# UNIT 8 DATA PREPROCESSING

**Structure**

## 8.0 INTRODUCTION

In the earlier unit we had studied the basic concepts of Data Mining. In this unit, we will study fundamental step in the data mining, known as data preprocessing. **Data preprocessing** is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms. Preprocessing of data is mainly to check the data quality. The quality of the data can be checked for its accuracy, completeness, consistency, timeliness, validity and interpretability. The major tasks include data cleaning, data integration, data reduction and data transformation.

We will focus on all these aspects in the following sections.

## 8.1 OBJECTIVES

After going through this unit, you should be able to:

- understand the definition of preprocessing of data;
- identify the purpose of preprocessing of data;
- describe the process of preprocessing;
- narrate the process of data reduction;
- list and discuss various methods in data reduction, and

- apply basic techniques for dealing with common problems with raw data including missing data inconsistent data, and data from multiple sources.

## 8.2 DATA PREPROCESSING : AN OVERVIEW

Data preprocessing is the process of transforming raw data into a useful, understandable format. Real-world or raw data usually has inconsistent formatting, human errors, and can also be incomplete. Data preprocessing resolves such issues and makes datasets more complete and efficient to perform data analysis.

It's a crucial process that can affect the success of data mining and machine learning projects. It makes knowledge discovery from datasets faster and can ultimately affect the performance of machine learning models.

In other words, data preprocessing is transforming data into a form that computers can easily work on. It makes data analysis or visualization easier and increases the accuracy and speed of the machine learning algorithms that train on the data.

As you know, a database is a collection of data points. Data points are also called observations, data samples, events, and records. Each sample is described using different characteristics, also known as features or attributes. Data preprocessing is essential to effectively build models with these features. Numerous problems can arise while collecting the data. You may have to aggregate data from different data sources, leading to mismatching data formats, such as integer and float. If you're aggregating data from two or more independent datasets, for example, the gender field may have two different values for men: man and male. Likewise, if you're aggregating data from ten different datasets, a field that's present in eight of them may be missing in the rest two.

By preprocessing data, we make it easier to interpret and use. This process eliminates inconsistencies or duplicates in data, which can otherwise negatively affect a model's accuracy. Data preprocessing also ensures that there aren't any incorrect or missing values due to human error or bugs. In short, employing data preprocessing techniques makes the database more complete and accurate.

### 8.2.1 Purpose of Data Preprocessing

Typical location properties in vast real-world datasets and databases are incomplete, chaotic and unfailing information. Unfinished data can arise for a number of reasons:

- Important attributes cannot always be usable.
- Valid data cannot be recorded because of misunderstanding or equipment failure.
- Data that disputed other recorded data should have been discarded.
- Missing data may be inferred, especially for tuples with missing values for certain attributes.
- Data collection techniques may be unreliable.
- At the moment of data entry, human or computer errors may have existed.
- Data processing failures can also occur.
- There might be technological disadvantages, such as narrow buffer sized for simultaneous data transfer and usage coordination.

- Data Routines for the cleaning of data are used by filling the missing values.
- Relieve noise effects, detect or remove outliers and address inconsistencies.
- Data integration is the hybrid process with many cubes or archives of databases. However, in multiple databases, those characteristics that define a specific may have separate titles, which lead to inconsistencies and redundancies.
- Data transformation is a process approach such as standardizations and consolidation that constitutes additional preprocessing processes that contribute to mining process results.
- Data reduction obtains a simpler data collection image, which is significantly smaller in duration, but provides the same analytical efficiency (or nearly the same). A multitude of methods for data reduction are in use. This includes:
- Data Aggregation (e.g., Creating a data cube).
- Attribute subset selection (By similarity, eliminating unnecessary attributes)
- Dimensionality Reduction (e.g., using encoding systems such as minimal encoding lengths or wavelets).
- Numerosity Reduction (e.g., "Replace" details by alternating, smaller representations such as clusters or parametric structures.
- Generalization (e.g., Data were minimized with the usage of the definition hierarchy).

### 8.2.2 Factors Contributing to Data Quality

Before looking at how data is preprocessed, let's look at some factors contributing to data quality as given below:

- **Accuracy:** Accuracy means that the information is correct. Outdated information, typos, and redundancies can affect a dataset's accuracy.
- **Consistency:** The data should have no contradictions. Inconsistent data may give you different answers to the same question.
- **Completeness:** The dataset shouldn't have incomplete fields or lack empty fields. This characteristic allows data scientists to perform accurate analyses as they have access to a complete picture of the situation the data describes.
- **Validity:** A dataset is considered valid if the data samples appear in the correct format, are within a specified range, and are of the right type. Invalid datasets are hard to organize and analyze.
- **Timeliness:** Data should be collected as soon as the event it represents occurs. As time passes, every dataset becomes less accurate and useful as it doesn't represent the current reality. Therefore, the topicality and relevance of data is a critical data quality characteristic.

It is not a simple and single step to do the data preprocessing and involves many stages which we will study in the next section.

## 8.3 DATA PREPROCESSING STAGES

An incomplete training set can lead to unintended consequences such as bias, leading to an unfair advantage or disadvantage for a particular group of people. Incomplete or inconsistent data can negatively affect the outcome of data mining projects as well. To resolve such problems, the process of data preprocessing is used.

There are four stages of data processing: data cleaning, data integration, data reduction, and data transformation as illustrated in Figure 1.

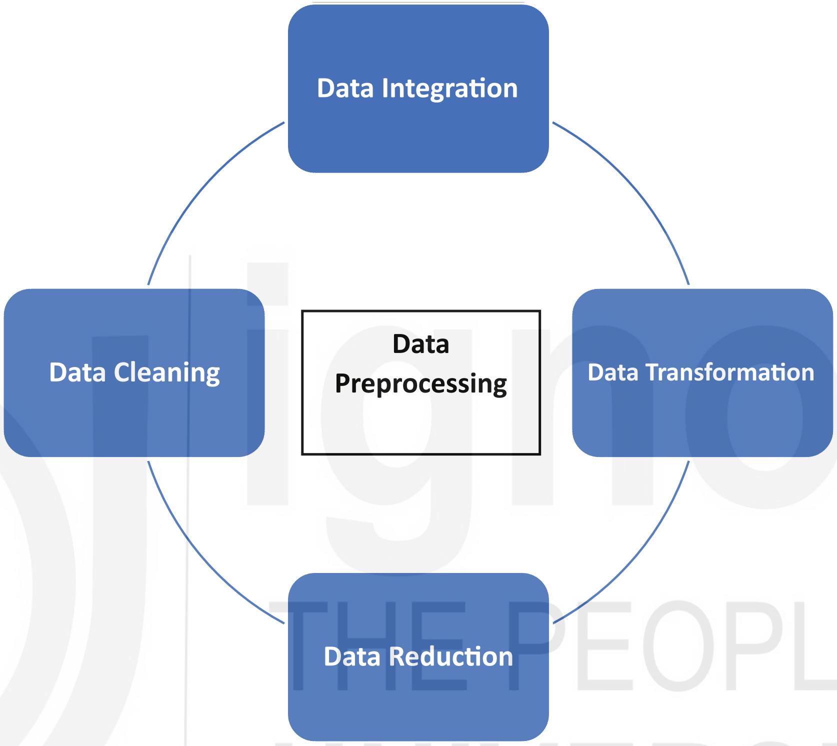

*Figure 1: Stages of Data Preprocessing*

### 8.3.1 Data Cleaning

Data cleaning or cleansing is the process of cleaning datasets by accounting for missing values, removing outliers, correcting inconsistent data points, and smoothing noisy data. In essence, the motive behind data cleaning is to offer complete and accurate samples for machine learning models.

The techniques used in data cleaning are specific to the data scientist's preferences and the problem they're trying to solve. Missing values and noisy data are issues that are solved during data cleaning and the techniques involved.

### 8.3.2 Data Integration

As data is collected from various sources, data integration is a crucial part of data preparation. Integration may lead to several inconsistent and redundant data points, ultimately leading to models with inferior accuracy. Following are some approaches to integrate data:

- **Data consolidation:** Data is physically brought together and stored in a single place. Having all data in one place increases efficiency and productivity. This step typically involves using data warehouse software.

- **Data virtualization:** In this approach, an interface provides a unified and real-time view of data from multiple sources. In other words, data can be viewed from a single point of view.
- **Data propagation:** Involves copying data from one location to another with the help of specific applications. This process can be synchronous or asynchronous and is usually event-driven.

### 8.3.3 Data Reduction

Data reduction is used to reduce the amount of data and thereby reduce the costs associated with data mining or data analysis. It offers a condensed representation of the dataset. Although this step reduces the volume, it maintains the integrity of the original data. This data preprocessing step is especially crucial when working with huge/large data. Dimensionality Reduction (which includes 2 segments feature selection and feature extraction), Feature subset selection and Numerosity Reduction are some of the techniques used for data reduction.

### 8.3.4 Data Transformation

Data Transformation is the process of converting data from one format to another. In essence, it involves methods for transforming data into appropriate formats that the computer can learn efficiently from. For example, the speed units can be miles per hour, meters per second, or kilometers per hour. Therefore a dataset may store values of the speed of a car in different units as such. Before feeding this data to an algorithm, we need to transform the data into the same unit. Smoothing, Aggregation, Discretization, Generalization, Feature Construction and concept hierarchy generation are some of the techniques used for data transformation.

More details on these stages can be studied in the following sections.

## 8.4 DATA CLEANING

The first stage of data preprocessing is Data cleaning which recognizes partial, incorrect, imprecise or inappropriate parts of the data from datasets. Data cleaning may eliminate typographical errors. It may ignore tuple contains missing values or alter values compared to a known list of entities. The data then becomes consistent with other data sets available in the system.

### 8.4.1 Missing Values

The problem of missing data values is quite common. It may happen during data collection or due to some specific data validation rule. In such cases, you need to collect additional data samples or look for additional datasets. The issue of missing values can also arise when you concatenate two or more datasets to form a bigger dataset. If not all fields are present in both datasets, it's better to delete such fields before merging. If 50% of values for any of the rows or columns in the database is missing, it's better to delete the entire row or column unless it's possible to fill the values using any of the above methods.

Some tuples have no significance mentioned for a range of attributes. We're going to fill up the vacant values. The following methods execute missing values over a number of attributes.

a) **Ignore the tuple**: This is usually done if the class label is missing (assuming the mining activity requires classification). This solution is not quite good since the tuple contains several attributes with missing values. It is really unfortunate since the percentage of values missed per variable varies greatly.

b) **Fill in the missing values manually:** This approach is time intensive and might not be feasible due to a wide range of data with many missing values.

c) **Use a global constant to fill in the missing values:** Replace all missing attributes with the same constant, such as a "unknown" or ∞.

d) **Use the attribute mean to fill in the missing value**

**e)** **Use the mean attribute for all samples belonging to the same class as the given tuple:** When the borrower is identified by credit danger, replace the missing number with the average income value in the same credit risk category as the tuple.

f) **Using the most likely meaning to fill in the missing value:** This can be computed using regression, inference-based tools using Bayesian formalism or the introduction of decision tree. For example, in the fixed Decision Tree the use of other market attributes is designed to approximate the missing revenue value.

### 8.4.2 Noisy Data

A large amount of meaningless data is called noise. More precisely, it's the random variance in a measured variable or data having incorrect attribute values. Noise includes duplicate or semi-duplicates of data points, data segments of no value for a specific research process, or unwanted information fields.

For example, if you need to predict whether a person can drive, information about their hair color, height, or weight will be irrelevant.

An outlier can be treated as noise, although some consider it a valid data point. Suppose you're training an algorithm to detect tortoises in pictures. The image dataset may contain images of turtles wrongly labeled as tortoises. This can be considered noise.

However, there can be a tortoise's image that looks more like a turtle than a tortoise. That sample can be considered an outlier and not necessarily noise. This is because we want to teach the algorithm all possible ways to detect tortoises, and so, deviation from the group is essential.

For numeric values, you can use a scatter plot or box plot to identify outliers.

Following are the methods for solving the problem of noisy data:

#### 8.4.2.1 Binning

Binning uses the "neighborhood" to smooth the storage value of the records. That's the value around it. The sorted values are split into "buckets" and "bins". Since binning methods consult the value community, local smoothing is carried out.

Stored price details (in dollars $): 2, 4, 6, 8, 10, 12, 14, 16, 18.

Example: Partition in bins (Equal-Frequency):

**Bin 1:** 2, 4, 6

**Bin 2:** 8, 10, 12

**Bin 3:** 14, 16, 18

In the above process, the price data is first sorted and then partitioned into equivalent frequency bins of size 3.

### Smoothing by bin mean

**Bin 1:** 4, 4, 4

**Bin 2:** 10, 10, 10

**Bin 3:** 16, 16, 16

The method assumes that any value in a bin is replaced with the mean value of the bin as it is smoothed by bin. In bin 1, for example the mean of definitions 2, 4 and 6 is 4 [(2+4+6)/3].

### Bin borders smoothing

**Bin 1:** 2, 2, 4

**Bin 2:** 8, 12, 12

**Bin 3:** 14, 14, 18

The maximum and minimum values are defined as the boundary values when the bin boundary is smoothed. Each bin value is then replaced by the closest limit value. The larger the diameter, the more smoothing impact is. Bins will also be of similar width when the range of values in each bin remains unchanged.

**Example:** Delete the noise with smoothing techniques from the following details:

4, 2, 6, 10, 8, 16, 12, 24, 22, 14, 26

**Stored price details (in dollars $):**

2, 4, 6, 8, 10, 12, 14, 16, 22, 24, 26

Partition in equal – frequency (equi-depth) bins:

**Bin 1:** 2, 4, 6, 8

**Bin 2:** 10, 10, 12, 14

**Bin 3:** 16, 22, 24, 26

**Smoothing by bin mean:**

**Bin 1:** 5, 5, 5, 5

**Bin 2:** 11, 11, 11, 11

**Bin 3:** 22, 22, 22, 22

**Smoothing by bin boundaries:**

**Bin 1:** 2, 2, 2, 8

**Bin 2:** 10, 10, 14, 14

**Bin 3:** 16, 16, 16, 26

### 8.4.2.2 Regression

By modifying data to perform, for instance for regression, data may be smoothed. Linear regression specifies that two attributes (or variables) be searched for the "right" line so one attribute can be used to predict the other.

Multiple linear regression is a linear regression extension, involving more than two attributes and outcomes that fit onto a multi-dimensional surface.

### 8.4.2.3 Clustering

Clustering Outliers may be detected if the same values are clustered in groups or clusters. Values that go outside the spectrum of clusters can intuitively be considered outliers.

### Check Your Progress 1:

1. What are missing values? How do you handle missing values?

   ……………………………………………………………………………
   ……………………………………………………………………………
   …………………………………………………………………………...

## 8.5 DATA INTEGRATION

Data Integration is the method of merging data derived from different sources of data into a consistent dataset. Data on the web is expanding in size and complexity, and is either unstructured or semi–structured. Integration of data is an extremely cumbersome and iterative process. The considerations during the integration process are mostly related to standards of heterogeneous data sources. Secondly, the process of integrating new data sources to the existing dataset is time–consuming, ultimately results in inappropriate consumption of valuable information. ELT (Extract–Transform–Load) tools are used to handle a larger volume of data; it integrates diverse sources into a single physical location, provides uniform conceptual schemas and provides querying capabilities.

Here are some approaches to integrate data:

- **Data consolidation:** Data is physically brought together and stored in a single place. Having all data in one place increases efficiency and productivity. This step typically involves using data warehouse software..

- **Data virtualization:** In this approach, an interface provides a unified and real-time view of data from multiple sources. In other words, data can be viewed from a single point of view.

- **Data propagation:** Involves copying data from one location to another with the help of specific applications. This process can be synchronous or asynchronous and is usually event-driven.

### 8.5.1 Data Integration Issues

During data migration, there are certain issues to be considered namely:

i. Schema Integration and Object Matching

ii. Redundancy

iii. Detection and Resolution of data value conflicts.

### i. Schema Integration and Object Matching:

It can be difficult because in different tables, different individuals can identify the same type. This is linked to the problem of individual identity. Metadata can be used to avoid errors in schema integration. Meta data can also be used for additional data conversation (e.g. where database codes "H" and "S" may be typed in one database, 1 & 2 in a separate database).

### ii Redundancy:

That is another serious issue, that if it can be "derived" from another attribute, an attribute (e.g. annual sales) can be redundant. Inconsistencies in the naming of dimensions can also add to the redundancies in the data collection arising from this. A correlation analysis can detect any redundancies, given the two attributes, the test determines how strongly the one attribute indicates the other, based on the data available. The relationship between two attributes, X and Y, can be tested for numerical attributes by calculating the correlation coefficient:

$$r_{X,Y}=(\sum(X-\bar{X})(Y-\bar{Y}))/(n-1)\sigma_A\,\sigma_B$$

where

n is the number of tuples

$\bar{X}$ means the value of X

$\bar{Y}$ means the value of Y

$\sigma_A$ Standard deviation of X

$\sigma_B$ Standard deviation of Y

$\overline{X}=\frac{\sum X}{n};\ \sigma_X=\sqrt{\sum}\sqrt{(X-\overline{X})^2)/(n-1)}$

$r_{X,Y}>0$ then X and Y are positively correlated

$r_{X,Y}<0$ then X and Y are negatively correlated

$r_{X,Y}=0$ then on correlation between X and Y

### iii. Detection and Resolution of Data Value Conflicts

The third major issue in the integration of data is the detection and resolution of conflicts over data. For instance, the significance of attributes from different sources can vary for the same real person – the world person. This could be connected to changes in image, size or encoding. Room prices in various cities covering not only different currencies, but also different services (such as free breakfast) and taxes for the hotel chain. A lower abstraction level of an attribute can be recorded on one system than the same attribute on another system. Semantic heterogeneity and structure of data face major data integration problems.

A thorough compilation of data from various sources can help to mitigate and avoid the resulting data collection of redundancies and anomalies. This helps to improve the accuracy and speed of the mining process.

## 8.6 DATA TRANSFORMATION

Raw data is usually transformed into a format suitable for analysis. Data can be normalized for instance transformation of the numerical variable to a common range. Categorical data can be transformed using aggregation which merges two or more attributes into a single attribute. Generalization can be applied on low–level attributes which are transformed to a higher level. Following are some of the strategies for data transformation:

- Smoothing
- Aggregation
- Discretization
- Attribute Construction
- Generalization
- Normalization

### 8.6.1 Smoothing

It is a mechanism used to remove data set noise using some algorithms to highlight main data set features. It helps to predict patterns and can be controlled when collecting data to eliminate or reduce any variance or noise.

The concept behind data smoothing is that simple changes that help to predict different trends and patterns can be detected. This helps investors or traders who want to look at a lot of data, which can be difficult to digest for patterns they would not see otherwise.

### 8.6.2 Aggregation

The compilation or grouping of data is a way of summer data collection and show. In order to integrate these data sources into the concept of data analysis, the data should be obtained from different data sources. This is a vital step because the accuracy of data analysis depends heavily on the volume and quality of the data used. Reliable high quality and sufficient amounts of data should be collected to achieve the necessary results. Data collection is useful in all aspects including judgments on the financing of commodity, pricing, processes and marketing strategies or business plans. For example, revenue can be aggregated to calculate the monthly and annual total.

### 8.6.3 Discretization

It is a way to transform continuous data into a number of short intervals. Many actual data mining activities have ongoing attributes. However, many of the existing data mining systems cannot handle these attributes.

While a data mining task can always handle a continuous attribute, a constant consistency attribute can significantly improve its efficiency by replacements of discrete values.

### 8.6.4 Attribute Construction

Where new attributes are created and added to support an attribute set in the mining procedure. This simplifies the original data and improves the efficiency of mining.

### 8.6.5 Generalization

Convert low - level data attributes with a hierarchical concept to high-level data attributes. For example, an age in numerical form initially (20, 64) is translated into a categorical sense (Young and Old).

Categorical attributes such as house addresses, for example, may be applied to higher-level definitions such as cities or areas.

### 8.6.6 Normalization

All data variables need to be translated into a given set. The standardization approach used is:

- Min – Max Normalization
- Z – Score Normalization
- Decimal Scaling Normalization

#### 8.6.6.1 Min – Max Normalization

- Which translates the initial data linearly.
- Suppose min X is the minimum and max X is the maximum limit of the variable.
- We've got the formula

$$\boldsymbol{v'} = \frac{\boldsymbol{v-minX}}{\boldsymbol{maxX-minX}}(\text{new\_}max_X\text{-new\_}min_X)\text{+ new\_}min_X$$

- where v is the value that you want to plot in the new set.
- v' is the fresh meaning you get when you normalize the old value.

**Example:** Assume the attribute revenue minimum and maximum values are 1000 and 16000 respectively. Diagram income in the [0.0, 1.0] range. For salaries a value of 14000 is converted into a minimum – maximum standardization.

$$= \frac{14000-1000}{16000-1000}(1.0-0)+0$$

$$=0.866$$

#### 8.6.6.2 Z-Score Normalization

- The value of a variable (X) in the Z-Score Normalization or Zero-Median Normalization is uniform according to the mean of the X and its standard deviation.
- The X value v of the X attribute is computer-standardized to V.

$$\boldsymbol{v'} = \frac{\boldsymbol{v-\overline{X}}}{\boldsymbol{\sigma\, X}}$$

Where$\overline{X}$ and σ are the mean and standard deviations respectively of the X attribute. This method is helpful if the true minimum and maximum attribute X are not certain or if outliers surpass the minimum – maximum normalization.

**Example:** Suppose that the average and standard deviation in attribute revenue number are respectively 22,000 and 7,000. In the case of Z-Score standardization, the revenue of 42,000 is transferred to Z-Score.

$$\frac{42000-22000}{7000} = 2.85$$

### 8.6.6.3 Decimal Scaling Normalization

- Standardizes the variable values by changing their decimal position.
- You can measure the number of points that the decimal point is passed by the absolute maximum value of the X attribute.

The value of v of the X attribute is computer-standardized to v^'.

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

J is such an integer that maximum (| v' |) < 1.

**Example:**

- Suppose the observed values differ between the lowest - 486 and the highest - 417.
- The X meaning square is 486. In order to normalize the calculation by decimal scaling, each value must be divided by 1000 so that -486 standardizes –0.486 and 417 to 0.417.

**Check Your Progress 2:**

1. What do you mean by Data Transformation? Mention some strategies to transform data.

   …………………………………………………………………………………
   …………………………………………………………………………………
   ……………………………………………………………………………..…

## 8.7 DATA REDUCTION

The last stage of data preprocessing is data reduction. Multifaceted exploration of huge data sources may consume considerable time or even be infeasible. When the number of predictor variables or the number of instances becomes large, mining algorithms suffer from dimensionality handling problems. Data reduction makes input data more effective in representation without loosening its integrity. Data reduction may or may not be lossless. The end database may contain all the information of the original database in well– organized format. Encoding techniques, hierarchy distribution data cube aggregation can be used to reduce the size of the dataset. Data reduction harmonizes feature selection process. Instance selection and Instance generation are two approaches used by data mining algorithm to reduce data size. Following are some of the strategies for Data Reduction:

1. **Data cube aggregation**, which applies aggregation of data during data cube construction.
2. **Set of sub-set attributes**, which may be detected and removed by obsolete, weakly significant or redundant measurements.
3. **Dimensionality reduction** where encoding approaches are used to reduce the data collection size to a minimum.
4. **Reduction of numbers when alternative**, smaller data representation replaces or predicts information.
5. **Discretization and category hierarchy generation** where raw data values

for the attributes are replaced by ranges or higher logical levels.

### 8.7.1 Data Cube Aggregation

Data Cube aggregation where a data cube is created using aggregation operations. Data cubes store multidimensional information aggregated. Each cell stores in a multi-dimensional space a data value corresponding to the data point.

For each attribute concept hierarchies may occur, allowing data to be evaluated at multiple abstraction levels. Data cubes provide quick access to pre-computed summary data; this will support both online analytical analysis and data mining.

### 8.7.2 Attribute Subset Selection

The selection of a sub-set attribute reduces data by removing redundant attributes or measurements irrelevant. The goal is to find a minimum set of attributes for the selection of a subset attribute. It decreases the number of attributes found in patterns, which promotes comprehension of patterns.

The data set will include a wide range of attributes. Some of them may be outdated or redundant, however. The objective of selecting attributes of sub-sets is to select a minimum number of attributes to lower data collection costs and to reduce loss of such unwanted attributes. The reduced data collection also enables the discovered pattern to be explained.

#### 8.7.2.1 Process of Attribute Subset Selection

The brute force approach can be very expensive, in which data with n attributes can be evaluated by any subset (2^n possible subsets). The simplest way to do this is to use metrics of predictive significance in order to recognise the best or worst attributes. The statistical significance test indicates that the characteristics are separate. This is a greedy approach where the statistically ideal value of the meaning level is determined by 5% and the models are tested time and time again until the p value of all the attributes is less than or equal to the selected meaning level. Attributes with a p-value above the value are discarded. This method is repeated time and time again until all attributes of the data set have a p-value that is less than or equal to their importance. This allows us to obtain reduced data with no irrelevant attributes.

#### 8.7.2.2 Attribute Subset Selection Methods

i. Stepwise Forward Selection

ii. Stepwise Backward Elimination

iii. Combination of Forward Selection and Backward Elimination

iv. Decision Tree Introduction

All of the above methods are greedy approaches to the selection of attribute subsets.

**1. Stepwise Forward Selection:**

The process starts with a minimum number of empty attributes. The most significant attributes with minimal p-value are selected and added to the minimum list. One attribute in each iteration is applied to the reduced collection.

**ii. Stepwise Backward Elimination**

All attributes are considered here in the initial collection of attributes. In each iteration, one attribute is omitted from the set of attributes of p significance.

### iii. Combination of Forward Selection and Backward Elimination

Phase by step, sorting and reverse exclusion are combined so that the necessary attributes are selected more effectively. This is the most common way to collect attributes.

### iv. Decision Tree Introduction

The introduction to the Decision Tree outlines a process flow diagram in which each internal node denotes an attribute test, every branch defines the possible test results and each leaf denotes the class forecast. For each node, the algorithm selects the best attribute to divide the data into different classes. From the data given, a tree is established. The tree contains a collection of attributes from the decreased subset attribute. The threshold measurement is used as a stop criterion.

## 8.7.3 Dimensionality Reduction

The final definition of machine learning problems is sometimes underlying too many factors. These factors are primarily variables known as features. The higher the number of features, the harder the training set can be imagined and worked on. Most of these features are frequently connected and therefore redundant. This is where algorithms of dimensional reduction come into play. The reduction in dimension is the method by obtaining a collection of main variables for reducing the number of random variables. It can be divided into feature selection and feature extraction.

### 8.7.3.1 Significance of Dimensionality Reduction

A simple explanation of reduced dimensionality is given by a basic e-mail classification problem, where we have to classify whether mail is spam or not. There is a wide range of features, including whether the e-mail has a general title, e-mail content, whether an e-mail uses a template, etc. However, both of these functions will overlap. In other cases, a question of moisture and rainfall classification would fall into only one basic characteristic, as the above two correlates strongly. Therefore, in such a problem, we can reduce the number of features. It is difficult to see a problem in 3-D classification if a 2-D can be mapped into a basic 2-D area with the problem of 1-D in a simple diagram. The following figure shows that a 3-D function space is divided into two 1-D function spaces, and the number of characteristics can be further decreased if they are associated.

### 8.7.3.2 Components of Dimensionality Reduction

Two-dimensional reduction components are available:

i. Feature Selection

ii. Feature Extraction

i **Feature Selection:** Here we try to fine-tune a subset of the original variables or features array to create a smaller subset that can be used for the query. It normally takes three directions namely – Filter, Wrapper and Embedded

**ii** **Feature Extraction:** This limits data to a lower dimension in a high dimensional space, i.e. space with a limited number of dimensions.

### 8.7.3.3 Dimensionality Reduction Methods

Following are some of the ways to perform Dimensionality Reduction:

- **Principal component analysis (PCA):** A statistical technique used to extract a new set of variables from a large set of variables. The newly extracted variables are called principal components. This method works only for features with numerical values.
- **High correlation filter:** A technique used to find highly correlated features and remove them; otherwise, a pair of highly correlated variables can increase the multicollinearity in the dataset.
- **Missing values ratio:** This method removes attributes having missing values more than a specified threshold.
- **Low variance filter:** Involves removing normalized attributes having variance less than a threshold value as minor changes in data translate to less information.
- **Random forest:** This technique is used to assess the importance of each feature in a dataset, allowing us to keep just the top most important features.

Other dimensionality reduction techniques include factor analysis, independent component analysis, and linear discriminant analysis (LDA).

### 8.7.3.4 Advantages and Disadvantages of Dimensionality Reduction

- It helps to compact data and also removes disc space.
- It reduces the time of calculation
- It also helps, if any, to eliminate redundant features.
- This can lead to a certain amount of loss of data.
- PCA continues to recognize ongoing correlations between variables that are often unwanted.
- If mean and covariance are not sufficient to classify datasets, PCA fails.
- We cannot see how many of the primary components for maintaining such thumb laws are applied.

## 8.7.4 Numerosity Reduction

It is a technique for data reduction that replaces the original information with a smaller kind of data representation. There are two ways to minimize numbers.

- Parametric Methods
- Non-Parametric Methods

### 8.7.4.1 Parametric Methods

Data is interpreted by any parametric system model. The model is used to simulate data, which means that only data parameters should be processed rather than actual data. To construct these models (i) regression and (ii) log-linear approaches are used.

i. **Regression**

Regression can be a simple linear regression or a multiple linear regression. If there is only one independent variable, the regression model is referred to as simple linear regression, and if there are more than one independent parameter, such models are called multiple linear regression.

The data is modelled on a straight line in linear regression. For example, the random variable y can be modelled as a linear function of another random variable x, with the equation y=a x+b where the line slopes and y – intercepts are determined by a and b (regression coefficients). Y is modelled as a linear function in multiple linear regression, with two or more (independent) prediction variables.

The problem with regression is where the output variable is a real or permanent feature, like "wage" or "weight." There are several different approaches, the simplest being linear regression. It tries to align the data with the right hyperplane travelling across the dots.

### Regression Analysis

It is a statistical way to estimate the interaction of dependent variables or criterion variables with one or more separate variables or predictors. The regression analysis explains adjustments for changes in the selected predictors in the parameters. Conditional predictor assumptions - parameters based on which the average value of the dependent variables is given when the individual variables are changed. Three main applications for the regression analysis are predictor strength, effect prediction and trend forecasting.

### Types of Regression

Given below are some of the types of Regression:

- Linear regression
- Logistic regression
- Polynomial regression
- Stepwise regression
- Ridge regression
- Lasso regression
- Elastic Net regression

## Linear Regression

This Linear Regression is used for statistical software. Linear regression is a linear approach for modelling an interaction with many predictors and explanatory variables between the criterion or scalar answer. Linear regression depends on the conditional distribution of the probability of the result given the predictor values. There is a chance that linear regression would overfit. The formulation of linear regression is:

$$y'=bX+A$$

## Logistic Regression

It shows that the dependent variable is assessed as a dichotomy. Logistic regression measures the parameters of the logistic model and defines a binomial regression.

Logistic regression would be used to discuss data with two possible parameters and the association of predictor criteria. The equation of logistic regression is:

$$\imath=\beta_0+\beta x_1+\beta_2 x_2$$

## Polynomial Regression

It is used for curvilinear information. The least square method is ideal for polynomial regression. The object of regression analysis is to model the expected value of the dependent variable y in relation to the independent variable X. The equation of polynomial regression is:

$$l=\beta_0+\beta_1 x_1+€$$

## Stepwise Regression

It is used with predictive models to balance regression models. It is done automatically. The variable is added or removed from the explanatory variables collection for each point. The methods of progressive regression are forward collection, retroactive exclusion and bi-directional elimination. The step-by-step regression is:

$$b_{j.std}=b_j S_x * (S_y^{-1})$$

## Ridge Regression

It is a method for studying the effects of multiple regressions. When there is a multi-linearity, the least square figures are unbiased. A certain degree of bias is added to the regression equations, which removes standard errors. The method of ridge regression is:

$$\beta = ( X^T X+ \lambda * I)^{-1}) X^T y$$

## Lasso Regression:

An approach to regression analysis that performs both variable selection and regularisation. It uses soft boundaries. It only selects a subset of covariates to use the final construct. The equation of Lasso regression is:

$$\boldsymbol{N^{-1} \sum_{i=1}^{N} \mathrm{f}\,(x_i, y_{I,} \alpha, \beta)}$$

## Elastic Regression

It is a regularized approach to regression that integrates lasso and ridge penalties linearly. It supports the creation of vector machines, metrics and portfolio optimization. The penalty's job is as follows:

$$\| \boldsymbol{\beta} \|_1 = \sum_{j=1}^{p} |\boldsymbol{\beta}_j$$

## Log-Linear Model

For a series of discrete attributes based on a smaller subset of dimensional combinations, a log-linear model can be used to predict the probability of each data point in a multidimensional space. This allows the development of higher dimensional data space from lower dimensional attributes.

The regression as well as the log - linear model on sparse data can be used, but their use can be limited.

### Non-Parametric Methods

These techniques are used to store reduced data representations such as histograms, clusters, samples, and aggregation of data cubes. Histograms, clustering, sampling and data cube aggregration falls under non-parametric methods.

### Histograms

Histogram is a frequency overview of the data. It uses binning to estimate the distribution of data that is a popular way of reducing data.

### Clustering

Clustering separates data into groups or clusters. This method divides all data into individual clusters. To keep specifics to a minimum, the representation of cluster data is used to bypass actual data. It also helps to locate outliers of data.

### Sampling

Sampling can be used for data reduction, as a wider range of data can be represented by a much smaller random sample (or subset).

### Data cube Aggregation:

Data Cube Aggregation means data transfer to a reduced number of measurements from a complex stage. The resulting data collection is smaller without the information required for the study mission.

## 8.7.5 Data Discretization and Concept Hierarchy Generation

Discretization techniques can be used to reduce the number of values for a given continuous attribute, by dividing the attribute into a range of intervals. Interval value labels can be used to replace actual data values. These methods are typically recursive, where a large amount of time is spent on sorting the data at each step. The smaller the number of distinct values to sort, the faster these methods should be.

Many discretization techniques can be applied recursively in order to provide a hierarchical or multiresolution partitioning of the attribute values known as concept hierarchy. A concept hierarchy for a given numeric attribute attribute defines a discretization of the attribute.

Concept hierarchies can be used to reduce the data y collecting and replacing low-level concepts (such as numeric value for the attribute age) by higher level concepts (such as young, middle-aged, or senior). Although detail is lost by such generalization, it becomes meaningful and it is easier to interpret. Manual definition of concept hierarchies can be tedious and time-consuming task for the user or domain expert. Fortunately, many hierarchies are implicit within the database schema and can be defined at schema definition level. Concept hierarchies often can be generated automatically or dynamically refined based on statistical analysis of the data distribution.

### 8.7.5.1 Data Discretization Categories

Following are the categories of Discretization:

- **Supervised Discretization:** Uses details regarding the class
- **Unsupervised Discretization or Splitting:** Does not use class knowledge

- **Top-Down Discretization or Splitting:** Here, the method begins by first finding one or a few points called break points or cut points to separate the whole set of attributes and then performs this recursively over the subsequent intervals.
- **Bottom – up Discretization or merging:** Here, the method starts by treating all continuous quantities as intervals and then applies this process recursively at the following intervals.

### 8.7.5.2 Discretization and Concept Hierarchy Generation for Numerical data

It is difficult and laborious for to specify concept hierarchies for numeric attributes due to the wide diversity of possible data ranges and the frequent updates if data values. Manual specification also could be arbitrary. Concept hierarchies for numeric attributes can be constructed automatically based on data distribution analysis. Five methods for concept hierarchy generation are defined below- binning histogram analysis entropy-based discretization and data segmentation by "natural partitioning".

i. Binning

ii. Histogram Analysis

iii. Entropy – Based Discretization

iv. Interval merging by X2 – Analysis

v. Cluster Analysis

vi. Discretization by Intuitive Partitioning

### i. Binning

It's a cutting technique that concentrates on the number of bins listed. Session techniques are also used as arbitrary methods to reduce numbers and generate hierarchical descriptions. These methods can be used recursively to generate Hierarchical definitions on the resulting partitions. It does not use class information and is therefore an arbitrary technique that is not tracked. It is adaptable to the number of containers identified by the customer as outliers.

### ii. Histogram Analysis

Like binning, histogram analysis is an unexpected discretization tool when class information is not used. Histograms partition the value for the disjoint range attribute A named buckets. The histogram analysis algorithm can be applied recursively to each partition to automatically produce a multi-level concept hierarchy that ends after a certain number of concept levels have been achieved.

### iii. Entropy – Based Discretization

It is one of the most commonly used discretization measures. Entropical discretization is a manual top-down separation technique. It addresses the knowledge of class distribution in its assessment and determination of divisions. The system selects the A value in order to discern the numerical function, which has the lowest entropy as a divisive point and corrects the resulting intervals to achieve a hierarchical discernment. This discretization forms a meaning hierarchy for A. Let D consist of a list of attributes and a data tuples defined for the label-class attributes. The simple way to distinguish an A attribute in the entropy set is as follows:

a. Each value of A can be seen as a potential interval or dividing point to separate the A spectrum. In other words, a split point for A divides tuples into two sub-sets in D, corresponding with conditions A <split point and A> and produces a binary disc recitation.

b. Suppose we divide the tuples into D with A and some split-points. Ideally, the partitioning of this would lead to an exact tuple classification. For instance, if we had two classes, we expect to have all class C1 tuples on one partition and all class C2 tuples on the other.

c. The method for the description of points of division is applied recursively before such stop thresholds are reached, for example when minimum information is fulfilled. The criterion is less than a small one, e or if the number of intervals is greater than a max interval should be required for all candidate split points.

### iv Interval merging by Chi-Square Analysis

- Chi blends the fastest way to pick the finest.
- Nearby loops and then mix recurrently in larger intervals. The method is regulated using classes.
- The Chi merge form is as follows:
- Initially, each independent value of a numerical attribute A is considered as one interval.
- For each adjacent pair of intervals, C2 monitoring is performed.
- Next intervals are combined with the lowest c2 values, since a couple of low c2 values show that the distributions are similar.
- A predefined stop criterion follows the fusion phase recursively.

### v. Cluster Analysis

Cluster analysis is a traditional technique of discretion. The A values can be split into clusters or groups to determine the numerical A attribute by a clustering algorithm. Clustering can be used to construct a concept hierarchy for A either by a top-down division technique or by a top-down approach in the form of a hierarchy node in each cluster. Each initial cluster or partition can be further broken down into several subclusters that form a lower hierarchical level. Repeated groups of neighboring clusters form clusters for the bottom-up approach to fusion.

### vi Discretization by Intuitive Partitioning:

Although the above discretization methods aid in numerical hierarchy, many users want numerical ranges divided into a very uniform, intuitive and natural interpretation of cycles.

The 3-4-5 rule can be used for a fairly uniform, natural appearance of numerical results. The rules generally divide a given data set by level according to the most important digit range into 3, 4 or 5 periods.

The fact is this:

- When a set of 3,6,7 or 9 is divided into three intervals on the largest digit (3 equal-width intervals for 3,6 and 9; and 3 intervals in the grouping of 2-3-2 for 7)

- If it covers 2, 4 or 8 separate values at the maximum digit, split the spectrum into 4 periods of equivalent duration.
- Split the spectrum into five equal-width increments if the largest digit spans 1,5 or 10 separate values.
- To establish a definition hierarchy for the numerical attribute assigned, the rule can be extended recursively for each interval.
- Real-world knowledge also includes extremely large positive or negative outer values based on minimum and maximum data values, which can distort any top-down discretionary process.

**Check Your Progress 3:**

1. How do you select the important features in your data?

…………………………………………………………………………………
…………………………………………………………………………………
……………………………………………………………………………...

## 8.8 SUMMARY

In this unit we had studied an important step in data mining i.e., *Data Preprocessing*. Data pre-processing consists of a series of steps to transform raw data derived from data extraction into a "clean" and "tidy" dataset prior to statistical analysis. Pre-processing aims at assessing and improving the quality of data to allow for reliable statistical analysis. We have studied several distinct stages those are involved in pre-processing data such as:

- *Data Cleaning:* This step deals with missing data, noise and duplicate or incorrect records while minimizing introduction of bias into the database.
- *Data Integration:* Extracted raw data can come from heterogeneous sources or be in separate datasets. This step reorganizes the various raw datasets into a single dataset that contain all the information required for the desired statistical analyses.
- *Data Transformation:* This step translates and/or scales variables stored in a variety of formats or units in the raw data into formats or units that are more useful for the statistical methods that the researcher wants to use.
- *Data Reduction:* After the dataset has been integrated and transformed, this step removes redundant records and variables, as well as reorganizes the data in an efficient and "tidy" manner for analysis.

Pre-processing is sometimes iterative and may involve repeating this series of steps until the data are satisfactorily organized for the purpose of statistical analysis. During pre-processing, one needs to take care not to accidentally introduce bias by modifying the dataset in ways that will impact the outcome of statistical analyses. Similarly, we must avoid reaching statistically significant results through "trial and error" analyses on differently pre-processed versions of a dataset.

## 8.9 SOLUTIONS/ANSWERS

### Check Your Progress 1:

1. In the data cleaning process, we can see there are lots of values that are missing or not filled or collected during the survey. We can handle such missing values using the following methods:

   - Ignoring such rows or dropping such records.
   - Fill values with mean, mode, and median.
   - you can also fill values using mean but for different classes, different means can be used.
   - You can also fill the most probable value using regression, Bayesian formula, or decision tree, KNN, and Prebuilt imputing libraries.
   - Fill with a constant value.
   - Fill values manually.

### Check Your Progress 2:

1. Data transformation consolidated or aggregate your data columns. It may impact your machine learning model performance. There are the following strategies to transform data:

   - Data Smoothing using binning or clustering
   - Aggregate your data
   - Scale or normalize your data for example scaling income column between 0 and 1 range.
   - Discretize your data for example convert continuous age column into the range 0–10, 11–20, and so on. Or we can also convert the continuous age column into conceptual labels such as youth, middle, and senior.

### Check Your Progress 3:

1. We can select the important features using random forest, or remove redundant features using recursive feature elimination. Let's all the categories of such methods.

   - **Filter Methods:** Pearson Correlation, Chi-Square, Anova, Information gain, and LDA.
   - **Wrapper Methods:** Forward Selection, backward elimination, Recursive feature elimination.
   - **Embedded Methods:** Ridge and Lasso Regression

## 8.10 FURTHER READINGS

1. Data Mining: Concepts and Techniques, 3rd Edition, Jiawei Han, Micheline Kamber, Jian Pei, Elsevier, 2012.
2. Data Mining, Charu C. Aggarwal, Springer, 2015.
3. Data Mining and Data Warehousing – Principles and Practical Techniques, Parteek Bhatia, Cambridge University Press, 2019.
4. Introduction to Data Mining, Pang Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar, Pearson, 2018.
5. Data Mining Techniques and Applications: An Introduction, Hongbo Du, Cengage Learning, 2013.
6. Data Mining : Vikram Pudi and P. Radha Krishna, Oxford, 2009.

ignou

THE PEOPLE'S

UNIVERSITY