R.M.K
GROUP OF ENGINEERING INSTITUTIONS

R.M.K GROUP OF INSTITUTIONS

R.M.K
GROUP OF
INSTITUTIONS

R.M.K GROUP OF INSTITUTIONS

Please read this disclaimer before proceeding:

This document is confidential and intended solely for the educational purpose of RMK Group of Educational Institutions. If you have received this document through email in error, please notify the system manager. This document contains proprietary information and is intended only to the respective group / learning community as intended. If you are not the addressee you should not disseminate, distribute or copy through e-mail. Please notify the sender immediately by e-mail if you have received this document by mistake and delete this document from your system. If you are not the intended recipient you are notified that disclosing, copying, distributing or taking any action in reliance on the contents of this information is strictly prohibited.

21AI502 MACHINE LEARNING

Department : CSE

Batch/Year: 2021-2025/III

Created by: Dr. P. VALARMATHIE

Dr. M. RAJA SUGUNA

Date: 06.01.2024

R.M.K GROUP OF INSTITUTIONS

1. TABLE OF CONTENTS

Course Objectives

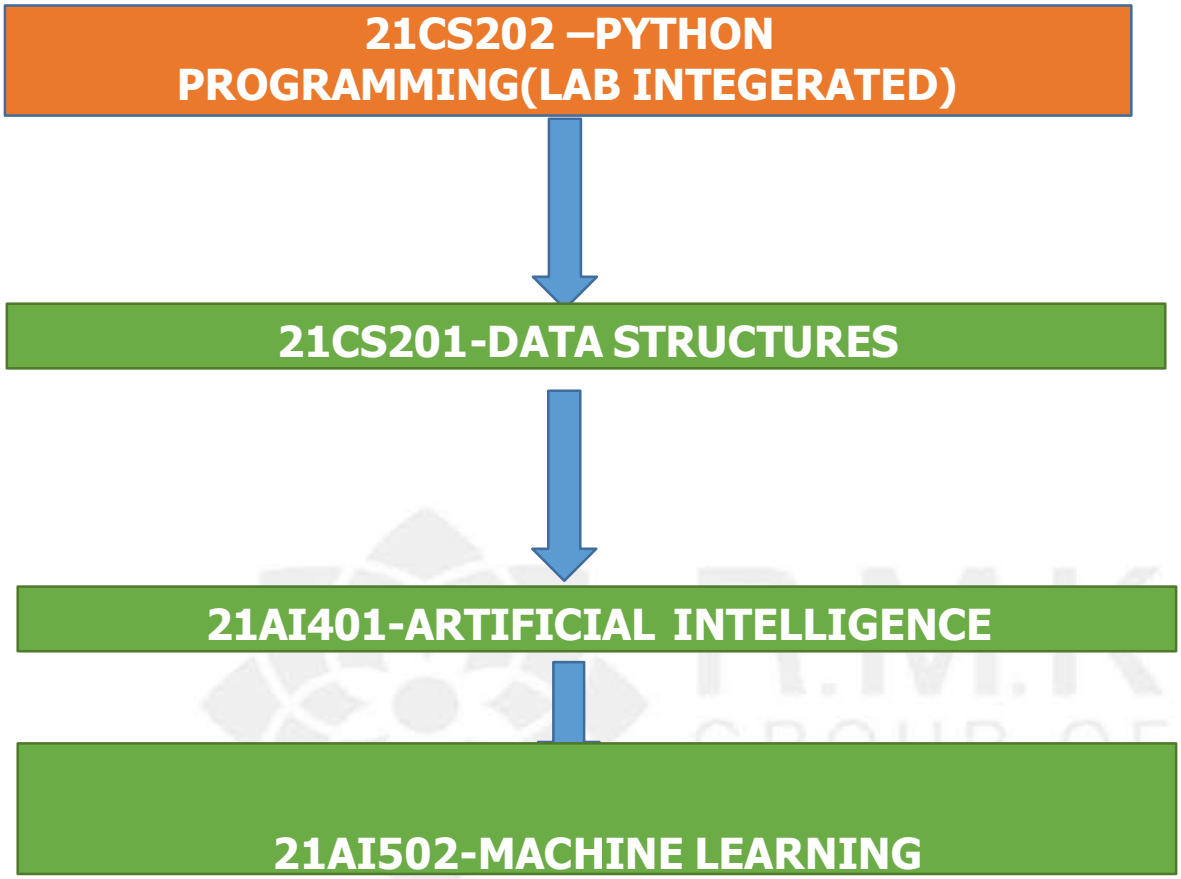
2. COURSE OBJECTIVES

- To discuss the basics of Machine Learning and Supervised Algorithms.
- To understand the various classification algorithms.
- To study dimensionality reduction techniques.
- To elaborate on unsupervised learning techniques.
- To discuss various Graphical models and understand thebasics of reinforcement learning.

PRE REQUISITES

3. PRE REQUISITES

PRE-REQUISITE CHART



Syllabus

R.M.K GROUP OF INSTITUTIONS

4. SYLLABUS

21AI502-MACHINE LEARNING

L T P C

3 0 0 3

UNIT-I INTRODUCTION 9

Machine Learning – Types – Applications – Preparing to Model – Activities – Data – Exploring structure of Data – Data Quality and Remediation – Data Pre-processing – Modelling and Evaluation: Selecting a Model -Training a Model – Model representation and Interpretability – Evaluating Performance of a Model – Improving Performance.

UNIT-II FEATURE ENGINEERING & DIMENSIONALITY REDUCTION 9

Feature Engineering – Feature Transformation – Feature Subset Selection - Principle Component Analysis – Feature Embedding – Factor Analysis – Singular value decomposition and Matrix Factorization – Multidimensional scaling – Linear Discriminant Analysis – Canonical Correlation Analysis – Isomap – Locally linear Embedding – Laplacian Eigenmaps.

Unit-III SUPERVISED LEARNING 9

Linear Regression -Relation between two variables – Steps – Evaluation – Logistic Regression – Decision Tree – Algorithms – Construction – Classification using Decision Tree – Issues – Rule-based Classification – Pruning the Rule Set – Support Vector Machines – Linear SVM – Optimal Hyperplane – Radial Basis Functions – Naïve Bayes Classifier – Bayesian Belief Networks.

UNIT-IV UNSUPERVISED LEARNING 9

Clustering – Types – Applications - Partitioning Methods – K-means Algorithm – K-Medoids – Hierarchical methods – Density based methods DBSCAN – Finding patterns using Association Rules – Hidden Markov Model.

UNIT-V NEURAL NETWORKS AND TYPES OF LEARNING 9

Biological Neuron – Artificial Neuron – Types of Activation function – Implementations of ANN – Architectures of Neural Networks – Learning Process in ANN – Back propagation – Deep Learning – Representation Learning – Active Learning – Instance based Learning – Association Rule Learning – Ensemble Learning Algorithm – Regularization Algorithm-Reinforcement Learning – Elements- Model-based- Temporal Difference Learning.

Course Outcomes

5. COURSE OUTCOME

Course Code	Course Outcome Statement	Cognitive / Affective Level of the Course Outcome	Course Outcome
Course Outcome Statements in Cognitive Domain			
21AI502	Explain the basics of Machine Learning and Supervised Algorithms	Apply K3	CO1
21AI502	Understand the various classification algorithms.	Apply K3	CO2
21AI502	Study dimensionality reduction techniques	Apply K3	CO3
21AI502	Elaborate on unsupervised learning techniques	Apply K4	CO4
21AI502	Understand various Graphical models and understand the basics of reinforcement learning	Apply K4	CO5

CO – PO/PSO Mapping

R.M.K GROUP OF INSTITUTIONS

6. CO-PO/PSO MAPPING

Correlation Matrix of the Course Outcomes to Programme Outcomes and Programme Specific Outcomes .

Course Outcomes (Cos)		Programme Outcomes (POs), Programme Specific Outcomes (PSOs)														
		PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
21AI502.1	K2	3	3	1	-	-	-	-	-	-	-	-	-	2	2	2
21AI502.2	K3	3	2	1	-	-	-	-	-	-	-	-	-	2	2	2
21AI502.3	K3	3	2	1	-	-	-	-	-	-	-	-	-	2	2	2
21AI502.4	K3	3	3	2	-	-	-	-	-	-	-	-	-	2	2	2
21AI502.5	K2	3	2	2	-	-	-	-	-	-	-	-	-	2	2	2

Lecture Plan

Unit I

LECTURE PLAN – UNIT I

	UNIT I INTRODUCTION						
Sl. No	TOPIC	NO OF PERIODS	PROPOSED LECTURE PEROID	ACTUAL LECTURE PERIOD	PERTAINING CO(s)	TAXONOMY LEVEL	MODE OF DELIVERY
1	Introduction to ML	1	03-01-2024		CO1	K2	MD1 , MD5
2	Preparing the model activities, data, exploring structure of data	1	04-01-2024		CO1	K1	MD1 , MD5
3	Data quality & remediation	1	05-01-2024		CO1	K2	MD1 , MD5
4	Data pre-processing	1	06-01-2024		CO1	K2	MD1 , MD5
5	Modelling & Evaluation, Selecting a model	1	08-01-2024		CO1	K2	MD1 , MD5
6	Training a model	1	09-01-2024		CO1	K2	MD1 , MD5
7	Model representation & interpretability	1	11-01-2024		CO1	K5	MD1 , MD5
8	Evaluating Performance of a model	1	12-01-2024		CO1	K2	MD1 , MD5
9	Improving Performance	1	22-01-2024		CO1	K2	MD1 , MD5

LECTURE PLAN – UNIT I

ASSESSMENT COMPONENTS

- AC 1. Unit Test
- AC 2. Assignment
- AC 3. Course Seminar
- AC 4. Course Quiz
- AC 5. Case Study
- AC 6. Record Work
- AC 7. Lab / Mini Project
- AC 8. Lab Model Exam
- AC 9. Project Review

MODE OF DELEIVERY

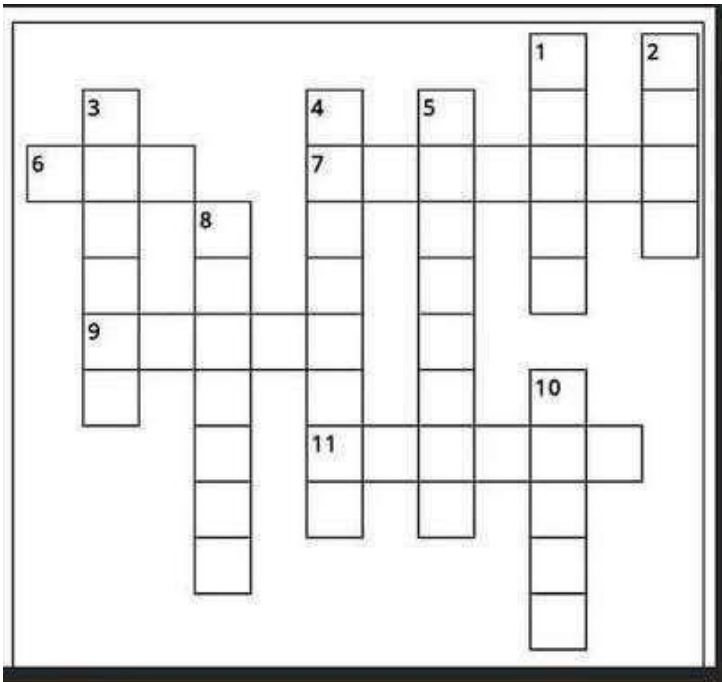
MD 1. Oral presentation
MD 2. Tutorial
MD 3. Seminar
MD 4 Hands On
MD 5. Videos
MD 6. Field Visit

Activity Based Learning

Unit I

8. ACTIVITY BASED LEARNING : UNIT – I

ACTIVITY 1: Cross Word puzzle



ACROSS

6. DEDICATED SILICON THAT IS DESIGNED FOR PROCESSING COMPUTER VISION MEDIA INCLUDING IMAGES AND VIDEO

7. A REGULARIZATION TECHNIQUE FOR NEURAL NETWORKS THAT PREVENTS OVERFITTING

9. A DEEP LEARNING FRAMEWORK DEVELOPED BY BERKELEY AI RESEARCH

11. PYTHON LIBRARY THAT ALLOWS YOU TO DEFINE, OPTIMIZE, AND EVALUATE MATHEMATICAL EXPRESSIONS

DOWN

1. AN UNSUPERVISED LEARNING ALGORITHM FOR OBTAINING VECTOR REPRESENTATIONS (EMBEDDINGS) FOR WORDS

2. INVENTED TO PREVENT THE VANISHING GRADIENT PROBLEM IN RECURRENT NEURAL NETWORKS BY USING A MEMORY GATING MECHANISM

3. A SOFTWARE LIBRARY THAT SPECIALIZES IN REAL-TIME COMPUTER VISION ALGORITHMS

4. A GRADIENT DESCENT BASED LEARNING ALGORITHM THAT ADAPTS THE LEARNING RATE PER PARAMETER OVER TIME

5. AN ALGORITHM AND TOOL TO LEARN WORD EMBEDDINGS BY TRYING TO PREDICT THE CONTEXT OF WORDS IN A DOCUMENT

8. USED TO CONVERT A VECTOR OF RAW SCORES INTO CLASS PROBABILITIES AT THE OUTPUT LAYER OF A NEURAL NETWORK USED FOR CLASSIFICATION

10. PERHAPS MOST COMMONLY USED IMAGE RECOGNITION DATASET

9. LECTURE NOTES : UNIT – I

INTRODUCTION

Syllabus:

Machine Learning – Types – Applications – Preparing to Model – Activities – Data – Exploring structure of Data – Data Quality and Remediation – Data Pre-processing – Modelling and Evaluation: Selecting a Model -Training a Model – Model representation and Interpretability – Evaluating Performance of a Model – Improving Performance.

Lecture Notes – Unit 1

R.M.K
GROUP OF
INSTITUTIONS

UNIT 1: INTRODUCTION

Introduction:

In the real world, we are surrounded by humans who can learn everything from their experiences with their learning capability, and we have computers or machines which work on our instructions. But can a machine also learn from experiences or past data like a human does? So here comes the role of **Machine Learning**.

Machine Learning is said as a subset of **artificial intelligence** that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own. The term machine learning was first introduced by **Arthur Samuel** in **1959**. We can define it in a summarized way as:

Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed.

With the help of sample historical data, which is known as **training data**, machine learning algorithms build a **mathematical model** that helps in making predictions or decisions without being explicitly programmed. Machine learning brings computer science and statistics together for creating predictive models. Machine learning constructs or uses the algorithms that learn from historical data. The more we will provide the information, the higher will be the performance.

A machine has the ability to learn if it can improve its performance by gaining more data.

FIG 1.1 Introduction

How does Machine Learning work?

A Machine Learning system **learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it**. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.

Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output. Machine learning has changed our way of thinking about the problem. The below block diagram explains the working of Machine Learning algorithm:

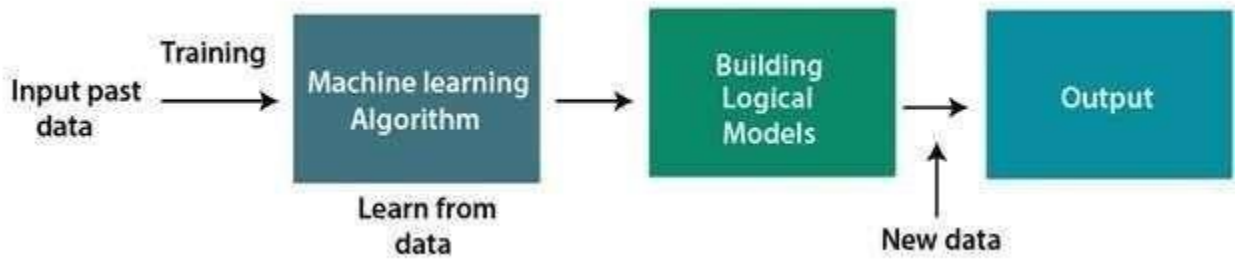


FIG 1.2 How it works

Features of Machine Learning:

- Machine learning uses data to detect various patterns in a given dataset.
- It can learn from past data and improve automatically.
- It is a data-driven technology.
- Machine learning is much similar to data mining as it also deals with the huge amount of the data.

1.1 TYPES OF MACHINE LEARNING

Machine learning can be classified into three broad categories:

Supervised learning – Also called predictive learning. A machine predicts the

class of unknown objects based on prior class- related information of similar objects.

Unsupervised learning – Also called descriptive learning. A machine finds patterns in unknown objects by grouping similar objects together.

Reinforcement learning – A machine learns to act on its own to achieve the given goals.

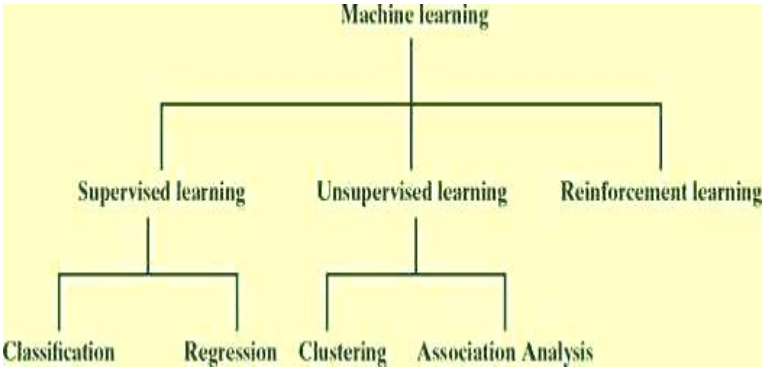


FIG. 1.3 Types of machine learning

1.1.1 Supervised learning

The major motivation of supervised learning is to learn from past information. So what kind of past information does the machine need for supervised learning? It is the information about the task which the machine has to execute. In context of the definition of machine learning, this past information is the experience. Let's try to understand it with an example.

Say a machine is getting images of different objects as input and the task is to segregate the images by either shape or colour of the object. If it is by shape, the images which are of round-shaped objects need to be separated from images of triangular-shaped objects, etc. If the segregation needs to happen based on colour, images of blue objects need to be separated from images of green objects. But how can the machine know what is round shape, or triangular shape? Same way, how can the machine distinguish image of an object based on whether it is blue or green in colour? A machine is very much likea little child whose parents or adults need to guide him with the basic information on shape and colour before he can start doing the task. A machine needs the basic information to be provided to it. This basic input, or the experience in the paradigm of machine learning, is given in the form of training data . Training data is the past information on a specific task. In context of the image segregation problem,

training data will have past data on different aspects or features on a number of images, along with a tag on whether the image is round or triangular, or blue or green in colour. The tag is called ' label' and we say that the training data is labelled in case of supervised learning.

Figure 1.4 is a simple depiction of the supervised learning process. Labelled training data containing past information comes as an input. Based on the training data, the machine builds a predictive model that can be used on test data to assign a label for each record in the test data.

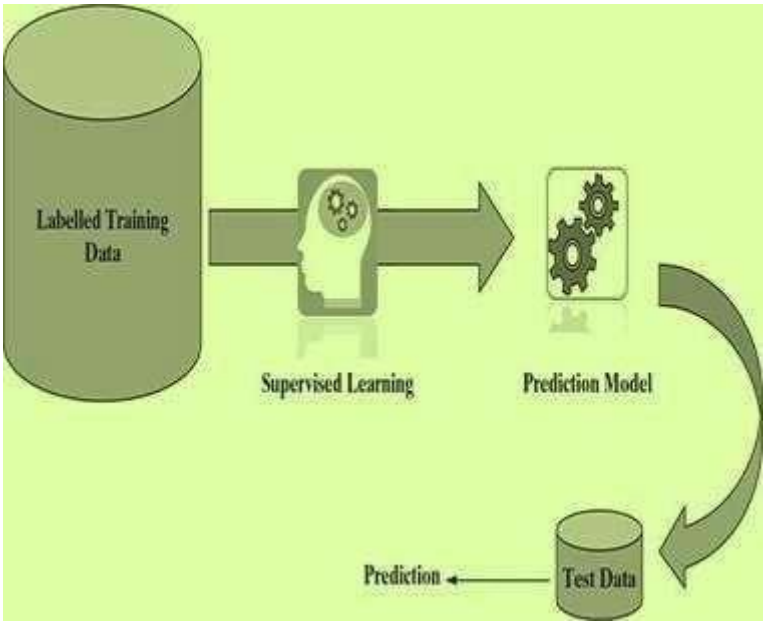


FIG. 1.4 Supervised learning

Some examples of supervised learning are

- Predicting the results of a game
- Predicting whether a tumour is malignant or benign Predicting the price of domains like real estate, stocks, etc.
- Classifying texts such as classifying a set of emails as spam or non- spam

Now, let's consider two of the above examples, say 'predicting whether a tumour is malignant or benign' and 'predicting price of domains such as real estate'. Are these two

problems same in nature? The answer is 'no'.Though both of them are prediction problems, in one case we are trying to predict which category or class an unknown data belongs to whereas in the other case we are trying to predict an absolute value and not a class. When we are trying to predict a categorical or nominal variable, the problem is known as a classification problem. Whereas when we are trying to predict a real- valued variable, the problem falls under the category of regression.

Let's try to understand these two areas of supervised learning, i.e. classification and regression in more details.

1.1.1.1 Classification

Let's discuss how to segregate the images of objects based on the shape . If the image is of a round object, it is put under one category, while if the image is of a triangular object, it is put under another category. In which category the machine should put an image of unknown category, also called a test data in machine learning parlance, depends on the information it gets from the past data, which we have called as training data. Since the training data has a label or category defined for each and every image, the machine has to map a new image or test data to a set of images to which it is similar to and assign the same label or category to the test data. So we observe that the whole problem revolves around assigning a label or category or class to a test data based on the label or category or class information that is imparted by the training data. Since the target objective is to assign a class label, this type of problem as classification problem. Figure 1.5 depicts the typical process of classification.

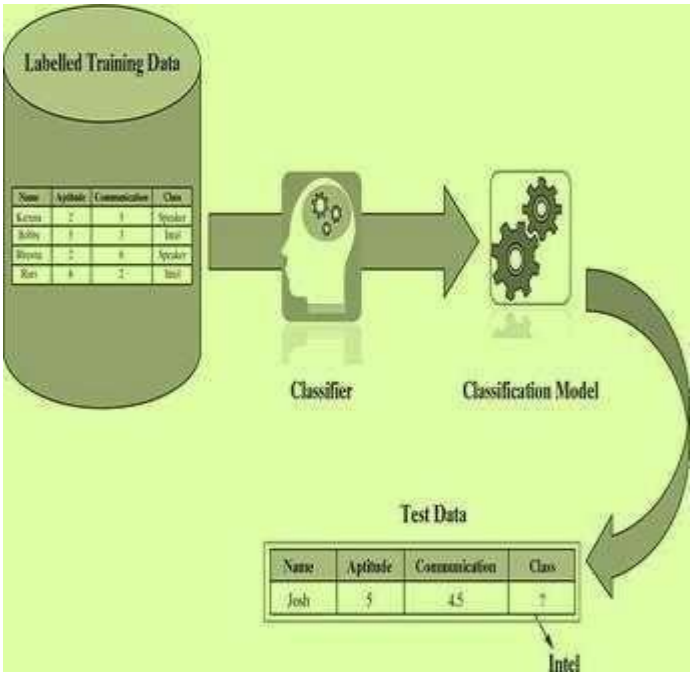


FIG. 1.5 Classification

There are number of popular machine learning algorithms which help in solving classification problems. To name a few, Naïve Bayes, Decision tree, and k- Nearest Neighbour algorithms are adopted by many machine learning practitioners.

A critical classification problem in context of banking domain is identifying potential fraudulent transactions. Since there are millions of transactions which have to be scrutinized and assured whether it might be a fraud transaction, it is not possible for any human being to carry out this task. Machine learning is effectively leveraged to do this task and this is a classic case of classification. Based on the past transaction data, specifically the ones labelled as fraudulent, all new incoming transactions are marked or labelled as normal or suspicious. The suspicious transactions are subsequently segregated for a closer review.

In summary, classification is a type of supervised learning where a target feature, which is of type categorical, is predicted for test data based on the information imparted by training data. The target categorical feature is known as class.

Some typical classification problems include:

1. Image classification Prediction of disease
2. Win–loss prediction of games Prediction of natural calamity like earthquake, flood, etc.
3. Recognition of handwriting

1.1.1.2. Regression

In linear regression, the objective is to predict numerical features like real estate or stock price, temperature, marks in an examination, sales revenue, etc. The underlying predictor variable and the target variable are continuous in nature. In case of linear regression, a straight-line relationship is 'fitted' between the predictor variables and the target variables, using the statistical concept of least squares method. As in the case of least squares method, the sum of square of error between actual and predicted values of the target variable is tried to be minimized. In case of simple linear regression, there is only one predictor variable whereas in case of multiple linear regression, multiple predictor variables can be included in the model.

Let's take the example of yearly budgeting exercise of the sales managers. They have to give sales prediction for the next year based on sales figure of previous years vis- à-vis investment being put in. Obviously, the data related to past as well as the data to be

predicted are continuous in nature. In a basic approach, a simple linear regression model can be applied with investment as predictor variable and sales revenue as the target variable.

Figure 1.6 shows a typical simple regression model, where regression line is fitted based on values of target variable with respect to different values of predictor variable. A typical linear regression model can be represented in the form –

$$y = \alpha + \beta x$$

where 'x' is the predictor variable and 'y' is the target variable.

The input data come from a famous multivariate data set named Iris introduced by the British statistician and biologist Ronald Fisher. The data set consists of 50 samples from each of three species of Iris – Iris setosa, Iris virginica, and Iris versicolor. Four features were measured for each sample – sepal length, sepal width, petal length, and petal width. These features can uniquely discriminate the different species of the flower.

The Iris data set is typically used as a training data for solving the classification problem of predicting the flower species based on feature values. However, we can also demonstrate regression using this data set, by predicting the value of one feature using another feature as predictor. In Figure 1.6, petal length is a predictor variable which, when fitted in the simple linear regression model, helps in predicting the value of the target variable sepal length.

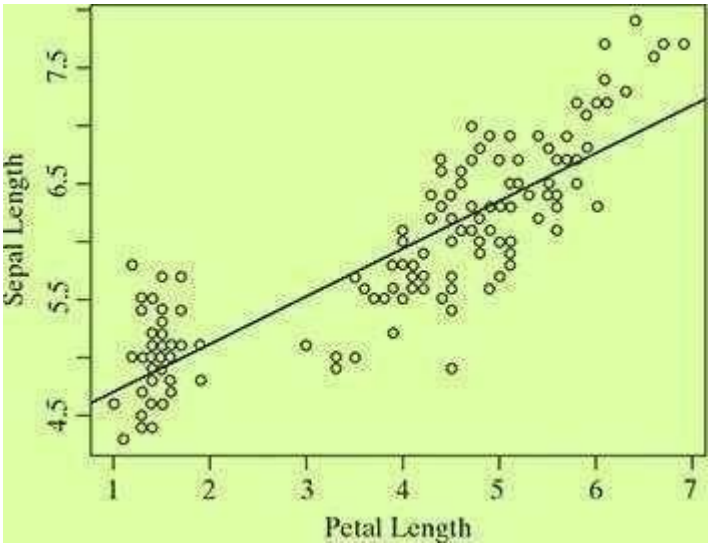


FIG. 1.6 Regression

Typical applications of regression can be seen in

- Demand forecasting in retails Sales prediction for managers
- Price prediction in real estate Weather forecast
- Skill demand forecast in job market

1.1.2 Unsupervised learning

Unlike supervised learning, in unsupervised learning, there is no labelled training data to learn from and no prediction to be made. In unsupervised learning, the objective is to take a dataset as input and try to find natural groupings or patterns within the data elements or records. Therefore, unsupervised learning is often termed as descriptive model and the process of unsupervised learning is referred as pattern discovery or knowledge discovery. One critical application of unsupervised learning is customer segmentation.

Clustering is the main type of unsupervised learning. It intends to group or organize similar objects together. For that reason, objects belonging to the same cluster are quite similar to each other while objects belonging to different clusters are quite dissimilar. Hence, the objective of clustering to discover the intrinsic grouping of unlabelled data and form clusters, as depicted in Figure 1.7. Different measures of similarity can be applied for clustering. One of the most commonly adopted similarity measure is distance. Two data items are considered as a part of the same cluster if the distance between them is less. In the same way, if the distance between the data items is high, the items do not generally belong to the same cluster. This is also known as distance-based clustering. Figure 1.8 depicts the process of clustering at a high level.

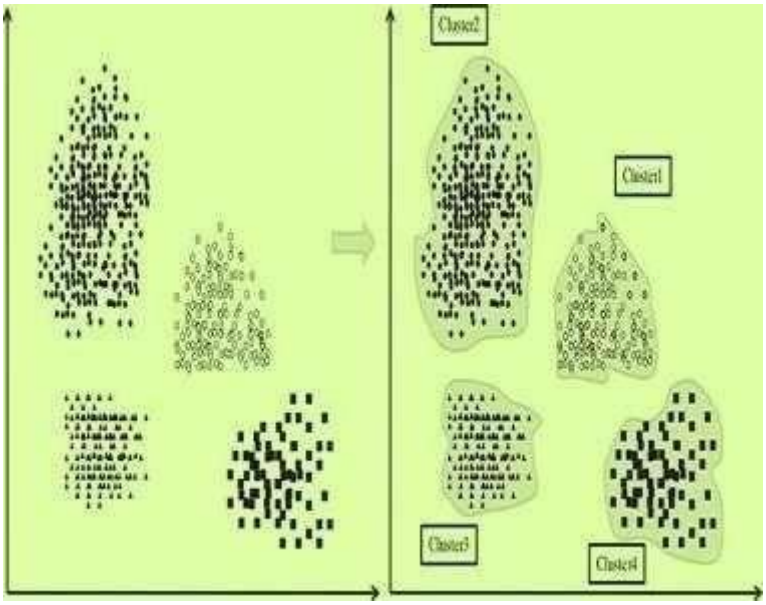


FIG. 1.7 Distance-based clustering

Other than clustering of data and getting a summarized view from it, one more variant of unsupervised learning is association analysis. As a part of association analysis, the association between data elements is identified.

Let's try to understand the approach of association analysis in context of one of the most common examples, i.e. market basket analysis as shown in Figure 1.9. From past transaction data in a grocery store, it may be observed that most of the customers who have bought item A, have also bought item B and item C or at least one of them. This means that there is a strong association of the event 'purchase of item A' with the event 'purchase of item B', or 'purchase of item C'. Identifying these sorts of associations is the goal of association analysis. This helps in boosting up sales pipeline, hence a critical input for the sales group.

Critical applications of association analysis include market basket analysis and recommender systems.

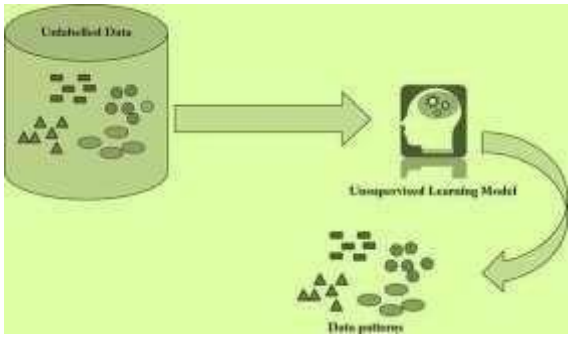


FIG. 1.8 Unsupervised learning

TransID	Items Bought
1	{Butter, Bread}
2	{Diaper, Bread, Milk, Beer}
3	{Milk, Chicken, Beer, Diaper}
4	{Bread, Diaper, Chicken, Beer}
5	{Diaper, Beer, Cookies, Ice cream}
...	...

Market Basket transactions
Frequent itemsets → {Diaper, Beer}
Possible association: Diaper → Beer

FIG. 1.9 Market basket analysis

1.1.3 Reinforcement learning

We have seen babies learn to walk without any prior knowledge of how to do it. Often we wonder how they really do it. They do it in a relatively simple way.

First they notice somebody else walking around, for example parents or anyone living around. They understand that legs have to be used, one at a time, to take a step. While walking, sometimes they fall down hitting an obstacle, whereas other times they are able to walk smoothly avoiding bumpy obstacles. When they are able to walk overcoming the obstacle, their parents are elated and appreciate the baby with loud claps / or may be a chocolates. When they fall down while circumventing an obstacle, obviously their parents do not give claps or chocolates. Slowly a time comes when the babies learn from mistakes and are able to walk with much ease.

In the same way, machines often learn to do tasks autonomously. Let's try to understand in context of the example of the child learning to walk. The action tried to be

achieved is walking, the child is the agent and the place with hurdles on which the child is trying to walk resembles the environment.

It tries to improve its performance of doing the task. When a sub-task is accomplished successfully, a reward is given. When a sub-task is not executed correctly, obviously no reward is given. This continues till the machine is able to complete execution of the whole task. This process of learning is known as reinforcement learning. Figure 1.10 captures the high-level process of reinforcement learning.

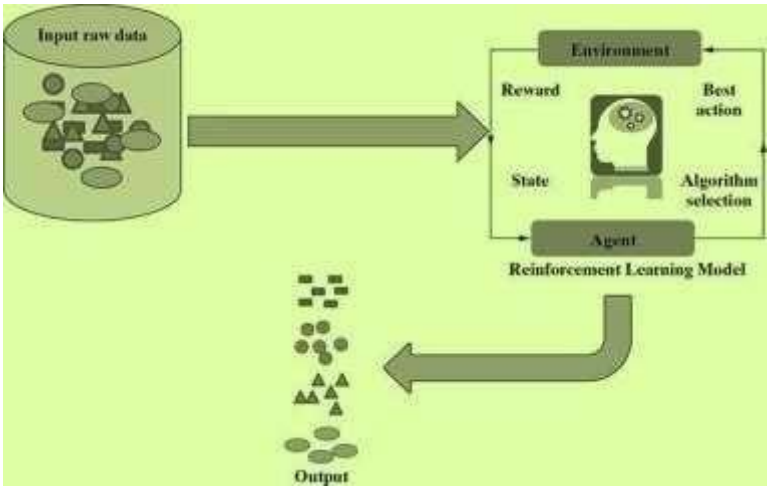


FIG. 1.10 Reinforcement learning

One contemporary example of reinforcement learning is self-driving cars. The critical information which it needs to take care of are speed and speed limit in different road segments, traffic conditions, road conditions, weather conditions, etc. The tasks that have to be taken care of are start/stop, accelerate/decelerate, turn to left / right.

2. APPLICATIONS OF MACHINE LEARNING

Wherever there is a substantial amount of past data, machine learning can be used to generate actionable insight from the data. Though machine learning is adopted in multiple forms in every business domain, we have covered below three major domains just to give some idea about what type of actions can be done using machine learning.

1. Banking and finance

In the banking industry, fraudulent transactions, especially the ones related to credit cards, are extremely prevalent. Since the volumes as well as velocity of the transactions

are extremely high, high performance machine learning solutions are implemented by almost all leading banks across the globe. The models work on a real-time basis, i.e. the fraudulent transactions are spotted and prevented right at the time of occurrence.
This helps in avoiding a lot of operational hassles in settling the disputes that customers will otherwise raise against those fraudulent transactions.

Customers of a bank are often offered lucrative proposals by other competitor banks. Proposals like higher bank interest, lower processing charge of loans, zero balance savings accounts, no overdraft penalty, etc. are offered to customers, with the intent that the customer switches over to the competitor bank. Also, sometimes customers get demotivated by the poor quality of services of the banks and shift to competitor banks. Machine learning helps in preventing or at least reducing the customer churn. Both descriptive and predictive learning can be applied for reducing customer churn. Using descriptive learning, the specific pockets of problem, i.e. a specific bank or a specific zone or a specific type of offering like car loan, may be spotted where maximum churn is happening. Quite obviously, these are troubled areas where further investigation needs to be done to find and fix the root cause. Using predictive learning, the set of vulnerable customers who may leave the bank very soon, can be identified. Proper action can be taken to make sure that the customers stay back.

2. Insurance

Insurance industry is extremely data intensive. For that reason, machine learning is extensively used in the insurance industry. Two major areas in the insurance industry where machine learning is used are risk prediction during new customer onboarding and claims management. During customer onboarding, based on the past information the risk profile of a new customer needs to be predicted. Based on the quantum of risk predicted, the quote is generated for the prospective customer.
When a customer claim comes for settlement, past information related to historic claims along with the adjustor notes are considered to predict whether there is any possibility of the claim to be fraudulent. Other than the past information related to the specific customer, information related to similar customers, i.e. customer belonging to the same geographical location, age group, ethnic group, etc., are also considered to formulate the model.

3. Healthcare

Wearable device data form a rich source for applying machine learning and predict the health conditions of the person real time. In case there is some health issue which is predicted by the learning model, immediately the person is alerted to take preventive

action. In case of some extreme problem, doctors or healthcare providers in the vicinity of the person can be alerted. Suppose an elderly person goes for a morning walk in a park close to his house. Suddenly, while walking, his blood pressure shoots up beyond a certain limit, which is tracked by the wearable. The wearable data is sent to a remote server and a machine learning algorithm is constantly analyzing the streaming data. It also has the history of the elderly person and persons of similar age group. The model predicts some fatality unless immediate action is taken. Alert can be sent to the person to immediately stop walking and take rest. Also, doctors and healthcare providers can be alerted to be on standby.

Machine learning along with computer vision also plays a crucial role in disease diagnosis from medical imaging.

3. Preparing to Model

1. MACHINE LEARNING ACTIVITIES

The first step in machine learning activity starts with data. In case of supervised learning, it is the labelled training data set followed by test data which is not labelled. In case of unsupervised learning, there is no question of labelled data but the task is to find patterns in the input data. A thorough review and exploration of the data is needed to understand the type of the data, the quality of the data and relationship between the different data elements. Based on that, multiple pre-processing activities may need to be done on the input data before we can go ahead with core machine learning activities.

Following are the typical preparation activities done once the input data comes into the machine learning system:

- Understand the type of data in the given input data set. Explore the data to understand the nature and quality.
- Explore the relationships amongst the data elements, e.g. inter- feature relationship.
- Find potential issues in data.
- Do the necessary remediation, e.g. impute missing data values, etc., if needed.
- Apply pre-processing steps, as necessary.
- Once the data is prepared for modelling, then the learning tasks start off. As a part of it, do the following activities:
- The input data is first divided into parts – the training data and the test data (called holdout). This step is applicable for supervised learning only.
- Consider different models or learning algorithms for selection.

- Train the model based on the training data for supervised learning problem and apply to unknown data. Directly apply the chosen unsupervised model on the input data for unsupervised learning problem.

After the model is selected, trained (for supervised learning), and applied on input data, the performance of the model is evaluated. Based on options available, specific actions can be taken to improve the performance of the model, if possible.

Figure 1.11 depicts the four-step process of machine learning.

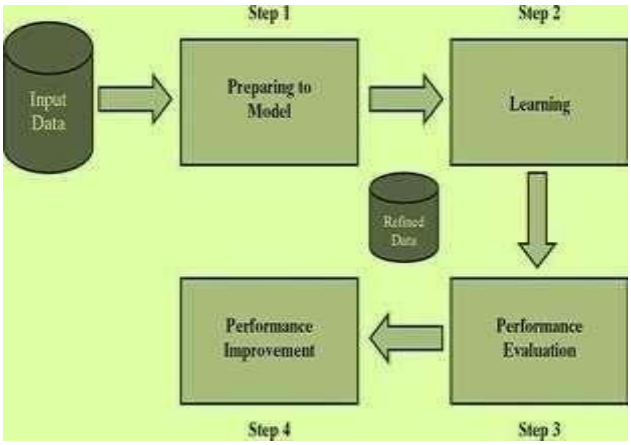


FIG. 1.11 Detailed process of machine learning

Table 1.1 contains a summary of steps and activities involved:

Step #	Step Name	Activities Involved
Step 1	Preparing to Model	• Understand the type of data in the given input data set • Explore the data to understand data quality • Explore the relationships amongst the data elements, e.g. inter-feature relationship • Find potential issues in data • Remediate data, if needed • Apply following pre-processing steps, as necessary: ✓ Dimensionality reduction ✓ Feature subset selection
Step 2	Learning	• Data partitioning/holdout • Model selection • Cross-validation
Step 3	Performance evaluation	• Examine the model performance, e.g. confusion matrix in case of classification • Visualize performance trade-offs using ROC curves
Step 4	Performance improvement	• Tuning the model • Ensembling • Bagging • Boosting

1.3.2 BASIC TYPES OF DATA IN MACHINE LEARNING

Before starting with types of data, let's first understand what a data set is and what are the elements of a data set. A data set is a collection of related information or records. The information may be on some entity or some subject area. For example (FIG 1.11), we may have a data set on students in which each record consists of information about a specific student. Again, we can have a data set on student performance which has records providing performance, i.e. marks on the individual subjects.

Each row of a data set is called a record. Each data set also has multiple attributes, each of which gives information on a specific characteristic. For example, in the data set on students, there are four attributes namely Roll Number, Name, Gender, and Age, each of which understandably is a specific characteristic about the student entity. Attributes can also be termed as feature, variable, dimension or field. Both the data sets, Student and Student Performance, are having four features or dimensions; hence they are told to have four-dimensional data space. A row or record represents a point in the four-dimensional data space as each row has specific values for each of the four attributes or features. Value of an attribute, quite understandably, may vary from record to record.

For example, if we refer to the first two records in the Student data set, the value of attributes Name, Gender, and Age are different (Fig.1.12).

Student data set:

Roll Number	Name	Gender	Age
129/011	Mihir Karmarkar	M	14
129/012	Geeta Iyer	F	15
129/013	Chanda Bose	F	14
129/014	Sreenu Subramanian	M	14
129/015	Pallav Gupta	M	16
129/016	Gajanan Sharma	M	15

Student performance data set:

Roll Number	Maths	Science	Percentage
129/011	89	45	89.33%
129/012	89	47	90.67%
129/013	68	29	64.67%
129/014	83	38	80.67%
129/015	57	23	53.33%
129/016	78	35	75.33%

FIG. 1.11 Examples of data set

Roll Number	Name	Gender	Age
129/011	Mihir Karmarkar	M	14
129/012	Geeta Iyer	F	15

FIG. 1.12 Data set records and attributes

Now that a context of data sets is given, let's try to understand the different types of data that we generally come across in machine learning problems. Data can broadly be divided into following two types:

- **Qualitative data**
- **Quantitative data**

Qualitative data provides information about the quality of an object or information which cannot be measured. For example, if we consider the quality of performance of students in terms of 'Good', 'Average', and 'Poor', it falls under the category of qualitative data. Also, name or roll number of students are information that cannot be measured using some scale of measurement. So they would fall under qualitative data. Qualitative data is also called categorical data.

Qualitative data can be further subdivided into two types as follows:

1. **Nominal data**
2. **Ordinal data**

Nominal data is one which has no numeric value, but a named value. It is used for assigning named values to attributes. Nominal values cannot be quantified.

Examples of nominal data are

1. Blood group: A, B, O, AB, etc.
2. Nationality: Indian, American, British, etc.
3. Gender: Male, Female, Other

It is obvious, mathematical operations such as addition, subtraction, multiplication, etc. cannot be performed on nominal data. For that reason, statistical functions such as mean, variance, etc. can also not be applied on nominal data. However, a basic count is possible. So mode, i.e. most frequently occurring value, can be identified for nominal data.

Ordinal data, in addition to possessing the properties of nominal data, can also be naturally ordered. This means ordinal data also assigns named values to attributes but unlike nominal data, they can be arranged in a sequence of increasing or decreasing value so that we can say whether a value is better than or greater than another value. Examples of ordinal data are

- Customer satisfaction: 'Very Happy', 'Happy', 'Unhappy', etc.
- Grades: A, B, C, etc.

- Hardness of Metal: 'Very Hard', 'Hard', 'Soft', etc.

Like nominal data, basic counting is possible for ordinal data. Hence, the mode can be identified. Since ordering is possible in case of ordinal data, median, and quartiles can be identified in addition. Mean can still not be calculated.

Quantitative data relates to information about the quantity of an object – hence it can be measured. For example, if we consider the attribute 'marks', it can be measured using a scale of measurement. Quantitative data is also termed as numeric data. There are two types of quantitative data:

- **Interval data**
- **Ratio data**

Interval data is numeric data for which not only the order is known, but the exact difference between values is also known. An ideal example of interval data is Celsius temperature. The difference between each value remains the same in Celsius temperature. For example, the difference between 12°C and 18°C degrees is measurable and is 6°C as in the case of difference between 15.5°C and 21.5°C. Other examples include date, time, etc.

For interval data, mathematical operations such as addition and subtraction are possible. For that reason, for interval data, the central tendency can be measured by mean, median, or mode. Standard deviation can also be calculated.

However, interval data do not have something called a 'true zero' value. For example, there is nothing called '0 temperature' or 'no temperature'. Hence, only addition and subtraction applies for interval data. The ratio cannot be applied. This means, we can say a temperature of 40°C is equal to the temperature of 20°C + temperature of 20°C. However, we cannot say the temperature of 40°C means it is twice as hot as in temperature of 20°C.

Ratio data represents numeric data for which exact value can be measured. Absolute zero is available for ratio data. Also, these variables can be added, subtracted, multiplied, or divided. The central tendency can be measured by mean, median, or mode and methods of dispersion such as standard deviation. Examples of ratio data include height, weight, age, salary, etc.
Figure 1.13 gives a summarized view of different types of data that we may find in a typical machine learning problem.

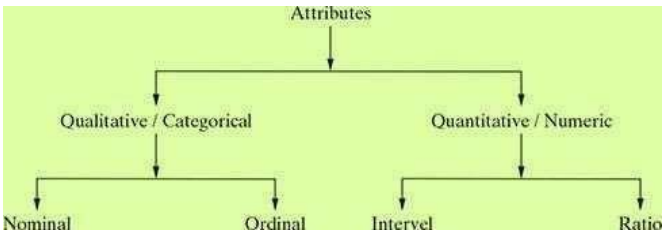


FIG. 1.13 Types of data

Apart from the approach detailed above, attributes can also be categorized into types based on a number of values that can be assigned. The attributes can be either discrete or continuous based on this factor.

Discrete attributes can assume a finite or countably infinite number of values. Nominal attributes such as roll number, street number, pin code, etc. can have a finite number of values whereas numeric attributes such as count, rank of students, etc. can have countably infinite values. A special type of discrete attribute which can assume two values only is called binary attribute.

Examples of binary attribute include male/ female, positive/negative, yes/no, etc.

Continuous attributes can assume any possible value which is a real number. Examples of continuous attribute include length, height, weight, price, etc.

1.3.3 EXPLORING STRUCTURE OF DATA

By now, we understand that in machine learning, we come across two basic data types – numeric and categorical. With this context in mind, we can delve deeper into understanding a data set. We need to understand that in a data set, which of the attributes are numeric and which are categorical in nature. This is because, the approach of exploring numeric data is different than the approach of exploring categorical data. In case of a standard data set, we may have the data dictionary available for reference. Data dictionary is a metadata repository, i.e. the repository of all information related to the structure of each data element contained in the data set. The data dictionary gives detailed information on each of the attributes – the description as well as the data type and other relevant details. In case the data dictionary is not available, we need to use standard library function of the machine learning tool that we are using and get the details. For the time being, let us move ahead with a standard data set from UCI machine learning repository.

The data set that we take as a reference is the Auto MPG data set available in the UCI repository. Figure 1.14 is a snapshot of the first few rows of the data set.

mpg	cylinder	displace-ment	horse-power	weight	accel-eration	model year	origin	car name
18	8	307	130	3504	12	70	1	Chevrolet chev-elle malibu
15	8	350	165	3693	11.5	70	1	Buick skylark 320
18	8	318	150	3436	11	70	1	Plymouth satellite
16	8	304	150	3433	12	70	1	Amc rebel sst
17	8	302	140	3449	10.5	70	1	Ford torino
15	8	429	198	4341	10	70	1	Ford galaxie 500
14	8	454	220	4354	9	70	1	Chevrolet impala
14	8	440	215	4312	8.5	70	1	Plymouth fury iii
14	8	455	225	4425	10	70	1	Pontiac catalina
15	8	390	190	3850	8.5	70	1	Amc acbassador dpl
15	8	383	170	3563	10	70	1	Dodge challenger se
14	8	340	160	3609	8	70	1	Plymouth ' cuda 340
15	8	400	150	3761	9.5	70	1	Chevrolet monte carlo
14	8	455	225	3086	10	70	1	Buick estate wagon (sw)
24	4	113	95	2372	15	70	3	Toyota corona mark ii
22	6	198	95	2933	15.5	70	1	Plymouth duster
18	6	199	97	2774	15.5	70	1	Amc hornet

FIG. 1.14 Auto MPG data set

As is quite evident from the data, the attributes such as 'mpg', 'cylinders', 'displacement', 'horsepower', 'weight', 'acceleration', 'model year', and 'origin' are all numeric. Out of these attributes, 'cylinders', 'model year', and 'origin' are discrete in nature as the only finite number of values can be assumed by these attributes. The remaining of the numeric attributes, i.e. 'mpg', 'displacement', 'horsepower', 'weight', and 'acceleration' can assume any real value.

Hence, these attributes are continuous in nature. The only remaining attribute 'car name' is of type categorical, or more specifically nominal. This data set is regarding prediction of fuel consumption in miles per gallon, i.e. the numeric attribute 'mpg' is the target attribute.

With this understanding of the data set attributes, we can start exploring the numeric and categorical attributes separately.

1. Exploring numerical data

There are two most effective mathematical plots to explore numerical data – box plot and histogram. We will explore all these plots one by one, starting with the most critical one, which is the box plot.

1. Understanding central tendency

To understand the nature of numeric variables, we can apply the measures of central tendency of data, i.e. mean and median. In statistics, measures of central tendency help us understand the central point of a set of data.

Mean, by definition, is a sum of all data values divided by the count of data elements. For example, mean of a set of observations – 21, 89, 34, 67, and 96 is calculated as below.

$$\text{Mean} = \frac{21 + 89 + 34 + 67 + 96}{5} = 61.4$$

If the above set of numbers represents marks of 5 students in a class, the mean marks, or the falling in the middle of the range is 61.4.

Median, on contrary, is the value of the element appearing in the middle of an ordered list of data elements. If we consider the above 5 data elements, the ordered list would be – 21, 34, 67, 89, and 96. Since there are 5 data elements, the 3rd element in the ordered list is considered as the median. Hence, the median value of this set of data is 67.

There might be a natural curiosity to understand why two measures of central tendency are reviewed. The reason is mean and median are impacted differently by data values appearing at the beginning or at the end of the range. Mean being calculated from the cumulative sum of data values, is impacted if too many data elements are having values closer to the far end of the range, i.e. close to the maximum or minimum values. It is especially sensitive to outliers, i.e. the values which are unusually high or low, compared to the other values.

Mean is likely to get shifted drastically even due to the presence of a small number of outliers. If we observe that for certain attributes the deviation between values of mean and median are quite high, we should investigate those attributes further and try to find out the root cause along with the need for remediation.

So, in the context of the Auto MPG data set, let's try to find out for each of the numeric attributes the values of mean and median. We can also find out if the deviation between these values is large. In Figure 1.15, the comparison between mean and median for all the attributes has been shown. We can see that for the attributes such as 'mpg', 'weight', 'acceleration', and 'model.year' the deviation between mean and median is not significant

which means the chance of these attributes having too many outlier values is less.
However, the deviation is significant for the attributes 'cylinders', 'displacement' and 'origin'. So, we need to further drill down and look at some more statistics for these attributes. Also, there is some problem in the values of the attribute 'horsepower' because of which the mean and median calculation isnot possible.

	mpg	cylin-ders	dis-place-ment	horse-power	weight	accel-eration	model year	origin
Median	23	4	148.5	?	2804	15.5	76	1
Mean	23.51	5.455	193.4	?	2970	15.57	76.01	1.573
Deviation	2.17	26.67%	23.22%		5.59%	0.45%	0.01%	36.43%
	Low	High	High		Low	Low	Low	High

FIG. 1.15 Mean vs. Median for Auto MPG

With a bit of investigation, we can find out that the problem is occurring because of the 6 data elements, as shown in FIG 1.16, do not have value for the attribute 'horsepower'.

mpg	cylinders	displace-ment	horse-power	weight	accel-eration	model year	origin	car name
25	4	98	?	2046	19	71	1	Ford pinto
21	6	200	?	2875	17	74	1	Ford maverick
40.9	4	85	?	1835	17.3	80	2	Renault lecar deluxe
23.6	4	140	?	2905	14.3	80	1	Ford mustang cobra
34.5	4	100	?	2320	15.8	81	2	Renault 18i
23	4	151	?	3035	20.5	82	1	Amc concord di

FIG. 1.16 Missing values of attribute 'horsepower' in Auto MPG

For that reason, the attribute 'horsepower' is not treated as a numeric. That's why the operations applicable on numeric variables, like mean or median, are failing. So we have to first remediate the missing values of the attribute 'horsepower' before being able to do any kind of exploration. However, we will cover the approach of remediation of missing values a little later.

2. Understanding data spread

Now that we have explored the central tendency of the different numeric attributes, we have a clear idea of which attributes have a large deviation between mean and median. Let's look closely at those attributes. To drill down more, we need to look at the entire range of values of the attributes, though not at the level of data elements as that may be too vast to review manually. So we will take a granular view of the data spread in the form of

- **Dispersionof data**
- **Position of the different data values**

➢ Measuring data dispersion

Consider the data values of two attributes

1. Attribute 1 values : 44, 46, 48, 45, and 47
2. Attribute 2 values : 34, 46, 59, 39, and 52

Both the set of values have a mean and median of 46.
However, the first set of values that is of attribute 1 is more concentrated or clustered around the mean/median value whereas the second set of values of attribute 2 is quite spread out or dispersed. To measure the extent of dispersion of a data, or to find out how much the different values of a data are spread out, the variance of the data is measured. The variance of a data is measured using the formula given below:

Variance $(x) = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$ where x is the variable or attribute whose variance is to be measured and n is the number of observations or values of variable x.

Standard deviation of a data is measured as follows:

$$\text{Standard deviation } (x) = \sqrt{\text{Variance } (x)}$$

Larger value of variance or standard deviation indicates more dispersion in the data and vice versa. In the above example, let's calculate the variance of attribute 1 and that of attribute 2.

For attribute 1,

$$\text{Variance} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

$$= \frac{44^2 + 46^2 + 48^2 + 45^2 + 47^2}{5} - \left(\frac{44 + 46 + 48 + 45 + 47}{5}\right)^2$$

$$= \frac{1936 + 2116 + 2304 + 2025 + 2209}{5} - \left(\frac{230}{5}\right)^2 = \frac{10590}{5} - (46)^2 = 2$$

For attribute 2,

So it is quite clear from the measure that attribute 1 values are quite concentrated around the mean while attribute 2 values are extremely spread out. Since this data was small, a visual inspection and understanding were possible and that matches with the measured value.

$$\text{Variance} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

$$= \frac{34^2 + 46^2 + 59^2 + 39^2 + 52^2}{5} - \left(\frac{34 + 46 + 59 + 39 + 52}{5}\right)^2$$

$$= \frac{1156 + 2116 + 3481 + 1521 + 2704}{5} - \left(\frac{230}{5}\right)^2 = \frac{10978}{5} - (46)^2 = 79.6$$

- **Measuring data value position**

When the data values of an attribute are arranged in an increasing order, we have seen earlier that median gives the central data value, which divides the entire data set into two halves. Similarly, if the first half of the data is divided into two halves so that each half consists of one- quarter of the data set, then that median of the first half is known as first quartile or Q1. In the same way, if the second half of the data is divided into two halves, then that median of the second half is known as third quartile or Q3. The overall median is also known as second quartile or Q2. So, any data set has five values -minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum.

Let's review these values for the attributes 'cylinders', 'displacement', and 'origin'. Figure 2.8

captures a summary of the range of statistics for the attributes. If we take the example of the attribute 'displacement', we can see that the difference between minimum value and Q1 is 36.2 and the difference between Q1 and median is 44.3. On the contrary, the difference between median and Q3 is 113.5 and Q3 and the maximum value is 193. In other words, the larger values are more spread out than the smaller ones. This helps in understanding why the value of mean is much higher than that of the median for the attribute 'displacement'. Similarly, in case of attribute 'cylinders', we can observe that the difference between minimum value and median is 1 whereas the difference between median and the maximum value is 4. For the attribute 'origin', the difference between minimum value and median is 0 whereas the difference between median and the maximum value is 2.

	cylinders	displacement	origin
Minimum	3	68	1
Q1	4	104.2	1
Median	4	148.5	1
Q3	8	262	2
Maximum	8	455	3

FIG. 1.17 Attribute value drill-down for Auto MPG

However, we still cannot ascertain whether there is any outlier present in the data. For that, we can better adopt some means to visualize the data. Box plot is an excellent visualization medium for numeric data.

2. Plotting and exploring numerical data

➢ Box plots

Now that we have a fairly clear understanding of the data set attributes in terms of spread and central tendency, let's try to make an attempt to visualize the whole thing as a box-plot. A box plot is an extremely effective mechanism to get a one-shot view and understand the nature of the data. But before we get to review the box plot for different attributes of Auto MPG data set, let's first try to understand a box plot in general and the interpretation of different aspects in a box plot. As we can see in Figure 2.9, the box plot (also called box and whisker plot) gives a standard visualization of the five- number summary statistics of a data, namely minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. Below is a detailed interpretation of a box plot.

The central rectangle or the box spans from first to third quartile (i.e. Q1 to Q3), thus giving

the inter-quartile range (IQR).

Median is given by the line or band within the box.

The lower whisker extends up to 1.5 times of the inter-quartile range (or IQR) from the bottom of the box, i.e. the first quartile or Q1.

However, the actual length of the lower whisker depends on the lowest data value that falls within (Q1 – 1.5 times of IQR). Let's try to understand this with an example. Say for a specific set of data, Q1 = 73, median = 76 and Q3 = 79. Hence, IQR will be 6 (i.e. Q3 – Q1).So, lower whisker can extend maximum till (Q1 – 1.5 × IQR) = 73 – 1.5 × 6 = 64. However, say there are lower range data values such as 70, 63, and 60. So, the lower whisker will come at 70 as this is the lowest data value larger than 64.

The upper whisker extends up to 1.5 as times of the inter-quartile range (or IQR) from the top of the box, i.e. the third quartile or Q3. Similar to lower whisker, the actual length of the upper whisker will also depend on the highest data value that falls within (Q3 + 1.5 times of IQR). Let's try to understand this with an example. For the same set of data mentioned in the above point, upper whisker can extend maximum till (Q3 + 1.5 × IQR) = 79 + 1.5 × 6 = 88. If there is higher range of data values like 82, 84, and 89. So, the upper whisker will come at 84 as this is the highest data value lower than 88.

The data values coming beyond the lower or upper whiskers are the ones which are of unusually low or high values respectively. These are the outliers, which may deserve special consideration.

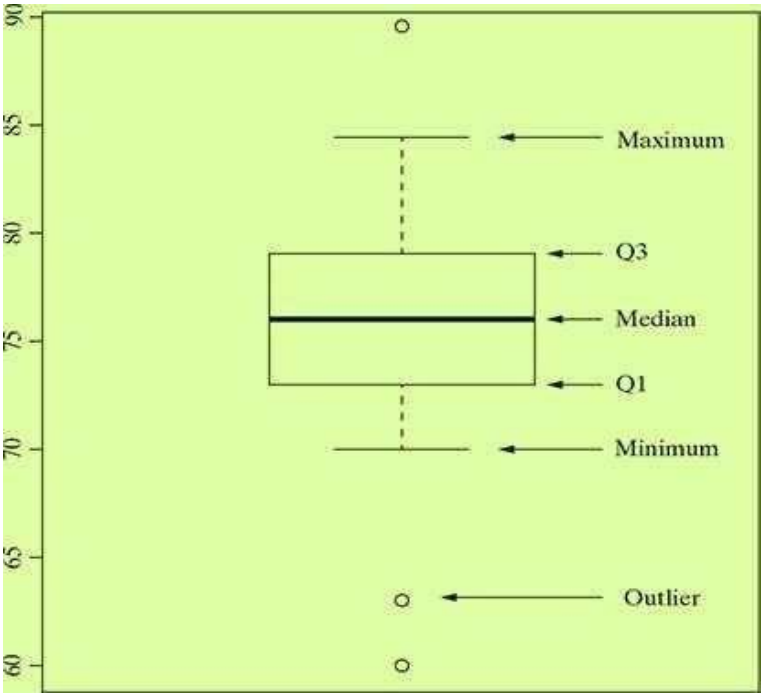


FIG. 1.18 Box plot

Let's visualize the box plot for the three attributes - 'cylinders', 'displacement', and 'origin'. We will also review the box plot of another attribute in which the deviation between mean and median is very little and see what the basic difference in the respective box plots is presents the respective box plots.

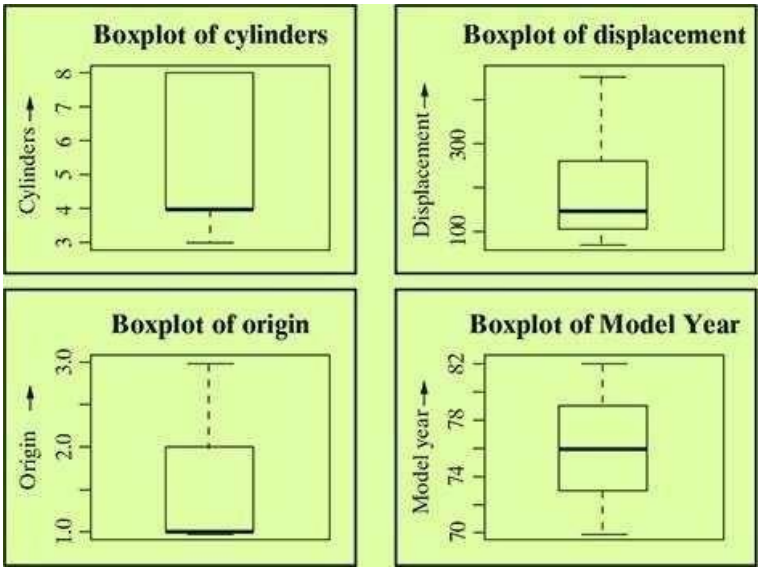


FIG. 1.19 Box plot of Auto MPG attributes

- **Analysing box plot for 'cylinders'**

The box plot for attribute 'cylinders' looks pretty weird in shape. The upper whisker is missing, the band for median falls at the bottom of the box, even the lower whisker is pretty small compared to the length of the box! Is everything right?

The answer is a big YES, and you can figure it out if you delve a little deeper into the actual data values of the attribute. The attribute 'cylinders' is discrete in nature having values from 3 to 8. Table 2.2 captures the frequency and cumulative frequency of it.

Table 2.2 Frequency of "Cylinders" Attribute

Cylinders	Frequency	Cumulative Frequency
3	4	4
4	204	208 (= 4 + 204)
5	3	211 (= 208 + 3)
6	84	295 (= 211 + 84)
7	0	295 (= 295 + 0)
8	103	398 (= 295 + 103)

As can be observed in the table, the frequency is extremely high for data value 4. Two other data values where the frequency is quite high are 6 and 8. So now if we try to find the quartiles, since the total frequency is 398, the first quartile (Q1), median (Q2), and third quartile (Q3) will be at a cumulative frequency 99.5 (i.e. average of 99th and 100th observation), 199 and 298.5 (i.e. average of 298th and 299th observation), respectively. This way Q1 = 4, median = 4 and Q3 = 8. Since there is no data value beyond 8, there is no upper whisker. Also, since both Q1 and median are 4, the band for median falls on the bottom of the box. Same way, though the lower whisker could have extended till -2 (Q1 – 1.5 × IQR = 4 – 1.5 × 4 = –2), in reality, there is no data value lower than 3. Hence, the lower whisker is also short. In any case, a value of cylinders less than 1 is not possible.

1.3.4 DATA QUALITY AND REMEDIATION

1.Data quality

Success of machine learning depends largely on the quality of data. A data which has the right quality helps to achieve better prediction accuracy, in case of supervised learning. However, it is not realistic to expect that the data will be flawless. We have already come across at least two types of problems:

1. Certain data elements without a value or data with a missing value.
2. Data elements having value surprisingly different from the other elements, which we term as outliers.

There are multiple factors which lead to these data quality issues. Following are some of them:

Incorrect sample set selection: The data may not reflect normal or regular quality due to incorrect selection of sample set. For example, if we are selecting a sample set of sales transactions from a festive period and trying to use that data to predict sales in future. In this case, the prediction will be far apart from the actual scenario, just because the sample set has been selected in a wrong time.

Similarly, if we are trying to predict poll results using a training data which doesn't comprise of a right mix of voters from different segments such as age, sex, ethnic diversities, etc., the prediction is bound to be a failure. It may also happen due to incorrect sample size. For example, a sample of small size may not be able to capture all aspects or information needed for right learning of the model.

Errors in data collection: resulting in outliers and missing values In many cases, a person or group of persons are responsible

for the collection of data to be used in a learning activity. In this manual process, there is the possibility of wrongly recording data either in terms of value (say 20.67 is wrongly recorded as 206.7 or 2.067) or in terms of a unit of measurement (say cm. is wrongly recorded as m. or mm.). This may result in data elements which have abnormally high or low value from other elements. Such records are termed as outliers.

It may also happen that the data is not recorded at all. In case of a survey conducted to collect data, it is all the more possible as survey responders may choose not to respond to a certain question. So the data value for that data element in that responder's record is missing.

5. Data remediation

The issues in data quality, as mentioned above, need to be remediated, if the right amount of efficiency has to be achieved in the learning activity. Out of the two major areas mentioned above, the first one can be remedied by proper sampling technique. This is a completely different area – covered as a specialized subject area in statistics. We will not cover that in this book. However, human errors are bound to happen, no matter whatever checks and balances we put in. Hence, proper remedial steps need to be taken for the second area mentioned above.

We will discuss how to handle outliers and missing values.

1. Handling outliers

Outliers are data elements with an abnormally high value which may impact prediction accuracy, especially in regression models. Once the outliers are identified and the decision has been taken to amend those values, you may consider one of the following approaches. However, if the outliers are natural, i.e. the value of the data element is surprisingly high or low because of a valid reason, then we should not amend it.

Remove outliers: If the number of records which are outliers is not many, a simple approach may be to remove them.

Imputation: One other way is to impute the value with mean or median or mode. The value of the most similar data element may also be used for imputation.

Capping: For values that lie outside the 1.5|×| IQR limits, we can cap them by replacing those observations below the lower limit with the value of 5th percentile and those that lie above the upper limit, with the value of 95th percentile.

If there is a significant number of outliers, they should be treated separately in the statistical model. In that case, the groups should be treated as two different groups, the model should be built for both groups and then the output can be combined.

2. Handling missing values

In a data set, one or more data elements may have missing values in multiple records. As discussed above, it can be caused by omission on part of the surveyor or a person who is

collecting sample data or by the responder, primarily due to his/her unwillingness to respond or lack of understanding needed to provide a response. It may happen that a specific question (based on which the value of a data element originates) is not applicable to a person or object with respect to which data is collected.

6. DATA PRE-PROCESSING

1.Dimensionality reduction

Till the end of the 1990s, very few domains were explored which included data sets with a high number of attributes or features. In general, the data sets used in machine learning used to be in few 10s. However, in the last two decades, there has been a rapid advent of computational biology like genome projects. These projects have produced extremely high-dimensional data sets with 20,000 or more features being very common.

Also, there has been a wide-spread adoption of social networking leading to a need for text classification for customer behaviour analysis.

High-dimensional data sets need a high amount of computational space and time. At the same time, not all features are useful – they degrade the performance of machine learning algorithms. Most of the machine learning algorithms perform better if the dimensionality of data set, i.e. the number of features in the data set, is reduced. Dimensionality reduction helps in reducing irrelevance and redundancy in features. Also, it is easier to understand a model if the number of features involved in the learning activity is less.

Dimensionality reduction refers to the techniques of reducing the dimensionality of a data set by creating new attributes by combining the original attributes. The most common approach for dimensionality reduction is known as Principal Component Analysis (PCA). PCA is a statistical technique to convert a set of correlated variables into a set of transformed, uncorrelated variables called principal components. The principal components are a linear combination of the original variables. They are orthogonal to each other. Since principal components are uncorrelated, they capture the maximum amount of variability in the data. However, the only challenge is that the original attributes are lost due to the transformation.

Another commonly used technique which is used for dimensionality reduction is Singular Value Decomposition (SVD).

- Feature subset selection

Feature subset selection or simply called feature selection, both for supervised as well as unsupervised learning, try to find out the optimal subset of the entire feature set which significantly reduces computational cost without any major impact on the learning accuracy. It

may seem that a feature subset may lead to loss of
useful information as certain features are going to be excluded from the final set of features used for learning. However, for elimination only features which are not relevant or redundant are selected.

A feature is considered as irrelevant if it plays an insignificant role (or contributes almost no information) in classifying or grouping together a set of data instances. All irrelevant features are eliminated while selecting the final feature subset. A feature is potentially redundant when the information contributed by the feature is more or less same as one or more other features. Among a group of potentially redundant features, a small number of features can be selected as a part of the final feature subset without causing any negative impact to learn model accuracy.

4. Modelling and Evaluation

1. SELECTING A MODEL

Now that you are familiar with the basic learning process and have understood model abstraction and generalization in that context, let's try to formalize it in context of a motivating example. Continuing the thread of the potential attack during the election campaign, New City Police department has succeeded in foiling the bid to attack the electoral candidate. However, this was a wake-up call for them and they want to take a proactive action to eliminate all criminal activities in the region.

They want to find the pattern of criminal activities in the recent past, i.e. they want to see whether the number of criminal incidents per month has any relation with an average income of the local population, weapon sales, the inflow of immigrants, and other such factors. Therefore, an association between potential causes of disturbance and criminal incidents has to be determined. In other words, the goal or target is to develop a model to infer how the criminal incidents change based on the potential influencing factors mentioned above.

In machine learning paradigm, the potential causes of disturbance, e.g. average income of the local population, weapon sales, the inflow of immigrants, etc. are input variables. They are also called predictors, attributes, features, independent variables, or simply variables. The number of criminal incidents is an output variable (also called response or dependent variable). Input variables can be denoted by X, while individual input variables are represented as X1, X2, X3, …, Xn and output variable by symbol Y. The relationship between X and Y is represented in the general form: $Y = f(X) + e$,

where 'f ' is the target function and 'e' is a random error term.

1. Predictive models

Models for supervised learning or predictive models, as is understandable from the name itself, try to predict certain value using the values in an input data set. The learning model attempts to establish a relation between the target feature, i.e. the feature being predicted, and the predictor features. The predictive models have a clear focus on what they want to learn and how they want to learn.

Predictive models, in turn, may need to predict the value of a category or class to which a data instance belongs to. Below are some examples:

1. Predicting win/loss in a cricket match
2. Predicting whether a transaction is fraud
3. Predicting whether a customer may move to another product

The models which are used for prediction of target features of categorical value are known as classification models. The target feature is known as a class and the categories to which classes are divided into are called levels. Some of the popular classification models include k-Nearest Neighbor (kNN), Naïve Bayes, and Decision Tree.

Predictive models may also be used to predict numerical values of the target feature based on the predictor features. Below are some examples:

1. Prediction of revenue growth in the succeeding year
2. Prediction of rainfall amount in the coming monsoon
3. Prediction of potential flu patients and demand for flu shots next winter

The models which are used for prediction of the numerical value of the target feature of a data instance are known as regression models. Linear Regression and Logistic Regression models are popular regression models.

2.Descriptivemodels

Models for unsupervised learning or descriptive models are used to describe a data set or gain insight from a data set. There is no target feature or single feature of interest in case of unsupervised learning. Based on the value of all features, interesting patterns or insights are derived about the data set.

Descriptive models which group together similar data instances, i.e. data instances having a similar value of the different features are called clustering models. Examples of

clustering include

1. Customer grouping or segmentation based on social, demographic, ethnic, etc. factors
2. Grouping of music based on different aspects like genre, language, time-period, etc.
3. Grouping of commodities in an inventory

The most popular model for clustering is k-Means.

Descriptive models related to pattern discovery is used for market basket analysis of transactional data. In market basket analysis, based on the purchase pattern available in the transactional data, the possibility of purchasing one product based on the purchase of another product is determined. For example, transactional data may reveal a pattern that generally a customer who purchases milk also purchases biscuit at the same time. This can be useful for targeted promotions or in-store set up. Promotions related to biscuits can be sent to customers of milk products or vice versa. Also, in the store products related to milk can be placed close to biscuits.

2. TRAINING A MODEL (FOR SUPERVISED LEARNING)

1. Holdout method

In case of supervised learning, a model is trained using the labelled input data. However, how can we understand the performance of the model? The test data may not be available immediately. Also, the label value of the test data is not known. That is the reason why a part of the input data is held back (that is how the name holdout originates) for evaluation of the model. This subset of the input data is used as the test data for evaluating the performance of a trained model. In general 70%–80% of the input data (which is obviously labelled) is used for model training.

The remaining 20%–30% is used as test data for validation of the performance of the model. However, a different proportion of dividing the input data into training and test data is also acceptable. To make sure that the data in both the buckets are similar in nature, the division is done randomly. Random numbers are used to assign data items to the partitions. This method of partitioning the input data into two parts – training and test data, which is by holding back a part of the input data for validating the trained model is known as holdout method.

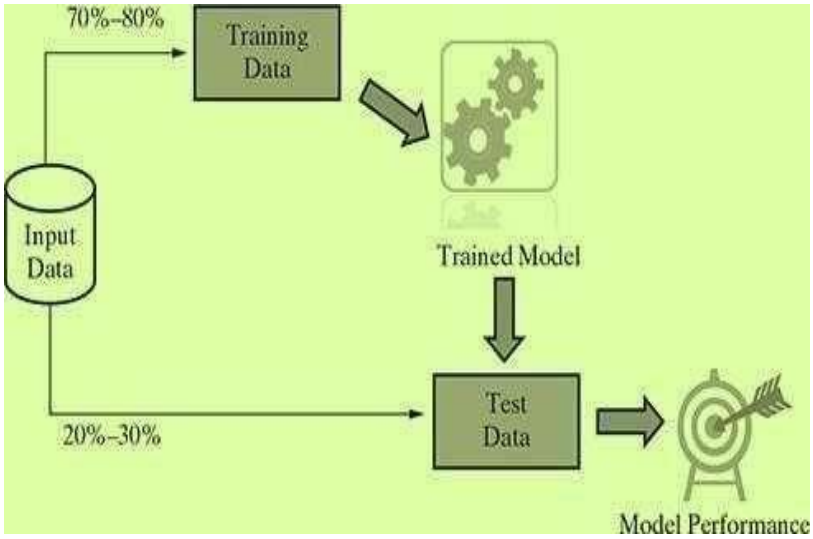


FIG. 1.20 Holdout method

Once the model is trained using the training data, the labels of the test data are predicted using the model's target function. Then the predicted value is compared with the actual value of the label. This is possible because the test data is a part of the input data with known labels. The performance of the model is in general measured by the accuracy of prediction of the label value.

In certain cases, the input data is partitioned into three portions – a training and a test data, and a third validation data. The validation data is used in place of test data, for measuring the model performance. It is used in iterations and to refine the model in each iteration. The test data is used only for once, after the model is refined and finalized, to measure and report the final performance of the model as a reference for future learning efforts.

An obvious problem in this method is that the division of data of different classes into the training and test data may not be proportionate. This situation is worse if the overall percentage of data related to certain classes is much less compared to other classes. This may happen despite the fact that random sampling is employed for test data selection. This problem can be addressed to some extent by applying stratified random sampling in place of sampling. In case of stratified random sampling, the whole data is broken into several homogenous groups or strata and a random sample is selected from each such stratum. This ensures that the generated random partitions have equal

proportions of each class.

1. K-fold Cross-validation method

Holdout method employing stratified random sampling approach still heads into issues in certain specific situations. Especially, the smaller data sets may have the challenge to divide the data of some of the classes proportionally amongst training and test data sets. A special variant of holdout method, called repeated holdout, is sometimes employed to ensure the randomness of the composed data sets. In repeated holdout, several random holdouts are used to measure the model performance. In the end, the average of all performances is taken. As multiple holdouts have been drawn, the training and test data (and also validation data, in case it is drawn) are more likely to contain representative data from all classes and resemble the original input data closely. This process of repeated holdout is the basis of k-fold cross-validation technique. In k-fold cross-validation, the data set is divided into k- completely distinct or non-overlapping random partitions called folds. Figure 1.21 depicts an overall approach for k-fold cross-validation.

The value of 'k' in k-fold cross-validation can be set to any number. However, there are two approaches which are extremely popular:

1. 10-fold cross-validation (10-fold CV)
2. Leave-one-out cross-validation (LOOCV)

10- fold cross-validation is by far the most popular approach. In this approach, for each of the 10-folds, each comprising of approximately 10% of the data, one of the folds is used as the test data for validating model performance trained based on the remaining 9 folds (or 90% of the data). This is repeated 10 times, once for each of the 10 folds being used as the test data and the remaining folds as the training data. The average performance across all folds is being reported. Figure 3.3 depicts the detailed approach of selecting the 'k' folds in k-fold cross-validation. As can be observed in the figure, each of the circles resembles a record in the input data set whereas the different colors indicate the different classes that the records belong to. The entire data set is broken into 'k' folds – out of which one fold is selected in each iteration as the test data set. The fold selected as test data set in each of the 'k' iterations is different. Also, note that though the circles resemble the records in the input data set, the contiguous circles represented as folds do not mean that they are subsequent records in the data set. This is more a virtual representation and not a physical representation. As already mentioned,

the records in a fold are drawn by using random sampling technique.

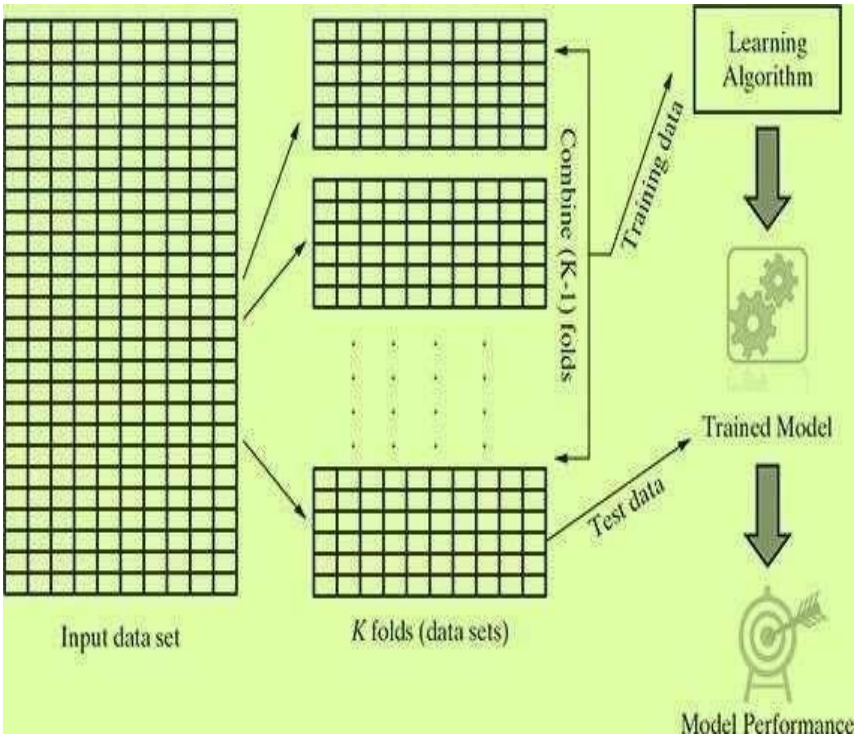


FIG. 1.21 Overall approach for K-fold cross-validation

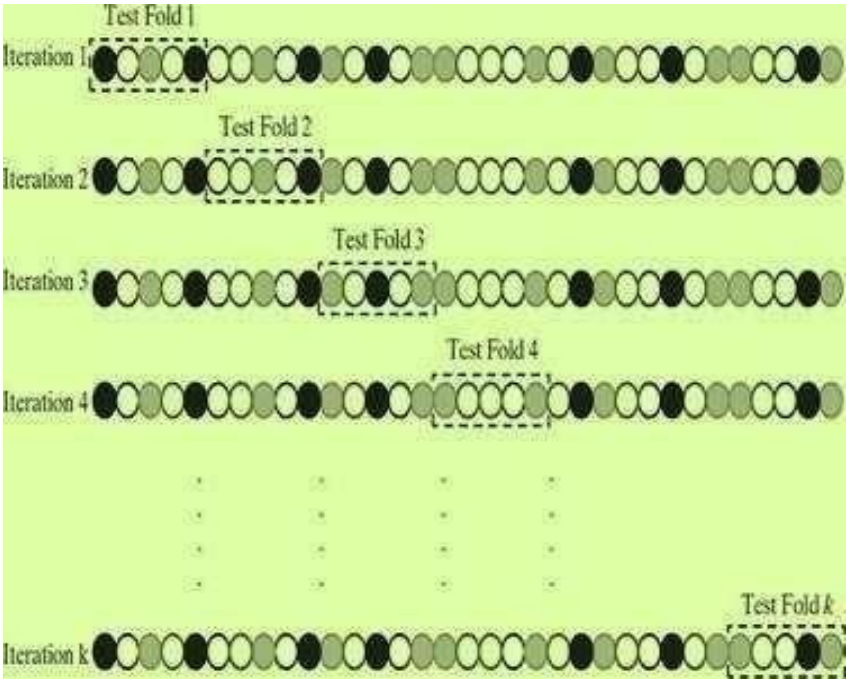


FIG. 1.22 Detailed approach for fold selection

Leave-one-out cross-validation (LOOCV) is an extreme case of k-fold cross-validation using one record or data instance at a time as a test data. This is done to maximize the count of data used to train the model. It is obvious

that the number of iterations for which it has to be run is equal to the total number of data in the input data set.

Hence, obviously, it is computationally very expensive and not used much in practice.

3. Bootstrap sampling

Bootstrap sampling or simply bootstrapping is a popular way to identify training and test data sets from the input data set. It uses the technique of Simple Random Sampling with Replacement (SRSWR), which is a well- known technique in sampling theory for drawing random samples. We have seen earlier that k-fold cross- validation divides the data into separate partitions – say 10 partitions in case of 10-fold cross-validation. Then it uses data instances from partition as test data and the remaining partitions as training data. Unlike this approach adopted in case of k-fold cross- validation, bootstrapping randomly picks data instances from the input data set, with the possibility of the same data instance to be picked multiple times. This essentially means that from the input data set having 'n' data instances, bootstrapping can create one or more training data sets having 'n' data instances, some of the data instances being repeated multiple times. Figure 3.4 briefly presents the approach followed in bootstrap sampling.

This technique is particularly useful in case of input data sets of small size, i.e. having very less number of data instances.

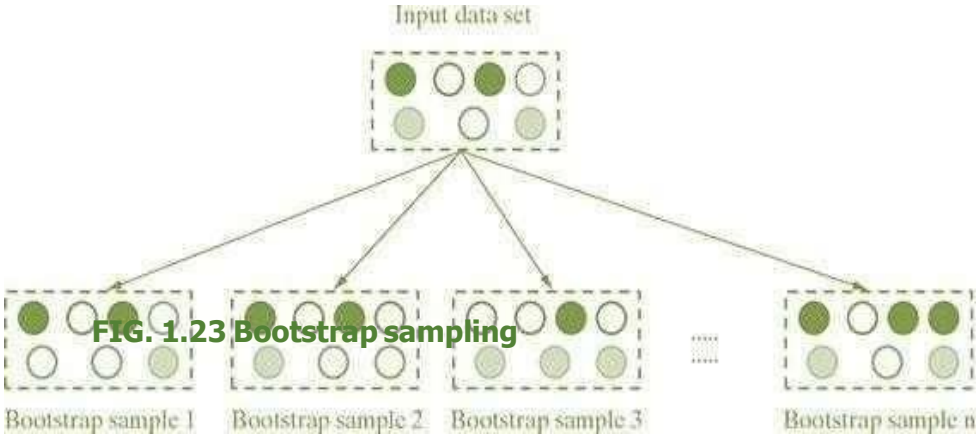


FIG. 1.23 Bootstrap sampling

4.Lazy vs. Eager learner

Eager learning follows the general principles of machine learning – it tries to construct a generalized, input- independent target function during the model training phase. It follows the typical steps of machine learning,

i.e. abstraction and generalization and comes up with a trained model at the end of the learning phase. Hence, when the test data comes in for classification, the eager learner is ready with the model and doesn't need to refer back to the training data. Eager learners take more time in the learning phase than the lazy learners. Some of the algorithms which adopt eager learning approach include Decision Tree, Support Vector Machine, Neural Network, etc.

Lazy learning, on the other hand, completely skips the abstraction and generalization processes, as explained in context of a typical machine learning process. In that respect, strictly speaking, lazy learner doesn't 'learn' anything. It uses the training data in exact, and uses the knowledge to classify the unlabelled test data. Since lazy learning uses training data as-is, it is also known as rote learning (i.e. memorization technique based on repetition). Due to its heavy dependency on the given training data instance, it is also known as instance learning. They are also called non-parametric learning. Lazy learners take very little time in training because not much of training actually happens. However, it takes quite some time in classification as for each tuple of test data, a comparison-based assignment of label happens. One of the most popular algorithm for lazy learning is k- nearest neighbor.

1.4.3 MODEL REPRESENTATION AND INTERPRETABILITY

We have already seen that the goal of supervised machine learning is to learn or derive a target function which can best determine the target variable from the set of input variables. A key consideration in learning the target function from the training data is the extent of generalization. This is because the input data is just a limited, specific view and the new, unknown data in the test data set may be differing quite a bit from the training data.

Fitness of a target function approximated by a learning algorithm determines how correctly it is able to classify a set of data it has never seen.

1. Underfitting

If the target function is kept too simple, it may not be able to capture the essential nuances and represent the underlying data well. A typical case of underfitting may occur when trying to represent a non-linear data with a linear model as demonstrated by both cases of underfitting shown. Many times underfitting happens due to unavailability of sufficient training data. Underfitting results in both poor performance with training data as well as poor generalization to test data. Underfitting can be avoided by

1. using more training data
2. reducing features by effective feature selection

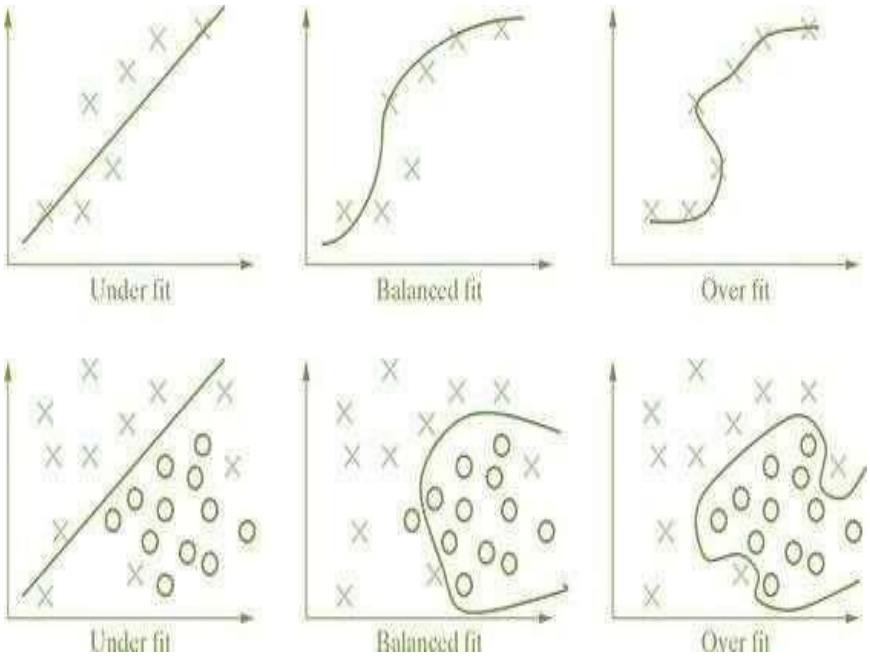


FIG. 1.24 Underfitting and Overfitting of models

2.Overfitting

Overfitting refers to a situation where the model has been designed in such a way that it emulates the training data too closely. In such a case, any specific deviation in the training data, like noise or outliers, gets embedded in the model. It adversely impacts the performance of the model on the test data. Overfitting, in many cases, occur as a result of trying to fit an excessively complex model to closely match the training data. This is represented with a sample data set in figure 3.5 . The target function, in these cases, tries to make sure all training data points are correctly partitioned by the decision

boundary.
However, more often than not, this exact nature is not replicated in the unknown test data set. Hence, the target function results in wrong classification in the test data set. Overfitting results in good performance with training data set, but poor generalization and hence poor performance with test data set. Overfitting can be avoided by

1. using re-sampling techniques like k-fold cross validation
2. hold back of a validation data set
3. remove the nodes which have little or no predictive power for the given machine learning problem.

1.4.4 EVALUATING PERFORMANCE OF A MODEL

3.5.1 Supervised learning - classification

In supervised learning, one major task is classification. The responsibility of the classification model is to assign class label to the target feature based on the value of the predictor features. For example, in the problem of predicting the win/loss in a cricket match, the classifier will assign a class value win/loss to target feature based on the values of other features like whether the team won the toss, number of spinners in the team, number of wins the team had in the tournament, etc. To evaluate the performance of the model, the number of correct classifications or predictions made by the model has to be recorded. A classification is said to be correct if, say for example in the given problem, it has been predicted by the model that the team will win and it has actually won.

Based on the number of correct and incorrect classifications or predictions made by a model, the accuracy of the model is calculated. If 99 out of 100 times the model has classified correctly, e.g. if in 99 out of 100 games what the model has predicted is same as what the outcome has been, then the model accuracy is said to be 99%. However, it is quite relative to say whether a model has performed well just by looking at the accuracy value. For example, 99% accuracy in case of a sports win predictor model may be reasonably good but the same number may not be acceptable as a good threshold when the learning problem deals with predicting a critical illness. In this case, even the 1% incorrect prediction may lead to loss of many lives. So the model performance needs to be evaluated in light of the learning problem in question. Also, in certain cases, erring on the side of caution may be preferred at the cost of overall accuracy. For that reason, we need to look more closely at the model accuracy and also at the same time look at

other measures of performance of a model like sensitivity, specificity, precision, etc. So, let's start with looking at model accuracy more closely. And let's try to understand it with an example.

There are four possibilities with regards to the cricket match win/loss prediction:

1. the model predicted win and the team won
2. the model predicted win and the team lost
3. the model predicted loss and the team won
4. the model predicted loss and the team lost

In this problem, the obvious class of interest is 'win'.

The first case, i.e. the model predicted win and the team won is a case where the model has correctly classified data instances as the class of interest. These cases are referred as True Positive (TP) cases.

The second case, i.e. the model predicted win and the team lost is a case where the model incorrectly classified data instances as the class of interest. These cases are referred as False Positive (FP) cases.

The third case, i.e. the model predicted loss and the team won is a case where the model has incorrectly classified as not the class of interest. These cases are referred as False Negative (FN) cases.

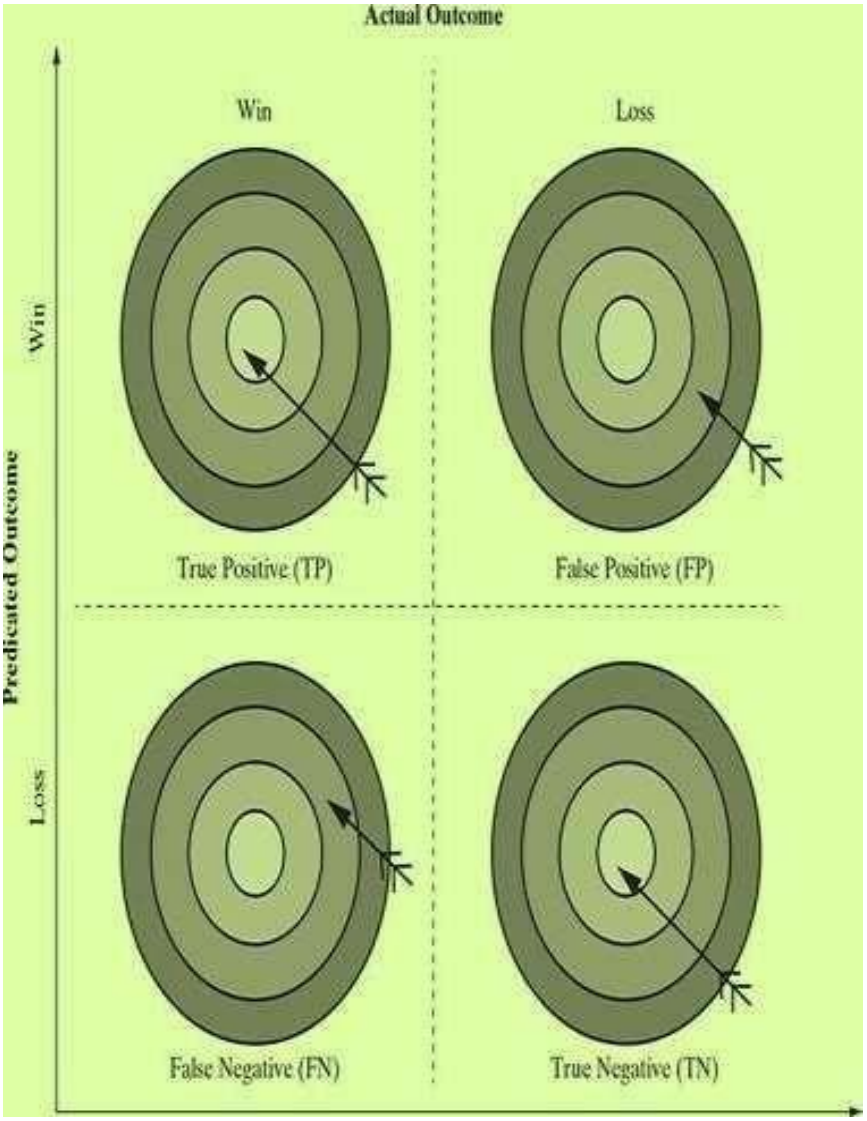


FIG. 1.25 Details of model classification

The fourth case, i.e. the model predicted loss and the team lost is a case where the model has correctly classified as not the class of interest. These cases are referred as True Negative (TN) cases. For any classification model, model accuracy is given by total number of correct classifications (either as the class of interest, i.e. True Positive or as not the class of interest, i.e. True Negative) divided by total number of classifications done.

$$\text{Model accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

A matrix containing correct and incorrect predictions in the form of TPs, FPs, FNs and TNs is known as confusion matrix. The win/loss prediction of cricket match has two classes of interest – win and loss. For that reason, it will generate a 2 × 2 confusion matrix. For a classification problem involving three classes, the confusion matrix would be 3 × 3, etc.

Let's assume the confusion matrix of the win/loss prediction of cricket match problem to be as below:

	ACTUAL WIN	ACTUAL LOSS
Predicted Win	85	4
Predicted Loss	2	9

In context of the above confusion matrix, total count of TPs = 85, count of FPs = 4, count of FNs = 2 and count of TNs = 9.

$$\therefore \text{Model accuracy} = \frac{TP + TN}{TP + FP + FN + TN} = \frac{85 + 9}{85 + 4 + 2 + 9} = \frac{94}{100} = 94\%$$

The percentage of misclassifications is indicated using error rate which is measured as

$$\text{Error rate} = \frac{FP + FN}{TP + FP + FN + TN}$$

In context of the above confusion matrix,

$$\text{Error rate} = \frac{FP + FN}{TP + FP + FN + TN} = \frac{4 + 2}{85 + 4 + 2 + 9} = \frac{6}{100} = 6\%$$
$$= 1 - \text{Model accuracy}$$

Sometimes, correct prediction, both TPs as well as TNs, may happen by mere coincidence. Since these occurrences boost model accuracy, ideally it should not happen. Kappa value of a model indicates the adjusted the model accuracy.

The sensitivity of a model measures the proportion of TP examples or positive cases which were correctly classified. It is measured as

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

In the context of the above confusion matrix for the cricket match win prediction problem,

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{85}{85 + 2} = \frac{85}{87} = 97.7\%$$

So, again taking the example of the malignancy prediction of tumours, class of interest is 'malignant'. Sensitivity measure gives the proportion of tumours which are actually malignant and have been predicted as malignant. It is quite obvious that for such problems the most critical measure of the performance of a good model is sensitivity. A high value of sensitivity is more desirable than a high value of accuracy.

Specificity is also another good measure to indicate a good balance of a model being excessively conservative or excessively aggressive. Specificity of a model measures the proportion of negative examples which have been correctly classified. In the context, of malignancy prediction of tumours, specificity gives the proportion of benign tumours which have been correctly classified. In the context of the above confusion matrix for the cricket match win prediction problem,

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{9}{9 + 4} = \frac{9}{13} = 69.2\%$$

A higher value of specificity will indicate a better model performance. However, it is quite understandable that a conservative approach to reduce False Negatives might actually push up the number of FPs. Reason for this is that the model, in order to reduce FNs, is going to classify more tumours as malignant. So the chance that benign tumours will be classified as malignant or FPs will increase.

There are two other performance measures of a supervised learning model which are similar to sensitivity and specificity. These are precision and recall. While precision gives the proportion of positive predictions which are truly positive, recall gives the proportion of TP cases over all actually positive cases.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision indicates the reliability of a model in predicting a class of interest. When the model is related to win / loss prediction of cricket, precision indicates how often it predicts the win correctly. In context of the above confusion matrix for the cricket match win prediction problem,

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{85}{85 + 4} = \frac{85}{89} = 95.5\%$$

It is quite understandable that a model with higher precision is perceived to be more reliable.

Recall indicates the proportion of correct prediction of positives to the total number of positives. In case of win/loss prediction of cricket, recall resembles what proportion of the total wins were predicted correctly.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

In the context of the above confusion matrix for the cricket match win prediction problem,

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{85}{85 + 2} = \frac{85}{87} = 97.7\%$$

1.F-measure

F-measure is another measure of model performance which combines the precision and recall. It takes the harmonic mean of precision and recall as calculated as

$$F\text{-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

In context of the above confusion matrix for the cricket match win prediction problem,

$$F\text{-measure} = \frac{2 \times 0.955 \times 0.977}{0.955 + 0.977} = \frac{1.866}{1.932} = 96.6\%$$

As a combination of multiple measures into one, F- score gives the right measure using which performance of different models can be compared. However, one assumption the calculation is based on is that precision and recall have equal weight, which may not always be true in reality. In certain problems, the disease prediction problems, e.g., precision may be given far more weightage. In that case, different weightages may be assigned to precision and recall. However, there may be a serious dilemma regarding what value to be adopted for each and what is the basis for the specific value adopted

2.Receiver operatingcharacteristic (ROC) curves

Receiver Operating Characteristic (ROC) curve helps in visualizing the performance of a classification model. It shows the efficiency of a model in the detection of true positives while avoiding the occurrence of false positives. To refresh our memory, true positives are those cases where the model has correctly classified data instances as the class of interest. For example, the model has correctly classified the tumours as malignant, in case of a tumour malignancy prediction problem. On the other hand, FPs are those cases where the model incorrectly classified data instances as the class of interest. Using the same example, in this case, the model has incorrectly classified the tumours as malignant, i.e. tumours which are actually benign have been classified as malignant.

$$\text{True Positive Rate TPR} = \frac{TP}{TP + FN}$$

$$\text{False Positive Rate FPR} = \frac{FP}{FP + TN}$$

5. IMPROVING PERFORMANCE OF A MODEL

Now we have almost reached the end of the journey of building learning models. We have got some idea about what modelling is, how to approach about it to solve a learning problem and how to measure the success of our model. Now comes a million dollar question. Can we improve the performance of our model? If so, then what are the levers for improving the performance? In fact, even before that comes the question of model selection – which model should be selected for which machine learning task? We have already discussed earlier that the model selection is done one several aspects:

1. Type of learning the task in hand, i.e. supervised or unsupervised
2. Type of the data, i.e. categorical or numeric
3. Sometimes on the problem domain

4. Above all, experience in working with different models to solve problems of diverse domains

One effective way to improve model performance is by tuning model parameter. Model parameter tuning is the process of adjusting the model fitting options. For example, in the popular classification model k-Nearest Neighbour (kNN), using different values of 'k' or the number of nearest neighbours to be considered, the model can be tuned. In the same way, a number of hidden layers can be adjusted to tune the performance in neural networks model. Most machine learning models have at least one parameter which can be tuned.

As an alternate approach of increasing the performance of one model, several models may be combined together. The models in such combination are complimentary to each other, i.e. one model may learn one type data sets well while struggle with another type of data set. Another model may perform well with the data set which the first one struggled with. This approach of combining different models with diverse strengths is known as ensemble

Ensemble helps in averaging out biases of the different underlying models and also reducing the variance.

Ensemble methods combine weaker learners to create stronger ones. A performance boost can be expected even if models are built as usual and then ensembled.

Following are the typical steps in ensemble process:

Build a number of models based on the training data

- For diversifying the models generated, the training data subset can be varied using the allocation function. Sampling techniques like bootstrapping may be used to generate unique training data sets. Alternatively, the same training data may be used but the models combined are quite varying, e.g, SVM, neural network, kNN, etc. The outputs from the different models are combined using a combination function. A very simple strategy of combining, say in case of a prediction task using ensemble, can be majority voting of the different models combined. For example, 3 out of 5 classes predict 'win' and 2 predict 'loss' – then the final outcome of the ensemble using majority vote would be a 'win'.

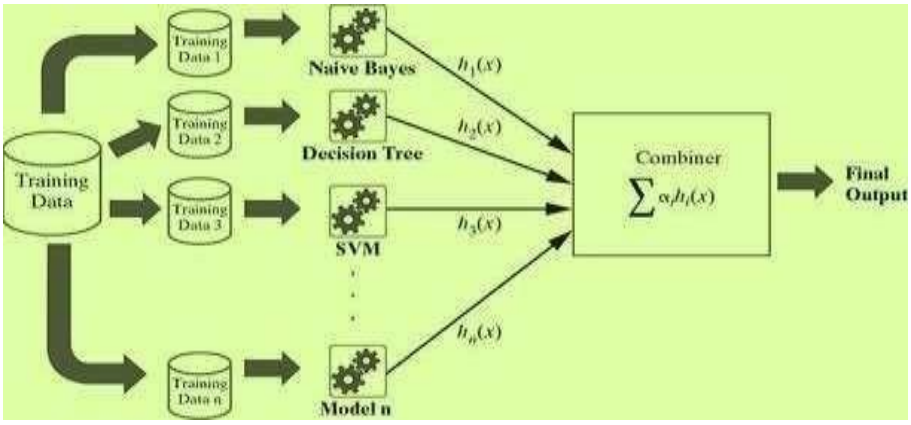


FIG. 1.26 Ensemble

One of the earliest and most popular ensemble models is bootstrap aggregating or bagging. Bagging uses bootstrap sampling method to generate multiple training data sets. These training data sets are used to generate (or train) a set of models using the same learning algorithm. Then the outcomes of the models are combined by majority voting (classification) or by average (regression). Bagging is a very simple ensemble technique which can perform really well for unstable learners like a decision tree, in which a slight change in data can impact the outcome of a model significantly.

Just like bagging, boosting is another key ensemble- based technique. In this type of ensemble, weaker learning models are trained on resampled data and the outcomes are combined using a weighted voting approach based on the performance of different models. Adaptive boosting or AdaBoost is a special variant of boosting algorithm. It is based on the idea of generating weak learners and slowly learning

Random forest is another ensemble-based technique. It is an ensemble of decision trees – hence the name random forest to indicate a forest of decision trees.

Assignments

10.ASSIGNMENT 1- UNIT 1

1. Identify and select a data set of your choice (UCI Repository datasets / Online Datasets/ WEKA datasets / R tool Datasets). Describe in a page about the data set stating the nature of the dataset, focus area, number of attributes, datatype of the attributes, volume of data, specific features that needs special attention to be paid etc. Clear Inference should be given at the end of each output. Submit the assignment as a printout. (Use any Programming language - Python/ R Programming for Analysis)

a) Measures of Central Tendency and Dispersion.
b) Analysis of Modality of all attributes.

2. Identify and select a data set of your choice (UCI Repository datasets / Online Datasets/ WEKA datasets / R tool Datasets). Describe in a page about the data set stating the nature of the dataset, focus area, number of attributes, datatype of the attributes, volume of data, specific features that needs special attention to be paid etc. Clear Inference should be given at the end of each output. Submit the assignment as a printout. (Use any Programming language - Python/ R Programming for Analysis)

a) Box plot analysis of all numerical attributes.
b) Histogram analysis of all numerical attributes.
c) Scatterplot Analysis and pairwise scatterplot of all numerical attributes of dataset.

3. Consider Breast Cancer Dataset to Apply PCA .The Breast Cancer data set is a real-valued multivariate data that consists of two classes, where each class signifies whether a patient has breast cancer or not. The two categories are: malignant and benign.The malignant class has 212 samples, whereas the benign class has 357 samples.It has 30 features shared across all classes: radius, texture, perimeter, area, smoothness, fractal dimension, etc

Note :
Data Set → https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

10.ASSIGNMENT 1- UNIT 1

4. Identify and select a data set of your choice (UCI Repository datasets / Online Datasets/ WEKA datasets / R tool Datasets). Describe in a page about the data set stating the nature of the dataset, focus area, number of attributes, datatype of the attributes, volume of data, specific features that needs special attention to be paid etc. Clear Inference should be given at the end of each output. Submit the assignment as a printout. (Use any Programming language - Python/ R Programming for Analysis)

a) Correlation analysis using contingency table and chi-square test for categorical attributes.

5. Consider iris flower Dataset to Apply SVD . The data is multivariate, with 150 measurements of 4 features (length and width cm of both sepal and petal) on 3 distinct Iris species. Of the 150 measurements, there are 50 measurements each for Iris setose, Iris versicolor, and Iris virginica.

Note :
DATASET- https://archive.ics.uci.edu/dataset/53/iris

Part A – Q & A

Unit - I

PART -A

S.No	Question and Answer	CO,K
1.	Define Machine Learning. Machine learning is a subfield of artificial intelligence, which is broadly defined as the capability of a machine to imitate intelligent human behavior.	CO1, K1
2.	List out Different Types of learning methods. Supervised Learning, Unsupervised Learning, reinforcement Learning	CO1, K1
3.	What is meant by supervised learning? Supervised learning, also known as supervised machine learning, is a subcategory of machine learning and artificial intelligence. It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately.	CO1, K1
4.	What is Unsupervised Learning? Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.	CO1, K1
5.	Differentiate supervised and unsupervised machine learning. In supervised machine learning, the machine is trained using labeled data. Then a new dataset is given into the learning model so that the algorithm provides a positive outcome by analyzing the labeled data. For example, we first require to label the data which is necessary to train the model while performing classification. In the unsupervised machine learning, the machine is not trained using labeled data and let the algorithms make the decisions without any corresponding output variables.	CO1, K1
6.	Define Reinforcement Learning? Reinforcement learning is a machine learning training method based on rewarding desired behaviors and/or punishing undesired ones. In general, a reinforcement learning agent is able to perceive and interpret its environment, take actions and learn through trial and error.	CO1, K1
7.	Where is supervised learning used? Linear regression is a supervised learning technique typically used in predicting, forecasting, and finding relationships between quantitative data.	CO1, K1

PART -A

S.No	Question and Answer	CO,K
8.	Give Example for Unsupervised Learning. ome examples of unsupervised learning algorithms include K-Means Clustering, Principal Component Analysis and Hierarchical Clustering.	CO1, K1
9.	Give Example for Reinforcement Learning. Reinforcement learning can be used in different fields such as healthcare, finance, recommendation systems, etc. Playing games like Go: Google has reinforcement learning agents that learn to solve problems by playing simple games like Go, which is a game of strategy	CO1, K1
10.	List Out real time application of ML. Image recognition Speech recognition. Medical diagnosis. Statistical arbitrage Predictive analytics	CO1, K1
11.	Define Data Science. ata science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions.	CO1, K1
12	What is the use of PCA in machine learning? or Write Applications of PCA in Machine Learning. It is used to reduce the number of dimensions in healthcare data. PCA can help resize an image. It can be used in finance to analyze stock data and forecast returns. PCA helps to find patterns in the high-dimensional datasets	CO1, K1

PART -A

S.No	Question and Answer	CO,K
14	Define Discriminant Analysis. Discriminant analysis is a versatile statistical method often used by market researchers to classify observations into two or more groups or categories. In other words, discriminant analysis is used to assign objects to one group among a number of known groups	CO1, K1
15.	Define Data preparation. ata preparation (also referred to as "data preprocessing") is the process of transforming raw data so that data scientists and analysts can run it through machine learning algorithms to uncover insights or make predictions.	CO1, K1
16.	What is data visualization and representation? Data visualization is the graphical representation of information and data in a pictorial or graphical format(Example: charts, graphs, and maps). Data visualization tools provide an accessible way to see and understand trends, patterns in data, and outliers. you must create a representation of the data to provide the model with a useful vantage point into the data's key qualities.	CO1, K1
17.	List two application of PCA. Some of the applications of Principal Component Analysis (PCA) are: Spike-triggered covariance analysis in Neuroscience. Quantitative Finance. Image Compression.	CO1, K1
18.	Write Disadvantage of supervised learning. Computation time is vast for supervised learning. Unwanted data downs efficiency. Pre-processing of data is no less than a big challenge. Always in need of updates. Anyone can overfit supervised algorithms easily.	CO1, K1
19.	What is classifier in machine learning? A classifier in a Machine Learning is a system that inputs a vector of discrete or continuous feature values and outputs a single discrete value, the class.	CO1, K1
20.	What is the main goal of supervised learning? The goal of Supervised Learning is to come up with, or infer, an approximate mapping function that can be applied to one or more input variables, and produce an output variable or result. The training process involves taking a supervised training data set with non features and a label.	CO1, K1

Part B – Questions

PART – B Q & A (WITH K LEVEL AND CO) UNIT 1

1. What is a model in context of machine learning? How can you train a model? CO1,K3
2. Explain "No Free Lunch" theorem in context of machine learning? CO1,K3
3. Explain in details the process of K-fold cross-validation? CO1,K3
4. Explain the bootstrap sampling? Why is it needed? CO1,K3
5. Why do we need to calculate Kappa value for a classification model? Show with a sample set of data, how to calculate Kappa value of a classification model. CO1,K3
6. Explain the process of ensemble of models. What role does in play in machine learning? CO1,K3
7. What is the main purpose of a descriptive model? State some real-world problems solved using descriptive models. CO1,K3
9. Explain the process of evaluating a linear regression model. CO1,K3
10. Differentiate (any two): CO1,K3
    1. Predictive vs. descriptive models
    2. Model underfitting vs. overfitting
    3. Cross-validation vs. bootstrapping
11. Write short notes on any two: CO1,K3
    1. LOOCV
    2. F-measure
    3. Silhouette width
    4. ROC curve

PART B Q s (WITH K LEVEL AND CO- CO1,K3) UNIT 1

1. What is a target function? Express target function in context of a real-life example. How Is the fitness of a target function measured? CO1,K3

2. What are predictive models? What are descriptive models? Give examples of both types of models. Explain the difference between these types of models. CO1,K3

3. Explain, in details, the process of evaluating the performance of a classification model. Explain the different parameters of measurement. CO1,K3

4. a. What is underfitting in context of machine learning models? What is he major cause of underfitting?

   b. What is overfitting? When does it happen?

   c. Explain bias-variance trade-off in context of model fitting. CO1,K3

5. Can the performance of a learning model be improved? If yes, explain how. CO1,K3

6. How would you evaluate the success of an unsupervised learning model? What are the most popular measures of performance for an unsupervised learning model? CO1,K3

7. Is there a way to use a classification model for a numerical data or a regression model on a categorical data? Explain your answer. CO1,K3

8. Describe the process of predictive modelling for numerical values. How is it different from predictive modelling for categorical values? CO1,K3

9. While predicting malignancy of tumour of a set of patients using a classification following are the data recorded: CO1,K3

   1. Correct predictions – 15 malignant, 75 benign
   2. Incorrect predictions – 3 malignant, 7 benign

   Calculate the error rate, Kappa value, sensitivity, precision, and F-measure of the model.

10. Write short notes on any two:
    1. Holdout method
    2. 10-fold cross-validation
    3. Parameter tuning
11. Write the difference between (any two):
    1. Purity vs. Silhouette width
    2. Bagging vs. Boosting
    3. Lazy vs. Eager learner

Supportive online Certification courses (NPTEL, Swayam, Coursera, Udemy, etc.,)

13. Supportive online Certification courses

NPTEL COURSE LINK

Introduction to Machine Learning

1. https://onlinecourses.nptel.ac.in/noc23_cs18/preview

COURSERA

Introduction to Machine Learning

2. https://www.coursera.org/specializations/mathematics-machine-learning

Real time Applications in day to day life and to Industry

14. Real time applications in day-to-day life and to Industry

Governments and the Public Sector

Governments have been key players in the advancement of computer technologies. The US Census Bureau has been analyzing data to understand population trends for decades. Governments now use predictive analytics like many other industries – to improve service and performance; detect and prevent fraud; and better understand consumer behavior. They are also using predictive analytics to enhance cybersecurity.

Health Insurance

In addition to detecting claims fraud, the health insurance industry is taking steps to identify patients most at risk of chronic disease and find what interventions are best. Express Scripts, a large pharmacy benefits company, uses analytics to identify those not adhering to prescribed treatments, resulting in a savings of $1,500 to $9,000 per patient.

Detecting fraud. Using multiple predictive analytics applications can improve, or even provide, pattern detection and catch criminal behavior. As cybersecurity is a growing concern, high-performance behavioral predictive analytics analyzes all behavioral patterns on a network in real-time to catch abnormalities that may indicate fraud, vulnerabilities and advanced persistent threats. Optimizing the safety of your systems and platform can not only reduce the legal work of punishing fraud cases, hire lesser analysts to conduct manual investigations, but predictive analytics also allows your firm to scale without human-labor limits and can provide an edge in your business and helping clients feel safe using your product and services.

Content Beyond Syllabus

Contents beyond the Syllabus

TYPES OF LOGISTIC REGRESSION MODELS

Logistic regression refers to any regression model in which the response variable is categorical.

There are three types of logistic regression models:

- **Binary logistic regression**: The response variable can only belong to one of two categories.
- **Multinomial logistic regression**: The response variable can belong to one of three or more categories and there is no natural ordering among the categories.
- **Ordinal logistic regression**: The response variable can belong to one of three or more categories and there is a natural ordering among the categories.

The following table summarizes these differences:

Types of Logistic Regression Models

	Binomial Logistic Regression	Multinomial Logistic Regression	Ordinal Logistic Regression
Number of Categories for Response Variable	2	3 or more	3 or more
Does Order of Categories Matter?	No	No	Yes

This tutorial provides a brief explanation of each type of logistic regression model along with examples of each.

Type #1: Binary Logistic Regression

Binary logistic regression models are a type of logistic regression in which the response variable can only belong to two categories.

Here are a couple examples:

Example 1: NBA Draft

Suppose a sports data scientist wants to use the predictor variables (1) points, (2) rebounds, and (3) assists to predict the probability that a given college basketball player gets drafted into the NBA.

Since there are only two possible outcomes (drafted or not drafted) for the response variable, the data scientist would use a binomial logistic regression model.

Example 2: Spam Detection

Suppose a business wants to use the predictor variables (1) word count and (2) country of origin to predict the probability that a given email is spam.

Since there are only two possible outcomes (spam or not spam) for the response variable, the business would use a binomial logistic regression model.

Type #2: Multinomial Logistic Regression

Multinomial logistic regression models are a type of logistic regression in which the response variable can belong to one of three or more categories and there is no natural ordering among the categories.

Here are a couple examples:

Example 1: Political Preference

Suppose a political scientist wants to use the predictor variables (1) annual income and (2) years of education to predict the probability that an individual will vote for one of four different presidential candidates.

Since there are more than two possible outcomes (there are four potential candidates) for the response variable and there is no natural ordering among the outcomes, the political scientist would use a multinomial logistic regression model.

Example 2: Sports Preference

Suppose a sports analyst wants to use the predictor variables (1) TV hours viewed per week and (2) age to predict the probability that an individual will pick either basketball, football, or baseball as their preferred sport.

Since there are more than two possible outcomes (there are three sports) for the response variable, the sports analyst would use a multinomial logistic regression model.

Type #3: Ordinal Logistic Regression

Ordinal logistic regression models are a type of logistic regression in which the response variable can belong to one of three or more categories and there is a natural ordering among the categories.

Here are a couple examples:

Example 1: School Ratings

Suppose an academic advisor wants to use the predictor variables (1) GPA, (2) ACT score, and (3) SAT score to predict the probability that an individual will get into a university that can be categorized into "bad", "mediocre", "good", or "great."

Assessment Schedule (Proposed Date & Actual Date)

R.M.K GROUP OF INSTITUTIONS

15. ASSESSMENT SCHEDULE

Tentative schedule for the Assessment During 2022-2023 EVEN semester

S.NO	Name of the Assessment	Start Date	Portions
2	IAT 1	12.02.2024	UNIT 1 & 2
4	IAT 2	01.04.2024	UNIT 3 & 4
7	Model	20.04.2024	ALL 5 UNITS

Prescribed Text Books & Reference

16. PRESCRIBED TEXT BOOKS & REFERENCE BOOKS

TEXT BOOKS:

1. Saikat Dutt, Subramanian Chandramouli, Amit Kumar Das, Machine Learning, Pearson, 2019. (Unit 1 – chap 1,2,3/ Unit 2 – Chap 4 / Unit 4 – 9 / Unit 5 – Chap 10, 11)
2. Ethem Alpaydin, Introduction to Machine Learning, Adaptive Computation and Machine Learning Series, Third Edition, MIT Press, 2014. (Unit 2 – Chap 6 / Unit 4 – chap 8.2.3 / Unit 5 – Chap 18)

REFERENCES:

1. Anuradha Srinivasaraghavan, Vincy Joseph, Machine Learning, First Edition, Wiley, 2019. (Unit 3 – Chap 7,8,9,10,11 / Unit 4 – 13, 11.4, 11.5,12)
2. Peter Harrington, "Machine Learning in Action", Manning Publications, 2012.
3. Stephen Marsland, "Machine Learning – An Algorithmic Perspective", Second Edition, Chapman and Hall/CRC Machine Learning and Pattern Recognition Series, 2014.
4. Tom M Mitchell, Machine Learning, First Edition, McGraw Hill Education, 2013.
5. Christoph Molnar, "Interpretable Machine Learning - A Guide for Making Black Box Models Explainable", Creative Commons License, 2020.

Mini Project Suggestions

17. MINI PROJECT SUGGESTION

1. Iris Flowers Classification Project

Project idea – The iris flowers have different species and you can distinguish them based on the length of petals and sepals.

2. Loan Prediction using Machine Learning

Project idea – The idea behind this ML project is to build a model that will classify how much loan the user can take.

3. Predict Taxi Fares with Random Forests

In the Predict Taxi Fares project, predict the location and time to earn the biggest fare using the New York taxi dataset. Use tidyverse for data processing and visualization. To predict location and time, experiment with a tree base model such as Decision Tree and Random Forest.

4. Cartoonify Image with Machine Learning

Project Idea: Transform images into its cartoon. Yes, the objective of this machine learning project is to CARTOONIFY the images.

5. Classify Song Genres from Audio Data

In the Classify Song Genres machine learning project, use the song dataset to classify songs into two categories: 'Hip-Hop' or 'Rock.' Check the correlation between features, normalize data using scikit-learn's StandardScaler, apply PCA (Principal Component Analysis) on scaled data, and visualize the results. After that, use the scikit-learn Logistic Regression and Decision Tree model to train and validate the results

Thank you