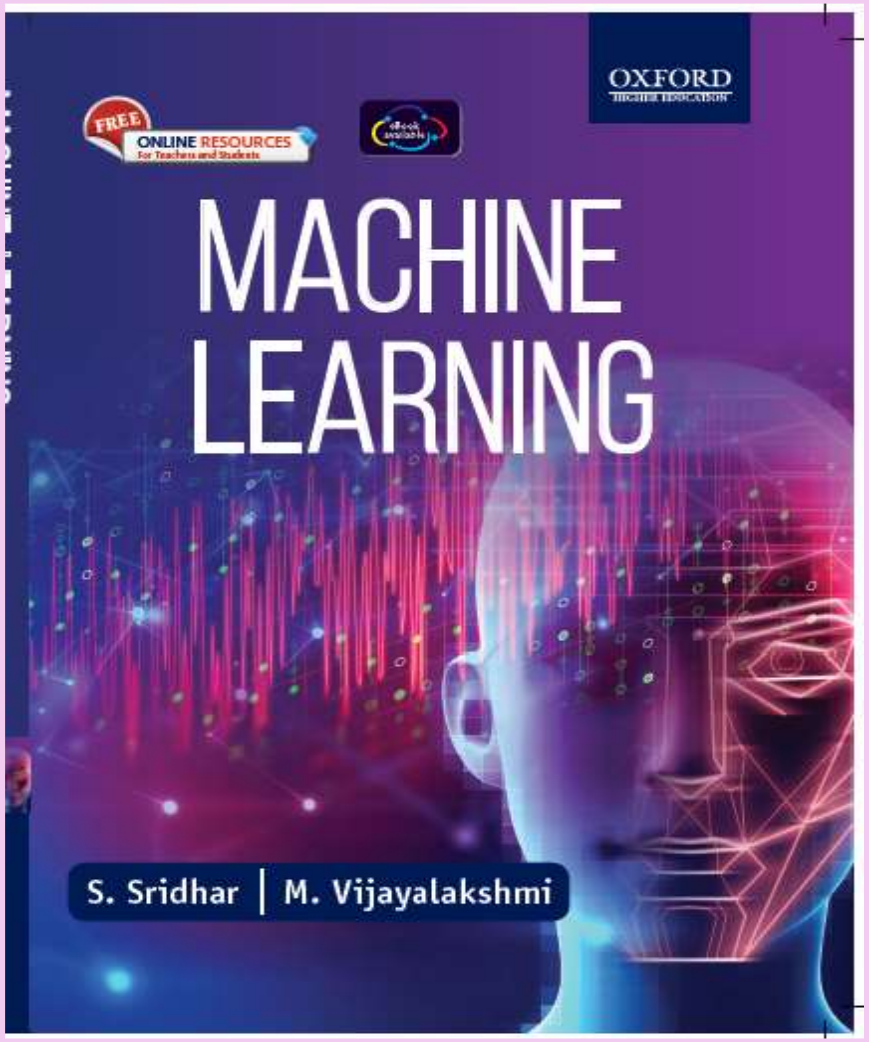


OXFORD
UNIVERSITY PRESS

Machine Learning

S. Sridhar and M. Vijayalakshmi

Chapter 2

Understanding of Data

What is Data?

- DATA ARE FACTS
- FACTS ARE IN THE FORM OF NUMBERS, AUDIO, VIDEO, IMAGE
- NEED TO ANALYZE DATA FOR TAKING DECISIONS

Characteristics of Big Data

1. Volume – Since there is a reduction in the cost of storing devices, there has been a tremendous growth of data. Small traditional data is measured in terms of gigabytes (GB) and terabytes (TB), but Big Data is measured in terms of petabytes (PB) and exabytes (EB). One exabyte is 1 million terabytes.
2. Velocity – The fast arrival speed of data and its increase in data volume is noted as velocity. The availability of IoT devices and Internet power ensures that the data is arriving at a faster rate. Velocity helps to understand the relative growth of big data and its accessibility by users, systems and applications.
3. Variety – The variety of Big Data includes:
   - Form – There are many forms of data. Data types range from text, graph, audio, video, to maps. There can be composite data too, where one media can have many other sources of data, for example, a video can have an audio song.
   - Function – These are data from various sources like human conversations, transaction records, and old archive data.
   - Source of data – This is the third aspect of variety. There are many sources of data. Broadly, the data source can be classified as open/public data, social media data and multimodal data. These are discussed in Section 2.3.1 of this chapter.

Characteristic of Data

4. Veracity of data – Veracity of data deals with aspects like conformity to the facts, truthfulness, believability, and confidence in data. There may be many sources of error such as technical errors, typographical errors, and human errors. So, veracity is one of the most important aspects of data.
5. Validity – Validity is the accuracy of the data for taking decisions or for any other goals that are needed by the given problem.
6. Value – Value is the characteristic of big data that indicates the value of the information that is extracted from the data and its influence on the decisions that are taken based on it.

Data Sources

A DATA SOURCE CAN BE ANYTHING –

- STRUCTURED DATA
- SEMI-STRUCTURED DATA
- UNSTRUCTURED DATA

Structured Data

A STRUCTURED DATA CAN BE ANY ONE OF THE FOLLOWING –

- RECORD DATA
- GRAPHICS DATA
- DATA MATRIX
- ORDERED DATA – SEQUENCE DATA, TIME SERIES DATA, TEMPORAL DATA

Unstructured Data

AN UNSTRUCTURED DATA CAN BE ANY ONE OF THE FOLLOWING –

- VIDEO, IMAGE, PROGRAMS
- BLOG DATA
- 80% OF ORGANIZATION DATA

SEMI-Structured Data

A SEMI-STRUCTURED DATA CAN BE ANY ONE OF THE FOLLOWING –

- XML/JSON OBJECTS
- RSS FEEDS
- HIERARCHICAL RECORDS

Data Storage

Flat Files These are the simplest and most commonly available data source. It is also the cheapest way of organizing the data. These flat files are the files where data is stored in plain ASCII or EBCDIC format. Minor changes of data in flat files affect the results of the data mining algorithms. Hence, flat file is suitable only for storing small dataset and not desirable if the dataset becomes larger.

Some of the popular spreadsheet formats are listed below:

- CSV files – CSV stands for comma–separated value files where the values are separated by commas. These are used by spreadsheet and database applications. The first row may have attributes and the rest of the rows represent the data.
- TSV files – TSV stands for Tab separated values files where values are separated by Tab.

Both CSV and TSV files are generic in nature and can be shared. There are many tools like Google Sheets and Microsoft Excel to process these files.

Data Storage

- DATABASE SYSTEMS
- TYPES ARE

1. TRANSACTIONAL DATABASE
2. TIME SERIES DATABASE
3. TEMPORAL DATABASE

Data Storage

- OTHER TYPES

World Wide Web (WWW) It provides a diverse, worldwide online information source. The objective of data mining algorithms is to mine interesting patterns of information present in WWW.

XML (eXtensible Markup Language) It is both human and machine interpretable data format that can be used to represent data that needs to be shared across the platforms.

Data Stream It is dynamic data, which flows in and out of the observing environment. Typical characteristics of data stream are huge volume of data, dynamic, fixed order movement, and real-time constraints.

RSS (Really Simple Syndication) It is a format for sharing instant feeds across services.

JSON (JavaScript Object Notation) It is another useful data interchange format that is often used for many machine learning algorithms.

Descriptive Analytics

Descriptive Analytics It is about describing the main features of the data. After data collection is done, descriptive analytics deals with the collected data and quantifies it. It is often stated that analytics is essentially statistics. There are two aspects of statistics – Descriptive and Inference. Descriptive analytics only focuses on the description part of the data and not the inference part.

Diagnostic Analytics

Diagnostic Analytics It deals with the question – 'Why?'. This is also known as causal analysis, as it aims to find out the cause and effect of the events. For example, if a product is not selling, diagnostic analytics aims to find out the reason. There may be multiple reasons and associated effects are analyzed as part of it.

Predictive Analytics

Predictive Analytics It deals with the future. It deals with the question – 'What will happen in future given this data?'. This involves the application of algorithms to identify the patterns to predict the future. The entire course of machine learning is mostly about predictive analytics and forms the core of this book.

Prescriptive Analytics

Prescriptive Analytics It is about the finding the best course of action for the business organizations. Prescriptive analytics goes beyond prediction and helps in decision making by giving a set of actions. It helps the organizations to plan better for the future and to mitigate the risks that are involved.

Data Analysis Framework

- FRAMEWORK

1. Date connection layer
2. Data management layer
3. Data analytics later
4. Presentation layer

Types of Processing

- CLOUD COMPUTING
- GRID COMPUTING
- H-COMPUTING

Good Data Characteristics

- GOD DATA SHOULD HAVE THESE CHARACTERISTICS

1. Timeliness – The data should be relevant and not stale or obsolete data.
2. Relevancy – The data should be relevant and ready for the machine learning or data mining algorithms. All the necessary information should be available and there should be no bias in the data.
3. Knowledge about the data – The data should be understandable and interpretable, and should be self-sufficient for the required application as desired by the domain knowledge engineer.

Open-Source Data

1. DIGITAL LIBRARIES
2. EXPERIMENTAL DATA LIKE GENOMIC AND BIOLOGICAL DATA
3. HEALTHCARE SYSTEMS LIKE PATIENT INSURANCE DATA

Social-Media Data

1. TWITTER DATA
2. FACEBOOK DATA
3. YOUTUBE VIDEOS
4. INSTAGRAM DATA

Multimodal Data

- IMAGE ARCHIVES WITH TEXT AND NUMERIC DATA
- WWW

Data Preprocessing

DATA THAT CAN CAUSE PROBLEMS

- INCOMPLETE DATA
- OUTLIER DATA
- INCONSISTENT DATA
- INACCURATE DATA
- MISSING VALUES
- DUPLICATE DATA

Missing Data

1. Ignore the tuple – A tuple with missing data, especially the class label, is ignored. This method is not effective when the percentage of the missing values increases.
2. Fill in the values manually – Here, the domain expert can analyse the data tables and carry out the analysis and fill in the values manually. But, this is time consuming and may not be feasible for larger sets.
3. A global constant can be used to fill in the missing attributes. The missing values may be 'Unknown' or be 'Infinity'. But, some data mining results may give spurious results by analysing these labels.
4. The attribute value may be filled by the attribute value. Say, the average income can replace a missing value.
5. Use the attribute mean for all samples belonging to the same class. Here, the average value replaces the missing values of all tuples that fall in this group.
6. Use the most possible value to fill in the missing value. The most probable value can be obtained from other methods like classification and decision tree prediction.

Noisy Data

BINNING TECHNIQUE

$S = \{12, 14, 19, 22, 24, 26, 28, 31, 34\}$

Bin 1 : 12 , 14, 19

Bin 2 : 22, 24, 26

Bin 3 : 28, 31, 32

By smoothing bins method, the bins are replaced by the bin means. This method results in:

Bin 1 : 15, 15, 15

Bin 2 : 24, 24, 24

Bin 3 : 30.3, 30.3, 30.3

Using smoothing by bin boundaries method, the bins' values would be like:

Bin 1 : 12, 12, 19

Bin 2 : 22, 22, 26

Bin 3 : 28, 32, 32

Data Normalization

MIN-MAX PROCEDURE

TRANSFORMS DATA TO THE RANGE 0-1

$$min\text{-}max = \frac{V - min}{max - min} \times (new\ max - new\ min) + new\ min$$

For marks 88,

$$min\text{-}max = \frac{88 - 88}{94 - 88} \times (1 - 0) + 0 = 0$$

Similarly, other marks can be computed as follows:

For marks 90,

$$min\text{-}max = \frac{90 - 88}{94 - 88} \times (1 - 0) + 0 = 0.33$$

For marks 92,

$$min\text{-}max = \frac{92 - 88}{94 - 88} \times (1 - 0) + 0 = \frac{4}{6} = 0.66$$

For marks 94,

$$min\text{-}max = \frac{94 - 88}{94 - 88} \times (1 - 0) + 0 = \frac{6}{6} = 1$$

So, it can be observed that the marks {88, 90, 92, 94} are mapped to the new range {0, 0.33, 0.66, 1}. Thus, the *Min-Max* normalization range is between 0 and 1.

Data Normalization

Z-SCORE

$$V* = V - \mu/\sigma$$

Here, σ is the standard deviation of the list V and μ is the mean of the list V.

Example 2.3: Consider the mark list $V = \{10, 20, 30\}$, convert the marks to z-score.

Solution: The mean and Sample Standard deviation (σ) values of the list V are 20 and 10, respectively. So the z-scores of these marks are calculated using Eq. (2.2) as:

$$z\text{-score of } 10 = \frac{10 - 20}{10} = -\frac{10}{10} = -1$$

$$z\text{-score of } 20 = \frac{20 - 20}{10} = \frac{0}{10} = 0$$

$$z\text{-score of } 30 = \frac{30 - 20}{10} = \frac{10}{10} = 1$$

Hence, the z-score of the marks 10, 20, 30 are −1, 0 and 1, respectively.

Types of Data

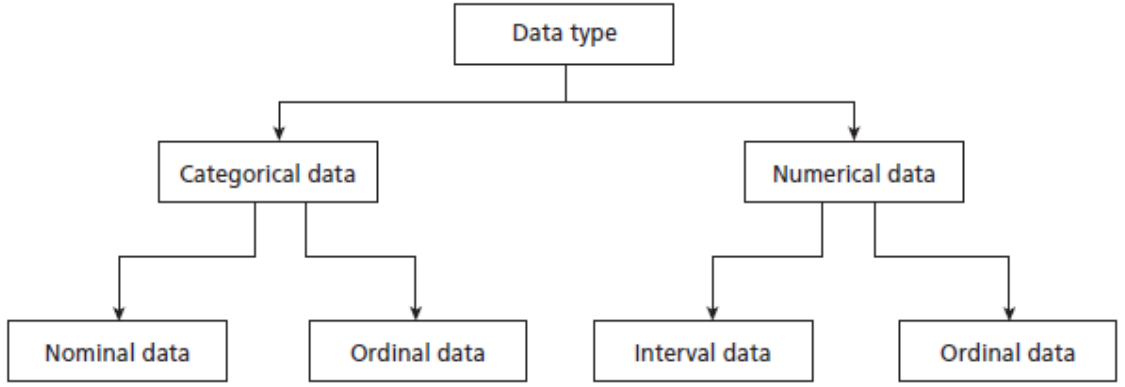


Figure 2.1: Types of Data

Nominal Data

Nominal Data – In Table 2.2, patient ID is nominal data. Nominal data are symbols and cannot be processed like a number. For example, the average of a patient ID does not make any statistical sense. Nominal data type provides only information but has no ordering among data. Only operations like (=, ≠) are meaningful for these data. For example, the patient ID can be checked for equality and nothing else.

Ordinal Data

Ordinal Data – It provides enough information and has natural order. For example, Fever = {Low, Medium, High} is an ordinal data. Certainly, low is less than medium and medium is less than high, irrespective of the value. Any transformation can be applied to these data to get a new value.

Numerical Data

Interval Data – Interval data is a numeric data for which the differences between values are meaningful. For example, there is a difference between 30 degree and 40 degree. Only the permissible operations are + and −.

Ratio Data – For ratio data, both differences and ratio are meaningful. The difference between the ratio and interval data is the position of zero in the scale. For example, take the Centigrade-Fahrenheit conversion. The zeroes of both scales do not match. Hence, these are interval data.

Types of Data

BASED ON VARIABLES

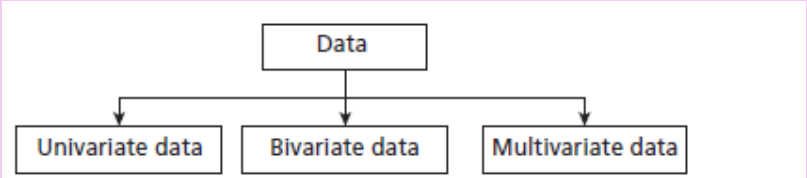
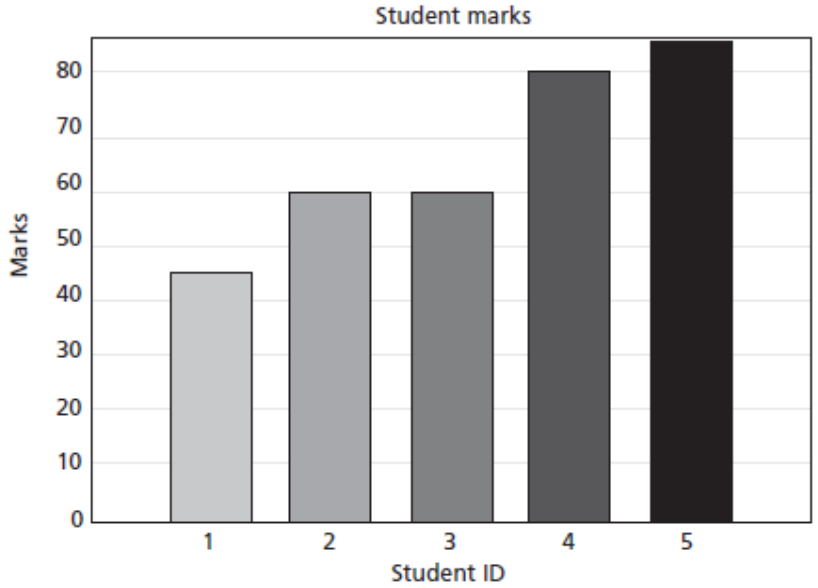


Figure 2.2: Types of Data Based on Variables

Data Visualization

Bar Chart A Bar chart (or Bar graph) is used to display the frequency distribution for variables. Bar charts are used to illustrate discrete data. The charts can also help to explain the counts of nominal data. It also helps in comparing the frequency of different groups.

The bar chart for students' marks {45, 60, 60, 80, 85} with Student ID = {1, 2, 3, 4, 5} is shown below in Figure 2.3.



Data Visualization

Pie Chart These are equally helpful in illustrating the univariate data. The percentage frequency distribution of students' marks {22, 22, 40, 40, 70, 70, 70, 85, 90, 90} is below in Figure 2.4.

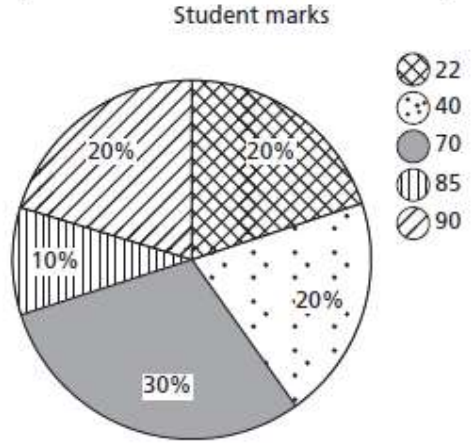


Figure 2.4: Pie Chart

It can be observed that the number of students with 22 marks are 2. The total number of students are 10. So, 2/10 × 100 = 20% space in a pie of 100% is allotted for marks 22 in Figure 2.4.

Data Visualization

Histogram It plays an important role in data mining for showing frequency distributions. The histogram for students' marks {45, 60, 60, 80, 85} in the group range of 0–25, 26–50, 51–75, 76–100 is given below in Figure 2.5. One can visually inspect from Figure 2.5 that the number of students in the range 76–100 is 2.

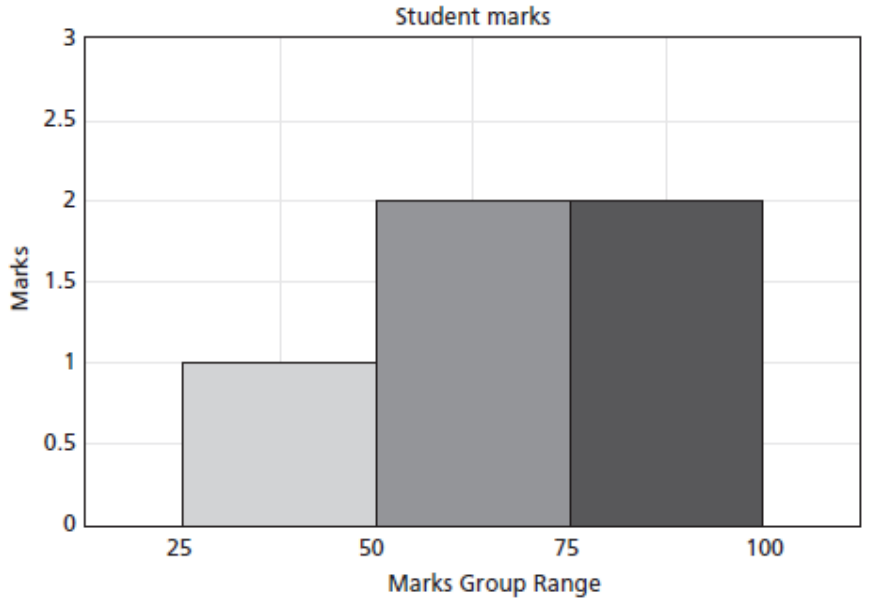


Figure 2.5: Sample Histogram of English Marks

Data Visualization

Dot Plots These are similar to bar charts. They are less clustered as compared to bar charts, as they illustrate the bars only with single points. The dot plot of English marks for five students with ID as {1, 2, 3, 4, 5} and marks {45, 60, 60, 80, 85} is given in Figure 2.6. The advantage is that by visual inspection one can find out who got more marks.

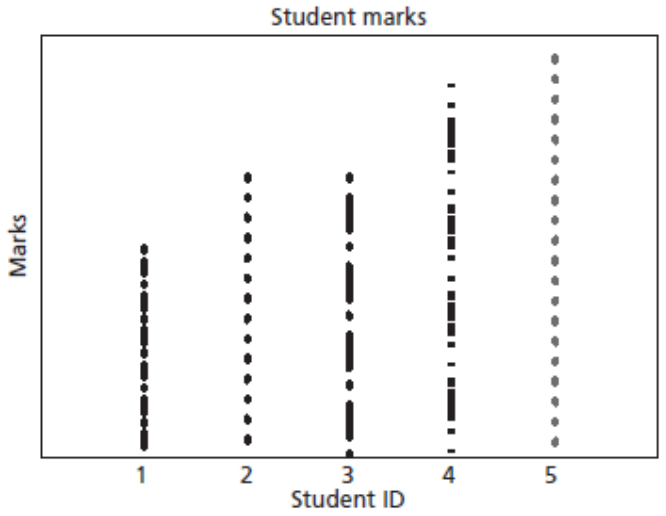
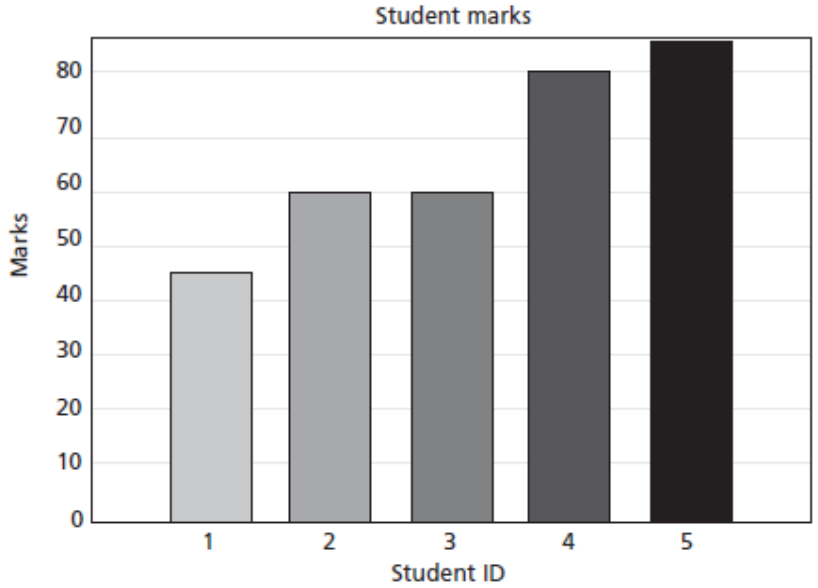


Figure 2.6: Dot Plots

Data Visualization

Bar Chart A Bar chart (or Bar graph) is used to display the frequency distribution for variables. Bar charts are used to illustrate discrete data. The charts can also help to explain the counts of nominal data. It also helps in comparing the frequency of different groups.

The bar chart for students' marks {45, 60, 60, 80, 85} with Student ID = {1, 2, 3, 4, 5} is shown below in Figure 2.3.



Central Tendency

MEAN OF DATA

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\text{Geometric mean} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{N}} = \sqrt[N]{x_1 \times x_2 \times \cdots \times x_N}$$

Central Tendency

MEDIAN OF DATA

Median – The middle value in the distribution is called median. If the total number of items in the distribution is odd, then the middle value is called median. If the numbers are even, then the average value of two items in the centre is the median. It can be observed that the median is the value where x_i is divided into two equal halves, with half of the values being lower than the median and half higher than the median. A median class is that class where $(N/2)^{th}$ item is present.

In the continuous case, the median is given by the formula:

$$\text{Median} = L_1 + \frac{\frac{N}{2} - cf}{f} \times i \qquad (2.7)$$

Central Tendency

MODE OF DATA

Mode – Mode is the value that occurs more frequently in the dataset. In other words, the value that has the highest frequency is called mode. Mode is only for discrete data and is not applicable for continuous data as there are no repeated values in continuous data.

DISPERSION

RANGE AND STANDARD DEVIATION

Range Range is the difference between the maximum and minimum of values of the given list of data.

Standard Deviation The mean does not convey much more than a middle point. For example, the following datasets {10, 20, 30} and {10, 50, 0} both have a mean of 20. The difference between these two sets is the spread of data.

Standard deviation is the average distance from the mean of the dataset to each point. The formula for sample standard deviation is given by:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}} \qquad (2.8)$$

Here, N is the size of the population, x_i is observation or value from the population and μ is the population mean. Often, $N-1$ is used instead of N in the denominator of Eq. (2.8). The reason is that for larger real-world, the division by $N-1$ gives an answer closer to the actual value.

DISPERSION

QUARTILES AND IQR

Quartiles and Inter Quartile Range It is sometimes convenient to subdivide the dataset using coordinates. Percentiles are about data that are less than the coordinates by some percentage of the total value. k^{th} percentile is the property that the k% of the data lies at or below X_i. For example, median is 50^{th} percentile and can be denoted as $Q_{0.50}$. The 25^{th} percentile is called first quartile (Q_1) and the 75^{th} percentile is called third quartile (Q_3).

Another measure that is useful to measure dispersion is Inter Quartile Range (IQR). The IQR is the difference between Q_3 and Q_1.

$$\text{Interquartile percentile} = Q_3 - Q_1 \tag{2.9}$$

Outliers are normally the values falling apart at least by the amount 1.5 × IQR above the third quartile or below the first quartile.

Interquartile is defined by $Q_{0.75} - Q_{0.25}$. (2.10)

Five-point summary

5-POINT SUMMARY

Five-point Summary and Box Plots The median, quartiles Q_1 and Q_3, and minimum and maximum written in the order < Minimum, Q_1, Median, Q_3, Maximum > is known as five-point summary.

Example 2.5: Find the 5-point summary of the list {13, 11, 2, 3, 4, 8, 9}.

Solution: The minimum is 2 and the maximum is 13. The Q_1, Q_2 and Q_3 are 3, 8 and 11, respectively. Hence, 5-point summary is {2, 3, 8, 11, 13}, that is, {minimum, Q_1, median, Q_3, maximum}.

Box plots are useful for describing 5-point summary. The Box plot for the set is given in Figure 2.7.

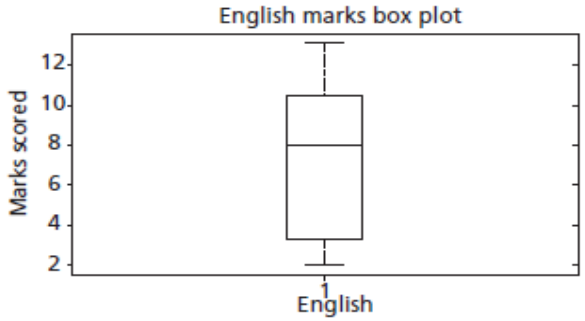


Figure 2.7: Box Plot for English Marks

Shape of Data

SKEWNESS AND KURTOSIS

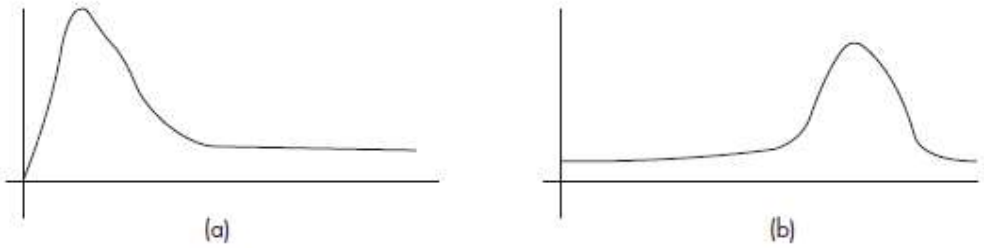


Figure 2.8: (a) Positive Skewed and (b) Negative Skewed Data

Also, the following measure is more commonly used to measure skewness. Let $X_1, X_2, \cdots, X_N$ be a set of 'N' values or observations then the skewness can be given as:

$$\frac{1}{N} \times \sum_{i=1}^{N} \frac{(x_i - \mu)^3}{\sigma^3} \qquad (2.13)$$

Shape of Data

KURTOSIS

Kurtosis also indicates the peaks of data. If the data is high peak, then it indicates higher kurtosis and vice versa.

$$\frac{\sum_{i=1}^{N}(x_i - \bar{x})^4 / N}{\sigma^4}$$

Shape of Data

MEAN ABSOLUTE DEVIATION AND COEFFICIENT OF VARIATION

Mean Absolute Deviation (MAD)

MAD is another dispersion measure and is robust to outliers. Normally, the outlier point is detected by computing the deviation from median and by dividing it by MAD. Here, the absolute deviation between the data and mean is taken. Thus, the absolute deviation is given as:

$$|x - \mu| \tag{2.15}$$

The sum of the absolute deviations is given as $\Sigma|x - \mu|$

Therefore, the mean absolute deviation is given as: $\frac{\Sigma|x - \mu|}{N}$ (2.16)

Coefficient of Variation (CV)

Coefficient of variation is used to compare datasets with different units. CV is the ratio of standard deviation and mean, and %CV is the percentage of coefficient of variations.

Stem-Leaf Plot

The stem and leaf plot for the English subject marks, say, {45, 60, 60, 80, 85} is given in Figure 2.9.

Stem	Leaf
4	5
5	
6	0 0
7	
8	0 5

Figure 2.9: Stem and Leaf Plot for English Marks

Q-Q Plot

QQ PLOT IS NORMALITY TEST. IF DATA CLOSER TO STRAIGHT LINE, THEN THE DISTRIBUTION IS NORMAL.

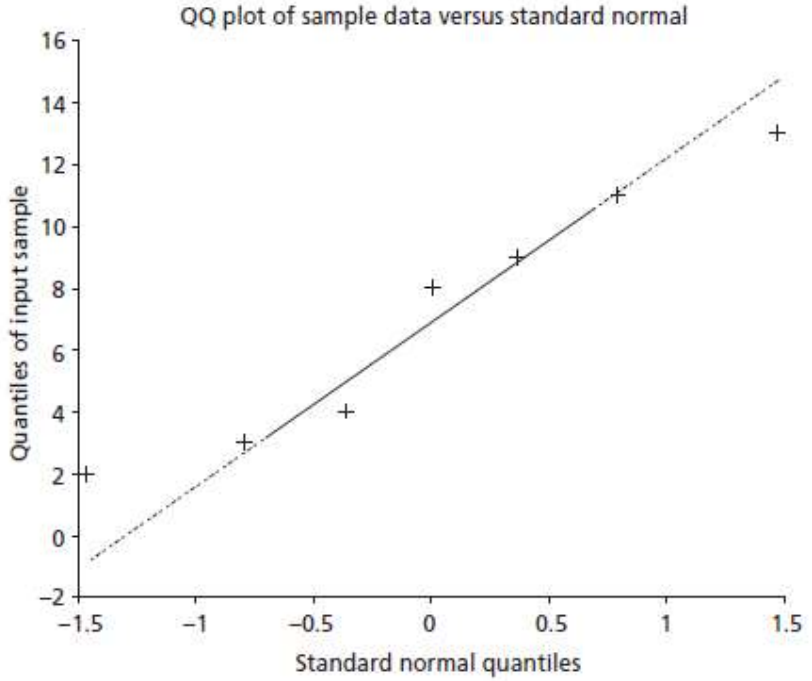


Figure 2.10: Normal *Q-Q* Plot

Bivariate Data

INVOLVES TWO VARIABLES

Table 2.3: Temperature in a Shop and Sales Data

Temperature (in centigrade)	Sales of Sweaters (in thousands)
5	200
10	150
15	140
20	75
22	60
23	55
25	20

Bivariate Data Visualization

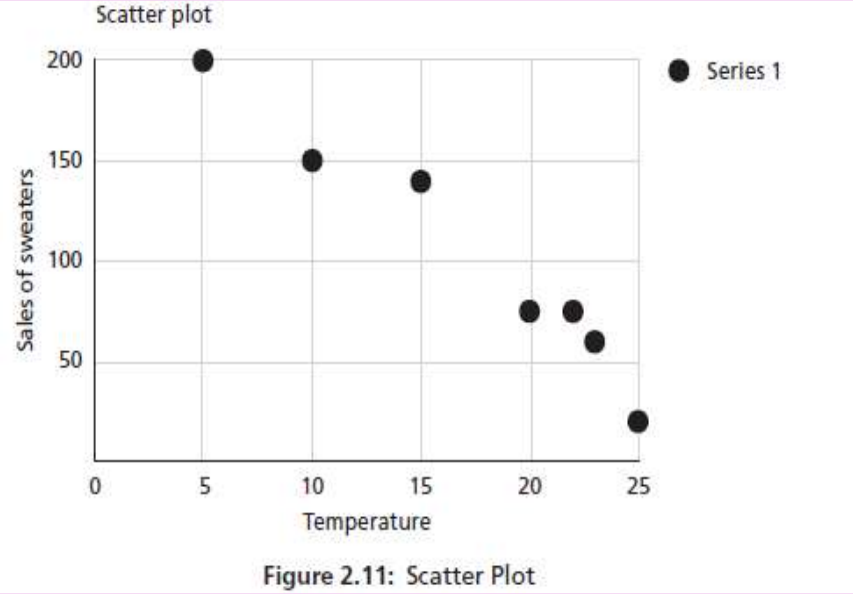


Figure 2.11: Scatter Plot

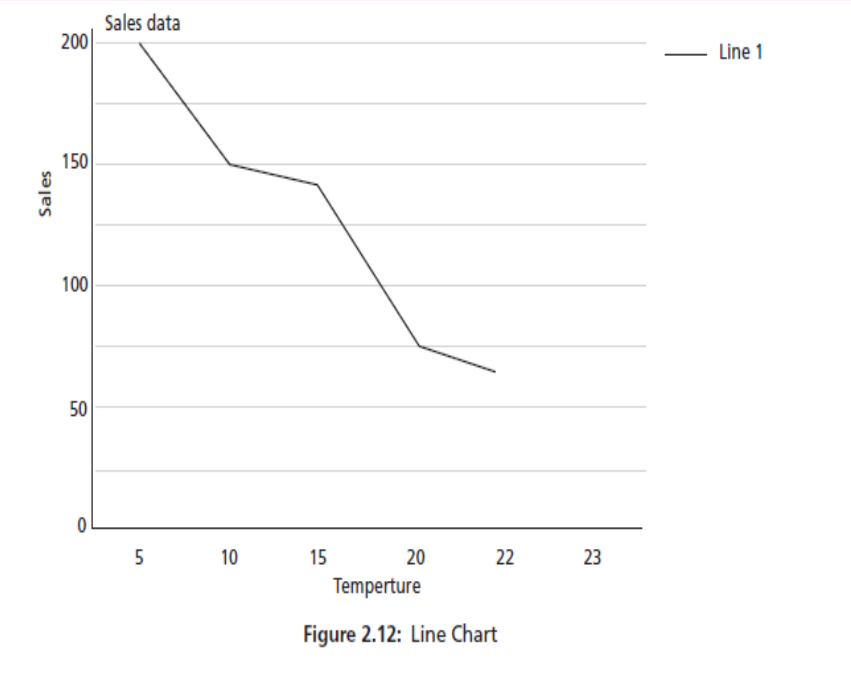


Figure 2.12: Line Chart

Bivariate Data – Covariance

$$\text{cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - E(X))(y_i - E(Y))$$

Example 2.6: Find the covariance of data $X = \{1, 2, 3, 4, 5\}$ and $Y = \{1, 4, 9, 16, 25\}$.

Solution: $\text{Mean}(X) = E(X) = \frac{15}{5} = 3$, $\text{Mean}(Y) = E(Y) = \frac{55}{5} = 11$. The covariance is computed using Eq. (2.17) as:

$$\frac{(1-3)(1-11) + (2-3)(4-11) + (3-30)(9-11) + (4-3)(16-11) + (5-3)(25-11)}{5} = 12$$

The covariance between X and Y is 12. It can be normalized to a value between −1 and +1. This is done by dividing it by the correlation of variables. This is called Pearson correlation coefficient. Sometimes, $N - 1$ is also can be used instead of N. In that case, the covariance is 60/4 = 15.

Bivariate Data – Correlation

If the given attributes are $X = (x_1, x_2, \cdots, x_N)$ and $Y = (y_1, y_2, \cdots, y_N)$, then the Pearson correlation coefficient, that is denoted as r, is given as:

$$r = \frac{COV(X,Y)}{\sigma_X \sigma_Y} \tag{2.18}$$

where, σ_X, σ_Y are the standard deviations of X and Y.

Bivariate Data – Correlation

Example 2.7: Find the correlation coefficient of data $X = \{1, 2, 3, 4, 5\}$ and $Y = \{1, 4, 9, 16, 25\}$.

Solution: The mean values of X and Y are $\frac{15}{5} = 3$ and $\frac{55}{5} = 11$. The standard deviations of X and Y are 1.41 and 8.6486, respectively. Therefore, the correlation coefficient is given as ratio of covariance (12 from the previous problem 2.5) and standard deviation of x and y as per Eq. (2.18) as:

$$r = \frac{12}{1.41 \times 8.6486} \approx 0.984$$

Multivariate Data Visualization

Heatmap

Heatmap is a graphical representation of 2D matrix. It takes a matrix as input and colours it. The darker colours indicate very large values and lighter colours indicate smaller values. The advantage of this method is that humans perceive colours well. So, by colour shaping, larger values can be perceived well. For example, in vehicle traffic data, heavy traffic regions can be differentiated from low traffic regions through heatmap.

Figure 2.13: Heatmap for Patient Data

Multivariate Data Visualization

Pairplot

Pairplot or scatter matrix is a data visual technique for multivariate data. A scatter matrix consists of several pair-wise scatter plots of variables of the multivariate data. All the results are presented in a matrix format. By visual examination of the chart, one can easily find relationships among the variables such as correlation between the variables.

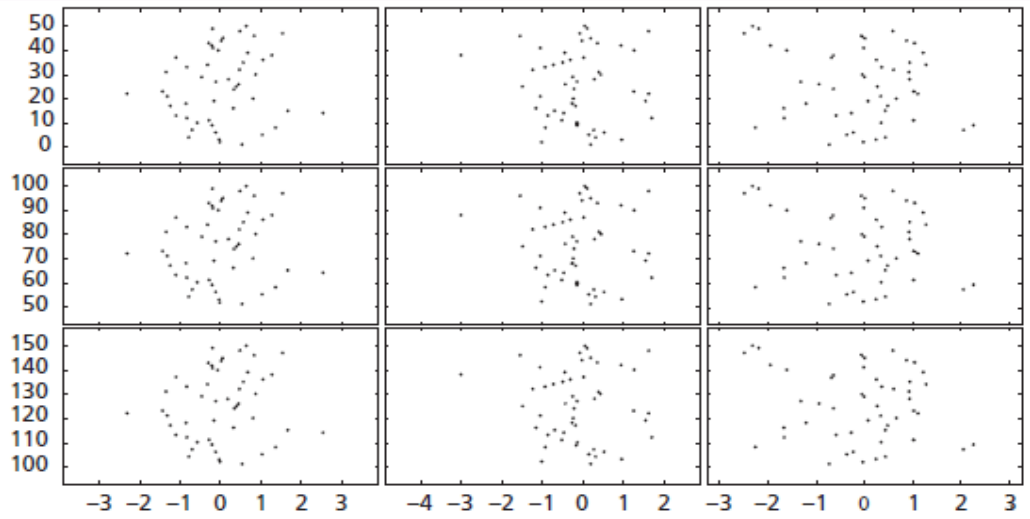


Figure 2.14: Pairplot for Random Data

Multivariate Essential Mathematics

1. GAUSSIAN ELIMINATION

To facilitate the application of Gaussian elimination method, the following row operations are applied:

1. Swapping the rows
2. Multiplying or dividing a row by a constant
3. Replacing a row by adding or subtracting a multiple of another row to it

These concepts are illustrated in Example 2.8.

Multivariate Essential Mathematics

1. GAUSSIAN ELIMINATION

Example 2.8: Solve the following set of equations using Gaussian Elimination method.

$$2x_1 + 4x_2 = 6$$
$$4x_1 + 3x_2 = 7$$

Solution: Rewrite this in matrix form as follows:

$$\begin{pmatrix} 2 & 4 & | & 6 \\ 4 & 3 & | & 7 \end{pmatrix}$$
$$\sim \begin{pmatrix} 2 & 4 & | & 6 \\ 4 & 3 & | & 7 \end{pmatrix} R_1 = \frac{R_1}{2}$$

Apply the transformation by dividing the row 1 by 2. There are no general guidelines of row operations other than reducing the given matrix to row echelon form. The operator ~ means reducing to. The above matrix can further be reduced as follows:

$$\sim \begin{pmatrix} 1 & 2 & | & 3 \\ 4 & 3 & | & 7 \end{pmatrix} R_2 = R_2 - 4R_1$$
$$\sim \begin{pmatrix} 1 & 2 & | & 3 \\ 0 & -5 & | & -5 \end{pmatrix} R_2 = R_2 / -5$$
$$\sim \begin{pmatrix} 1 & 2 & | & 3 \\ 0 & 1 & | & 1 \end{pmatrix} R_1 = R_1 - 2R_2$$
$$\sim \begin{pmatrix} 1 & 0 & | & 1 \\ 0 & 1 & | & 1 \end{pmatrix}$$

Therefore, in the reduced echelon form, it can be observed that:

$$x_2 = 1$$
$$x_1 = 1$$

Multivariate Essential Mathematics

1. MATRIX DECOMPOSITION

LU Decomposition

One of the simplest matrix decompositions is *LU* decomposition where the matrix *A* can be decomposed matrices:

$$A = LU$$

Here, *L* is the lower triangular matrix and *U* is the upper triangular matrix. The decomposition can be done using Gaussian elimination method as discussed in the previous section. First, an identity matrix is augmented to the given matrix. Then, row operations and Gaussian elimination is applied to reduce the given matrix to get matrices *L* and *U*.

Multivariate Essential Mathematics

1. MATRIX DECOMPOSITION

Example 2.9: Find LU decomposition of the given matrix:

$$A = \begin{pmatrix} 1 & 2 & 4 \\ 3 & 3 & 2 \\ 3 & 4 & 2 \end{pmatrix}$$

Solution: First, augment an identity matrix and apply Gaussian elimination. The steps are as shown in:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 & 4 \\ 3 & 3 & 2 \\ 3 & 4 & 2 \end{bmatrix} \quad \boxed{\text{Initial Matrix}}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 & 4 \\ 0 & -3 & -10 \\ 3 & 4 & 2 \end{bmatrix} \quad \boxed{R_2 = R_2 - 3R_1}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 & 4 \\ 0 & -3 & -10 \\ 0 & -2 & -10 \end{bmatrix} \quad \boxed{R_3 = R_3 - 3R_1}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 3 & \frac{2}{3} & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 & 4 \\ 0 & -3 & -10 \\ 0 & 0 & \frac{-10}{3} \end{bmatrix} \quad \boxed{R_3 = R_3 - \frac{2}{3}R_2}$$

Now, it can be observed that the first matrix is L as it is the lower triangular matrix whose values are the determiners used in the reduction of equations above such as 3, 3 and 2/3. The second matrix is U, the upper triangular matrix whose values are the values of the reduced matrix because of Gaussian elimination.

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 3 & \frac{2}{3} & 1 \end{pmatrix} \text{ and } U = \begin{pmatrix} 1 & 2 & 4 \\ 0 & -3 & -10 \\ 0 & 0 & -\frac{10}{3} \end{pmatrix}.$$

Multivariate Essential Mathematics

1. DISTRIBUTIONS

PDF of the normal distribution is given as:

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Rectangular Distribution – This is also known as uniform distribution. It has equal probabilities for all values in the range a, b. The uniform distribution is given as follows:

$$P(X = x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{Otherwise} \end{cases} \quad (2.25)$$

Multivariate Essential Mathematics

EXPONENTIAL DISTRIBUTION

The PDF is given as follows:

$$f(x, \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad (\lambda > 0)$$

Multivariate Essential Mathematics

BINOMIAL DISTRIBUTION

The binomial distribution function is given as follows, where p is the probability of success and probability of failure is $(1 - p)$. The probability of success in a certain number of trials is given as:

$$p^k(1-p)^{n-k} \text{ or } p^k q^{n-k} \tag{2.28}$$

Combining both, one gets PDF of binomial distribution as:

$$\binom{n}{k} p^k (1-p)^{n-k} \tag{2.29}$$

Multivariate Essential Mathematics

POSSON AND BERNOULLI DISTRIBUTION

The PDF of Poisson distribution is given as follows:

$$f(X = x; \lambda) = Pr[X = x] = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

Bernoulli Distribution – This distribution models an experiment whose outcome is binary. The outcome is positive with p and negative with $1 - p$. The PMF of this distribution is given as:

$$f(k;p) = \begin{cases} q = 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1. \end{cases} \quad (2.34)$$

The mean is p and variance is $p(1 - p) = q$

Density Estimation

Generally, there can be many unspecified distributions with different set of parameters. The EM algorithm has two stages:

1. Expectation (E) Stage – In this stage, the expected PDF and its parameters are estimated for each latent variable.
2. Maximization (M) stage – In this, the parameters are optimized using the MLE function.

Hypothesis Testing

Z-TEST

$$Z = \frac{X - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

Example 2.10: Let 12 be the population mean (μ) with the population variance (σ^2) of 2. Consider the sample $X = \{1, 2, 3, 4, 5\}$. Apply z-test and show whether the result is significant.

Solution: The sample mean of $X = \frac{15}{5} = 3$ and the number of samples is $n = 5$. Substituting in Eq. (2.48) gives:

$$Z = \frac{X - \mu}{\frac{\sigma}{\sqrt{n}}} = -10.06$$

By checking the critical value at significance 0.05, one can find that the null hypothesis H0 is rejected.

Hypothesis Testing

PAIRED T-TEST

One Sample Test In this test, the mean of one group is checked against the set average that can be either theoretical value or population mean. So, the procedure is:

- Select a group
- Compute average
- Compare it with theoretical value and compute *t*-statistic:

$$t = \frac{m - \mu}{\frac{s}{\sqrt{n}}} \qquad (2.49)$$

Here, t is t-statistic, m is the mean of the group, μ is the theoretical value or population mean, s is the standard deviation, and n is the group size or sample size.

Hypothesis Testing

PAIRED T-TEST

Independent Two Sample *t*-test *t*-statistic for two groups *A* and *B* is computed as follows:

$$t = \frac{mean(A) - mean(B)}{\sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} \quad (2.50)$$

Hypothesis Testing

PAIRED T-TEST

$$t = \frac{m - \mu}{\frac{s}{\sqrt{n}}}$$

Here, t is t-statistic, m is the mean of the group, μ is the theoretical value or population mean, s is the standard deviation, and n is the group size or sample size.

Hypothesis Testing

CHI-SQUARE TEST

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad (2.51)$$

Here, E is the expected frequency, O is the observed frequency and the degree of freedom is $C - 1$, where, C is number of categories. The Chi-Square test allows us to detect the duplication of data and helps to remove the redundancy of values.

Hypothesis Testing

CHI-SQUARE TEST

Example 2.11: Consider the following Table 2.4, where the machine learning course registration is done by both boys and girls. There are 50 boys and 50 girls in the class and the registration of the course is given in the table. Apply Chi-Square test and find out whether any differences exist between boys and girls for course registration.

Table 2.4: Observed Data

Gender	Registered	Not Registered	Total
Boys	35	15	50
Girls	25	25	50
Total	60	40	100

Hypothesis Testing

CHI-SQUARE TEST

Table 2.5: Expected Data

Gender	Registered	Not Registered	Total
Boys	$\frac{50 \times 60}{100} = 30$	$\frac{50 \times 40}{100} = 20$	50
Girls	$\frac{50 \times 60}{100} = 30$	$\frac{50 \times 40}{100} = 20$	50
Total	60	40	100

The Chi-Statistic is obtained using Eq. (2.51) as follows:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = \frac{(35-30)^2}{30} + \frac{(15-20)^2}{20} + \frac{(25-30)^2}{30} + \frac{(25-20)^2}{20} = 4.166$$ for degree of freedom = number of categories −1 = 2 − 1 = 1. The p value for this statistic is 0.0412. This is less than 0.05. Therefore, the result is significant.

Feature Engineering

Features are attributes. Feature engineering is about determining the subset of features that form an important part of the input that improves the performance of the model, be it classification or any other model in machine learning.

Feature Engineering

- FEATURE TRANSFORMATION
- FEATURE SELECTIONS

Characteristics of Good Features

- FEATURES ARE REMOVED USING RELEVANCY
- FEATURES ARE REMOVED BASED ON REDUNDANCY

FEATURE SELECTION

FORWARD SELECTION

This procedure starts with an empty set of attributes. Every time, an attribute is tested for statistical significance for best quality and is added to the reduced set. This process is continued till a good reduced set of attributes is obtained.

FEATURE SELECTION

BACKWARD SELECTION

This procedure starts with a complete set of attributes. At every stage, the procedure removes the worst attribute from the set, leading to the reduced set.

Principal Component Analysis

Consider a group of random vectors of the form:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

The *mean vector* of the set of random vectors is defined as:

$$m_x = E\{x\}$$

Principal Component Analysis

Compute Covariance matrix as

$$C = E\{(x - m_x)(x - m_x)^T\}$$

Compute Eigen values and Eigen vectors and matrix A as a set of eigen vectors

$$A = \frac{1}{M}\sum_{i=1}^{M} x_i x_i^T - m_x m_x^T$$

Principal Component Analysis

Compute PCA as

$$y = A(x - m_x)$$

The original Data can be recovered as

$$x = A^T y + m_x$$

If K largest eigen values are used, the recovered information would be:

$$x = A_K^T y + m_x$$

PCA Algorithm

1. The target dataset x is obtained
2. The mean is subtracted from the dataset. Let the mean be m. Thus, the adjusted dataset is $X - m$. The objective of this process is to transform the dataset with zero mean.
3. The covariance of dataset x is obtained. Let it be C.

4. Eigen values and eigen vectors of the covariance matrix are calculated.
5. The eigen vector of the highest eigen value is the principal component of the dataset. The eigen values are arranged in a descending order. The feature vector is formed with these eigen vectors in its columns.

   Feature vector = {eigen vector$_1$, eigen vector$_2$, ... , eigen vector$_n$}
6. Obtain the transpose of feature vector. Let it be A.
7. PCA transform is $y = A \times (x - m)$, where x is the input dataset, m is the mean, and A is the transpose of the feature vector.

$$\text{Original data } (f) = \{(A)^{-1} \times y\} + m$$
$$= \{(A)^{T} \times y\} + m$$

PCA Example

Example 2.12: Let the data points be $\begin{pmatrix}2\\6\end{pmatrix}$ and $\begin{pmatrix}1\\7\end{pmatrix}$. Apply PCA and find the transformed data.

Again, apply the inverse and prove that PCA works.

Solution: One can combine two vectors into a matrix as follows:

The mean vector can be computed as Eq. (2.53) as follows:

$$\mu = \begin{pmatrix}\frac{2+1}{2}\\ \frac{6+7}{2}\end{pmatrix} = \begin{pmatrix}1.5\\6.5\end{pmatrix}$$

PCA Example

As part of PCA, the mean must be subtracted from the data to get the adjusted data:

$$x_1 = \begin{pmatrix} 2-1.5 \\ 6-6.5 \end{pmatrix} = \begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix}$$

$$x_2 = \begin{pmatrix} 1-1.5 \\ 7-6.5 \end{pmatrix} = \begin{pmatrix} -0.5 \\ 0.5 \end{pmatrix}$$

One can find the covariance for these data vectors. The covariance can be obtained using Eq. (2.54):

$$m_1 = \begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix} \begin{pmatrix} 0.5 & -0.5 \end{pmatrix} = \begin{pmatrix} 0.25 & -0.25 \\ -0.25 & 0.25 \end{pmatrix}$$

$$m_2 = \begin{pmatrix} -0.5 \\ 0.5 \end{pmatrix} \begin{pmatrix} -0.5 & 0.5 \end{pmatrix} = \begin{pmatrix} 0.25 & -0.25 \\ -0.25 & 0.25 \end{pmatrix}$$

The final covariance matrix is obtained by adding these two matrices as:

$$C = \begin{pmatrix} 0.5 & -0.5 \\ -0.5 & 0.5 \end{pmatrix}$$

PCA Example

The eigen values and eigen vectors of matrix C can be obtained (left as an exercise) as $\lambda_1 = 1$, $\lambda_1 = 0$. The eigen vectors are $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$. The matrix A can be obtained by packing the eigen vector of these eigen values (after sorting it) of matrix C. For this problem, $A = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}$. The transpose of A, $A^T = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}$ is also the same matrix as it is an orthogonal matrix. The matrix can be normalized by diving each elements of the vector, by the norm of the vector to get:

$$A = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

PCA Example

The transformed matrix y using Eq. (2.55) is given as:

$$y = A \times (x - m)$$

Recollect that $(x-m)$ is the adjusted matrix.

$$y = A(x-m) = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 0.5 & -0.5 \\ -0.5 & 0.5 \end{pmatrix}$$

$$= \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \left(\text{for convenience } 0.5 = \frac{1}{2} \right)$$

$$= \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 0 \end{pmatrix}$$

PCA Example

The eigen values and eigen vectors of matrix C can be obtained (left as an exercise) as $\lambda_1 = 1$, $\lambda_1 = 0$. The eigen vectors are $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$. The matrix A can be obtained by packing the eigen vector of these eigen values (after sorting it) of matrix C. For this problem, $A = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}$. The transpose of A, $A^T = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}$ is also the same matrix as it is an orthogonal matrix. The matrix can be normalized by diving each elements of the vector, by the norm of the vector to get:

$$A = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

Verification

One can check the original matrix can be retrieved from this matrix as:

$$\{(A)^T \times y\} + m$$

$$x = A^T y + m = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 1.5 \\ 6.5 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} + \begin{pmatrix} 1.5 \\ 6.5 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 6 & 7 \end{pmatrix}$$

Therefore, one can infer the original is obtained without any loss of information.

LDA Algorithm

The aim of LDA is to optimize the function:

$$J(V) = \frac{V^T \sigma_B V}{V^T \sigma_W V}$$

$$\sigma_B = N_1(\mu_1 - \mu)(\mu_1 - \mu)^T + N_2(\mu_2 - \mu)(\mu_2 - \mu)^T$$

$$\sigma_W = \sum_{x_i \in c_1} (x_i - \mu_1)(x_i - \mu_1)^T + \sum_{x_i \in c_2} (x_i - \mu_2)(x_i - \mu_2)^T$$

LDA Algorithm

The maximization of $J(V)$ should satisfy the equation:

$$\sigma_B V = \lambda \sigma_W V \text{ or } \sigma_W^{-1} \sigma_B V = \lambda V \tag{2.63}$$

As $\sigma_B V$ is always in the direction of $(\mu_1 - \mu_2)$, V can be given as:

$$V = \sigma_W^{-1}(\mu_1 - \mu_2) \tag{2.64}$$

Let $V = \{v_1, v_2, \cdots, v_d\}$ be the generalized eigen vectors of σ_B and σ_W, where, d is the largest eigen values as in PCA. The transformation of x is then given as:

$$y = V_d^T x \tag{2.65}$$

Like in PCA, the largest eigen values can be retained to have projections.

SVD Algorithm

$$A = USV^T \tag{2.66}$$

Here, A is the given matrix of dimension $m \times n$, U is the orthogonal matrix whose dimension is $m \times n$, S is the diagonal matrix of dimension $n \times n$, and V is the orthogonal matrix. The procedure for finding decomposition matrix is given as follows:

eigen values & eigen vectors

SVD Algorithm

1. For a given matrix, find AA^T
2. Find eigen values of AA^T
3. Sort the eigen values in a descending order. Pack the eigen vectors as a matrix U.
4. Arrange the square root of the eigen values in diagonal. This matrix is diagonal matrix, S.
5. Find eigen values and eigen vectors for A^TA. Find the eigen value and pack the eigen vector as a matrix called V.

SVD Example

Example 2.13: Find SVD of the matrix:

$$A = \begin{pmatrix} 1 & 2 \\ 4 & 9 \end{pmatrix}$$

Solution: The first step is to compute:

$$AA^T = \begin{pmatrix} 1 & 2 \\ 4 & 9 \end{pmatrix}\begin{pmatrix} 1 & 4 \\ 2 & 9 \end{pmatrix} = \begin{pmatrix} 5 & 22 \\ 22 & 97 \end{pmatrix}$$

The eigen value and eigen vector of this matrix can be calculated to get U. The eigen values of this matrix are 0.0098 and 101.9902.

SVD Example

The eigen vectors of this matrix are:

$$u_1 = \begin{pmatrix} 0.2268 \\ 1 \end{pmatrix}$$

$$u_2 = \begin{pmatrix} -4.4086 \\ 1 \end{pmatrix}$$

These vectors are normalized to get the vectors respectively as:

$$u_1 = \begin{pmatrix} 0.2212 \\ 0.9752 \end{pmatrix}$$

$$u_2 = \begin{pmatrix} -0.9752 \\ 0.2212 \end{pmatrix}$$

The matrix U can be obtained by concatenating the above vector as:

$$U = [u_1, u_2] = \begin{pmatrix} 0.2212 & -0.9752 \\ 0.9752 & 0.2212 \end{pmatrix}$$

SVD Example

The matrix V can be obtained by finding A^TA. It is $\begin{pmatrix} 17 & 38 \\ 38 & 85 \end{pmatrix}$. The eigen values are 0.0098 and 101.9902. The eigen vectors can be found as follows:

$$v_1 = \begin{pmatrix} 0.447 \\ 1 \end{pmatrix} \text{ when } \lambda = 101.99$$

$$v_2 = \begin{pmatrix} -2.236 \\ 1 \end{pmatrix} \text{ when } \lambda = 0.0098$$

The above can be normalized as follows:

$$v_1 = \begin{pmatrix} 0.4082 \\ 0.9129 \end{pmatrix}$$

$$v_2 = \begin{pmatrix} -0.9129 \\ 0.4082 \end{pmatrix}$$

The matrix V can be obtained by concatenating the above vector as:

$$V = [v_1 \; v_2] = \begin{pmatrix} 0.4081 & -0.9129 \\ 0.9129 & 0.4082 \end{pmatrix}$$

SVD Example

The matrix S can be found as the diagonal matrix as:

$$S = \begin{pmatrix} \sqrt{101.9902} & 0 \\ 0 & \sqrt{0.0098} \end{pmatrix} = \begin{pmatrix} 10.099 & 0 \\ 0 & 0.099 \end{pmatrix}$$

Therefore, the matrix decomposition $A = U\,SV^T$ is complete.

Summary

1. Data is fact. Data must be converted to information.
2. Data analysis is an operation that converts data to information. Data analytics is a general area encompassing data analysis.
3. Exploratory Data analysis aims to understand the data better.
4. Descriptive statistics aims to summarize the data and data visualization aims to understand the data using charts.
5. Data types are qualitative and quantitative. The qualitative data types include nominal and ordinal. The quantitative data types include interval and ratio.
6. Data also can be classified as univariate data, bivariate data, and multivariate data based on the variables used.
7. The measures of the central tendency include mean, median, mode, and midrange.
8. The most common dispersion measures are range, 5-number summary, interquartile range and standard deviation.
9. Skewness and kurtosis are shape measures. Skewness, kurtosis, mean absolute deviation, and coefficient of variation help in assessing the shape.

Summary

10. Visualization is an important aspect of data mining which helps the user to recognize and interpret the results quickly.
11. Bivariate data involves two variables and hence the focus is on finding relationships among two variables.
12. Multivariate data has more variables. The focus is on finding the relations, cause and effects, and statistical inference.
13. Covariance is a measure of joint probability of random variables, say x and y. It is defined as $cov(X, Y)$. Covariance is used to measure variance between two dimensions. Correlation measures the strength of linear relationship among variables.
14. The idea of feature reduction is to transform a given set of measurements to a new set of features, so that the features exhibit high information packing properties. This leads to a reduced and compact set of features.
15. Sampling is a statistical procedure of data reduction Statistics starts by identifying the group for carrying out a study.
16. The assumption of an experiment is called hypothesis. Statistical methods are used to confirm or reject the hypothesis.
17. Z-test, t-tests, and Chi-Square tests are used to evaluate the hypothesis.