N tuples from D ($s < N$), where the probability of drawing any tuple in D is $1/N$, that is, all tuples are equally likely to be sampled.
- Simple random sample with replacement (SRSWR) of size s : This is similar to

- SRSWOR, except that each time a tuple is drawn from D, it is recorded and then replaced. That is, after a tuple is drawn, it is placed back in D so that it may be drawn again.



2.3.3 Data Transformation

Data Transformation strategies overview:

- The data are transformed or consolidated into forms appropriate for mining is called Transformation.

It involves-

- **Smoothing:** remove noise from data, techniques are binning, clustering and regression.
- **Aggregation:** summarization or aggregation operations are applied to the data. Ex: daily sales data may be aggregated. So as to compute monthly and annual total amounts. This is used in constructing data cube.

- **Generalization:** low level or primitive(raw) data are replaced by high-level concepts through the use of concept hierarchy. Ex: categorical attribute street can be generalized to higher-level concepts like city or country.
- **Normalization:** attribute data are scaled. So as to fall within a small, specified range, such as -1.0 to 1.0 or 0.0 to 1.0

2.3.4 Data Transformation by normalization:-

Normalization is useful for classification algorithms involving neural networks, or distance measurements such as nearest neighbor classification & clustering. Methods for data normalization are min-max normalization, z-score normalization and normalization by decimal scaling.

a) Min-max normalization : It performs a linear transformation on the original data. $\min A$ and $\max A$ are the minimum and maximum values of an attribute, A . Min-max normalization maps a value, v , of A to v_0 in the range $[\text{new min } A; \text{new max } A]$ by computing

Min-max normalization preserves the relationships among the original data values.

b) z-score normalization :(or zero-mean normalization), the values for an attribute, A , are normalized based on the mean and standard deviation of A . A value, v , of A is normalized to v_0 by computing

$$v' = \frac{v - \bar{A}}{\sigma_A},$$

where \bar{A} and σ_A are the mean and standard deviation, respectively, of attribute A.

This method of normalization is useful when the actual minimum and maximum of attribute A are unknown or when there are outliers that dominate the min-max normalization

c) Normalization by decimal scaling : Normalizes by moving the decimal point of values of attribute A. The number of decimal points moved depends on the maximum absolute value of A.

A value, v , of A is normalized to v' by computing.

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

where j is the smallest integer such that

$$\text{Max}(|v'|) < 1.$$

➤ **Attribute construction** : New attributes constructed from the given attributes and added to improve accuracy and to understanding of structure of high dimensional data.

Ex : add the attribute area based on the attributes height and width.

UNIT-II

Assignment-Cum-Tutorial Questions

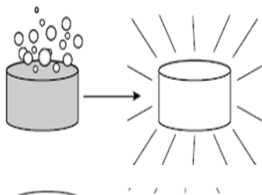
SECTION-A

Objective Questions

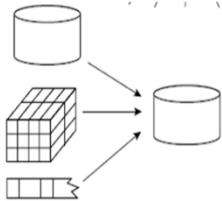
1. Real world data may contains _____ and _____ data.
2. When to apply the data preprocessing techniques for mining the data
 - A) Before mining.
 - B) During mining.
 - C) After mining.
 - D) All of the time.
3. Match the following :

$-2, 32, 100, 59, 48 \rightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

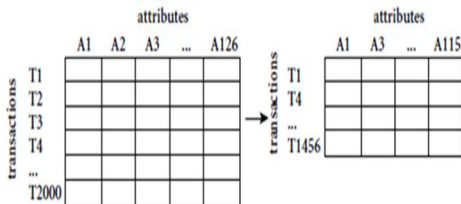
Data reduction



Data integration



Data transformation



Data cleaning

4. Use the attribute mean to fill the missing value of data []
 1,2,3,4,5,6,__,7,8,9,10.
 A) 2.0 B) 3.0 C) 5.5 D) 5.0
5. Data for Attendance : 50,55,60,65,70,75,80,85,90,95
 Partition the above attendance data into equidepth bins of depth 5. []

- A) Bin 1:50,55,60,65,70 Bin 2: 75,80,85,90,95
 B) Bin 1:50,55,60,65 Bin 2: 70,75,80,85,90,95
 C) Bin 1:50,55,60,65,70,75 Bin 2: ,75,80,85,90,95
 D) Bin 1:50,55,60 Bin 2:65,70,75,80,85,90,95
6. For the above attendance apply bin means smoothing technique []
 A) Bin 1: 65,65,65,65,65 Bin2 : 85,85,85,85,85
 B) Bin 1: 60,60,60,60,60 Bin2 : 85,85,85,85,85
 C) Bin 1: 65,65,65,65,65 Bin2 : 80,80,80,80,80
 D) Bin 1: 75,75,75,75,75 Bin2 : 85,85,85,85,85
7. For the above attendance apply bin medians smoothing technique.
 A) Bin 1: 60,60,60,60,60 Bin2 : 85,85,85,85,85 []
 B) Bin 1: 65,65,65,65,65 Bin2 : 85,85,85,85,85
 C) Bin 1: 65,65,65,65,65 Bin2 : 80,80,80,80,80
 D) Bin 1: 75,75,75,75,75 Bin2 : 85,85,85,85,85
8. Data for Attendance : 4,8,15 Smoot by bin boundaries []
 A) 4,4,15 B) 4,15,15 C) 4,4,4 D) 15,15,15
9. Data Reduction is the process of reduced representation of data in size not in values. [T/F]
10. Reducing the number of attributes to solve the high dimensionality problem is called as _____. []
 A) Curse of dimensionality. B) Dimensionality reduction.
 C) Cleaning. D) Over fitting.
11. _____ and _____ are the popular and effective methods of lossy data compression technique.
12. ____is the method of fitting the data values into a fixed model []
 A) Clustering. B) Regression. C. Smoothing. D) Aggregation.
13. Use min-max normalization transformation technique for finding transformed income value of \$10000 with min_income=1000,

max_income=50000 and mapping range of income [0.0,1.0] The Transformed income value=_____.

- A) 0.225 B) 0.325 C) 0.425 D) 0.525

SECTION-B

SUBJECTIVE QUESTIONS

1. Illustrate the need for data preprocessing. List and explain various data preprocessing techniques.
2. What is data cleaning? Describe the approaches to fill missing values.
3. Define noisy data. Describe various techniques for smoothing noisy data.
4. Discuss the issues to be considered for data integration.
5. What is data normalization? Explain any two Normalization methods.
6. Outline about Data Cube Aggregation as a data reduction technique.
7. Elaborate different attribute subset selection methods with examples
8. What is a concept hierarchy? Explain different techniques used to generate concept hierarchy for categorical data.
9. Write short notes on Sampling in Numerosity Reduction.
10. Write short notes on Histograms in Numerosity Reduction.
11. Explain different sampling approaches used in data Reduction

Problems

12. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
13. Question: Use smoothing by bin means to smooth the data, using a bin depth of 3. Illustrate your steps.
14. Apply the min-max normalization to transform the value 35 into the range [0.0, 1.0] using the data for age given in question 2.
15. Apply z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years. Using the data for age given in question 2.

16. Use these methods to *normalize* the following group of data:

200, 300, 400, 600, 1000

- (a) min-max normalization by setting *min* D 0 and *max* D 1
- (b) z-score normalization
- (c) z-score normalization using the mean absolute deviation instead of standard deviation
- (d) normalization by decimal scaling

17. Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result.

<i>age</i>	23	23	27	27	39	41	47	49	50
<i>%fat</i>	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
<i>age</i>	52	54	54	56	57	58	58	60	61
<i>%fat</i>	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

Calculate the mean and standard deviation of *age* and *%fat*.