Quantum Series

ENGINEERING

QUANTUM *Series*

Semester - 5 CS & IT

Machine Learning Techniques

- Topic-wise coverage of entire syllabus in Question-Answer form.
- Short Questions (2 Marks)

Session 2024-25 Odd Semester

Includes solution of following AKTU Question Papers

2020-21 • 2021-22 • 2022-23 • 2023-24

सजग सूचना

छात्रों, डीलरों, दुकानदारों, कॉपियर्स और जिनसे यह संबंधित हो सकता है क्वांटम पुस्तकों के कॉपीराइट उल्लंघन के संबंध में दिल्ली के माननीय उच्च न्यायालय का आदेश

यह छात्रों, डीलरों, दुकानदारों, [illegible]र्स और आम जनता को सूचित करना है कि क्वांटम पेज प्राइवेट [illegible]स्तकों की क्वांटम श्रृंखला में कॉपीराइट का मालिक है।

क्वांटम पुस्तकों (चाहे सॉफ्ट [illegible] हार्ड कॉ[illegible] की कोई भी अनधिकृत प्रतिलिपि, स्कैनिंग, पुनरुत्पादन, [illegible]रण या [illegible], क्वांटम पेज प्राइवेट लिमिटेड के कॉपीराइट और ट्रेडमार्क अधिकारों का उल्लंघन के बराबर है, जो दिवानी अपराध के साथ-साथ दांडिक अपराध है, जिनकी सजा कारावास तक है।

क्वांटम पेज प्राइवेट लिमिटेड ने दिल्ली के माननीय उच्च न्यायालय के समक्ष क्वांटम पुस्तकों के उल्लंघनकर्ताओं के खिलाफ **क्वांटम पेज प्राइवेट लिमिटेड बनाम टेलीग्राम एफजेड एलएलसी और अन्य, सीएस (कॉम) ९२१/२०२२** नामक एक मुकदमा शुरू किया है। 23 दिसंबर, 2022 के अपने आदेश में, माननीय उच्च न्यायालय ने माना है कि क्वांटम पुस्तकों का अनधिकृत पुनरुत्पादन कॉपीराइट उल्लंघन के बराबर है और टेलीग्राम चैनलों सहित विभिन्न स्रोतों से क्वांटम पुस्तकों की उल्लंघनकारी प्रतियों को हटाने का निर्देश दिया है।

छात्रों, डीलरों, दुकानदारों, कॉपीयरों और आम जनता को एतद्द्वारा आगाह किया जाता है कि वे क्वांटम पुस्तकों (चाहे सॉफ्ट कॉपी या हार्ड कॉपी में) की कोई भी अनधिकृत नकल, स्कैनिंग, पुनरुत्पादन, वितरण और वितरण न करें।

क्वांटम पुस्तकों के ऐसे किसी भी अनधिकृत उपयोग से उल्लंघनकर्ताओं पर क्वांटम पेज प्राइवेट लिमिटेड द्वारा दीवानी और आपराधिक कार्यवाही शुरू की जाएगी।

QUANTUM SERIES

For

B.Tech Students of Third Year
of All Engineering Colleges Affiliated to
Dr. A.P.J. Abdul Kalam Technical University,
Uttar Pradesh, Lucknow
(Formerly Uttar Pradesh Technical University)

Machine Learning Techniques

By

Kanika Dhama

Quantum Page
QUANTUM PAGE PVT. LTD.
Ghaziabad ■ New Delhi

1 UNIT

Introduction

CONTENTS

PART-1

Learning, Types of Learning.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.1. Define the term learning. What are the components of a learning system ?

Answer

1. Learning refers to the change in a subject's behaviour to a given situation brought by repeated experiences in that situation, provided that the behaviour changes cannot be explained on the basis of native response tendencies, matriculation or temporary states of the subject.
2. Learning agent can be thought of as containing a performance element that decides what actions to take and a learning element that modifies the performance element so that it makes better decisions.
3. The design of a learning element is affected by three major issues :
   a. Components of the performance element.
   b. Feedback of components.
   c. Representation of the components.

The important components of learning are :

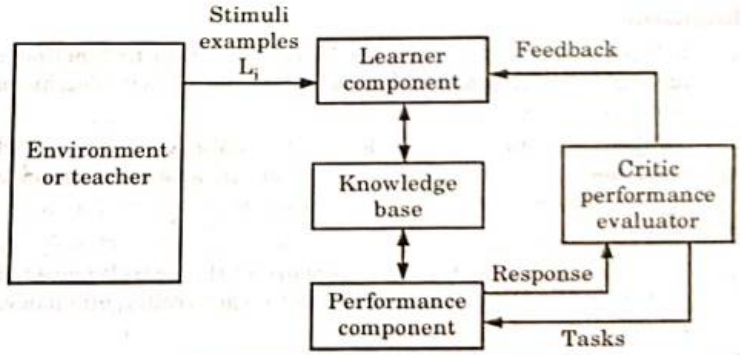


Fig. 1.1.1. General learning model.

1. **Acquisition of new knowledge :**
   a. One component of learning is the acquisition of new knowledge.

b. Simple data acquisition is easy for computers, even though it is difficult for people.

2. **Problem solving :**

The other component of learning is the problem solving that is required for both to integrate into the system, new knowledge that is presented to it and to deduce new information when required facts are not been presented.

Que 1.2. **Write down the performance measures for learning.**

Answer

Following are the performance measures for learning are :

1. **Generality :**
   a. The most important performance measure for learning methods is the generality or scope of the method.
   b. Generality is a measure of the case with which the method can be adapted to different domains of application.
   c. A completely general algorithm is one which is a fixed or self adjusting configuration that can learn or adapt in any environment or application domain.
2. **Efficiency :**
   a. The efficiency of a method is a measure of the average time required to construct the target knowledge structures from some specified initial structures.
   b. Since this measure is often difficult to determine and is meaningless without some standard comparison time, a relative efficiency index can be used instead.
3. **Robustness :**
   a. Robustness is the ability of a learning system to function with unreliable feedback and with a variety of training examples, including noisy ones.
   b. A robust system must be able to build tentative structures which are subjected to modification or withdrawal if later found to be inconsistent with statistically sound structures.
4. **Efficacy :**
   a. The efficacy of a system is a measure of the overall power of the system. It is a combination of the factors generality, efficiency, and robustness.
5. **Ease of implementation :**
   a. Ease of implementation relates to the complexity of the programs and data structures, and the resources required to develop the given learning system.

b. Lacking good complexity metrics, this measure will often be somewhat subjective.

Que 1.3. **Discuss supervised and unsupervised learning.**

Answer

Supervised learning :

1. Supervised learning is also known as associative learning, in which the network is trained by providing it with input and matching output patterns.
2. Supervised training requires the pairing of each input vector with a target vector representing the desired output.
3. The input vector together with the corresponding target vector is called training pair.

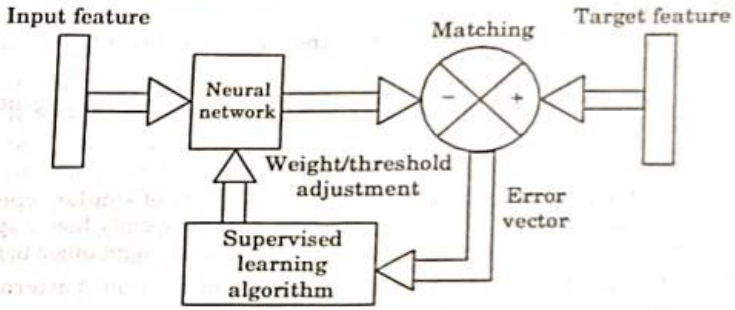


Fig. 1.3.1.

4. During the training session an input vector is applied to the network, and it results in an output vector.
5. This response is compared with the target response.
6. If the actual response differs from the target response, the network will generate an error signal.
7. This error signal is then used to calculate the adjustment that should be made in the synaptic weights so that the actual output matches the target output.
8. The error minimization in this kind of training requires a supervisor or teacher.
9. These input-output pairs can be provided by an external teacher, or by the system which contains the neural network (self-supervised).
10. Supervised training methods are used to perform non-linear mapping in pattern classification networks, pattern association networks and multilayer neural networks.

11. Supervised learning generates a global model that maps input objects to desired outputs.
12. In some cases, the map is implemented as a set of local models such as in case-based reasoning or the nearest neighbour algorithm.
13. In order to solve problem of supervised learning following steps are considered :
    i. Determine the type of training examples.
    ii. Gathering a training set.
    iii. Determine the input feature representation of the learned function.
    iv. Determine the structure of the learned function and corresponding learning algorithm.
    v. Complete the design.

Unsupervised learning :

1. It is a learning in which an output unit is trained to respond to clusters of pattern within the input.
2. Unsupervised training is employed in self-organizing neural networks.
3. This training does not require a teacher.
4. In this method of training, the input vectors of similar types are grouped without the use of training data to specify how a typical member of each group looks or to which group a member belongs.
5. During training the neural network receives input patterns and organizes these patterns into categories.
6. When new input pattern is applied, the neural network provides an output response indicating the class to which the input pattern belongs.
7. If a class cannot be found for the input pattern, a new class is generated.
8. Though unsupervised training does not require a teacher, it requires certain guidelines to form groups.
9. Grouping can be done based on color, shape and any other property of the object.
10. It is a method of machine learning where a model is fit to observations.
11. It is distinguished from supervised learning by the fact that there is no priori output.
12. In this, a data set of input objects is gathered.
13. It treats input objects as a set of random variables. It can be used in conjunction with Bayesian inference to produce conditional probabilities.

14. Unsupervised learning is useful for data compression and clustering.

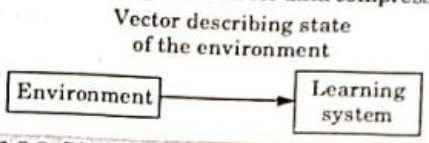


Fig. 1.3.2. Block diagram of unsupervised learning.

15. In unsupervised learning, system is supposed to discover statistically salient features of the input population.
16. Unlike the supervised learning paradigm, there is not a priori set of categories into which the patterns are to be classified; rather the system must develop its own representation of the input stimuli.

Que 1.4. Describe briefly reinforcement learning ?

Answer

1. Reinforcement learning is the study of how artificial system can learn to optimize their behaviour in the face of rewards and punishments.
2. Reinforcement learning algorithms have been developed that are closely related to methods of dynamic programming which is a general approach to optimal control.
3. Reinforcement learning phenomena have been observed in psychological studies of animal behaviour, and in neurobiological investigations of neuromodulation and addiction.

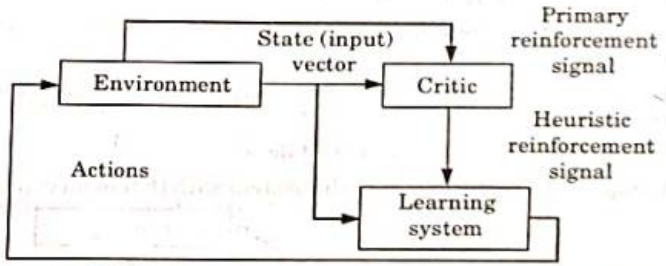


Fig. 1.4.1. Block diagram of reinforcement learning.

4. The task of reinforcement learning is to use observed rewards to learn an optimal policy for the environment.
5. An optimal policy is a policy that maximizes the expected total reward.
6. Without some feedback about what is good and what is bad, the agent will have no grounds for deciding which move to make.
7. The agents need to know that something good has happened when it wins and that something bad has happened when it loses.
8. This kind of feedback is called a reward or reinforcement.

9. Reinforcement learning is very valuable in the field of robotics, where the tasks to be performed are frequently complex enough to defy encoding as programs and no training data is available.
10. The robot's task consists of finding out, through trial and error (or success), which actions are good in a certain situation and which are not.
11. In many cases humans learn in a very similar way.
12. For example, when a child learns to walk, this usually happens without instruction, rather simply through reinforcement.
13. Successful attempts at working are rewarded by forward progress, and unsuccessful attempts are penalized by often painful falls.
14. Positive and negative reinforcement are also important factors in successful learning in school and in many sports.
15. In many complex domains, reinforcement learning is the only feasible way to train a program to perform at high levels.

Que 1.5. **What are the steps used to design a learning system ?**

Answer

Steps used to design a learning system are :

1. Specify the learning task.
2. Choose a suitable set of training data to serve as the training experience.
3. Divide the training data into groups or classes and label accordingly.
4. Determine the type of knowledge representation to be learned from the training experience.
5. Choose a learner classifier that can generate general hypotheses from the training data.
6. Apply the learner classifier to test data.
7. Compare the performance of the system with that of an expert human.

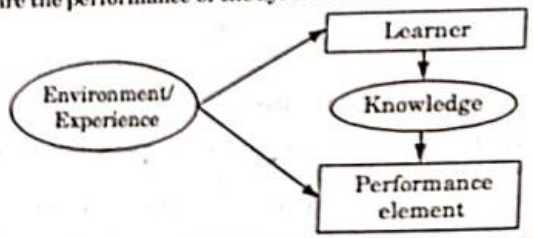


Fig. 1.5.1.

PART-2

Well Defined Learning Problems, Designing a Learning System.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.6. **Write short note on well defined learning problem with example.**

Answer

Well defined learning problem :

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Three features in learning problems :

1. The class of tasks (T)
2. The measure of performance to be improved (P)
3. The source of experience (E)

For example :

1. **A checkers learning problem :**
   a. **Task (T) :** Playing checkers.
   b. **Performance measure (P) :** Percent of games won against opponents.
   c. **Training experience (E) :** Playing practice games against itself.
2. **A handwriting recognition learning problem :**
   a. **Task (T) :** Recognizing and classifying handwritten words within images.
   b. **Performance measure (P) :** Percent of words correctly classified.
   c. **Training experience (E) :** A database of handwritten words with given classifications.
3. **A robot driving learning problem :**
   a. **Task (T) :** Driving on public four-lane highways using vision sensors.
   b. **Performance measure (P) :** Average distance travelled before an error (as judged by human overseer).
   c. **Training experience (E) :** A sequence of images and steering commands recorded while observing a human driver.

Que 1.7. **Describe well defined learning problems role's in machine learning.**

Answer

Well defined learning problems role's in machine learning :

1. **Learning to recognize spoken words :**
   a. Successful speech recognition systems employ machine learning in some form.
   b. For example, the SPHINX system learns speaker-specific strategies for recognizing the primitive sounds (phonemes) and words from the observed speech signal.
   c. Neural network learning methods and methods for learning hidden Markov models are effective for automatically customizing to individual speakers, vocabularies, microphone characteristics, background noise, etc.
2. **Learning to drive an autonomous vehicle :**
   a. Machine learning methods have been used to train computer controlled vehicles to steer correctly when driving on a variety of road types.
   b. For example, the ALYINN system has used its learned strategies to drive unassisted at 70 miles per hour for 90 miles on public highways among other cars.
3. **Learning to classify new astronomical structures :**
   a. Machine learning methods have been applied to a variety of large databases to learn general regularities implicit in the data.
   b. For example, decision tree learning algorithms have been used by NASA to learn how to classify celestial objects from the second Palomar Observatory Sky Survey.
   c. This system is used to automatically classify all objects in the Sky Survey, which consists of three terabytes of image data.
4. **Learning to play world class backgammon :**
   a. The most successful computer programs for playing games such as backgammon are based on machine learning algorithms.
   b. For example, the world's top computer program for backgammon, TD-GAMMON learned its strategy by playing over one million practice games against itself.

PART-3

History of ML, Introduction of Machine Learning Approaches - (Artificial Neural Network, Clustering, Reinforcement Learning, Decision Tree Learning, Bayesian Network, Support Vector Machine, Genetic Algorithm).

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.8. **Describe briefly the history of machine learning.**

Answer

A. Early history of machine learning :

1. In 1943, neurophysiologist Warren McCulloch and mathematician Walter Pitts wrote a paper about neurons, and how they work. They created a model of neurons using an electrical circuit, and thus the neural network was created.
2. In 1952, Arthur Samuel created the first computer program which could learn as it ran.
3. Frank Rosenblatt designed the first artificial neural network in 1958, called Perceptron. The main goal of this was pattern and shape recognition.
4. In 1959, Bernard Widrow and Marcian Hoff created two models of neural network. The first was called ADELINE, and it could detect binary patterns. For example, in a stream of bits, it could predict what the next one would be. The second was called MADELINE, and it could eliminate echo on phone lines.

B. 1980s and 1990s :

1. In 1982, John Hopfield suggested creating a network which had bidirectional lines, similar to how neurons actually work.
2. Use of back propagation in neural networks came in 1986, when researchers from the Stanford psychology department decided to extend an algorithm created by Widrow and Hoff in 1962. This allowed multiple layers to be used in a neural network, creating what are known as 'slow learners', which will learn over a long period of time.
3. In 1997, the IBM computer Deep Blue, which was a chess-playing computer, beat the world chess champion.
4. In 1998, research at AT&T Bell Laboratories on digit recognition resulted in good accuracy in detecting handwritten postcodes from the US Postal Service.

C. 21st Century :

1. Since the start of the 21st century, many businesses have realised that machine learning will increase calculation potential. This is why they are researching more heavily in it, in order to stay ahead of the competition.

2. Some large projects include :
   i. GoogleBrain (2012)
   ii. AlexNet (2012)
   iii. DeepFace (2014)
   iv. DeepMind (2014)
   v. OpenAI (2015)
   vi. ResNet (2015)
   vii. U-net (2015)

Que 1.9. Explain briefly the term machine learning.

Answer

1. Machine learning is an application of Artificial Intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
2. Machine learning focuses on the development of computer programs that can access data.
3. The primary aim is to allow the computers to learn automatically without human intervention or assistance and adjust actions accordingly.
4. Machine learning enables analysis of massive quantities of data.
5. It generally delivers faster and more accurate results in order to identify profitable opportunities or dangerous risks.
6. Combining machine learning with AI and cognitive technologies can make it even more effective in processing large volumes of information.

Que 1.10. What are the applications of machine learning ?

Answer

Following are the applications of machine learning :

1. **Image recognition :**
   a. Image recognition is the process of identifying and detecting an object or a feature in a digital image or video.
   b. This is used in many applications like systems for factory automation, toll booth monitoring, and security surveillance.
2. **Speech recognition :**
   a. Speech Recognition (SR) is the translation of spoken words into text.
   b. It is also known as Automatic Speech Recognition (ASR), computer speech recognition, or Speech To Text (STT).

   c. In speech recognition, a software application recognizes spoken words.
3. **Medical diagnosis :**
   a. ML provides methods, techniques, and tools that can help in solving diagnostic and prognostic problems in a variety of medical domains.
   b. It is being used for the analysis of the importance of clinical parameters and their combinations for prognosis.
4. **Statistical arbitrage :**
   a. In finance, statistical arbitrage refers to automated trading strategies that are typical of a short-term and involve a large number of securities.
   b. In such strategies, the user tries to implement a trading algorithm for a set of securities on the basis of quantities such as historical correlations and general economic variables.
5. **Learning associations :** Learning association is the process for discovering relations between variables in large data base.
6. **Extraction :**
   a. Information Extraction (IE) is another application of machine learning.
   b. It is the process of extracting structured information from unstructured data.

Que 1.11. What are the advantages and disadvantages of machine learning ?

Answer

Advantages of machine learning are :

1. **Easily identifies trends and patterns :**
   a. Machine learning can review large volumes of data and discover specific trends and patterns that would not be apparent to humans.
   b. For an e-commerce website like Flipkart, it serves to understand the browsing behaviours and purchase histories of its users to help cater to the right products, deals, and reminders relevant to them.
   c. It uses the results to reveal relevant advertisements to them.
2. **No human intervention needed (automation) :** Machine learning does not require physical force *i.e.*, no human intervention is needed.
3. **Continuous improvement :**
   a. ML algorithms gain experience, they keep improving in accuracy and efficiency.
   b. As the amount of data keeps growing, algorithms learn to make accurate predictions faster.

4. **Handling multi-dimensional and multi-variety data :**
   a. Machine learning algorithms are good at handling data that are multi-dimensional and multi-variety, and they can do this in dynamic or uncertain environments.

Disadvantages of machine learning are :

1. **Data acquisition :**
   a. Machine learning requires massive data sets to train on, and these should be inclusive/unbiased, and of good quality.
2. **Time and resources :**
   a. ML needs enough time to let the algorithms learn and develop enough to fulfill their purpose with a considerable amount of accuracy and relevancy.
   b. It also needs massive resources to function.
3. **Interpretation of results :**
   a. To accurately interpret results generated by the algorithms. We must carefully choose the algorithms for our purpose.
4. **High error-susceptibility :**
   a. Machine learning is autonomous but highly susceptible to errors.
   b. It takes time to recognize the source of the issue, and even longer to correct it.

Que 1.12. What are the advantages and disadvantages of different types of machine learning algorithm ?

Answer

Advantages of supervised machine learning algorithm :

1. Classes represent the features on the ground.
2. Training data is reusable unless features change.

Disadvantages of supervised machine learning algorithm :

1. Classes may not match spectral classes.
2. Varying consistency in classes.
3. Cost and time are involved in selecting training data.

Advantages of unsupervised machine learning algorithm :

1. No previous knowledge of the image area is required.
2. The opportunity for human error is minimised.
3. It produces unique spectral classes.
4. Relatively easy and fast to carry out.

Disadvantages of unsupervised machine learning algorithm :

1. The spectral classes do not necessarily represent the features on the ground.
2. It does not consider spatial relationships in the data.
3. It can take time to interpret the spectral classes.

Advantages of semi-supervised machine learning algorithm :

1. It is easy to understand.
2. It reduces the amount of annotated data used.
3. It is stable, fast convergent.
4. It is simple.
5. It has high efficiency.

Disadvantages of semi-supervised machine learning algorithm :

1. Iteration results are not stable.
2. It is not applicable to network level data.
3. It has low accuracy.

Advantages of reinforcement learning algorithm :

1. Reinforcement learning is used to solve complex problems that cannot be solved by conventional techniques.
2. This technique is preferred to achieve long-term results which are very difficult to achieve.
3. This learning model is very similar to the learning of human beings. Hence, it is close to achieving perfection.

Disadvantages of reinforcement learning algorithm :

1. Too much reinforcement learning can lead to an overload of states which can diminish the results.
2. Reinforcement learning is not preferable for solving simple problems.
3. Reinforcement learning needs a lot of data and a lot of computation.
4. The curse of dimensionality limits reinforcement learning for real physical systems.

Que 1.13. Write short note on Artificial Neural Network (ANN).

Answer

1. **Artificial Neural Networks (ANN) or neural networks are computational algorithms that intended to simulate the behaviour of biological systems composed of neurons.**

2. ANNs are computational models inspired by an animal's central nervous systems.
3. It is capable of machine learning as well as pattern recognition.
4. A neural network is an oriented graph. It consists of nodes which in the biological analogy represent neurons, connected by arcs.
5. It corresponds to dendrites and synapses. Each arc associated with a weight at each node.
6. A neural network is a machine learning algorithm based on the model of a human neuron. The human brain consists of millions of neurons.
7. It sends and process signals in the form of electrical and chemical signals.
8. These neurons are connected with a special structure known as synapses. Synapses allow neurons to pass signals.
9. An Artificial Neural Network is an information processing technique. It works like the way human brain processes information.
10. ANN includes a large number of connected processing units that work together to process information. They also generate meaningful results from it.

Que 1.14. Write short note on clustering.

Answer

1. Clustering is a division of data into groups of similar objects.
2. Each group or cluster consists of objects that are similar among themselves and dissimilar to objects of other groups as shown in Fig. 1.14.1.

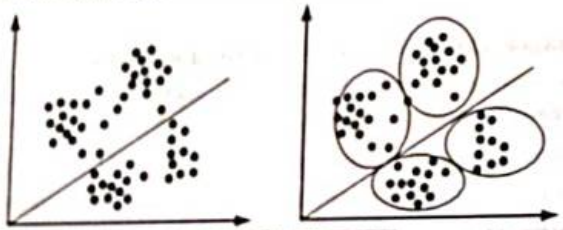


Fig. 1.14.1. Clusters.

3. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the object in the other cluster.
4. Clusters may be described as connected regions of a multidimensional space containing relatively high density points, separated from each other by a region containing a relatively low density points.
5. From the machine learning perspective, clustering can be viewed as unsupervised learning of concepts.
6. Clustering analyzes data objects without help of known class label.

7. In clustering, the class labels are not present in training data simply because they are not known to cluster the data objects.
8. Hence, it is the type of unsupervised learning.
9. For this reason, clustering is a form of learning by observation rather than learning by examples.
10. There are certain situations where clustering is useful. These include :
    a. The collection and classification of training data can be costly and time consuming. Therefore it is difficult to collect a training data set. A large number of training samples are not all labelled. Then it is useful to train a supervised classifier with a small portion of training data and then use clustering procedures to tune the classifier based on the large, unclassified dataset.
    b. For data mining, it can be useful to search for grouping among the data and then recognize the cluster.
    c. The properties of feature vectors can change over time. Then, supervised classification is not reasonable. Because the test feature vectors may have completely different properties.
    d. The clustering can be useful when it is required to search for good parametric families for the class conditional densities, in case of supervised classification.

Que 1.15. What are the applications of clustering ?

Answer

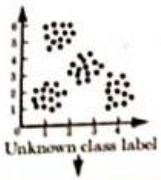
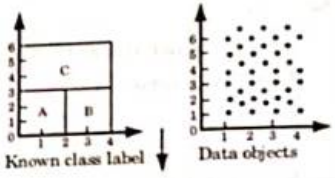
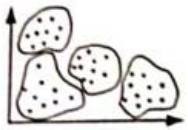
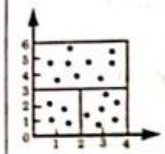
Following are the applications of clustering :

1. **Data reduction :**
    a. In many cases, the amount of available data is very large and its processing becomes complicated.
    b. Cluster analysis can be used to group the data into a number of clusters and then process each cluster as a single entity.
    c. In this way, data compression is achieved.
2. **Hypothesis generation :**
    a. In this case, cluster analysis is applied to a data set to infer hypothesis that concerns about the nature of the data.
    b. Clustering is used here to suggest hypothesis that must be verified using other data sets.
3. **Hypothesis testing :** In this context, cluster analysis is used for the verification of the validity of a specific hypothesis.
4. **Prediction based on groups :**
    a. In this case, cluster analysis is applied to the available data set and then the resulting clusters are characterized based on the characteristics of the patterns by which they are formed.

b. In this sequence, if an unknown pattern is given, we can determine the cluster to which it is more likely to belong and characterize it based on the characterization of the respective cluster.

Que 1.16. Differentiate between clustering and classification.

Answer

S.No.	Clustering	Classification
1.	Clustering analyzes data objects without known class label.	In classification, data are grouped by analyzing the data objects whose class label is known.
2.	There is no prior knowledge of the attributes of the data to form clusters.	There is some prior knowledge of the attributes of each classification.
3.	It is done by grouping only the input data because output is not predefined.	It is done by classifying output based on the values of the input data.
4.	The number of clusters is not known before clustering. These are identified after the completion of clustering. 	The number of classes is known before classification as there is predefined output based on input data. 
5.		
6.	It is considered as unsupervised learning because there is no prior knowledge of the class labels.	It is considered as the supervised learning because class labels are known before.

Que 1.17. What are the various clustering techniques ?

Answer

1. Clustering techniques are used for combining observed examples into clusters or groups which satisfy two following main criteria :
   a. Each group or cluster is homogeneous *i.e.*, examples belong to the same group are similar to each other.
   b. Each group or cluster should be different from other clusters *i.e.*, examples that belong to one cluster should be different from the examples of the other clusters.
2. Depending on the clustering techniques, clusters can be expressed in different ways :
   a. Identified clusters may be exclusive, so that any example belongs to only one cluster.
   b. They may be overlapping *i.e.*, an example may belong to several clusters.
   c. They may be probabilistic *i.e.*, an example belongs to each cluster with a certain probability.
   d. Clusters might have hierarchical structure.

Major classifications of clustering techniques are :

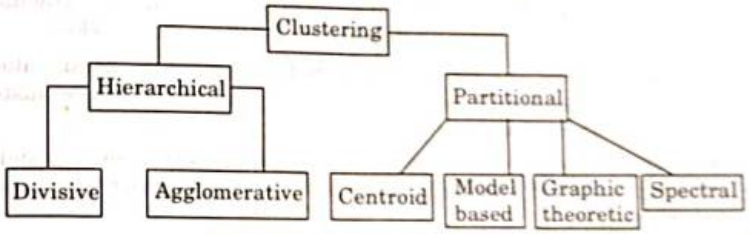


Fig. 1.17.1. Types of clustering.

   a. Once a criterion function has been selected, clustering becomes a well-defined problem in discrete optimization. We find those partitions of the set of samples that extremize the criterion function.
   c. The sample set is finite, there are only a finite number of possible partitions.
   d. The clustering problem can always be solved by exhaustive enumeration.

1. **Hierarchical clustering :**
   a. This method works by grouping data object into a tree of clusters.
   b. This method can be further classified depending on whether the hierarchical decomposition is formed in bottom up (merging) or top down (splitting) fashion.

Following are the two types of hierarchical clustering :

a. **Agglomerative hierarchical clustering :** This bottom up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster.

b. **Divisive hierarchical clustering :**

i. This top down strategy does the reverse of agglomerative strategy by starting with all objects in one cluster.

ii. It subdivides the cluster into smaller and smaller pieces until each object forms a cluster on its own.

2. **Partitional clustering :**

a. This method first creates an initial set of number of partitions where each partition represents a cluster.

b. The clusters are formed to optimize an objective partition criterion such as a dissimilarity function based on distance so that the objects within a cluster are similar whereas the objects of different clusters are dissimilar.

Following are the types of partitioning methods :

a. **Centroid based clustering :**

i. In this, it takes the input parameter and partitions a set of object into a number of clusters so that resulting intracluster similarity is high but the intercluster similarity is low.

ii. Cluster similarity is measured in terms of the mean value of the objects in the cluster, which can be viewed as the cluster's centroid or center of gravity.

b. **Model-based clustering :** This method hypothesizes a model for each of the cluster and finds the best fit of the data to that model.

Que 1.18. Describe reinforcement learning.

Answer

1. Reinforcement learning is the study of how animals and artificial systems can learn to optimize their behaviour in the face of rewards and punishments.
2. Reinforcement learning algorithms related to methods of dynamic programming which is a general approach to optimal control.
3. Reinforcement learning phenomena have been observed in psychological studies of animal behaviour, and in neurobiological investigations of neuromodulation and addiction.
4. The task of reinforcement learning is to use observed rewards to learn an optimal policy for the environment. An optimal policy is a policy that maximizes the expected total reward.

Que 1.19. Explain decision tree in detail.

Answer

1. A decision tree is a flowchart structure in which each internal node represents a test on a feature, each leaf node represents a class label and branches represent conjunctions of features that lead to those class labels.
2. The paths from root to leaf represent classification rules.
3. Fig 1.19.1, illustrate the basic flow of decision tree for decision making with labels (Rain(Yes), Rain(No)).

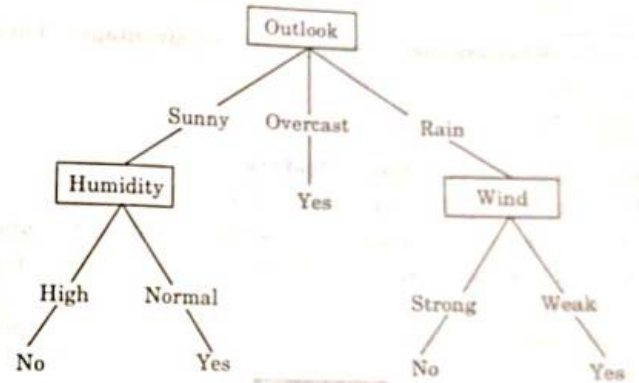


Fig. 1.19.1.

4. Decision tree is the predictive modelling approach used in statistics, data mining and machine learning.
5. Decision trees are constructed via an algorithmic approach that identifies the ways to split a data set based on different conditions.
6. Decision trees are a non-parametric supervised learning method used for both classification and regression tasks.
7. Classification trees are the tree models where the target variable can take a discrete set of values.
8. Regression trees are the decision trees where the target variable can take continuous set of values.

Que 1.20. What are the steps used for making decision tree ?

Answer

Steps used for making decision tree are :

1. Get list of rows (dataset) which are taken into consideration for making decision tree (recursively at each node).

2. Calculate uncertainty of our dataset or Gini impurity or how much our data is mixed up etc.
3. Generate list of all question which needs to be asked at that node.
4. Partition rows into True rows and False rows based on each question asked.
5. Calculate information gain based on Gini impurity and partition of data from previous step.
6. Update highest information gain based on each question asked.
7. Update question based on information gain (higher information gain).
8. Divide the node on question. Repeat again from step 1 until we get pure node (leaf nodes).

Que 1.21. What are the advantages and disadvantages of decision tree method ?

Answer

Advantages of decision tree method are :

1. Decision trees are able to generate understandable rules.
2. Decision trees perform classification without requiring computation.
3. Decision trees are able to handle both continuous and categorical variables.
4. Decision trees provide a clear indication for the fields that are important for prediction or classification.

Disadvantages of decision tree method are :

1. Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
2. Decision trees are prone to errors in classification problems with many class and relatively small number of training examples.
3. Decision tree are computationally expensive to train. At each node, each candidate splitting field must be sorted before its best split can be found.
4. In decision tree algorithms, combinations of fields are used and a search must be made for optimal combining weights. Pruning algorithms can also be expensive since many candidate sub-trees must be formed and compared.

Que 1.22. Write short note on Bayesian belief networks.

Answer

1. Bayesian belief networks specify joint conditional probability distributions.
2. They are also known as belief networks, Bayesian networks, or probabilistic networks.

3. A Belief Network allows class conditional independencies to be defined between subsets of variables.
4. It provides a graphical model of causal relationship on which learning can be performed.
5. We can use a trained Bayesian network for classification.
6. There are two components that define a Bayesian belief network :

a. Directed acyclic graph :

i. Each node in a directed acyclic graph represents a random variable.

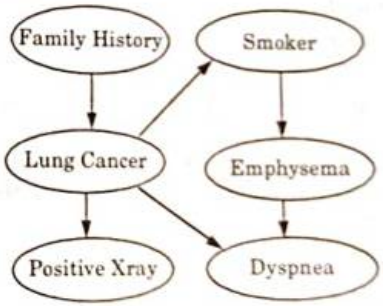
ii. These variable may be discrete or continuous valued.

iii. These variables may correspond to the actual attribute given in the data.

Directed acyclic graph representation : The following diagram shows a directed acyclic graph for six Boolean variables.

i. The arc in the diagram allows representation of causal knowledge.

ii. For example, lung cancer is influenced by a person's family history of lung cancer, as well as whether or not the person is a smoker.



iii. It is worth noting that the variable Positive X-ray is independent of whether the patient has a family history of lung cancer or that the patient is a smoker, given that we know the patient has lung cancer.

b. Conditional probability table :

The conditional probability table for the values of the variable LungCancer (LC) showing each possible combination of the values of its parent nodes, FamilyHistory (FH), and Smoker (S) is as follows :

	FH,S	FH,-S	-FH,S	-FH,S
LC	0.8	0.5	0.7	0.1
-LC	0.2	0.5	0.3	0.9

Que 1.23. **Write a short note on support vector machine.**

Answer

1. A Support Vector Machine (SVM) is machine learning algorithm that analyzes data for classification and regression analysis.
2. SVM is a supervised learning method that looks at data and sorts it into one of two categories.
3. An SVM outputs a map of the sorted data with the margins between the two as far apart as possible.
4. Applications of SVM :
   i. Text and hypertext classification
   ii. Image classification
   iii. Recognizing handwritten characters
   iv. Biological sciences, including protein classification

Que 1.24. **Explain genetic algorithm with flow chart.**

Answer

Genetic algorithm (GA) :

1. The genetic algorithm is a method for solving both constrained and unconstrained optimization problems that is based on natural selection.
2. The genetic algorithm repeatedly modifies a population of individual solutions.
3. At each step, the genetic algorithm selects individuals at random from the current population to be parents and uses them to produce the children for the next generation.
4. Over successive generations, the population evolves toward an optimal solution.

Flow chart : The genetic algorithm uses three main types of rules at each step to create the next generation from the current population :

a. **Selection rule :** Selection rules select the individuals, called parents, that contribute to the population at the next generation.

b. **Crossover rule :** Crossover rules combine two parents to form children for the next generation.

c. **Mutation rule :** Mutation rules apply random changes to individual parents to form children.

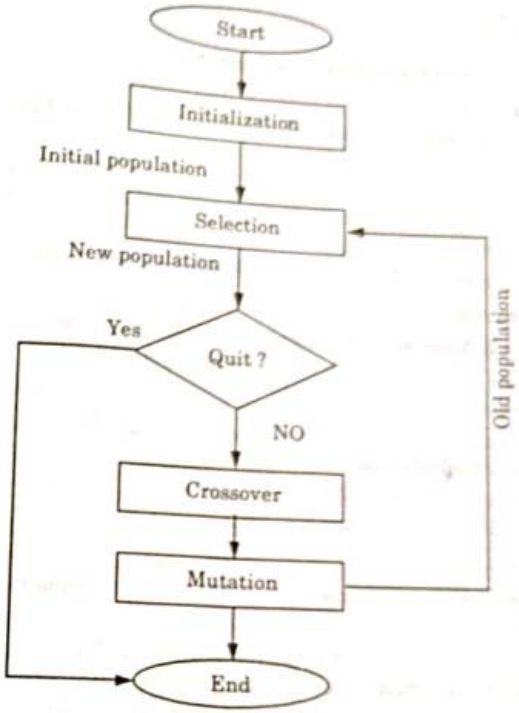


Fig. 1.24.1.

PART-4

Issues in Machine Learning and Data Science Vs. Machine Learning.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.25. **Briefly explain the issues related with machine learning.**

Answer

Issues related with machine learning are :

1. **Data quality :**
   a. It is essential to have good quality data to produce quality ML algorithms and models.
   b. To get high-quality data, we must implement data evaluation, integration, exploration, and governance techniques prior to developing ML models.
   c. Accuracy of ML is driven by the quality of the data.
2. **Transparency :**
   a. It is difficult to make definitive statements on how well a model is going to generalize in new environments.
3. **Manpower :**
   a. Manpower means having data and being able to use it. This does not introduce bias into the model.
   b. There should be enough skill sets in the organization for software development and data collection.
4. **Other :**
   a. The most common issue with ML is people using it where it does not belong.
   b. Every time there is some new innovation in ML, we see overzealous engineers trying to use it where it's not really necessary.
   c. This used to happen a lot with deep learning and neural networks.
   d. Traceability and reproduction of results are two main issues.

Que 1.26. **What are the classes of problem in machine learning ?**

Answer

Common classes of problem in machine learning :

1. **Classification :**
   a. In classification data is labelled *i.e.*, it is assigned a class, for example, spam/non-spam or fraud/non-fraud.
   b. The decision being modelled is to assign labels to new unlabelled pieces of data.
   c. This can be thought of as a discrimination problem, modelling the differences or similarities between groups.
2. **Regression :**
   a. Regression data is labelled with a real value rather than a label.
   b. The decision being modelled is what value to predict for new unpredicted data.

3. **Clustering :**
   a. In clustering data is not labelled, but can be divided into groups based on similarity and other measures of natural structure in the data.
   b. For example, organising pictures by faces without names, where the human user has to assign names to groups, like iPhoto on the Mac.
4. **Rule extraction :**
   a. In rule extraction, data is used as the basis for the extraction of propositional rules.
   b. These rules discover statistically supportable relationships between attributes in the data.

Que 1.27. **Differentiate between data science and machine learning.**

Answer

S. No.	Data science	Machine learning
1.	Data science is a concept used to tackle big data and includes data cleansing, preparation, and analysis.	Machine learning is defined as the practice of using algorithms to use data, learn from it and then forecast future trends for that topic.
2.	It includes various data operations.	It includes subset of Artificial Intelligence.
3.	Data science works by sourcing, cleaning, and processing data to extract meaning out of it for analytical purposes.	Machine learning uses efficient programs that can use data without being explicitly told to do so.
4.	SAS, Tableau, Apache, Spark, MATLAB are the tools used in data science.	Amazon Lex, IBM Watson Studio, Microsoft Azure ML Studio are the tools used in ML.
5.	Data science deals with structured and unstructured data.	Machine learning uses statistical models.
6.	Fraud detection and healthcare analysis are examples of data science.	Recommendation systems such as Spotify and Facial Recognition are examples of machine learning.

☺☺☺

2 UNIT

Regression and Bayesian Learning

CONTENTS

PART-1

Regression, Linear Regression and Logistic Regression.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 2.1. **Define the term regression with its type.**

Answer

1. Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).
2. Regression helps investment and financial managers to value assets and understand the relationships between variables, such as commodity prices and the stocks of businesses dealing in those commodities.

There are two type of regression :

a. Simple linear regression : It uses one independent variable to explain or predict the outcome of dependent variable Y.

$$Y = a + bX + u$$

b. Multiple linear regression : It uses two or more independent variables to predict outcomes.

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_tX_t + u$$

Where :

Y = The variable we you are trying to predict (dependent variable).

X = The variable that we are using to predict Y (independent variable).

a = The intercept.

b = The slope.

u = The regression residual.

Que 2.2. **Describe briefly linear regression.**

Answer

1. Linear regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope.
2. It is used to predict values within a continuous range, (for example : sales, price) rather than trying to classify them into categories (for example : cat, dog).

3. Following are the types of linear regression :

a. **Simple regression :**

i. Simple linear regression uses traditional slope-intercept form to produce accurate prediction, $y = mx + b$

where, m and b are the variables,

x represents our input data and y represents our prediction.

b. **Multivariable regression :**

i. A multi-variable linear equation is given below, where w represents the coefficients, or weights :

$$f(x, y, z) = w_1x + w_2y + w_3z$$

ii. The variables x, y, z represent the attributes, or distinct pieces of information that, we have about each observation.

iii. For sales predictions, these attributes might include a company's advertising spend on radio, TV, and newspapers.

$$\text{Sales} = w_1 \text{ Radio} + w_2 \text{ TV} + w_3 \text{ Newspapers}$$

Que 2.3. Explain logistics regression.

Answer

1. Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable.
2. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.
3. The dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).
4. A logistic regression model predicts $P(Y = 1)$ as a function of X. It is one of the simplest *ML* algorithms that can be used for various classification problems such as spam detection, diabetes prediction, cancer detection etc.

Que 2.4. What are the types of logistics regression ?

Answer

Logistics regression can be divided into following types :

1. **Binary (Binomial) Regression :**
   a. In this classification, a dependent variable will have only two possible types either 1 and 0.
   b. For example, these variables may represent success or failure, yes or no, win or loss etc.
2. **Multinomial regression :**
   a. In this classification, dependent variable can have three or more possible unordered types or the types having no quantitative significance.
   b. For example, these variables may represent "Type *A*" or "Type *B*" or "Type *C*".

3. **Ordinal regression :**
   a. In this classification, dependent variable can have three or more possible ordered types or the types having a quantitative significance.
   b. For example, these variables may represent "poor" or "good", "very good", "Excellent" and each category can have the scores like 0, 1, 2, 3.

Que 2.5. Differentiate between linear regression and logistics regression.

Answer

S. No.	Linear regression	Logistics regression
1.	Linear regression is a supervised regression model.	Logistic regression is a supervised classification model.
2.	In Linear regression, we predict the value by an integer number.	In Logistic regression, we predict the value by 1 or 0.
3.	No activation function is used.	Activation function is used to convert a linear regression equation to the logistic regression equation.
4.	A threshold value is added.	No threshold value is needed.
5.	It is based on the least square estimation.	The dependent variable consists of only two categories.
6.	Linear regression is used to estimate the dependent variable in case of a change in independent variables.	Logistic regression is used to calculate the probability of an event.
7.	Linear regression assumes the normal or gaussian distribution of the dependent variable.	Logistic regression assumes the binomial distribution of the dependent variable.

PART-2

Bayesian Learning, Bayes Theorem, Concept Learning, Bayes Optimal Classifier, Naive Bayes Classifier, Bayesian Belief Networks, EM Algorithm.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 2.6. Explain Bayesian learning. Explain two category classification.

Answer

Bayesian learning :

1. Bayesian learning is a fundamental statistical approach to the problem of pattern classification.
2. This approach is based on quantifying the tradeoffs between various classification decisions using probability and costs that accompany such decisions.
3. Because the decision problem is solved on the basis of probabilistic terms, hence it is assumed that all the relevant probabilities are known.
4. For this we define the state of nature of the things present in the particular pattern. We denote the state of nature by ω.

Two category classification :

1. Let ω_1, ω_2 be the two classes of the patterns. It is assumed that the a priori probabilities $p(\omega_1)$ and $p(\omega_2)$ are known.
2. Even if they are not known, they can easily be estimated from the available training feature vectors.
3. If N is total number of available training patterns and N_1, N_2 of them belong to ω_1 and ω_2, respectively then $p(\omega_1) = N_1/N$ and $p(\omega_2) = N_2/N$.
4. The conditional probability density functions $p(x \mid \omega_i)$, $i = 1, 2$ is also assumed to be known which describes the distribution of the feature vectors in each of the classes.
5. The feature vectors can take any value in the l-dimensional feature space.
6. Density functions $p(x \mid \omega_i)$ become probability and will be denoted by $p(x \mid \omega_i)$ when the feature vectors can take only discrete values.
7. Consider the conditional probability,

$$p(\omega_i \mid x) = \frac{p(x \mid \omega_i) p(\omega_i)}{p(x)} \quad \ldots(2.6.1)$$

where $p(x)$ is the probability density function of x and for which we have

$$p(x) = \sum_{i=1}^{2} p(x \mid \omega_i) p(\omega_i) \quad \ldots(2.6.2)$$

8. Now, the Baye's classification rule can be defined as :
   a. If $p(\omega_1 \mid x) > p(\omega_2 \mid x)$ x is classified to ω_1
   b. If $p(\omega_1 \mid x) < p(\omega_2 \mid x)$ x is classified to ω_2 ...(2.6.3)
9. In the case of equality the pattern can be assigned to either of the two classes. Using equation (2.6.1), decision can equivalently be based on the inequalities :
   a. $p(x \mid \omega_1) p(\omega_1) > p(x \mid \omega_2) p(\omega_2)$
   b. $p(x \mid \omega_1) p(\omega_1) < p(x \mid \omega_2) p(\omega_2)$ ...(2.6.4)
10. Here $p(x)$ is not taken because it is same for all classes and it does not affect the decision.
11. Further, if the priori probabilities are equal, *i.e.*,
    a. $p(\omega_1) = p(\omega_2) = 1/2$ then Eq. (2.6.4) becomes,
    b. $p(x \mid \omega_1) > p(x \mid \omega_2)$
    c. $p(x \mid \omega_1) < p(x \mid \omega_2)$
12. For example, in Fig. 2.6.1, two equiprobable classes are presented which shows the variations of $p(x \mid \omega_i)$, $i = 1, 2$ as functions of x for the simple case of a single feature ($l = 1$).
13. The dotted line at x_0 is a threshold which partitions the space into two regions, R_1 and R_2. According to Baye's decisions rule, for all value of x in R_1 the classifier decides ω_1 and for all values in R_2 it decides ω_2.
14. From the Fig. 2.6.1, it is obvious that the errors are unavoidable. There is a finite probability for an x to lie in the R_2 region and at the same time to belong in class ω_1. Then there is error in the decision.

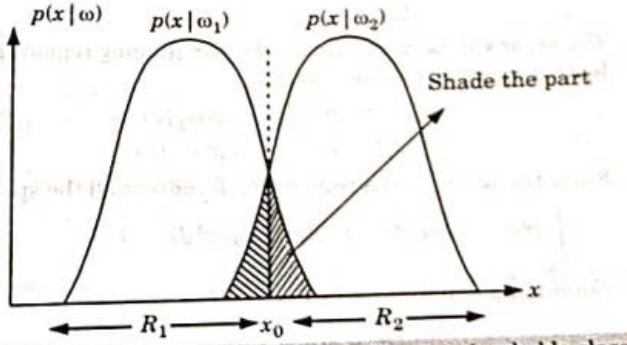


Fig. 2.6.1. Bayesian classifier for the case of two equiprobable classes.

15. The total probability, P of committing a decision error for two equiprobable classes is given by,

$$P_e = \frac{1}{2} \int_{-\infty}^{x_0} p(x \mid \omega_2) dx + \frac{1}{2} \int_{x_0}^{1-\infty} p(x \mid \omega_1) dx$$

which is equal to the total shaded area under the curves in Fig. 2.6.1.

Que 2.7. Explain how the decision error for Bayesian classification can be minimized.

Answer

1. Bayesian classifier can be made optimal by minimizing the classification error probability.
2. In Fig. 2.7.1, it is observed that when the threshold is moved away from x_0, the corresponding shaded area under the curves always increases.
3. Hence, we have to decrease this shaded area to minimize the error.
4. Let R_1 be the region of the feature space for ω_1 and R_2 be the corresponding region for ω_2.
5. Then an error will be occurred if, $x \in R_1$ although it belongs to ω_2 or if $x \in R_2$ although it belongs to ω_1 i.e.,

$$P_e = p(x \in R_2, \omega_1) + p(x \in R_1, \omega_2) \quad \ldots(2.7.1)$$

6. P_e can be written as,

$$P_e = p(x \in R_2 | \omega_1) p(\omega_1) + p(x \in R_1 | \omega_2) p(\omega_2)$$

$$= P(\omega_1) \int_{R_2} p(x | \omega_1) dx + p(\omega_2) \int_{R_1} p(x | \omega_2) dx \quad \ldots(2.7.2)$$

7. Using the Baye's rule,

$$= P \int_{R_2} p(\omega_1 | x) p(x) dx + \int_{R_1} p(\omega_2 | x) p(x) dx \quad \ldots(2.7.3)$$

8. The error will be minimized if the partitioning regions R_1 and R_2 of the feature space are chosen so that

$$R_1 : p(\omega_1 | x) > p(\omega_2 | x)$$

$$R_2 : p(\omega_2 | x) > p(\omega_1 | x) \quad \ldots(2.7.4)$$

9. Since the union of the regions R_1, R_2 covers all the space, we have

$$\int_{R_1} p(\omega_1 | x) p(x) dx + \int_{R_2} p(\omega_1 | x) p(x) dx = 1 \quad \ldots(2.7.5)$$

10. Combining equation (2.7.3) and (2.7.5), we get,

$$P_e = p(\omega_1) \int_{R_1} (p(\omega_1 | x) - p(\omega_2 | x)) p(x) dx \quad \ldots(2.7.6)$$

11. Thus, the probability of error is minimized if R_1 is the region of space in which $p(\omega_1 | x) > p(\omega_2 | x)$. Then R_2 becomes region where the reverse is true.

12. In a classification task with M classes, $\omega_1, \omega_2, \ldots, \omega_M$ an unknown pattern, represented by the feature vector x, is assigned to class ω_i if $p(\omega_i | x) > p(\omega_j | x) \; \forall \; j \neq i$.

Que 2.8. Consider the Bayesian classifier for the uniformly distributed classes, where :

$$P(x/w_1) = \begin{cases} \dfrac{1}{a_2 - a_1}, & x \in [a_1, a_2] \\ 0, & \text{muullion} \end{cases}$$

$$P(x/w_2) = \begin{cases} \dfrac{1}{b_2 - b_1}, & x \in [b_1, b_2] \\ 0, & \text{muullion} \end{cases}$$

Show the classification results for some values for a and b ("muullion" means "otherwise").

Answer

Typical cases are presented in the Fig. 2.8.1.

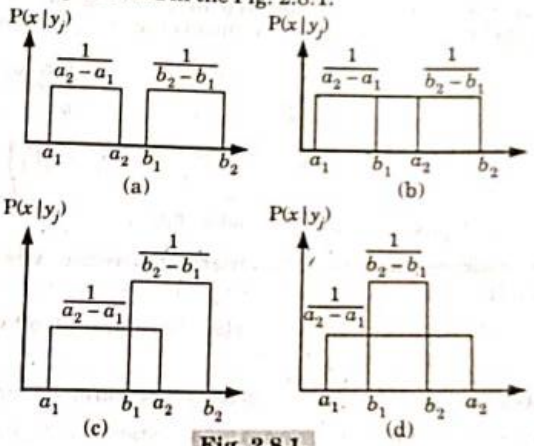


Fig. 2.8.1.

Que 2.9. Define Bayes classifier. Explain how classification is done by using Bayes classifier.

Answer

1. A Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem (from Bayesian statistics) with strong (Naive) independence assumptions.

2. A Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature.
3. Depending on the precise nature of the probability model, Naive Bayes classifiers can be trained very efficiently in a supervised learning.
4. In many practical applications, parameter estimation for Naive Bayes models uses the method of maximum likelihood; in other words, one can work with the Naive Bayes model without believing in Bayesian probability or using any Bayesian methods.
5. An advantage of the Naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification.
6. The perceptron bears a certain relationship to a classical pattern classifier known as the Bayes classifier.
7. When the environment is Gaussian, the Bayes classifier reduces to a linear classifier.

In the Bayes classifier, or Bayes hypothesis testing procedure, we minimize the average risk, denoted by R. For a two-class problem, represented by classes C_1 and C_2, the average risk is defined :

$$R = C_{11}P_1\int_{H_1} P_x(x/C_1)dx + C_{22}P_2\int_{H_2} P_x(x/C_2)dx + C_{21}P_1\int_{H_2} P_x(x/C_1)dx + C_{12}P_2\int_{H_1} P_x(x/C_2)dx$$

where the various terms are defined as follows :

P_i = Prior probability that the observation vector x is drawn from subspace H_i, with $i = 1, 2$, and $P_1 + P_2 = 1$

C_{ij} = Cost of deciding in favour of class C_i represented by subspace H_i when class C_j is true, with $i, j = 1, 2$

$P_x(x/C_i)$ = Conditional probability density function of the random vector X

8. Fig. 2.9.1(*a*) depicts a block diagram representation of the Bayes classifier. The important points in this block diagram are two fold :
   a. The data processing in designing the Bayes classifier is confined entirely to the computation of the likelihood ratio $\wedge(x)$.
   b. This computation is completely invariant to the values assigned to the prior probabilities and involved in the decision-making process. These quantities merely affect the values of the threshold x.
   c. From a computational point of view, we find it more convenient to work with logarithm of the likelihood ratio rather than the likelihood ratio itself.

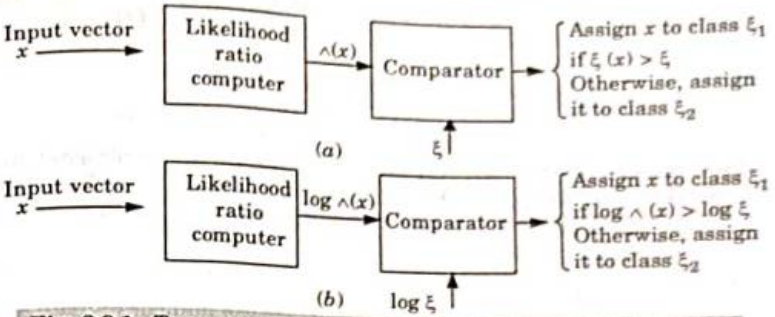


Fig. 2.9.1. Two equivalent implementations of the Bayes classifier : (*a*) Likelihood ratio test, (*b*) Log-likelihood ratio test

Que 2.10. **Discuss Bayes classifier using some example in detail.**

Answer

Bayes classifier : Refer Q. 2.9, Page 2–8L, Unit-2.

For example :

1. Let D be a training set of features and their associated class labels. Each feature is represented by an n-dimensional attribute vector $X = (x_1, x_2, ..., x_n)$ depicting n measurements made on the feature from n attributes, respectively $A_1, A_2, ..., A_n$.
2. Suppose that there are m classes, $C_1, C_2, ..., C_m$. Given a feature X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, classifier predicts that X belongs to class C_i if and only if,

   $$p(C_i|X) > p(C_j|X) \text{ for } 1 \le j \le m,\ j \ne i$$

   Thus, we maximize $p(C_i|X)$. The class C_i for which $p(C_i|X)$ is maximized is called the maximum posterior hypothesis. By Bayes theorem,

   $$p(C_i|X) = \frac{p(X|C_i)p(C_i)}{p(X)}$$

3. As $p(X)$ is constant for all classes, only $P(X|C_i)$ $P(C_i)$ need to be maximized. If the class prior probabilities are not known then it is commonly assumed that the classes are equally likely *i.e.*, $p(C_1) = p(C_2) = ... p(C_m)$ and therefore $p(X|C_i)$ is maximized. Otherwise $p(X|C_i)\,p(C_i)$ is maximized.
4. i. Given data sets with many attributes, the computation of $p(X|C_i)$ will be extremely expensive.
   ii. To reduce computation in evaluating $p(X|C_i)$, the assumption of class conditional independence is made.

iii. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the feature.

Thus, $p(X|C_i) = \prod_{k=1}^{n} p(x_k|C_i)$

$= p(x_1|C_2) \times p(x_2|C_2)x.... \times p(x_n|C_i)$

iv. The probabilities $p(x_1|C_i), p(x_2|C_i),...., p(x_n|C_i)$ are easily estimated from the training feature. Here x_k refers to the value of attribute A_k for each attribute, it is checked whether the attribute is categorical or continuous valued.

v. For example, to compute $p(X|C_i)$ we consider,

a. If A_k is categorical then $p(x_k|C_i)$ is the number of feature of class C_i in D having the value x_k for A_k divided by $|C_i, D|$, the number of features of class C_i in D.

b. If A_k is continuous valued then continuous valued attribute is typically assumed to have a Gaussian distribution with a mean μ and standard deviation σ, defined by,

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\left[\frac{-1(x-\mu)^2}{2 \quad \sigma^2}\right]}$$

so that $p(x_k|C_i) = g(x_k)$.

vi. There is a need to compute the mean μ and the standard deviation σ of the value of attribute A_k for training set of class C_i. These values are used to estimate $p(x_k|C_i)$.

vii. For example, let X = (35, Rs. 40,000) where A_1 and A_2 are the attributes age and income, respectively. Let the class label attribute be buys-computer.

viii. The associated class label for X is yes (*i.e.*, buys-computer = yes). Let's suppose that age has not been discretized and therefore exists as a continuous valued attribute.

ix. Suppose that from the training set, we find that customer in D who buy a computer are 38 ± 12 years of age. In other words, for attribute age and this class, we have $\mu = 38$ and $\sigma = 12$.

5. In order to predict the class label of X, $p(X|C_i)\,p(C_i)$ is evaluated for each class C_i. The classifier predicts that the class label of X is the class C_i, if and only if

$p(X|C_i)\,P(C_i) > p(X|C_j)\,p(C_j)$ for $1 \le j \le m, j \ne i$,

The predicted class label is the class C_i for which $p(X|C_i)\,P(C_i)$ is the maximum.

Que 2.11. Let blue, green, and red be three classes of objects with prior probabilities given by P(blue) = 1/4, P(green) = 1/2, P(red) = 1/4. Let there be three types of objects pencils, pens, and paper. Let the class-conditional probabilities of these objects be given as follows. Use Bayes classifier to classify pencil, pen and paper.

P(pencil/green) = 1/3	P(pen/green) = 1/2	P(paper/green) = 1/6
P(pencil/blue) = 1/2	P(pen/blue) = 1/6	P(paper/blue) = 1/3
P(pencil/red) = 1/6	P(pen/red) = 1/3	P(paper/red) = 1/2

Answer

As per Bayes rule :

$$\text{P(green/pencil)} = \frac{\text{P(pencil/ green) P(green)}}{\text{(P(pencil/ green) P(green) + P(pencil/ blue) P(blue) + P(pencil/ red) P(red)}}$$

$$= \frac{\frac{1}{3} \times \frac{1}{2}}{\left(\frac{1}{3} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{4} + \frac{1}{6} \times \frac{1}{4}\right)} = \frac{\frac{1}{6}}{0.33} = 0.5050$$

$$\text{P(blue/pencil)} = \frac{\text{P(pencil/ blue) P(blue)}}{\text{(P(pencil/ green) P(green) + P(pencil/ blue) P(blue) + P(pencil/ red) P(red)}}$$

$$= \frac{\frac{1}{2} \times \frac{1}{4}}{0.33} = 0.378$$

$$\text{P(red/pencil)} = \frac{\text{P(pencil/ red) P(red)}}{\text{(P(pencil/ red) P(red) + P(pencil/ blue) P(blue) + P(pencil/ green) P(green)}}$$

$$= \frac{\frac{1}{6} \times \frac{1}{4}}{0.33} = \frac{\frac{1}{24}}{0.33} = 0.126$$

Since, P(green/pencil) has the highest value therefore pencil belongs to class green.

$$\text{P(green/pen)} = \frac{\text{P(pen/ green) P(green)}}{\text{P(pen/ green) P(green) + P(pen/ blue) P(blue) + P(pen/ red) P(red)}}$$

$$= \frac{\frac{1}{2} \times \frac{1}{2}}{\frac{1}{2} \times \frac{1}{2} + \frac{1}{6} \times \frac{1}{4} + \frac{1}{3} \times \frac{1}{4}} = \frac{\frac{1}{4}}{0.375} = 0.666$$

$$P(\text{blue/pen}) = \frac{P(\text{pen/ blue})P(\text{blue})}{P(\text{pen/ green})\,P(\text{green}) + P(\text{pen/ blue})\,P(\text{blue}) + P(\text{pen/ red})\,P(\text{red})}$$

$$= \frac{\frac{1}{6}\times\frac{1}{4}}{0.375} = \frac{\frac{1}{24}}{0.375} = 0.111$$

$$P(\text{red/pen}) = \frac{P(\text{pen/ red})\,P(\text{red})}{P(\text{pen/ green})\,P(\text{green}) + P(\text{pen/ blue})\,P(\text{blue}) + P(\text{pen/ red})\,P(\text{red})}$$

$$= \frac{\frac{1}{3}\times\frac{1}{4}}{0.375} = \frac{\frac{1}{12}}{0.375} = 0.222$$

Since P(green/pen) has the highest value therefore, pen belongs to class green.

$$P(\text{green/paper}) = \frac{P(\text{paper/ green})\,P(\text{green})}{P(\text{paper/ green})\,P(\text{green}) + P(\text{paper/ blue})\,P(\text{blue}) + P(\text{paper/ red})\,P(\text{red})}$$

$$= \frac{\frac{1}{6}\times\frac{1}{2}}{\frac{1}{6}\times\frac{1}{2}+\frac{1}{3}\times\frac{1}{4}+\frac{1}{2}\times\frac{1}{4}} = \frac{\frac{1}{12}}{\frac{1}{12}+\frac{1}{12}+\frac{1}{8}}$$

$$= \frac{\frac{1}{12}}{0.291} = 0.286$$

$$P(\text{blue/paper}) = \frac{P(\text{paper/ blue})\,P(\text{blue})}{P(\text{paper/ green})\,P(\text{green}) + P(\text{paper/ blue})\,P(\text{blue}) + P(\text{paper/ red})\,P(\text{red})}$$

$$= \frac{\frac{1}{3}\times\frac{1}{4}}{0.291} = \frac{\frac{1}{12}}{0.291} = 0.286$$

$$P(\text{red/paper}) = \frac{P(\text{paper/ red})\,P(\text{red})}{P(\text{paper/ green})\,P(\text{green}) + P(\text{paper/ blue})\,P(\text{blue}) + P(\text{paper/ red})\,P(\text{red})}$$

$$= \frac{\frac{1}{2}\times\frac{1}{4}}{0.291} = \frac{\frac{1}{8}}{0.291} = 0.429$$

Since, P(red/paper) has the highest value therefore, paper belongs to class red.

Que 2.12. **Explain Naive Bayes classifier.**

Answer

1. Naive Bayes model is the most common Bayesian network model used in machine learning.
2. Here, the class variable C is the root which is to be predicted and the attribute variables X_i are the leaves.
3. The model is Naive because it assumes that the attributes are conditionaily independent of each other, given the class.

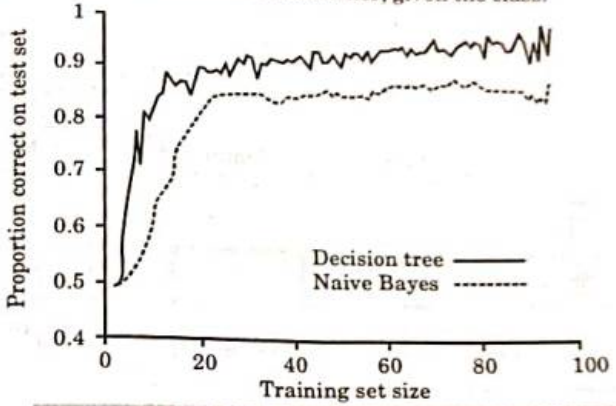


Fig. 2.12.1. The learning curve for Naive Bayes learning.

4. Assuming Boolean variables, the parameters are :

$$\theta = P(C = \text{true}),\ \theta_{i1} = P(X_i = \text{true} \mid C = \text{true}),$$

$$\theta_{i2} = P(X_i = \text{true} \mid C = \text{False})$$

5. Naive Bayes models can be viewed as Bayesian networks in which each X_i has C as the sole parent and C has no parents.
6. A Naive Bayes model with gaussian $P(X_i \mid C)$ is equivalent to a mixture of gaussians with diagonal covariance matrices.
7. While mixtures of gaussians are used for density estimation in continuous domains, Naive Bayes models used in discrete and mixed domains.
8. Naive Bayes models allow for very efficient inference of marginal and conditional distributions.
9. Naive Bayes learning has no difficulty with noisy data and can give more appropriate probabilistic predictions.

Que 2.13. **Consider a two-class (Tasty or non-Tasty) problem with the following training data. Use Naive Bayes classifier to classify the pattern :**
"Cook = Asha, Health-Status = Bad, Cuisine = Continental".

Cook	Health-Status	Cuisine	Tasty
Asha	Bad	Indian	Yes
Asha	Good	Continental	Yes
Sita	Bad	Indian	No
Sita	Good	Indian	Yes
Usha	Bad	Indian	Yes
Usha	Bad	Continental	No
Sita	Bad	Continental	No
Sita	Good	Continental	Yes
Usha	Good	Indian	Yes
Usha	Good	Continental	No

Answer

Cook			Health-status			Cuisine		
	Yes	No		Yes	No		Yes	No
Asha	2	0	Bad	2	3	Indian	4	1
Sita	2	2	Good	4	1	Continental	2	3
Usha	2	2						

Tasty	
Yes	No
6	4

Cook			Health-status			Cuisine		
	Yes	No		Yes	No		Yes	No
Asha	2/6	0	Bad	2/6	3/4	Indian	4/6	1/4
Sita	2/6	2/4	Good	4/6	1/4	Continental	2/6	3/4
Usha	2/6	2/4						

Tasty	
Yes	No
6/10	4/10

Likelihood of yes = $\frac{2}{6} \times \frac{2}{6} \times \frac{2}{6} \times \frac{6}{10} = 0.023$

Likelihood of no = $0 \times \frac{3}{4} \times \frac{3}{4} \times \frac{4}{10} = 0$

Therefore, the prediction is tasty.

Que 2.14. Explain EM algorithm with steps.

Answer

1. The Expectation-Maximization (EM) algorithm is an iterative way to find maximum-likelihood estimates for model parameters when the data is incomplete or has missing data points or has some hidden variables.
2. EM chooses random values for the missing data points and estimates a new set of data.
3. These new values are then recursively used to estimate a better first data, by filling up missing points, until the values get fixed.
4. These are the two basic steps of the EM algorithm :
   - **a. Estimation Step :**
     - i. Initialize μ_k, Σ_k and π_k by random values, or by K means clustering results or by hierarchical clustering results.
     - ii. Then for those given parameter values, estimate the value of the latent variables (*i.e.*, γ_k).
   - **b. Maximization Step :** Update the value of the parameters (*i.e.*, μ_k, Σ_k and π_k) calculated using ML method :
     - i. Initialize the mean μ_k, the covariance matrix Σ_k and the mixing coefficients π_k by random values, (or other values).
     - ii. Compute the π_k values for all k.
     - iii. Again estimate all the parameters using the current π_k values.
     - iv. Compute log-likelihood function.
     - v. Put some convergence criterion.
     - vi. If the log-likelihood value converges to some value (or if all the parameters converge to some values) then stop, else return to Step 2.

Que 2.15. **Describe the usage, advantages and disadvantages of EM algorithm.**

Answer

Usage of EM algorithm :

1. It can be used to fill the missing data in a sample.
2. It can be used as the basis of unsupervised learning of clusters.
3. It can be used for the purpose of estimating the parameters of Hidden Markov Model (HMM).
4. It can be used for discovering the values of latent variables.

Advantages of EM algorithm are :

1. It is always guaranteed that likelihood will increase with each iteration.
2. The E-step and M-step are often pretty easy for many problems in terms of implementation.
3. Solutions to the M-steps often exist in the closed form.

Disadvantages of EM algorithm are :

1. It has slow convergence.
2. It makes convergence to the local optima only.
3. It requires both the probabilities, forward and backward (numerical optimization requires only forward probability).

Que 2.16. **Write a short note on Bayesian network.**

OR

Explain Bayesian network by taking an example. How is the Bayesian network powerful representation for uncertainty knowledge ?

Answer

1. A Bayesian network is a directed acyclic graph in which each node is annotated with quantitative probability information.
2. The full specification is as follows :
   - i. A set of random variables makes up the nodes of the network variables may be discrete or continuous.
   - ii. A set of directed links or arrows connects pairs of nodes. If there is an arrow from x to node y, x is said to be a parent of y.
   - iii. Each node x_i has a conditional probability distribution $P(x_i \mid \text{parent}(x_i))$ that quantifies the effect of parents on the node.
   - iv. The graph has no directed cycles (and hence is a directed acyclic graph or DAG).

3. A Bayesian network provides a complete description of the domain. Every entry in the full joint probability distribution can be calculated from the information in the network.
4. Bayesian networks provide a concise way to represent conditional independence relationships in the domain.
5. A Bayesian network is often exponentially smaller than the full joint distribution.

For example :

1. Suppose we want to determine the possibility of grass getting wet or dry due to the occurrence of different seasons.
2. The weather has three states : Sunny, Cloudy, and Rainy. There are two possibilities for the grass : Wet or Dry.
3. The sprinkler can be on or off. If it is rainy, the grass gets wet but if it is sunny, we can make grass wet by pouring water from a sprinkler.
4. Suppose that the grass is wet. This could be contributed by one of the two reasons - Firstly, it is raining. Secondly, the sprinklers are turned on.
5. Using the Baye's rule, we can deduce the most contributing factor towards the wet grass.

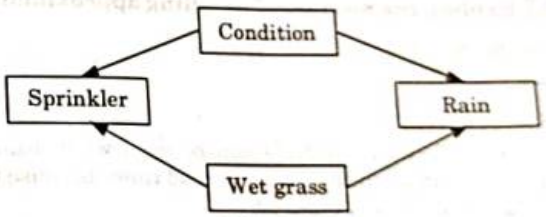


Fig. 2.16.1.

Bayesian network possesses the following merits in uncertainty knowledge representation :

1. Bayesian network can conveniently handle incomplete data.
2. Bayesian network can learn the casual relation of variables. In data analysis, casual relation is helpful for field knowledge understanding, it can also easily lead to precise prediction even under much interference.
3. The combination of bayesian network and bayesian statistics can take full advantage of field knowledge and information from data.
4. The combination of bayesian network and other models can effectively avoid over-fitting problem.

Que 2.17. Explain the role of prior probability and posterior probability in Bayesian classification.

Answer

Role of prior probability :

1. The prior probability is used to compute the probability of the event before the collection of new data.
2. It is used to capture our assumptions / domain knowledge and is independent of the data.
3. It is the unconditional probability that is assigned before any relevant evidence is taken into account.

Role of posterior probability :

1. Posterior probability is used to compute the probability of an event after collection of data.
2. It is used to capture both the assumptions / domain knowledge and the pattern in observed data.
3. It is the conditional probability that is assigned after the relevant evidence or background is taken into account.

Que 2.18. Explain the method of handling approximate inference in Bayesian networks.

Answer

1. Approximate inference methods can be used when exact inference methods lead to unacceptable computation times because the network is very large or densely connected.
2. Methods handling approximate inference :
   - i. **Simulation methods :** This method use the network to generate samples from the conditional probability distribution and estimate conditional probabilities of interest when the number of samples is sufficiently large.
   - ii. **Variational methods :** This method express the inference task as a numerical optimization problem and then find upper and lower bounds of the probabilities of interest by solving a simplified version of this optimization problem.

PART-3

Support Vector Machine, Introduction, Types of Support Vector Kernel - (Linear Kernel Polynomial Kernel, and Gaussian Kernel), Hyperplane : (Decision Surface), Properties of SVM, and Issues in SVM.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 2.19. Write short note on support vector machine.

Answer

Refer Q. 1.23, Page 1–23L, Unit-1.

Que 2.20. What are the types of support vector machine ?

Answer

Following are the types of support vector machine :

1. **Linear SVM :** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
2. **Non-linear SVM :** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

Que 2.21. What is polynomial kernel ? Explain polynomial kernel using one dimensional and two dimensional.

Answer

1. The polynomial kernel is a kernel function used with Support Vector Machines (SVMs) and other kernelized models, that represents the

similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.

2. Polynomial kernel function is given by the equation :

$$(a \times b + r)^d$$

where, a and b are two different data points that we need to classify.

r determines the coefficients of the polynomial.

d determines the degree of the polynomial.

3. We perform the dot products of the data points, which gives us the high dimensional coordinates for the data.
4. When $d = 1$, the polynomial kernel computes the relationship between each pair of observations in 1-Dimension and these relationships help to find the support vector classifier.
5. When $d = 2$, the polynomial kernel computes the 2-Dimensional relationship between each pair of observations which help to find the support vector classifier.

Que 2.22. Describe Gaussian Kernel (Radial Basis Function).

Answer

1. RBF kernel is a function whose value depends on the distance from the origin or from some point.
2. Gaussian Kernel is of the following format :

$$K(X_1, X_2,) = \text{exponent}\,(-\gamma \| X_1 - X_2 \|^2)$$

$\| X_1 - X_2 \|$ = Euclidean distance between X_1 and X_2

Using the distance in the original space we calculate the dot product (similarity) of X_1 and X_2.

3. Following are the parameters used in Gaussain Kernel:

a. **C :** Inverse of the strength of regularization.

Behavior : As the value of 'c' increases the model gets overfits.

As the value of 'c' decreases the model underfits.

b. **γ :** Gamma (used only for RBF kernel)

Behavior : As the value of 'γ' increases the model gets overfits.

As the value of 'γ' decreases the model underfits.

Que 2.23. Write short note on hyperplane (Decision surface).

Answer

1. A hyperplane in an n-dimensional Euclidean space is a flat, n-1 dimensional subset of that space that divides the space into two disconnected parts.
2. For example let's assume a line to be one dimensional Euclidean space.
3. Now pick a point on the line, this point divides the line into two parts.
4. The line has 1 dimension, while the point has 0 dimensions. So a point is a hyperplane of the line.
5. For two dimensions we saw that the separating line was the hyperplane.
6. Similarly, for three dimensions a plane with two dimensions divides the 3d space into two parts and thus act as a hyperplane.
7. Thus for a space of n dimensions we have a hyperplane of n-1 dimensions separating it into two parts.

Que 2.24. What are the advantages and disadvantags of SVM ?

Answer

Advantages of SVM are :

1. **Guaranteed optimality :** Owing to the nature of Convex Optimization, the solution will always be global minimum, not a local minimum.
2. **The abundance of implementations :** We can access it conveniently.
3. SVM can be used for linearly separable as well as non-linearly separable data. Linearly separable data pases hard margin whereas non-linearly separable data poses a soft margin.
4. SVMs provide compliance to the semi-supervised learning models. It can be used in areas where the data is labeled as well as unlabeled. It only requires a condition to the minimization problem which is known as the transductive SVM.
5. Feature Mapping used to be quite a load on the computational complexity of the overall training performance of the model. However, with the help of Kernel Trick, SVM can carry out the feature mapping using the simple dot product.

Disadvantages of SVM :

1. SVM does not give the best performance for handling text structures as compared to other algorithms that are used in handling text data. This leads to loss of sequential information and thereby, leading to worse performance.

2. SVM cannot return the probabilistic confidence value that is similar to logistic regression. This does not provide much explanation as the confidence of prediction is important in several applications.
3. The choice of the kernel is perhaps the biggest limitation of the support vector machine. Considering so many kernels present, it becomes difficult to choose the right one for the data.

Que 2.25. Explain the properties of SVM.

Answer

Following are the properties of SVM :

1. **Flexibility in choosing a similarity function :** Sparseness of solution when dealing with large data sets only support vectors are used to specify the separating hyperplane
2. **Ability to handle large feature spaces :** complexity does not depend on the dimensionality of the feature space
3. **Overfitting can be controlled by soft margin approach :** A simple convex optimization problem which is guaranteed to converge to a single global solution

Que 2.26. What are the parameters used in support vector classifier ?

Answer

Parameters used in support vector classifier are :

1. **Kernel :**
   a. Kernel, is selected based on the type of data and also the type of transformation.
   b. By default, the kernel is Radial Basis Function Kernel (RBF).
2. **Gamma :**
   a. This parameter decides how far the influence of a single training example reaches during transformation, which in turn affects how tightly the decision boundaries end up surrounding points in the input space.
   b. If there is a small value of gamma, points farther apart are considered similar.
   c. So, more points are grouped together and have smoother decision boundaries (may be less accurate).
   d. Larger values of gamma cause points to be closer together (may cause overfitting).

3. **The 'C' parameter :**
   a. This parameter controls the amount of regularization applied on the data.
   b. Large values of C mean low regularization which in turn causes the training data to fit very well (may cause overfitting).
   c. Lower values of C mean higher regularization which causes the model to be more tolerant of errors (may lead to lower accuracy).

3 UNIT Decision Tree Learning

CONTENTS

PART-1

Decision Tree Learning, Decision Tree Learning Algorithm, Inductive Bias, Inductive Inference with Decision Trees.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.1. **Describe the basic terminology used in decision tree.**

Answer

Basic terminology used in decision trees are :

1. **Root node :** It represents entire population or sample and this further gets divided into two or more homogeneous sets.
2. **Splitting :** It is a process of dividing a node into two or more sub-nodes.
3. **Decision node :** When a sub-node splits into further sub-nodes, then it is called decision node.

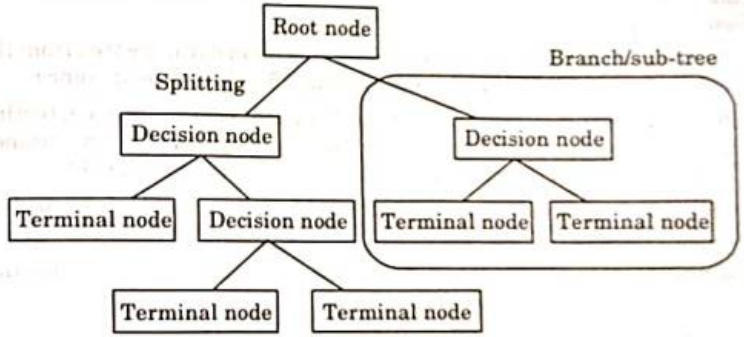


Fig. 3.1.1.

4. **Leaf / Terminal node :** Nodes that do not split is called leaf or terminal node.
5. **Pruning :** When we remove sub-nodes of a decision node, this process is called pruning. This process is opposite to splitting process.
6. **Branch / sub-tree :** A sub section of entire tree is called branch or sub-tree.
7. **Parent and child node :** A node which is divided into sub-nodes is called parent node of sub-nodes where as sub-nodes are the child of parent node.

Que 3.2. **Why do we use decision tree ?**

Answer

1. Decision trees can be visualized, simple to understand and interpret.
2. They require less data preparation whereas other techniques often require data normalization, the creation of dummy variables and removal of blank values.
3. The cost of using the tree (for predicting data) is logarithmic in the number of data points used to train the tree.
4. Decision trees can handle both categorical and numerical data whereas other techniques are specialized for only one type of variable.
5. Decision trees can handle multi-output problems.
6. Decision tree is a white box model *i.e.*, the explanation for the condition can be explained easily by Boolean logic because there are two outputs. For example yes or no.
7. Decision trees can be used even if assumptions are violated by the dataset from which the data is taken.

Que 3.3. **How can we express decision trees ?**

Answer

1. Decision trees classify instances by sorting them down the tree from the root to leaf node, which provides the classification of the instance.
2. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute as shown in Fig. 3.3.1.
3. This process is then repeated for the subtree rooted at the new node.
4. The decision tree in Fig. 3.3.1 classifies a particular morning according to whether it is suitable for playing tennis and returning the classification associated with the particular leaf.
5. For example, the instance

   (Outlook = Rain, Temperature = Hot, Humidity = High, Wind = Strong)

   would be sorted down the left most branch of this decision tree and would therefore be classified as a negative instance.
6. In other words, decision tree represent a disjunction of conjunctions of constraints on the attribute values of instances.

   (Outlook = Sunny $\wedge$ Humidity = Normal) $\vee$ (Outlook = Overcast) $\vee$ (Outlook = Rain $\wedge$ Wind = Weak)

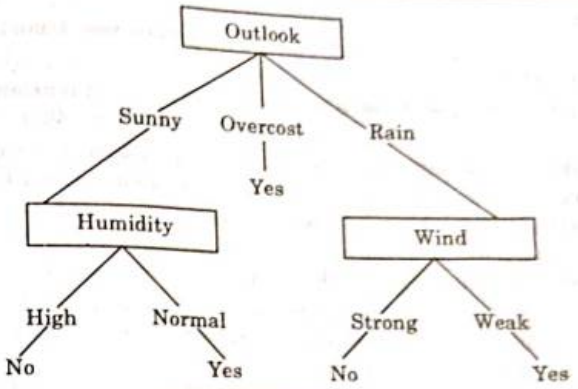


Fig. 3.3.1.

Que 3.4. **Explain various decision tree learning algorithms.**

Answer

Various decision tree learning algorithms are :

1. **ID3 (Iterative Dichotomiser 3) :**
   - i. ID3 is an algorithm used to generate a decision tree from a dataset.
   - ii. To construct a decision tree, ID3 uses a top-down, greedy search through the given sets, where each attribute at every tree node is tested to select the attribute that is best for classification of a given set.
   - iii. Therefore, the attribute with the highest information gain can be selected as the test attribute of the current node.
   - iv. In this algorithm, small decision trees are preferred over the larger ones. It is a heuristic algorithm because it does not construct the smallest tree.
   - v. For building a decision tree model, ID3 only accepts categorical attributes. Accurate results are not given by ID3 when there is noise and when it is serially implemented.
   - vi. Therefore data is preprocessed before constructing a decision tree.
   - vii. For constructing a decision tree information gain is calculated for each and every attribute and attribute with the highest information gain becomes the root node. The rest possible values are denoted by arcs.
   - viii. All the outcome instances that are possible are examined whether they belong to the same class or not. For the instances of the same class, a single name is used to denote the class otherwise the instances are classified on the basis of splitting attribute.

2. **C4.5 :**
   i. C4.5 is an algorithm used to generate a decision tree. It is an extension of ID3 algorithm.
   ii. C4.5 generates decision trees which can be used for classification and therefore C4.5 is referred to as statistical classifier.
   iii. It is better than the ID3 algorithm because it deals with both continuous and discrete attributes and also with the missing values and pruning trees after construction.
   iv. C5.0 is the commercial successor of C4.5 because it is faster, memory efficient and used for building smaller decision trees.
   v. C4.5 performs by default a tree pruning process. This leads to the formation of smaller trees, simple rules and produces more intuitive interpretations.
3. **CART (Classification And Regression Trees) :**
   i. CART algorithm builds both classification and regression trees.
   ii. The classification tree is constructed by CART through binary splitting of the attribute.
   iii. Gini Index is used for selecting the splitting attribute.
   iv. The CART is also used for regression analysis with the help of regression tree.
   v. The regression feature of CART can be used in forecasting a dependent variable given a set of predictor variable over a given period of time.
   vi. CART has an average speed of processing and supports both continuous and nominal attribute data.

Que 3.5. What are the advantages and disadvantages of different decision tree learning algorithm ?

Answer

Advantages of ID3 algorithm :

1. The training data is used to create understandable prediction rules.
2. It builds short and fast tree.
3. ID3 searches the whole dataset to create the whole tree.
4. It finds the leaf nodes thus enabling the test data to be pruned and reducing the number of tests.
5. The calculation time of ID3 is the linear function of the product of the characteristic number and node number.

Disadvantages of ID3 algorithm :

1. For a small sample, data may be overfitted or overclassified.

2. For making a decision, only one attribute is tested at an instant thus consuming a lot of time.
3. Classifying the continuous data may prove to be expensive in terms of computation, as many trees have to be generated to see where to break the continuous sequence.
4. It is overly sensitive to features when given a large number of input values.

Advantages of C4.5 algorithm :

1. C4.5 is easy to implement.
2. C4.5 builds models that can be easily interpreted.
3. It can handle both categorical and continuous values.
4. It can deal with noise and missing value attributes.

Disadvantages of C4.5 algorithm :

1. A small variation in data can lead to different decision trees when using C4.5.
2. For a small training set, C4.5 does not work very well.

Advantages of CART algorithm :

1. CART can handle missing values automatically using proxy splits.
2. It uses combination of continuous/discrete variables.
3. CART automatically performs variable selection.
4. CART can establish interactions among variables.
5. CART does not vary according to the monotonic transformation of predictive variable.

Disadvantages of CART algorithm :

1. CART has unstable decision trees.
2. CART splits only by one variable.
3. It is non-parametric algorithm.

PART-2

Entropy and Information Theory, Information Gain, ID-3 Algorithm, Issues in Decision Tree Learning.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.6. Explain attribute selection measures used in decision tree.

Answer

Attribute selection measures used in decision tree are :

1. **Entropy :**
   i. Entropy is a measure of uncertainty associated with a random variable.
   ii. The entropy increases with the increase in uncertainty or randomness and decreases with a decrease in uncertainty or randomness.
   iii. The value of entropy ranges from 0-1.

$$\text{Entropy}(D) = \sum_{i=1}^{r} - p_i \log_2(p_i)$$

   where p_i is the non-zero probability that an arbitrary tuple in D belongs to class C and is estimated by $|C_i, D|/|D|$.
   iv. A log function of base 2 is used because the entropy is encoded in bits 0 and 1.
2. **Information gain :**
   i. ID3 uses information gain as its attribute selection measure.
   ii. Information gain is the difference between the original information gain requirement (*i.e.* based on the proportion of classes) and the new requirement (*i.e.* obtained after the partitioning of A).

$$\text{Gain}(D, A) = \text{Entropy}(D) - \sum_{j=1}^{v} \frac{|D_j|}{|D|} \text{Entropy}(D_j)$$

   Where,

   D : A given data partition

   A : Attribute

   V : Suppose we partition the tuples in D on some attribute A having V distinct values

   iii. D is split into V partition or subsets, $\{D_1, D_2, ..D_j\}$ where D_j contains those tuples in D that have outcome a_j, of A.
   iv. The attribute that has the highest information gain is chosen.
3. **Gain ratio :**
   i. The information gain measure is biased towards tests with many outcomes.
   ii. That is, it prefers to select attributes having a large number of values.
   iii. As each partition is pure, the information gain by partitioning is maximal. But such partitioning cannot be used for classification.
   iv. C4.5 uses this attribute selection measure which is an extension to the information gain.

   v. Gain ratio differs from information gain, which measures the information with respect to a classification that is acquired based on some partitioning.
   vi. Gain ratio applies kind of information gain using a split information value defined as :

$$\text{SplitInfo}_A = - \sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2 \left(\frac{|D_j|}{|D|} \right)$$

   vii. The gain ratio is then defined as :

$$\text{Gain ratio } (A) = \frac{\text{Gain } (A)}{\text{SplitInfo}_A (D)}$$

   viii. A splitting attribute is selected which is the attribute having the maximum gain ratio.

Que 3.7. Explain applications of decision tree in various areas of data mining.

Answer

The various decision tree applications in data mining are :

1. **E-Commerce :** It is used widely in the field of e-commerce, decision tree helps to generate online catalog which is an important factor for the success of an e-commerce website.
2. **Industry :** Decision tree algorithm is useful for producing quality control (faults identification) systems.
3. **Intelligent vehicles :** An important task for the development of intelligent vehicles is to find the lane boundaries of the road.
4. **Medicine :**
   a. Decision tree is an important technique for medical research and practice. A decision tree is used for diagnostic of various diseases.
   b. Decision tree is also used for hard sound diagnosis.
5. **Business :** Decision trees find use in the field of business where they are used for visualization of probabilistic business models, used in CRM (Customer Relationship Management) and used for credit scoring for credit card users and for predicting loan risks in banks.

Que 3.8. Explain procedure of ID3 algorithm.

Answer

ID3 (Examples, Target Attribute, Attributes) :

1. Create a Root node for the tree.
2. If all Examples are positive, return the single-node tree root, with label = +

3. If all Examples are negative, return the single-node tree root, with label = –
4. If Attributes is empty, return the single-node tree root, with label = most common value of target attribute in examples.
5. Otherwise begin
   a. $A \leftarrow$ the attribute from Attributes that best classifies Examples
   b. The decision attribute for Root $\leftarrow A$
   c. For each possible value, V_i, of A,
      i. Add a new tree branch below root, corresponding to the test A = V_i
      ii. Let Example V_i be the subset of Examples that have value V_i for A
      iii. If Example V_i is empty
         a. Then below this new branch add a leaf node with label = most common value of TargetAttribute in Examples
         b. Else below this new branch add the sub-tree ID3 (Example V_i, TargetAttribute, Attributes–{A})
6. End
7. Return root.

Que 3.9. Explain inductive bias with inductive system.

Answer

Inductive bias :

1. Inductive bias refers to the restrictions that are imposed by the assumptions made in the learning method.
2. For example, assuming that the solution to the problem of road safety can be expressed as a conjunction of a set of eight concepts.
3. This does not allow for more complex expressions that cannot be expressed as a conjunction.
4. This inductive bias means that there are some potential solutions that we cannot explore, and not contained within the version space we examine.
5. Order to have an unbiased learner, the version space would have to contain every possible hypothesis that could possibly be expressed.
6. The solution that the learner produced could never be more general than the complete set of training data.
7. In other words, it would be able to classify data that it had previously seen (as the rote learner could) but would be unable to generalize in order to classify new, unseen data.

8. The inductive bias of the candidate elimination algorithm is that it is only able to classify a new piece of data if all the hypotheses contained within its version space give data the same classification.
9. Hence, the inductive bias does impose a limitation on the learning method.

Inductive system :

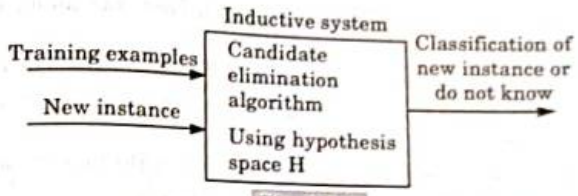


Fig. 3.9.1.

Que 3.10. Explain inductive learning algorithm.

Answer

Inductive learning algorithm :

Step 1 : Divide the table '*T*' containing *m* examples into *n* sub-tables (*t*1, *t*2, ... *tn*). One table for each possible value of the class attribute (repeat steps 2-8 for each sub-table).

Step 2 : Initialize the attribute combination count *j* = 1.

Step 3 : For the sub-table on which work is going on, divide the attribute list into distinct combinations, each combination with *j* distinct attributes.

Step 4 : For each combination of attributes, count the number of occurrences of attribute values that appear under the same combination of attributes in unmarked rows of the sub-table under consideration, and at the same time, not appears under the same combination of attributes of other sub-tables. Call the first combination with the maximum number of occurrences the max-combination MAX.

Step 5 : If MAX = = null, increase *j* by 1 and go to Step 3.

Step 6 : Mark all rows of the sub-table where working, in which the values of MAX appear, as classified.

Step 7 : Add a rule (IF attribute = "XYZ" → THEN decision is YES/ NO) to R (rule set) whose left-hand side will have attribute names of the MAX with their values separated by AND, and its right hand side contains the decision attribute value associated with the sub-table.

Step 8 : If all rows are marked as classified, then move on to process another sub-table and go to Step 2, else, go to Step 4. If no sub-tables are available, exit with the set of rules obtained till then.

Que 3.11. **Which learning algorithms are used in inductive bias ?**

Answer

Learning algorithm used in inductive bias are :

1. **Rote-learner :**
   a. Learning corresponds to storing each observed training example in memory.
   b. Subsequent instances are classified by looking them up in memory.
   c. If the instance is found in memory, the stored classification is returned.
   d. Otherwise, the system refuses to classify the new instance.
   e. **Inductive bias :** There is no inductive bias.
2. **Candidate-elimination :**
   a. New instances are classified only in the case where all members of the current version space agree on the classification.
   b. Otherwise, the system refuses to classify the new, instance.
   c. **Inductive bias :** The target concept can be represented in its hypothesis space.
3. **FIND-S :**
   a. This algorithm, finds the most specific hypothesis consistent with the training examples.
   b. It then uses this hypothesis to classify all subsequent instances.
   c. **Inductive bias :** The target concept can be represented in its hypothesis space, and all instances are negative instances unless the opposite is entailed by its other knowledge.

Que 3.12. **Discuss the issues related to the applications of decision trees.**

Answer

Issues related to the applications of decision trees are :

1. **Missing data :**
   a. When values have gone unrecorded, or they might be too expensive to obtain.
   b. Two problems arise :
      i. To classify an object that is missing from the test attributes.
      ii. To modify the information gain formula when examples have unknown values for the attribute.

2. **Multi-valued attributes :**
   a. When an attribute has many possible values, the information gain measure gives an inappropriate indication of the attribute's usefulness.
   b. In the extreme case, we could use an attribute that has a different value for every example.
   c. Then each subset of examples would be a singleton with a unique classification, so the information gain measure would have its highest value for this attribute, the attribute could be irrelevant or useless.
   d. One solution is to use the gain ratio.
3. **Continuous and integer valued input attributes :**
   a. Height and weight have an infinite set of possible values.
   b. Rather than generating infinitely many branches, decision tree learning algorithms find the split point that gives the highest information gain.
   c. Efficient dynamic programming methods exist for finding good split points, but it is still the most expensive part of real world decision tree learning applications.
4. **Continuous-valued output attributes :**
   a. If we are trying to predict a numerical value, such as the price of a work of art, rather than discrete classifications, then we need a regression tree.
   b. Such a tree has a linear function of some subset of numerical attributes, rather than a single value at each leaf.
   c. The learning algorithm must decide when to stop splitting and begin applying linear regression using the remaining attributes.

PART-3

Instance-based Learning.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.13. **Write short note on instance-based learning.**

Answer

1. Instance-Based Learning (IBL) is an extension of nearest neighbour or K-NN classification algorithms.
2. IBL algorithms do not maintain a set of abstractions of model created from the instances.
3. The K-NN algorithms have large space requirement.
4. They also extend it with a significance test to work with noisy instances, since a lot of real-life datasets have training instances and K-NN algorithms do not work well with noise.
5. Instance-based learning is based on the memorization of the dataset.
6. The number of parameters is unbounded and grows with the size of the data.
7. The classification is obtained through memorized examples.
8. The cost of the learning process is 0, all the cost is in the computation of the prediction.
9. This kind learning is also known as lazy learning.

Que 3.14. Explain instance-based learning representation.

Answer

Following are the instance based learning representation :

Instance-based representation (1) :

1. The simplest form of learning is plain memorization.
2. This is a completely different way of representing the knowledge extracted from a set of instances : just store the instances themselves and operate by relating new instances whose class is unknown to existing ones whose class is known.
3. Instead of creating rules, work directly from the examples themselves.

Instance-based representation (2) :

1. Instance-based learning is lazy, deferring the real work as long as possible.
2. In instance-based learning, each new instance is compared with existing ones using a distance metric, and the closest existing instance is used to assign the class to the new one. This is also called the nearest-neighbour classification method.
3. Sometimes more than one nearest neighbour is used, and the majority class of the closest k-nearest neighbours is assigned to the new instance. This is termed the k-nearest neighbour method.

Instance-based representation (3) :

1. When computing the distance between two examples, the standard Euclidean distance may be used.

2. A distance of 0 is assigned if the values are identical, otherwise the distance is 1.
3. Some attributes will be more important than others. We need some kinds of attribute weighting. To get suitable attribute weights from the training set is a key problem.
4. It may not be necessary, or desirable, to store all the training instances.

Instance-based representation (4) :

1. Generally some regions of attribute space are more stable with regard to class than others, and just a few examples are needed inside stable regions.
2. An apparent drawback to instance-based representation is that they do not make explicit the structures that are learned.

Fig. 3.14.1.

Que 3.15. What are the performance dimensions used for instance-based learning algorithm ?

Answer

Performance dimension used for instance-based learning algorithm are :

1. **Generality :**
   a. This is the class of concepts that describe the representation of an algorithm.
   b. IBL algorithms can pac-learn any concept whose boundary is a union of a finite number of closed hyper-curves of finite size.
2. **Accuracy :** This concept describes the accuracy of classification.
3. **Learning rate :**
   a. This is the speed at which classification accuracy increases during training.
   b. It is a more useful indicator of the performance of the learning algorithm than accuracy for finite-sized training sets.
4. **Incorporation costs :**
   a. These are incurred while updating the concept descriptions with a single training instance.
   b. They include classification costs.

5. **Storage requirement :** This is the size of the concept description for IBL algorithms, which is defined as the number of saved instances used for classification decisions.

Que 3.16. **What are the functions of instance-based learning ?**

Answer

Functions of instance-based learning are :

1. **Similarity function :**
   a. This computes the similarity between a training instance *i* and the instances in the concept description.
   b. Similarities are numeric-valued.
2. **Classification function :**
   a. This receives the similarity function's results and the classification performance records of the instances in the concept description.
   b. It yields a classification for *i*.
3. **Concept description updater :**
   a. This maintains records on classification performance and decides which instances to include in the concept description.
   b. Inputs include *i*, the similarity results, the classification results, and a current concept description. It yields the modified concept description.

Que 3.17. **What are the advantages and disadvantages of instance-based learning ?**

Answer

Advantages of instance-based learning :

1. Learning is trivial.
2. Works efficiently.
3. Noise resistant.
4. Rich representation, arbitrary decision surfaces.
5. Easy to understand.

Disadvantages of instance-based learning :

1. Need lots of data.
2. Computational cost is high.
3. Restricted to $x \in R^n$.
4. Implicit weights of attributes (need normalization).
5. Need large space for storage *i.e.*, require large memory.
6. Expensive application time.

PART-4

K-Nearest Neighbour Learning, Locally Weighted Regression, Radial Basis Function Networks.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.18. **Describe *K*-Nearest Neighbour algorithm with steps.**

Answer

1. The *KNN* classification algorithm is used to decide the new instance should belong to which class.
2. When $K = 1$, we have the nearest neighbour algorithm.
3. *KNN* classification is incremental.
4. *KNN* classification does not have a training phase, all instances are stored. Training uses indexing to find neighbours quickly.
5. During testing, *KNN* classification algorithm has to find *K*-nearest neighbours of a new instance. This is time consuming if we do exhaustive comparison.
6. K-nearest neighbours use the local neighborhood to obtain a prediction.

   Algorithm : Let m be the number of training data samples. Let p be an unknown point.

1. Store the training samples in an array of data points array. This means each element of this array represents a tuple (x, y).
2. For $i = 0$ to m :

   Calculate Euclidean distance $d(\text{arr}[i], p)$.
3. Make set S of K smallest distances obtained. Each of these distances corresponds to an already classified data point.
4. Return the majority label among S.

Que 3.19. **What are the advantages and disadvantages of *K*-nearest neighbour algorithm ?**

Answer

Advantages of KNN algorithm :

1. **No training period :**
   a. KNN is called lazy learner (Instance-based learning).

b. It does not learn anything in the training period. It does not derive any discriminative function from the training data.

c. In other words, there is no training period for it. It stores the training dataset and learns from it only at the time of making real time predictions.

d. This makes the KNN algorithm much faster than other algorithms that require training for example, SVM, Linear Regression etc.

2. Since the KNN algorithm requires no training before making predictions, new data can be added seamlessly which will not impact the accuracy of the algorithm.

3. KNN is very easy to implement. There are only two parameters required to implement KNN *i.e.*, the value of K and the distance function (for example, Euclidean).

Disadvantages of KNN :

1. **Does not work well with large dataset :** In large datasets, the cost of calculating the distance between the new point and each existing points is huge which degrades the performance of the algorithm.

2. **Does not work well with high dimensions :** The KNN algorithm does not work well with high dimensional data because with large number of dimensions, it becomes difficult for the algorithm to calculate the distance in each dimension.

3. **Need feature scaling :** We need to do feature scaling (standardization and normalization) before applying KNN algorithm to any dataset. If we do not do so, KNN may generate wrong predictions.

4. **Sensitive to noisy data, missing values and outliers :** KNN is sensitive to noise in the dataset. We need to manually represent missing values and remove outliers.

Que 3.20. Explain locally weighted regression.

Answer

1. Model-based methods, such as neural networks and the mixture of Gaussians, use the data to build a parameterized model.

2. After training, the model is used for predictions and the data are generally discarded.

3. In contrast, memory-based methods are non-parametric approaches that explicitly retain the training data, and use it each time a prediction needs to be made.

4. Locally Weighted Regression (LWR) is a memory-based method that performs a regression around a point using only training data that are local to that point.

5. LWR was suitable for real-time control by constructing an LWR-based system that learned a difficult juggling task.

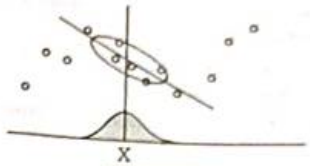


Fig. 3.20.1.

6. The LOESS (Locally Estimated Scatterplot Smoothing) model performs a linear regression on points in the data set, weighted by a kernel centered at x.

7. The kernel shape is a design parameter for which the original LOESS model uses a tricubic kernel :

$$h_i(x) = h(x - x_i) = \exp(-k(x - x_i)^2),$$

where k is a smoothing parameter.

8. For brevity, we will drop the argument x for $h_i(x)$, and define $n = \Sigma_i h_i$. We can then write the estimated means and covariances as :

$$\mu_x = \frac{\Sigma_i h_i x_i}{n}, \sigma_x^2 = \frac{\Sigma_i h_i (x_i - \mu_x)^2}{n}, \sigma_{xy} = \frac{\Sigma_i h_i (x_i - \mu_x)(y_i - \mu_y)}{n}$$

$$\mu_y = \frac{\Sigma_i h_i y_i}{n}, \sigma_y^2 = \frac{\Sigma_i h_i (y_i - \mu_y)^2}{n}, \sigma_{y/x}^2 = \sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2}$$

9. We use the data covariances to express the conditional expectations and their estimated variances :

$$\hat{y} = \mu_y + \frac{\sigma_{xy}}{\sigma_x^2}(x - \mu_x), \frac{\sigma_{y/x}^2}{n^2}\left(\sum_i h_i^2 + \frac{(x - \mu_x)^2}{\sigma_x^2}\sum_i h_i^2 \frac{(x_i - \mu_x)^2}{\sigma_x^2}\right)$$

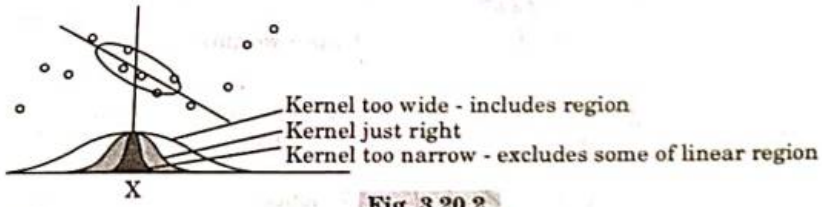


Fig. 3.20.2.

Que 3.21. Explain Radial Basis Function (RBF).

Answer

1. A Radial Basis Function (RBF) is a function that assigns a real value to each input from its domain (it is a real-value function), and the value produced by the RBF is always an absolute value *i.e.*, it is a measure of distance and cannot be negative.

2. Euclidean distance (the straight-line distance) between two points in Euclidean space is used.
3. Radial basis functions are used to approximate functions, such as neural networks acts as function approximators.
4. The following sum represents a radial basis function network :

$$y(x) = \sum_{i=1}^{N} w_i \, \phi(\| x - x_i \|),$$

5. The radial basis functions act as activation functions.
6. The approximant $y(x)$ is differentiable with respect to the weights which are learned using iterative update methods common among neural networks.

Que 3.22. Explain the architecture of a radial basis function network.

Answer

1. Radial Basis Function (RBF) networks have three layers : an input layer, a hidden layer with a non-linear RBF activation function and a linear output layer.
2. The input can be modeled as a vector of real numbers $x \in R^n$.
3. The output of the network is then a scalar function of the input vector, $\phi : R^n \to R$, and is given by

$$\phi(x) = \sum_{i=1}^{N} a_i \, \rho(\| x - c_i \|)$$

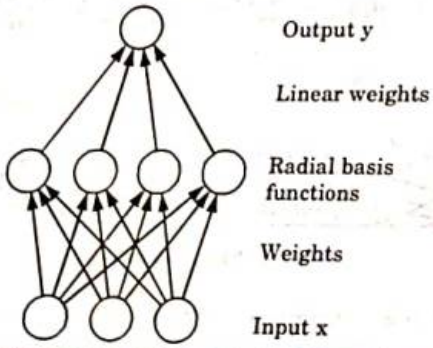


Fig. 3.22.1. Architecture of a radial basis function network. An input vector x is used as input to all radial basis functions, each with different parameters. The output of the network is a linear combination of the outputs from radial basis functions.

where n is the number of neurons in the hidden layer, c_i is the center vector for neuron i and a_i is the weight of neuron i in the linear output neuron.

4. Functions that depend only on the distance from a center vector are radially symmetric about that vector.
5. In the basic form all inputs are connected to each hidden neuron.
6. The radial basis function is taken to be Gaussian

$$\rho(\| x - c_i \|) = \exp\left[-\beta \| x - c_i \|^2\right]$$

7. The Gaussian basis functions are local to the center vector in the sense that

$$\lim_{\|x\| \to \infty} \rho(\| x - c_i \|) = 0$$

i.e., changing parameters of one neuron has only a small effect for input values that are far away from the center of that neuron.

8. Given certain mild conditions on the shape of the activation function, RBF networks are universal approximators on a compact subset of R^n.
9. This means that an RBF network with enough hidden neurons can approximate any continuous function on a closed, bounded set with arbitrary precision.
10. The parameters a_i, c_i, ρ, and β are determined in a manner that optimizes the fit between ϕ and the data.

PART-5

Case-based Learning.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.23. Write short note on case-based learning algorithm.

Answer

1. Case-Based Learning (CBL) algorithms contain an input as a sequence of training cases and an output concept description, which can be used to generate predictions of goal feature values for subsequently presented cases.

2. The primary component of the concept description is case-base, but almost all CBL algorithms maintain additional related information for the purpose of generating accurate predictions (for example, settings for feature weights).
3. Current CBL algorithms assume that cases are described using a feature-value representation, where features are either predictor or goal features.
4. CBL algorithms are distinguished by their processing behaviour.

Disadvantages of case-based learning algorithm :

1. They are computationally expensive because they save and compute similarities to all training cases.
2. They are intolerant of noise and irrelevant features.
3. They are sensitive to the choice of the algorithm's similarity function.
4. There is no simple way they can process symbolic valued feature values.

Que 3.24. What are the functions of case-based learning algorithm ?

Answer

Functions of case-based learning algorithm are :

1. **Pre-processor :** This prepares the input for processing (for example, normalizing the range of numeric-valued features to ensure that they are treated with equal importance by the similarity function, formatting the raw input into a set of cases).
2. **Similarity :**
   a. This function assesses the similarities of a given case with the previously stored cases in the concept description.
   b. Assessment may involve explicit encoding and/or dynamic computation.
   c. CBL similarity functions find a compromise along the continuum between these extremes.
3. **Prediction :** This function inputs the similarity assessments and generates a prediction for the value of the given case's goal feature (*i.e.*, a classification when it is symbolic-valued).
4. **Memory updating :** This updates the stored case-base, such as by modifying or abstracting previously stored cases, forgetting cases presumed to be noisy, or updating a feature's relevance weight setting.

Que 3.25. Describe case-based learning cycle with different schemes of CBL.

Answer

Case-based learning algorithm processing stages are :

1. **Case retrieval :** After the problem situation has been assessed, the best matching case is searched in the case-base and an approximate solution is retrieved.
2. **Case adaptation :** The retrieved solution is adapted to fit better in the new problem.

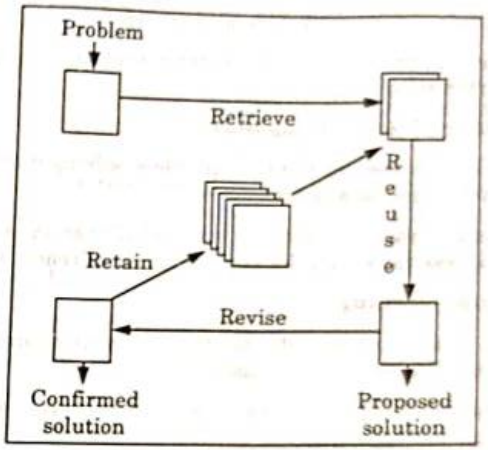


Fig. 3.25.1. The CBL cycle.

3. **Solution evaluation :**
   a. The adapted solution can be evaluated either before the solution is applied to the problem or after the solution has been applied.
   b. In any case, if the accomplished result is not satisfactory, the retrieved solution must be adapted again or more cases should be retrieved.
4. **Case-base updating :** If the solution was verified as correct, the new case may be added to the case base.

Different scheme of the CBL working cycle are :

1. Retrieve the most similar case.
2. Reuse the case to attempt to solve the current problem.
3. Revise the proposed solution if necessary.

4. Retain the new solution as a part of a new case.

Que 3.26. What are the benefits of CBL as a lazy problem solving method ?

Answer

The benefits of CBL as a lazy Problem solving method are :

1. **Ease of knowledge elicitation :**
   a. Lazy methods can utilise easily available case or problem instances instead of rules that are difficult to extract.
   b. So, classical knowledge engineering is replaced by case acquisition and structuring.
2. **Absence of problem-solving bias :**
   a. Cases can be used for multiple problem-solving purposes, because they are stored in a raw form.
   b. This in contrast to eager methods, which can be used merely for the purpose for which the knowledge has already been compiled.
3. **Incremental learning :**
   a. A CBL system can be put into operation with a minimal set solved cases furnishing the case base.
   b. The case base will be filled with new cases increasing the system's problem-solving ability.
   c. Besides augmentation of the case base, new indexes and clusters categories can be created and the existing ones can be changed.
   d. This in contrast requires a special training period whenever informatics extraction (knowledge generalisation) is performed.
   e. Hence, dynamic on-line adaptation a non-rigid environment is possible.
4. **Suitability for complex and not-fully formalised solution spaces :**
   a. CBL systems can applied to an incomplete model of problem domain, implementation involves both to identity relevant case features and to furnish, possibly a partial case base, with proper cases.
   b. Lazy approaches are appropriate for complex solution spaces than eager approaches, which replace the presented data with abstractions obtained by generalisation.

5. **Suitability for sequential problem solving :**
   a. Sequential tasks, like these encountered reinforcement learning problems, benefit from the storage of history in the form of sequence of states or procedures.
   b. Such a storage is facilitated by lazy approaches.
6. **Ease of explanation :**
   a. The results of a CBL systém can be justified based upon the similarity of the current problem to the retrieved case.
   b. CBL are easily traceable to precedent cases, it is also easier to analyse failures of the system.
7. **Ease of maintenance :** This is particularly due to the fact that CBL systems can adapt to many changes in the problem domain and the relevant environment, merely by acquiring.

Que 3.27. What are the limitations of CBL ?

Answer

Limitations of CBL are :

1. **Handling large case bases :**
   a. High memory / storage requirements and time-consuming retrieval accompany CBL systems utilising large case bases.
   b. Although the order of both is linear with the number of cases, these problems usually lead to increased construction costs and reduced system performance.
   c. These problems are less significant as the hardware components become faster and cheaper.
2. **Dynamic problem domains :**
   a. CBL systems may have difficulties in handling dynamic problem domains, where they may be unable to follow a shift in the way problems are solved, since they are strongly biased towards what has already worked.
   b. This may result in an outdated case base.
3. **Handling noisy data :**
   a. Parts of the problem situation may be irrelevant to the problem itself.

b. Unsuccessful assessment of such noise present in a problem situation currently imposed on a CBL system may result in the same problem being unnecessarily stored numerous times in the case base because of the difference due to the noise.

c. In turn this implies inefficient storage and retrieval of cases.

4. **Fully automatic operation :**

a. In a CBL system, the problem domain is not fully covered.

b. Hence, some problem situations can occur for which the system has no solution.

c. In such situations, CBL systems expect input from the user.

Que 3.28. **What are the applications of CBL ?**

Answer

Applications of CBL :

1. **Interpretation :** It is a process of evaluating situations / problems in some context (For example, HYPO for interpretation of patent laws KICS for interpretation of building regulations, LISSA for interpretation of non-destructive test measurements).
2. **Classification :** It is a process of explaining a number of encountered symptoms (For example, CASEY for classification of auditory impairments, CASCADE for classification of software failures, PAKAR for causal classification of building defects, ISFER for classification of facial expressions into user defined interpretation categories.
3. **Design :** It is a process of satisfying a number of posed constraints (For example, JULIA for meal planning, CLAVIER for design of optimal layouts of composite airplane parts, EADOCS for aircraft panels design).
4. **Planning :** It is a process of arranging a sequence of actions in time (For example, BOLERO for building diagnostic plans for medical patients, TOTLEC for manufacturing planning).
5. **Advising :** It is a process of resolving diagnosed problems (For example, DECIDER for advising students, HOMER).

Que 3.29. **What are major paradigms of machine learning ?**

Answer

Major paradigms of machine learning are :

1. **Rote Learning :**
   a. There is one-to-one mapping from inputs to stored representation.
   b. Learning by memorization.

c. There is Association-based storage and retrieval.

2. **Induction :** Machine learning use specific examples to reach general conclusions.
3. **Clustering :** Clustering is a task of grouping a set of objects in such a way that objects in the same group are similar to each other than to those in other group.
4. **Analogy :** Determine correspondence between two different representations.
5. **Discovery :** Unsupervised *i.e.*, specific goal not given.
6. **Genetic algorithms :**
   a. Genetic algorithms are stochastic search algorithms which act on a population of possible solutions.
   b. They are probabilistic search methods means that the states which they explore are not determined solely by the properties of the problems.
7. **Reinforcement :**
   a. In reinforcement only feedback (positive or negative reward) given at end of a sequence of steps.
   b. Requires assigning reward to steps by solving the credit assignment problem which steps should receive credit or blame for a final result.

Que 3.30. **Briefly explain the inductive learning problem.**

Answer

Inductive learning problem are :

1. **Supervised versus unsupervised learning :**
   a. We want to learn an unknown function $f(x) = y$, where x is an input example and y is the desired output.
   b. Supervised learning implies we are given a set of (x, y) pairs by a teacher.
   c. Unsupervised learning means we are only given the xs.
   d. In either case, the goal is to estimate f.
2. **Concept learning :**
   a. Given a set of examples of some concept/class/category, determine if a given example is an instance of the concept or not.
   b. If it is an instance, we call it a positive example.
   c. If it is not, it is called a negative example.

3. **Supervised concept learning by induction :**
   a. Given a training set of positive and negative examples of a concept, construct a description that will accurately classify whether future examples are positive or negative.
   b. That is, learn some good estimate of function f given a training set $\{(x1, y1), (x2, y2), ..., (xn, yn)\}$ where each y_i is either + (positive) or – (negative).

4 UNIT

Artificial Neural Network and Deep Learning

CONTENTS

PART-1

Artifical Neural Network, Perceptron's Multilayer Perceptron, Gradient Descent and the Delta Rule.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.1. Describe Artificial Neural Network (ANN) with different layers.

Answer

Artificial Neural Network : *Refer Q. 1.13, Page 1–14L, Unit-1.*

A neural network contains the following three layers :

a. **Input layer :** The activity of the input units represents the raw information that can feed into the network.

b. **Hidden layer :**

i. Hidden layer is used to determine the activity of each hidden unit.

ii. The activities of the input units and the weights depend on the connections between the input and the hidden units.

iii. There may be one or more hidden layers.

c. **Output layer :** The behaviour of the output units depends on the activity of the hidden units and the weights between the hidden and output units.

Que 4.2. What are the advantages and disadvantage of Artificial Neural Network ?

Answer

Advantages of Artificial Neural Networks (ANN) :

1. Problems in ANN are represented by attribute-value pairs.
2. ANNs are used for problems having the target function, output may be discrete-valued, real-valued, or a vector of several real or discrete-valued attributes.
3. ANNs learning methods are quite robust to noise in the training data. The training examples may contain errors, which do not affect the final output.

4. It is used where the fast evaluation of the learned target function required.
5. ANNs can bear long training times depending on factors such as the number of weights in the network, the number of training examples considered, and the settings of various learning algorithm parameters.

Disadvantages of Artificial Neural Networks (ANN) :

1. **Hardware dependence :**
   a. Artificial neural networks require processors with parallel processing power, by their structure.
   b. For this reason, the realization of the equipment is dependent.
2. **Unexplained functioning of the network :**
   a. This is the most important problem of ANN.
   b. When ANN gives a probing solution, it does not give a clue as to why and how.
   c. This reduces trust in the network.
3. **Assurance of proper network structure :**
   a. There is no specific rule for determining the structure of artificial neural networks.
   b. The appropriate network structure is achieved through experience and trial and error.
4. **The difficulty of showing the problem to the network :**
   a. ANNs can work with numerical information.
   b. Problems have to be translated into numerical values before being introduced to ANN.
   c. The display mechanism to be determined will directly influence the performance of the network.
   d. This is dependent on the user's ability.
5. **The duration of the network is unknown :**
   a. The network is reduced to a certain value of the error on the sample means that the training has been completed.
   b. This value does not give us optimum results.

Que 4.3. What are the characteristics of Artificial Neural Network ?

Answer

Characteristics of Artificial Neural Network are :

1. It is neurally implemented mathematical model.
2. It contains large number of interconnected processing elements called neurons to do all the operations.

3. Information stored in the neurons is basically the weighted linkage of neurons.
4. The input signals arrive at the processing elements through connections and connecting weights.
5. It has the ability to learn, recall and generalize from the given data by suitable assignment and adjustment of weights.
6. The collective behaviour of the neurons describes its computational power, and no single neuron carries specific information.

Que 4.4. Explain the application areas of artificial neural network.

Answer

Application areas of artificial neural network are :

1. **Speech recognition :**
   a. Speech occupies a prominent role in human-human interaction.
   b. Therefore, it is natural for people to expect speech interfaces with computers.
   c. In the present era, for communication with machines, humans still need sophisticated languages which are difficult to learn and use.
   d. To ease this communication barrier, a simple solution could be communication in a spoken language that is possible for the machine to understand.
   e. Hence, ANN is playing a major role in speech recognition.
2. **Character recognition :**
   a. It is a problem which falls under the general area of Pattern Recognition.
   b. Many neural networks have been developed for automatic recognition of handwritten characters, either letters or digits.
3. **Signature verification application :**
   a. Signatures are useful ways to authorize and authenticate a person in legal transactions.
   b. Signature verification technique is a non-vision based technique.
   c. For this application, the first approach is to extract the feature or rather the geometrical feature set representing the signature.
   d. With these feature sets, we have to train the neural networks using an efficient neural network algorithm.
   e. This trained neural network will classify the signature as being genuine or forged under the verification stage.
4. **Human face recognition :**
   a. It is one of the biometric methods to identify the given face.

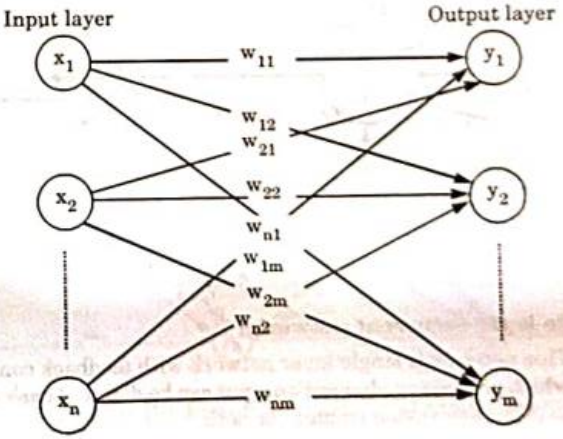
   b. It is a typical task because of the characterization of "non-face" images.
   c. However, if a neural network is well trained, then it can be divided into two classes namely images having faces and images that do not have faces.

Que 4.5. Explain different types of neuron connection wi'' architecture.

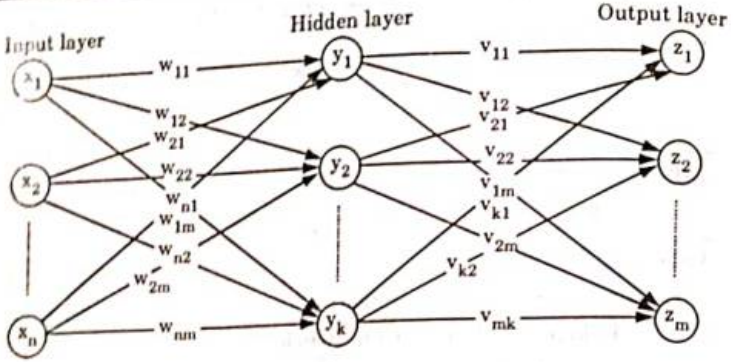
Answer

Different types of neuron connection are :

1. **Single-layer feed forward network :**
   a. In this type of network, we have only two layers *i.e.*, input layer and output layer but input layer does not count because no computation is performed in this layer.
   b. Output layer is formed when different weights are applied on input nodes and the cumulative effect per node is taken.
   c. After this the neurons collectively give the output layer to compute the output signals.



2. **Multilayer feed forward network :**
   a. This layer has hidden layer which is internal to the network and has no direct contact with the external layer.
   b. Existence of one or more hidden layers enables the network to be computationally stronger.
   c. There are no feedback connections in which outputs of the model are fed back into itself.



3. Single node with its own feedback :

a. When outputs can be directed back as inputs to the same layer or preceding layer nodes, then it results in feedback networks.

b. Recurrent networks are feedback networks with closed loop. Fig. 4.5.1 shows a single recurrent network having single neuron with feedback to itself.

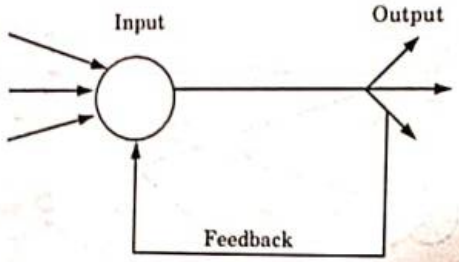


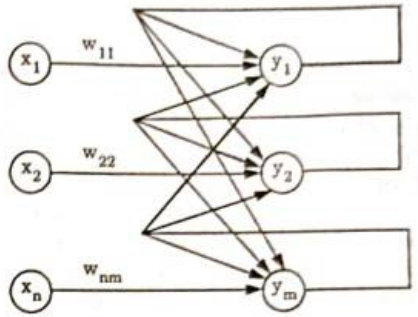
Fig. 4.5.1.

4. Single-layer recurrent network :

a. This network is single layer network with feedback connection in which processing element's output can be directed back to itself or to other processing element or both.

b. Recurrent neural network is a class of artificial neural network where connections between nodes form a directed graph along a sequence.

c. This allows it to exhibit dynamic temporal behaviour for a time sequence. Unlike feed forward neural networks, RNNs can use their internal state (memory) to process sequences of inputs.

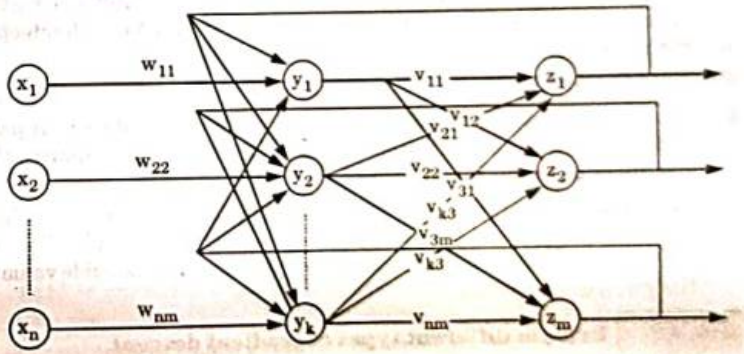


5. Multilayer recurrent network :

a. In this type of network, processing element output can be directed to the processing element in the same layer and in the preceding layer forming a multilayer recurrent network.

b. They perform the same task for every element of a sequence, with the output being depended on the previous computations. Inputs are not needed at each time step.

c. The main feature of a multilayer recurrent neural network is its hidden state, which captures information about a sequence.



Que 4.6. Discuss the benefits of artificial neural network.

Answer

1. Artificial neural networks are flexible and adaptive.
2. Artificial neural networks are used in sequence and pattern recognition systems, data processing, robotics, modeling, etc.
3. ANN acquires knowledge from their surroundings by adapting to internal and external parameters and they solve complex problems which are difficult to manage.

4. It generalizes knowledge to produce adequate responses to unknown situations.
5. Artificial neural networks are flexible and have the ability to learn, generalize and adapts to situations based on its findings.
6. This function allows the network to efficiently acquire knowledge by learning. This is a distinct advantage over a traditionally linear network that is inadequate when it comes to modelling non-linear data.
7. An artificial neuron network is capable of greater fault tolerance than a traditional network. Without the loss of stored data, the network is able to regenerate a fault in any of its components.
8. An artificial neuron network is based on adaptive learning.

Que 4.7. Write short note on gradient descent.

Answer

1. Gradient descent is an optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient.
2. A gradient is the slope of a function, the degree of change of a parameter with the amount of change in another parameter.
3. Mathematically, it can be described as the partial derivatives of a set of parameters with respect to its inputs. The more the gradient, the steeper the slope.
4. Gradient Descent is a convex function.
5. Gradient Descent can be described as an iterative method which is used to find the values of the parameters of a function that minimizes the cost function as much as possible.
6. The parameters are initially defined a particular value and from that, Gradient Descent run in an iterative fashion to find the optimal values of the parameters, using calculus, to find the minimum possible value of the given cost function.

Que 4.8. Explain different types of gradient descent.

Answer

Different types of gradient descent are :

1. **Batch gradient descent :**
   a. This is a type of gradient descent which processes all the training examples for each iteration of gradient descent.
   b. When the number of training examples is large, then batch gradient descent is computationally very expensive. So, it is not preferred.
   c. Instead, we prefer to use stochastic gradient descent or mini-batch gradient descent.

2. **Stochastic gradient descent :**
   a. This is a type of gradient descent which processes single training example per iteration.
   b. Hence, the parameters are being updated even after one iteration in which only a single example has been processed.
   c. Hence, this is faster than batch gradient descent. When the number of training examples is large, even then it processes only one example which can be additional overhead for the system as the number of iterations will be large.
3. **Mini-batch gradient descent :**
   a. This is a mixture of both stochastic and batch gradient descent.
   b. The training set is divided into multiple groups called batches.
   c. Each batch has a number of training samples in it.
   d. At a time, a single batch is passed through the network which computes the loss of every sample in the batch and uses their average to update the parameters of the neural network.

Que 4.9. What are the advantages and disadvantages of batch gradient descent ?

Answer

Advantages of batch gradient descent :

1. Less oscillations and noisy steps taken towards the global minima of the loss function due to updating the parameters by computing the average of all the training samples rather than the value of a single sample.
2. It can benefit from the vectorization which increases the speed of processing all training samples together.
3. It produces a more stable gradient descent convergence and stable error gradient than stochastic gradient descent.
4. It is computationally efficient as all computer resources are not being used to process a single sample rather are being used for all training samples.

Disadvantages of batch gradient descent :

1. Sometimes a stable error gradient can lead to a local minima and unlike stochastic gradient descent no noisy steps are there to help to get out of the local minima.
2. The entire training set can be too large to process in the memory due to which additional memory might be needed.
3. Depending on computer resources it can take too long for processing all the training samples as a batch.

Que 4.10. What are the advantages and disadvantages of stochastic gradient descent ?

Answer

Advantages of stochastic gradient descent :

1. It is easier to fit into memory due to a single training sample being processed by the network.
2. It is computationally fast as only one sample is processed at a time.
3. For larger datasets it can converge faster as it causes updates to the parameters more frequently.
4. Due to frequent updates the steps taken towards the minima of the loss function have oscillations which can help getting out of local minimums of the loss function (in case the computed position turns out to be the local minimum).

Disadvantages of stochastic gradient descent :

1. Due to frequent updates the steps taken towards the minima are very noisy. This can often lead the gradient descent into other directions.
2. Also, due to noisy steps it may take longer to achieve convergence to the minima of the loss function.
3. Frequent updates are computationally expensive due to using all resources for processing one training sample at a time.
4. It loses the advantage of vectorized operations as it deals with only a single example at a time.

Que 4.11. Explain delta rule. Explain generalized delta learning rule (error backpropagation learning rule).

Answer

Delta rule :

1. The delta rule is specialized version of backpropagation's learning rule that uses single layer neural networks.
2. It calculates the error between calculated output and sample output data, and uses this to create a modification to the weights, thus implementing a form of gradient descent.

Generalized delta learning rule (Error backpropagation learning) :

In generalized delta learning rule (error backpropagation learning). We are given the training set :

$$(x^1, y^1), \ldots, (x^k, y^k)$$

where $x^k = [x^k{}_1, \ldots x_n{}^k]$ and $y^k \in R$, $k = 1, \ldots, K$.

Step 1 : $\eta > 0$, $E_{max} > 0$ are chosen.

Step 2 : Weights w are initialized at small random values, $k = 1$, and the running error E is set to 0.

Step 3 : Input x^k is presented, $x := x^k$, $y := y^k$, and output O is computed as :

$$O = \frac{1}{1 + \exp(- W^T O)}$$

where O_t is the output vector of the hidden layer :

$$O_t = \frac{1}{1 + \exp(- W_t^T x)}$$

Step 4 : Weights of the output unit are updated

$$W := W + \eta \delta o$$

where $\delta = (y - O)O(1 - O)$

Step 5 : Weights of the hidden units are updated

$$w_t = w_t + \eta \delta W_t O_t (1 - O_t) x, \; l = 1, \ldots, L$$

Step 6 : Cumulative cycle error is computed by adding the present error to E

$$E := E + 1/2(y - O)^2$$

Step 7 : If $k < K$ then $k := k + 1$ and we continue the training by going back to step 2, otherwise we go to step 8.

Step 8 : The training cycle is completed. For $E < E_{max}$ terminate the training session. If $E > E_{max}$ then $E := 0$, $k := 1$ and we initiate a new training cycle by going back to step 3.

PART-2

Multilayer Network, Derivation of Back Propagation Algorithm, Generalization.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.12. Write short note on backpropagation algorithm.

Answer

1. Backpropagation is an algorithm used in the training of feedforward neural networks for supervised learning.
2. Backpropagation efficiently computes the gradient of the loss function with respect to the weights of the network for a single input-output example.
3. This makes it feasible to use gradient methods for training multi-layer networks, updating weights to minimize loss, we use gradient descent or variants such as stochastic gradient descent.

4. The back propagation algorithm works by computing the gradient of the loss function with respect to each weight by the chain rule, iterating backwards one layer at a time from the last layer to avoid redundant calculations of intermediate terms in the chain rule; this is an example of dynamic programming.
5. The term backpropagation refers only to the algorithm for computing the gradient, but it is often used loosely to refer to the entire learning algorithm, also including how the gradient is used, such as by stochastic gradient descent.
6. Backpropagation generalizes the gradient computation in the delta rule, which is the single-layer version of backpropagation, and is in turn generalized by automatic differentiation, where backpropagation is a special case of reverse accumulation (reverse mode).

Que 4.13. Explain perceptron with single flow graph.

Answer

1. The perceptron is the simplest form of a neural network used for classification of patterns said to be linearly separable.
2. It consists of a single neuron with adjustable synaptic weights and bias.
3. The perceptron build around a single neuron is limited for performing pattern classification with only two classes.
4. By expanding the output layer of perceptron to include more than one neuron, more than two classes can be classified.
5. Suppose, a perceptron have synaptic weights denoted by $w_1, w_2, w_3, \ldots w_m$.
6. The input applied to the perceptron are denoted by $x_1, x_2, \ldots x_m$.
7. The externally applied bias is denoted by b.

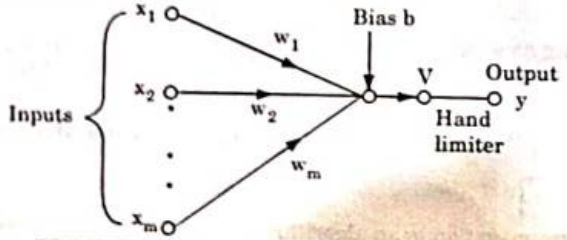


Fig. 4.13.1. Signal flow graph of the perceptron.

8. From the model, we find that the hard limiter input or induced local field of the neuron as

$$V = \sum_{i=1}^{m} w_i x_i + b$$

9. The goal of the perceptron is to correctly classify the set of externally applied input $x_1, x_2, \ldots x_m$ into one of two classes G_1 and G_2.
10. The decision rule for classification is that if output y is +1 then assign the point represented by input $x_1, x_2, \ldots x_m$ to class G_1 else y is –1 then assign to class G_2.
11. In Fig. 4.13.2, if a point (x_1, x_2) lies below the boundary lines is assigned to class G_2 and above the line is assigned to class G_1. Decision boundary is calculated as :

$$w_1x_1 + w_2x_2 + b = 0$$

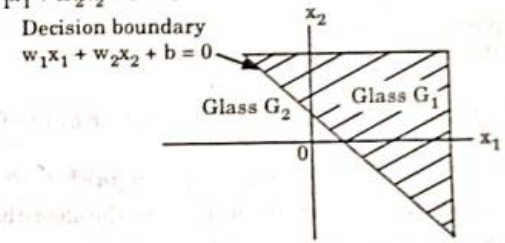


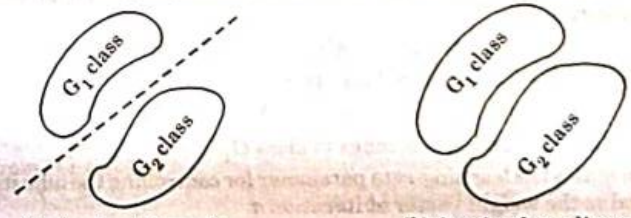
Fig. 4.13.2.

12. There are two decision regions separated by a hyperplane defined as :

$$\sum_{i=1}^{m} w_i x_i + b = 0$$

The synaptic weights $w_1, w_2, \ldots w_m$ of the perceptron can be adapted on an iteration by iteration basis.

13. For the adaption, an error-correction rule known as perceptron convergence algorithm is used.
14. For a perceptron to function properly, the two classes G_1 and G_2 must be linearly separable.
15. Linearly separable means, the pattern or set of inputs to be classified must be separated by a straight line.
16. Generalizing, a set of points in n-dimensional space are linearly separable if there is a hyperplane of $(n-1)$ dimensions that separates the sets.



(a) A pair of linearly separable patterns

(b) A pair of non-linearly separable patterns

Fig. 4.13.3.

Que 4.14. **State and prove perceptron convergence theorem.**

Answer

Statement : The Perceptron convergence theorem states that for any data set which is linearly separable the Perceptron learning rule is guaranteed to find a solution in a finite number of steps.

Proof :

1. To derive the error-correction learning algorithm for the perceptron.
2. The perceptron convergence theorem used the synaptic weights $w_1, w_2, \ldots w_m$ of the perceptron can be adapted on an iteration by iteration basis.
3. The bias $b(n)$ is treated as a synaptic weight driven by fixed input equal to + 1.

$$x(n) = [+1, x_1(n), x_2(n), \ldots x_m(n)]^T$$

Where n denotes the iteration step in applying the algorithm.

4. Correspondingly, we define the weight vector as

$$w(n) = [b(n), w_1(n), w_2(n) \ldots\ldots, w_m(n)]^T$$

Accordingly, the linear combiner output is written in the compact form :

$$v(n) = \sum_{i=0}^{n} w_i(n)\, x_i(n) = w^T(n)\, x(n)$$

The algorithm for adapting the weight vector is stated as :

1. If the nth member of input set $x(n)$, is correctly classified into linearly separable classes, by the weight vector $w(n)$ (that is output is correct) then no adjustment of weights are done.

$$w(n+1) = w(n)$$

if $w^T x(n) > 0$ and $x(n)$ belongs to class G_1.

$$w(n+1) = w(n)$$

if $w^T x(n) \leq 0$ and $x(n)$ belongs to class G_2.

2. Otherwise, the weight vector of the perceptron is updated in accordance with the rule :

$$w(n+1) = w(n) - \eta(n)\, x(n)$$

if $w^T(n)\, x(n) > 0$ and $x(n)$ belongs to class G_2.

$$w(n+1) = w(n) - \eta(n)\, x(n)$$

if $w^T(n)\, x(n) \leq 0$ and $x(n)$ belongs to class G_1.

where $\eta(n)$ is the learning-rate parameter for controlling the adjustment applied to the weight vector at iteration n.

Also small α leads to slow learning and large α leads to fast learning. For a constant α, the learning algorithm is termed as fixed increment algorithm.

Que 4.15. **Explain multilayer perceptron with its architecture and characteristics.**

Answer

Multilayer perceptron :

1. The perceptrons which are arranged in layers are called multilayer perceptron. This model has three layers : an input layer, output layer and hidden layer.
2. For the perceptrons in the input layer, the linear transfer function used and for the perceptron in the hidden layer and output layer, the sigmoidal or squashed-S function is used.
3. The input signal propagates through the network in a forward direction.
4. On a layer by layer basis, in the multilayer perceptron bias $b(n)$ is treated as a synaptic weight driven by fixed input equal to +1.

$$x(n) = [+1, x_1(n), x_2(n), \ldots\ldots\ldots x_m(n)]^T$$

where n denotes the iteration step in applying the algorithm.

Correspondingly, we define the weight vector as :

$$w(n) = [b(n), w_1(n), w_2(n)\ldots\ldots\ldots, w_m(n)]^T$$

5. Accordingly, the linear combiner output is written in the compact form :

$$V(n) = \sum_{i=0}^{m} w_i(n) x_i(n) = w^T(n) \times x(n)$$

The algorithm for adapting the weight vector is stated as :

1. If the nth number of input set $x(n)$, is correctly classified into linearly separable classes, by the weight vector $w(n)$ (that is output is correct) then no adjustment of weights are done.

$$w(n+1) = w(n)$$

if $w^T x(n) > 0$ and $x(n)$ belongs to class G_1.

$$w(n+1) = w(n)$$

if $w^T x(n) \leq 0$ and $x(n)$ belongs to class G_2.

2. Otherwise, the weight vector of the perceptron is updated in accordance with the rule.

Architecture of multilayer perceptron :

1. Fig. 4.15.1 shows architectural graph of multilayer perceptron with two hidden layer and an output layer.
2. Signal flow through the network progresses in a forward direction, from the left to right and on a layer-by-layer basis.
3. Two kinds of signals are identified in this network :
   a. **Functional signals :** Functional signal is an input signal and propagates forward and emerges at the output end of the network as an output signal.

b. **Error signals:** Error signal originates at an output neuron and propagates backward through the network.

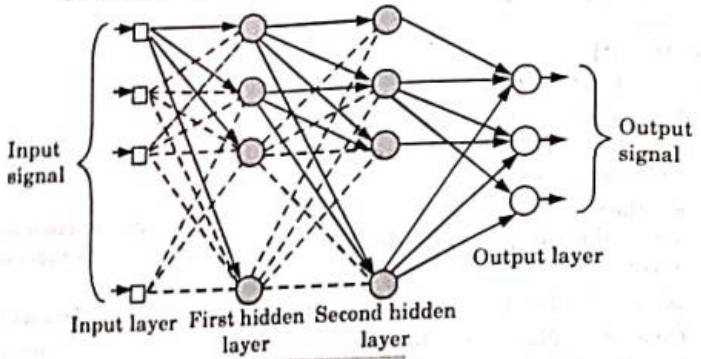


Fig. 4.15.1.

4. Multilayer perceptrons have been applied successfully to solve some difficult and diverse problems by training them in a supervised manner with highly popular algorithm known as the error backpropagation algorithm.

Characteristics of multilayer perceptron :

1. In this model, each neuron in the network includes a non-linear activation function (non-linearity is smooth). Most commonly used non-linear function is defined by :

$$y_j = \frac{1}{1+\exp(-v_j)}$$

where v_j is the induced local field (*i.e.*, the sum of all weights and bias) and y is the output of neuron j.

2. The network contains hidden neurons that are not a part of input or output of the network. Hidden layer of neurons enabled network to learn complex tasks.
3. The network exhibits a high degree of connectivity.

Que 4.16. How tuning parameters effect the backpropagation neural network ?

Answer

Effect of tuning parameters of the backpropagation neural network :

1. **Momentum factor :**
   a. The momentum factor has a significant role in deciding the values of learning rate that will produce rapid learning.
   b. It determines the size of change in weights or biases.

   c. If momentum factor is zero, the smoothening is minimum and the entire weight adjustment comes from the newly calculated change.
   d. If momentum factor is one, new adjustment is ignored and previous one is repeated.
   e. Between 0 and 1 is a region where the weight adjustment is smoothened by an amount proportional to the momentum factor.
   f. The momentum factor effectively increases the speed of learning without leading to oscillations and filters out high frequency variations of the error surface in the weight space.

2. **Learning coefficient :**
   a. A formula to select learning coefficient is :

$$h = \frac{1.5}{(N_1^2 + N_2^2 + ... + N_m^2)}$$

   Where N_1 is the number of patterns of type 1 and m is the number of different pattern types.
   b. The small value of learning coefficient less than 0.2 produces slower but stable training.
   c. The largest value of learning coefficient *i.e.*, greater than 0.5, the weights are changed drastically but this may cause optimum combination of weights to be overshot resulting in oscillations about the optimum.
   d. The optimum value of learning rate is 0.6 which produce fast learning without leading to oscillations.

3. **Sigmoidal gain :**
   a. If sigmoidal function is selected, the input-output relationship of the neuron can be set as

$$O = \frac{1}{(1 + e^{-\lambda(1+\theta)})} \quad ...(4.16.1)$$

   where λ is a scaling factor known as sigmoidal gain.
   b. As the scaling factor increases, the input-output characteristic of the analog neuron approaches that of the two state neuron or the activation function approaches the (Satisifiability) function.
   c. It also affects the backpropagation. To get graded output, as the sigmoidal gain factor is increased, learning rate and momentum factor have to be decreased in order to prevent oscillations.

4. **Threshold value :**
   a. 0 in eq. (4.16.1) is called as threshold value or the bias or the noise factor.

b. A neuron fires or generates an output if the weighted sum of the input exceeds the threshold value.

c. One method is to simply assign a small value to it and not to change it during training.

d. The other method is to initially choose some random values and change them during training.

Que 4.17. Discuss selection of various parameters in Backpropagation Neural Network (BPN).

Answer

Selection of various parameters in BPN :

1. **Number of hidden nodes :**
   a. The guiding criterion is to select the minimum nodes in the first and third layer, so that the memory demand for storing the weights can be kept minimum.
   b. The number of separable regions in the input space M, is a function of the number of hidden nodes H in BPN and H = M – 1.
   c. When the number of hidden nodes is equal to the number of training patterns, the learning could be fastest.
   d. In such cases, BPN simply remembers training patterns losing all generalization capabilities.
   e. Hence, as far as generalization is concerned, the number of hidden nodes should be small compared to the number of training patterns with help of Vapnik Chervonenkis dimension (VCdim) of probability theory.
   f. We can estimate the selection of number of hidden nodes for a given number of training patterns as number of weights which is equal to $I_1 * I_2 + I_2 * I_3$, where I_1 and I_3 denote input and output nodes and I_2 denote hidden nodes.
   g. Assume the training samples T to be greater than VCdim. Now if we accept the ratio 10 : 1

$$10 * T = \frac{I_2}{(I_1 + I_3)}$$

$$I_2 = \frac{10T}{(I_1 + I_3)}$$

   Which yields the value for I_2.

2. **Momentum coefficient α :**
   a. To reduce the training time we use the momentum factor because it enhances the training process.
   b. The influences of momentum on weight change is

$$[\Delta W]^{n+1} = -\eta \frac{\partial E}{\partial W} + \alpha [\Delta W]^n$$

   c. The momentum also overcomes the effect of local minima.
   d. The use of momentum term will carry a weight change process through one or local minima and get it into global minima.

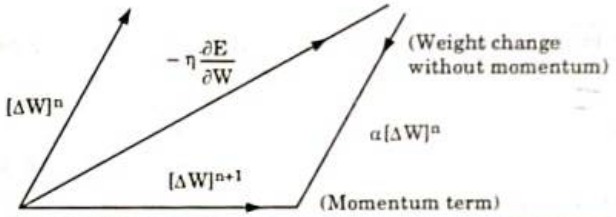


Fig. 4.17.1. Influence of momentum term on weight change.

3. **Sigmoidal gain λ :**
   a. When the weights become large and force the neuron to operate in a region where sigmoidal function is very flat, a better method of coping with network paralysis is to adjust the sigmoidal gain.
   b. By decreasing this scaling factor, we effectively spread out sigmoidal function on wide range so that training proceeds faster.

4. **Local minima :**
   a. One of the most practical solutions involves the introduction of a shock which changes all weights by specific or random amounts.
   b. If this fails, then the most practical solution is to rerandomize the weights and start the training all over.

PART-3

Unspervised Learning, SOM Algorithm and its Variants.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.18. Write short note on unsupervised learning.

Answer

1. Unsupervised learning is the training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance.

2. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.
3. Unlike supervised learning, no teacher is provided that means no training will be given to the machine.
4. Therefore machine is restricted to find the hidden structure in unlabeled data by our-self.

Que 4.19. Classify unsupervised learning into two categories of algorithm.

Answer

Classification of unsupervised learning algorithm into two categories :

1. **Clustering :** A clustering problem is where we want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
2. **Association :** An association rule learning problem is where we want to discover rules that describe large portions of our data, such as people that buy *X* also tend to buy *Y*.

Que 4.20. What are the applications of unsupervised learning ?

Answer

Following are the application of unsupervised learning :

1. Unsupervised learning automatically split the dataset into groups base on their similarities.
2. Anomaly detection can discover unusual data points in our dataset. It is useful for finding fraudulent transactions.
3. Association mining identifies sets of items which often occur together in our dataset.
4. Latent variable models are widely used for data preprocessing. Like reducing the number of features in a dataset or decomposing the dataset into multiple components.

Que 4.21. What is Self-Organizing Map (SOM) ?

Answer

1. Self-Organizing Map (SOM) provides a data visualization technique which helps to understand high dimensional data by reducing the dimensions of data to a map.
2. SOM also represents clustering concept by grouping similar data together.

3. A self-Organizing Map (SOM) or Self-Organizing Feature Map (SOFM) is a type of Artificial Neural Network (ANN) that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map, and is therefore a method to do dimensionality reduction.
4. Self-organizing maps differ from other artificial neural networks as they apply competitive learning as opposed to error-correction learning (such as backpropagation with gradient descent), and in the sense that they use a neighborhood function to preserve the topological properties of the input space.

Que 4.22. Write the steps used in SOM algorithm.

Answer

Following are the steps used in SOM algorithm :

1. Each node's weights are initialized.
2. A vector is chosen at random from the set of training data.
3. Every node is examined to calculate which one's weights are most like the input vector. The winning node is commonly known as the Best Matching Unit (BMU).
4. Then the neighbourhood of the BMU is calculated. The amount of neighbors decreases over time.
5. The winning weight is rewarded with becoming more like the sample vector. The neighbours also become more like the sample vector. The closer a node is to the BMU, the more its weights get altered and the farther away the neighbor is from the BMU, the less it learns.
6. Repeat step 2 for N iterations.

Que 4.23. What are the basic processes used in SOM ? Also explain stages of SOM algorithm.

Answer

Basics processes used in SOM algorithm are :

1. **Initialization :** All the connection weights are initialized with small random values.
2. **Competition :** For each input pattern, the neurons compute their respective values of a discriminant function which provides the basis for competition. The particular neuron with the smallest value of the discriminant function is declared the winner.

3. **Cooperation :** The winning neuron determines the spatial location of a topological neighbourhood of excited neurons, thereby providing the basis for cooperation among neighbouring neurons.
4. **Adaptation :** The excited neurons decrease their individual values of the discriminant function in relation to the input pattern through suitable adjustment of the associated connection weights, such that the response of the winning neuron to the subsequent application of a similar input pattern is enhanced.

Stages of SOM algorithm are :

1. **Initialization :** Choose random values for the initial weight vectors w_j.
2. **Sampling :** Draw a sample training input vector x from the input space.
3. **Matching :** Find the winning neuron $I(x)$ that has weight vector closest to the input vector, *i.e.*, the minimum value of $d_j(x) = \sum_{i=1}^{D} (x_i - w_{ji})^2$.
4. **Updating :** Apply the weight update equation

$$Dw_{ji} = h(t)\, T_{j,I(x)}(t)\, (x_i - w_{ji})$$

where $T_{j,I(x)}(t)$ is a Gaussian neighbourhood and $h(t)$ is the learning rate.

5. **Continuation :** Keep returning to step 2 until the feature map stops changing.

PART-4

Deep Learning, Introduction, Concept of Convolutional Neural Network, Types of Layers, (Convolutional Layers, Activation Function, Pooling, Fully Connected).

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.24. **What do you understand by deep learning ?**

Answer

1. Deep learning is the subfield of artificial intelligence that focuses on creating large neural network models that are capable of making accurate data-driven decisions.

2. Deep learning is used where the data is complex and has large datasets.
3. Facebook uses deep learning to analyze text in online conversations. Google and Microsoft all use deep learning for image search and machine translation.
4. All modern smart phones have deep learning systems running on them. For example, deep learning is the standard technology for speech recognition, and also for face detection on digital cameras.
5. In the healthcare sector, deep learning is used to process medical images (X-rays, CT, and MRI scans) and diagnose health conditions.
6. Deep learning is also at the core of self-driving cars, where it is used for localization and mapping, motion planning and steering, and environment perception, as well as tracking driver state.

Que 4.25. **Describe different architecture of deep learning.**

Answer

Different architecture of deep learning are :

1. **Deep Neural Network :** It is a neural network with a certain level of complexity (having multiple hidden layers in between input and output layers). They are capable of modeling and processing non-linear relationships.
2. **Deep Belief Network (DBN) :** It is a class of Deep Neural Network. It is multi-layer belief networks. Steps for performing DBN are :
   a. Learn a layer of features from visible units using Contrastive Divergence algorithm.
   b. Treat activations of previously trained features as visible units and then learn features of features.
   c. Finally, the whole DBN is trained when the learning for the final hidden layer is achieved.
3. **Recurrent (perform same task for every element of a sequence) Neural Network :** Allows for parallel and sequential computation. Similar to the human brain (large feedback network of connected neurons). They are able to remember important things about the input they received and hence enable them to be more precise.

Que 4.26. **What are the advantages, disadvantages and limitation of deep learning ?**

Answer

Advantages of deep learning :

1. Best in-class performance on problems.
2. Reduces need for feature engineering.
3. Eliminates unnecessary costs.
4. Identifies defects easily that are difficult to detect.

Disadvantages of deep learning :

1. Large amount of data required.
2. Computationally expensive to train.
3. No strong theoretical foundation.

Limitations of deep learning :

1. Learning through observations only.
2. The issue of biases.

Que 4.27. What are the various applications of deep learning ?

Answer

Following are the application of deep learning :

1. **Automatic text generation :** Corpus of text is learned and from this model new text is generated, word-by-word or character-by-character. Then this model is capable of learning how to spell, punctuate, form sentences, or it may even capture the style.
2. **Healthcare :** Helps in diagnosing various diseases and treating it.
3. **Automatic machine translation :** Certain words, sentences or phrases in one language is transformed into another language (Deep Learning is achieving top results in the areas of text, images).
4. **Image recognition :** Recognizes and identifies peoples and objects in images as well as to understand content and context. This area is already being used in Gaming, Retail, Tourism, etc.
5. **Predicting earthquakes :** Teaches a computer to perform viscoelastic computations which are used in predicting earthquakes.

Que 4.28. Define convolutional networks.

Answer

1. Convolutional networks also known as Convolutional Neural Networks (CNNs) are a specialized kind of neural network for processing data that has a known, grid-like topology.
2. Convolutional neural network indicates that the network employs a mathematical operation called convolution.
3. Convolution is a specialized kind of linear operation.
4. Convolutional networks are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers.
5. CNNs, (ConvNets), are quite similar to regular neural networks.
6. They are still made up of neurons with weights that can be learned from data. Each neuron receives some inputs and performs a dot product.
7. They still have a loss function on the last fully connected layer.
8. They can still use a non-linearity function a regular neural network receives input data as a single vector and passes through a series of hidden layers.

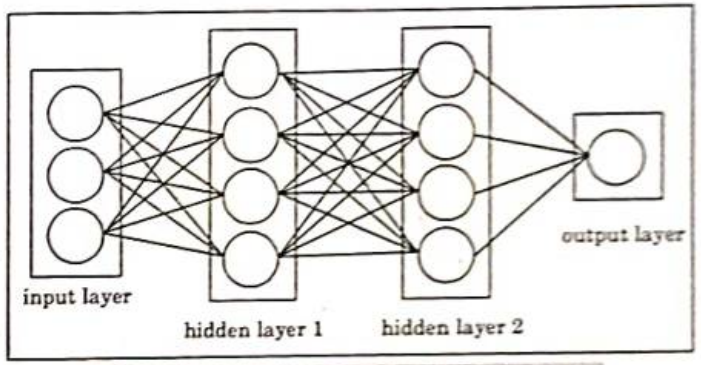


Fig. 4.28.1. A regular three-layer neural network.

9. Every hidden layer consists of neurons, wherein every neuron is fully connected to all the other neurons in the previous layer.
10. Within a single layer, each neuron is completely independent and they do not share any connections.
11. The fully connected layer, (the output layer), contains class scores in the case of an image classification problem. There are three main layers in a simple ConvNet.

Que 4.29. **Write short note on convolutional layer.**

Answer

1. Convolutional layers are the major building blocks used in convolutional neural networks.
2. A convolution is the simple application of a filter to an input that results in an activation.
3. Repeated application of the same filter to an input results in a map of activations called a feature map, indicating the locations and strength of a detected feature in an input, such as an image.
4. The innovation of convolutional neural networks is the ability to automatically learn a large number of filters in parallel specific to a training dataset under the constraints of a specific predictive modeling problem, such as image classification.
5. The result is highly specific features that can be detected anywhere on input images.

Que 4.30. **Describe briefly activation function, pooling and fully connected layer.**

Answer

Activation function :

1. An activation function is a function that is added into an artificial neural network in order to help the network learn complex patterns in the data.
2. When comparing with a neuron-based model that is in our brains, the activation function is at the end deciding what is to be fired to the next neuron.
3. That is exactly what an activation function does in an ANN as well.
4. It takes in the output signal from the previous cell and converts it into some form that can be taken as input to the next cell.

Pooling layer :

1. A pooling layer is a new layer added after the convolutional layer. Specifically, after a non-linearity (for example ReLU) has been applied to the feature maps output by a convolutional layer, for example, the layers in a model may look as follows :
   a. Input image
   b. Convolutional layer
   c. Non-linearity
   d. Pooling layer
2. The addition of a pooling layer after the convolutional layer is a common pattern used for ordering layers within a convolutional neural network that may be repeated one or more times in a given model.
3. The pooling layer operates upon each feature map separately to create a new set of the same number of pooled feature maps.

Fully connected layer :

1. Fully connected layers are an essential component of Convolutional Neural Networks (CNNs), which have been proven very successful in recognizing and classifying images for computer vision.
2. The CNN process begins with convolution and pooling, breaking down the image into features, and analyzing them independently.
3. The result of this process feeds into a fully connected neural network structure that drives the final classification decision.

PART-5

Concept of Convolution (1D and 2D) Layers, Training of Network, Case Study of CNN for eg on Diabetic Retinopathy, Building a Smart Speaker, Self Deriving Car etc.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.31. **Explain 1D and 2D convolutional neural network.**

Answer

1D convolutional neural network :

1. Convolutional Neural Network (CNN) models were developed for image classification, in which the model accepts a two-dimensional input representing an image's pixels and color channels, in a process called feature learning.

2. This same process can be applied to one-dimensional sequences of data.
3. The model extracts features from sequences data and maps the internal features of the sequence.
4. A 1D CNN is very effective for deriving features from a fixed-length segment of the overall dataset, where it is not so important where the feature is located in the segment.
5. 1D Convolutional Neural Networks work well for :
   a. Analysis of a time series of sensor data.
   b. Analysis of signal data over a fixed-length period, for example, an audio recording.
   c. Natural Language Processing (NLP), although Recurrent Neural Networks which leverage Long Short Term Memory (LSTM) cells are more promising than CNN as they take into account the proximity of words to create trainable patterns.

2D convolutional neural network :

1. In a 2D convolutional network, each pixel within the image is represented by its x and y position as well as the depth, representing image channels (red, green, and blue).
2. It moves over the images both horizontally and vertically.

Que 4.32. How we trained a network ? Explain.

Answer

1. Once a network has been structured for a particular application, that network is ready to be trained.
2. To start this process the initial weights are chosen randomly. Then, the training, or learning begins.
3. There are two approaches to training :
   a. In supervised training, both the inputs and the outputs are provided. The network then processes the inputs and compares its resulting outputs against the desired outputs.
   b. Errors are then propagated back through the system, causing the system to adjust the weights which control the network. This process occurs over and over as the weights are continually tweaked.
   c. The set of data which enables the training is called the "training set." During the training of a network the same set of data is processed many times as the connection weights are ever refined.

   d. The other type of training is called unsupervised training. In unsupervised training, the network is provided with inputs but not with desired outputs.
   e. The system itself must then decide what features it will use to group the input data. This is often referred to as self-organization or adaption.

Que 4.33. Describe diabetic retinopathy on the basis of deep learning.

Answer

1. Diabetic Retinopathy (DR) is one of the major causes of blindness in the western world. Increasing life expectancy, indulgent lifestyles and other contributing factors mean the number of people with diabetes is projected to continue rising.
2. Regular screening of diabetic patients for DR has been shown to be a cost-effective and important aspect of their care.
3. The accuracy and timing of this care is of significant importance to both the cost and effectiveness of treatment.
4. If detected early enough, effective treatment of DR is available; making this a vital process.
5. Classification of DR involves the weighting of numerous features and the location of such features. This is highly time consuming for clinicians.
6. Computers are able to obtain much quicker classifications once trained, giving the ability to aid clinicians in real-time classification.
7. The efficacy of automated grading for DR has been an active area of research in computer imaging with encouraging conclusions.
8. Significant work has been done on detecting the features of DR using automated methods such a support vector machines and k-NN classifiers.
9. The majority of these classification techniques are on two class classification for DR or no DR.

Que 4.34. Using artificial neural network how we recognize speaker.

Answer

1. With the technology advancements in smart home sector, voice control and automation are key components that can make a real difference in people's lives.

2. The voice recognition technology market continues to involve rapidly as almost all smart home devices are providing speaker recognition capability today.
3. However, most of them provide cloud-based solutions or use very deep Neural Networks for speaker recognition task, which are not suitable models to run on smart home devices.
4. Here, we compare relatively small Convolutional Neural Networks (CNN) and evaluate effectiveness of speaker recognition using these models on edge devices. In addition, we also apply transfer learning technique to deal with a problem of limited training data.
5. By developing solution suitable for running inference locally on edge devices, we eliminate the well-known cloud computing issues, such as data privacy and network latency, etc.
6. The preliminary results proved that the chosen model adapts the benefit of computer vision task by using CNN and spectrograms to perform speaker classification with precision and recall ~ 84 % in time less than 60 ms on mobile device with Atom Cherry Trail processor.

Que 4.35. Artificial intelligence plays important role in self-driving car explain.

Answer

1. The rapid development of the Internet economy and Artificial Intelligence (AI) has promoted the progress of self-driving cars.
2. The market demand and economic value of self-driving cars are increasingly prominent. At present, more and more enterprises and scientific research institutions have invested in this field. Google, Tesla, Apple, Nissan, Audi, General Motors, BMW, Ford, Honda, Toyota, Mercedes, and Volkswagen have participated in the research and development of self-driving cars.
3. Google is an Internet company, which is one of the leaders in self-driving cars, based on its solid foundation in artificial intelligence.
4. In June 2015, two Google self-driving cars were tested on the road. So far, Google vehicles have accumulated more than 3.2 million km of tests, becoming the closest to the actual use.
5. Another company that has made great progress in the field of self-driving cars is Tesla. Tesla was the first company to devote self-driving technology to production.

6. Followed by the Tesla models series, its "auto-pilot" technology has made major breakthroughs in recent years.
7. Although the Tesla's autopilot technology is only regarded as Level 2 stage by the National Highway Traffic Safety Administration (NHTSA), Tesla shows us that the car has basically realized automatic driving under certain conditions.

☺☺☺

5 UNIT

Reinforcement Learning and Genetic Algorithm

CONTENTS

PART-1

Introduction to Reinforcement Learning.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.1. Describe reinforcement learning.

Answer

1. Reinforcement learning is the study of how animals and artificial systems can learn to optimize their behaviour in the face of rewards and punishments.
2. Reinforcement learning algorithms related to methods of dynamic programming which is a general approach to optimal control.
3. Reinforcement learning phenomena have been observed in psychological studies of animal behaviour, and in neurobiological investigations of neuromodulation and addiction.
4. The task of reinforcement learning is to use observed rewards to learn an optimal policy for the environment. An optimal policy is a policy that maximizes the expected total reward.
5. Without some feedback about what is good and what is bad, the agent will have no grounds for deciding which move to make.
6. The agents needs to know that something good has happened when it wins and that something bad has happened when it loses.
7. This kind of feedback is called a reward or reinforcement.
8. Reinforcement learning is valuable in the field of robotics, where the tasks to be performed are frequently complex enough to defy encoding as programs and no training data is available.
9. In many complex domains, reinforcement learning is the only feasible way to train a program to perform at high levels.

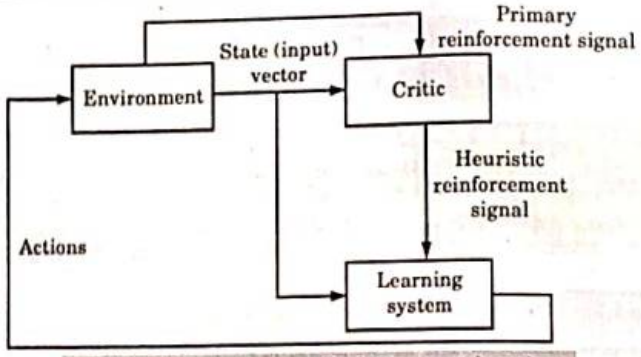


Fig. 5.1.1. Block diagram of reinforcement learning.

Que 5.2. **Differentiate between reinforcement and supervised learning.**

Answer

S. No.	Reinforcement learning	Supervised learning
1.	Reinforcement learning is all about making decisions sequentially. In simple words we can say that the output depends on the state of the current input and the next input depends on the output of the previous input.	In supervised learning, the decision is made on the initial input or the input given at the start.
2.	In reinforcement learning decision is dependent. So, we give labels to sequences of dependent decisions.	Supervised learning decisions are independent of each other so labels are given to each decision.
3.	**Example :** Chess game.	**Example :** Object recognition.

Que 5.3. **What is reinforcement learning ? Explain passive reinforcement learning and active reinforcement learning.**

Answer

Reinforcement learning : Refer Q. 5.1, Page 5–2L, Unit-5.

Passive reinforcement learning :

1. In passive learning, the agent's policy π is fixed. In state s, it always executes the action $\pi(s)$.
2. Its goal is simply to learn how good the policy is – that is, to learn the utility function $U^{\pi}(s)$.
3. Fig. 5.3.1 shows a policy for the world and the corresponding utilities.
4. In Fig. 5.3.1(a) the policy happens to be optimal with rewards of $R(s) = -0.04$ in the non-terminal states and no discounting.
5. Passive learning agent does not know the transition model $T(s, a, s')$, which specifies the probability of reaching state s' from state s after doing action a; nor does it know the reward function $R(s)$ which specifies the reward for each state.
6. The agent executes a set of trials in the environment using its policy π.
7. In each trial, the agent starts in state (1, 1) and experiences a sequence of state transitions until it reaches one of the terminal states, (4, 2) or (4, 3).
8. Its percepts supply both the current state and the reward received in that state. Typical trials might look like this.

$(1,1)_{-0.04} \rightarrow (1,2)_{-0.04} \rightarrow (1,3)_{-0.04} \rightarrow (1,2)_{-0.04} \rightarrow (1,3)_{-0.04} \rightarrow (2,3)_{-0.04} \rightarrow (3,3)_{-0.04} \rightarrow (4,3)_{+1}$
$(1,1)_{-0.04} \rightarrow (1,2)_{-0.04} \rightarrow (1,3)_{-0.04} \rightarrow (2,3)_{-0.04} \rightarrow (3,3)_{-0.04} \rightarrow (3,2)_{-0.04} \rightarrow (3,3)_{-0.04} \rightarrow (4,3)_{+1}$
$(1,1)_{-0.04} \rightarrow (2,1)_{-0.04} \rightarrow (3,1)_{-0.04} \rightarrow (3,2)_{-0.04} \rightarrow (4,2)_{-1}$

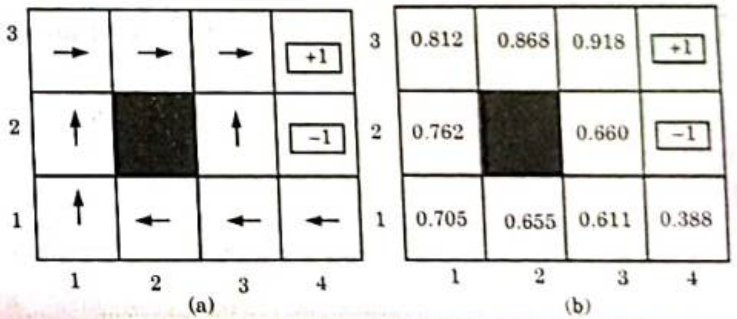


Fig. 5.3.1. (a) A policy π for the 4 × 3 world;
(b) The utilities of the states in the 4 × 3 world, given policy π.

9. Each state percept is subscripted with the reward received. The object is to use the information about rewards to learn the expected utility $U^{\pi}(s)$ associated with each non-terminal state s.
10. The utility is defined to be the expected sum of (discounted) rewards obtained if policy π is followed :

$$U^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \mid \pi, s_0 = s\right]$$

where γ is a discount factor, for the 4×5 world we set $\gamma = 1$.

Active reinforcement learning :

1. An active agent must decide what actions to take.
2. First, the agent will need to learn a complete model with outcome probabilities for all actions, rather than just model for the fixed policy.
3. We need to take into account the fact that the agent has a choice of actions.
4. The utilities it needs to learn are those defined by the optimal policy, they obey the Bellman equations :

$$U(S) = R(S) + \gamma \max_{a} \sum_{s'} T(s, a, s') U(s')$$

5. These equations can be solved to obtain the utility function U using the value iteration or policy iteration algorithms.
6. A utility function U is optimal for the learned model, the agent can extract an optimal action by one-step look-ahead to maximize the expected utility.
7. Alternatively, if it uses policy iteration, the optimal policy is already available, so it should simply execute the action the optimal policy recommends.

Que 5.4. What are the different types of reinforcement learning ? Explain.

Answer

Types of reinforcement learning :

1. **Positive reinforcement learning :**
   a. Positive reinforcement learning is defined as when an event, occurs due to a particular behaviour, increases the strength and the frequency of the behaviour.
   b. In other words, it has a positive effect on the behaviour.
   c. Advantages of positive reinforcement learning are :
      i. Maximizes performance.
      ii. Sustain change for a long period of time.
   d. Disadvantages of positive reinforcement learning :
      i. Too much reinforcement can lead to overload of states which can diminish the results.
2. **Negative reinforcement learning :**
   a. Negative reinforcement is defined as strengthening of behaviour because a negative condition is stopped or avoided.

   b. Advantages of negative reinforcement learning :
      i. Increases behaviour.
      ii. It provide defiance to minimum standard of performance.
   c. Disadvantages of negative reinforcement learning :
      i. It only provides enough to meet up the minimum behaviour.

Que 5.5. What are the elements of reinforcement learning ?

Answer

Elements of reinforcement learning :

1. **Policy (π) :**
   a. It defines the behaviour of the agent which action to take in a given state to maximize the received reward in the long term.
   b. It stimulus-response rules or associations.
   c. It could be a simple lookup table or function, or need more extensive computation (for example, search).
   d. It can be probabilistic.
2. **Reward function (r) :**
   a. It defines the goal in a reinforcement learning problem, maps a state or action to a scalar number, the reward (or reinforcement).
   b. The RL agent's objective is to maximize the total reward it receives in the long run.
   c. It defines good and bad events.
   d. It cannot be altered by the agent but may inform change of policy.
   e. It can be probabilistic (expected reward).
3. **Value function (V) :**
   a. It defines the total amount of reward an agent can expect to accumulate over the future, starting from that state.
   b. A state may yield a low reward but have a high value (or the opposite). For example, immediate pain/pleasure vs. long term happiness.
4. **Transition model (M) :**
   a. It defines the transitions in the environment action a taken in the states, will lead to state s^2.
   b. It can be probabilistic.

PART-2

Learning Task, Example of Reinforcement Learning in Practice.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.6. Describe briefly learning task used in machine learning.

Answer

1. A machine learning task is the type of prediction or inference being made, based on the problem or question that is being asked, and the available data.
2. For example, the classification task assigns data to categories, and the clustering task groups data according to similarity.
3. Machine learning tasks rely on patterns in the data rather than being explicitly programmed.
4. A supervised machine learning task that is used to predict which of two classes (categories) an instance of data belongs to.
5. The input of a classification algorithm is a set of labeled examples, where each label is an integer of either 0 or 1.
6. The output of a binary classification algorithm is a classifier, which we can use to predict the class of new unlabeled instances.
7. An unsupervised machine learning task that is used to group instances of data into clusters that contain similar characteristics.
8. Clustering can also be used to identify relationships in a dataset that we might not logically derive by browsing or simple observation.
9. The inputs and outputs of a clustering algorithm depend on the methodology chosen.

Que 5.7. Explain different machine learning task.

Answer

Following are most common machine learning tasks :

1. **Data preprocessing :** Before starting training the models, it is important to prepare data appropriately. As part of data preprocessing following is done :
   a. Data cleaning
   b. Handling missing data
2. **Exploratory data analysis :** Once data is preprocessed, the next step is to perform exploratory data analysis to understand data distribution and relationship between / within the data.

3. **Feature engineering / selection :** Feature selection is one of the critical tasks which would be used when building machine learning models. Feature selection is important because selecting right features would not only help build models of higher accuracy but also help achieve objectives related to building simpler models, reduce overfitting etc.
4. **Regression :** Regression tasks deal with estimation of numerical values (continuous variables). Some of the examples include estimation of housing price, product price, stock price etc.
5. **Classification :** Classification task is related with predicting a category of a data (discrete variables). Most common example is predicting whether or not an email is spam or not, whether a person is suffering from a particular disease or not, whether a transaction is fraud or not, etc.
6. **Clustering :** Clustering tasks are all about finding natural groupings of data and a label associated with each of these groupings (clusters). Some of the common example includes customer segmentation, product features identification for product roadmap.
7. **Multivariate querying :** Multivariate querying is about querying or finding similar objects.
8. **Density estimation :** Density estimation problems are related with finding likelihood or frequency of objects.
9. **Dimension reduction :** Dimension reduction is the process of reducing the number of random variables under consideration, and can be divided into feature selection and feature extraction.
10. **Model algorithm / selection :** Many a times, there are multiple models which are trained using different algorithms. One of the important task is to select most optimal models for deploying them in production.
11. **Testing and matching :** Testing and matching tasks relates to comparing data sets.

Que 5.8. Explain reinforcement learning with the help of an example.

Answer

1. Reinforcement learning (RL) is learning concerned with how software agents ought to take actions in an environment in order to maximize the notion of cumulative reward.
2. The software agent is not told which actions to take, but instead must discover which actions yield the most reward by trying them.

   For example,

   Consider the scenario of teaching new tricks to a cat :

1. As cat does not understand English or any other human language, we cannot tell her directly what to do. Instead, we follow a different strategy.

2. We emulate a situation, and the cat tries to respond in many different ways. If the cat's response is the desired way, we will give her fish.
3. Now whenever the cat is exposed to the same situation, the cat executes a similar action even more enthusiastically in expectation of getting more reward (food).
4. That's like learning that cat gets from "what to do" from positive experiences.
5. At the same time, the cat also learns what not do when faced with negative experiences.

Working of reinforcement learning :

1. In this case, the cat is an agent that is exposed to the environment (In this case, it is your house). An example of a state could be our cat sitting, and we use a specific word in for cat to walk.
2. Our agent reacts by performing an action transition from one "state" to another "state."
3. For example, the cat goes from sitting to walking.
4. The reaction of an agent is an action, and the policy is a method of selecting an action given a state in expectation of better outcomes.
5. After the transition, they may get a reward or penalty in return.

PART-3

Learning Models for Reinforcement (Markov Decision Process, Q Learning, Q Learning Function, Q Learning Algorithm), Application of Reinforcement Learning.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.9. **Describe important term used in reinforcement learning method.**

Answer

Following are the terms used in reinforcement learning :

Agent : It is an assumed entity which performs actions in an environment to gain some reward.

i. **Environment (e) :** A scenario that an agent has to face.

ii. **Reward (R) :** An immediate return given to an agent when he or she performs specific action or task.

iii. **State (s) :** State refers to the current situation returned by the environment.

iv. **Policy (π) :** It is a strategy which applies by the agent to decide the next action based on the current state.

v. **Value (V) :** It is expected long-term return with discount, as compared to the short-term reward.

vi. **Value Function :** It specifies the value of a state that is the total amount of reward. It is an agent which should be expected beginning from that state.

vii. **Model of the environment :** This mimics the behavior of the environment. It helps you to make inferences to be made and also determine how the environment will behave.

viii. **Model based methods :** It is a method for solving reinforcement learning problems which use model-based methods.

ix. **Q value or action value (Q) :** Q value is quite similar to value. The only difference between the two is that it takes an additional parameter as a current action.

Que 5.10. **Explain approaches used to implement reinforcement learning algorithm.**

Answer

There are three approaches used implement a reinforcement learning algorithm :

1. **Value-Based :**

a. In a value-based reinforcement learning method, we should try to maximize a value function V(s). In this method, the agent is expecting a long-term return of the current states under policy π.

2. **Policy-based :**

a. In a policy-based RL method, we try to come up with such a policy that the action performed in every state helps you to gain maximum reward in the future.

b. Two types of policy-based methods are :

i. **Deterministic :** For any state, the same action is produced by the policy π.

ii. **Stochastic :** Every action has a certain probability, which is determined by the following equation stochastic policy :

$$n(a/s) = P/A = a/S = S$$

3. **Model-Based :**

a. In this Reinforcement Learning method, we need to create a virtual model for each environment.

b. The agent learns to perform in that specific environment.

Que 5.11. Describe learning models of reinforcement learning.

Answer

1. Reinforcement learning is defined by a specific type of problem, and all its solutions are classed as reinforcement learning algorithms.
2. In the problem, an agent is supposed to decide the best action to select based on his current state.
3. When this step is repeated, the problem is known as a Markov Decision Process.
4. A Markov Decision Process (MDP) model contains :

a. A State is a set of tokens that represent every state that the agent can be in.

b. A Model (sometimes called Transition Model) gives an action's effect in a state. In particular, T(S, a, S') defines a transition T where being in state S and taking an action 'a' takes us to state S' (S and S' may be same).

c. An Action A is set of all possible actions. A(s) defines the set of actions that can be taken being in state S.

d. A Reward is a real-valued reward function. R(s) indicates the reward for simply being in the state S. R(S,a) indicates the reward for being in a state S and taking an action 'a'. R(S,a,S') indicates the reward for being in a state S, taking an action 'a' and ending up in a state S'.

e. A Policy is a solution to the Markov Decision Process. A policy is a mapping from S to a. It indicates the action 'a' to be taken while in state S.

Que 5.12. What are the application of reinforcement learning and why we use reinforcement learning ?

Answer

Following are the applications of reinforcement learning :

1. Robotics for industrial automation.
2. Business strategy planning.
3. Machine learning and data processing.
4. It helps us to create training systems that provide custom instruction and materials according to the requirement of students.
5. Aircraft control and robot motion control.

Following are the reasons for using reinforcement learning :

1. It helps us to find which situation needs an action.
2. Helps us to discover which action yields the highest reward over the longer period.

3. Reinforcement Learning also provides the learning agent with a reward function.
4. It also allows us to figure out the best method for obtaining large rewards.

Que 5.13. When not to use reinforcement learning ? What are the challenges of reinforcement learning ?

Answer

We cannot apply reinforcement learning model is all the situation. Following are the conditions when we should not use reinforcement learning model.

1. When we have enough data to solve the problem with a supervised learning method.
2. When the action space is large reinforcement learning is computing heavy and time-consuming.

Challenges we will face while doing reinforcement learning are :

1. Feature/reward design which should be very involved.
2. Parameters may affect the speed of learning.
3. Realistic environments can have partial observability.
4. Too much reinforcement may lead to an overload of states which can diminish the results.
5. Realistic environments can be non-stationary.

Que 5.14. Explain the term Q-learning.

Answer

1. Q-learning is a model-free reinforcement learning algorithm.
2. Q-learning is a values-based learning algorithm. Value based algorithms updates the value function based on an equation (particularly Bellman equation).
3. Whereas the other type, policy-based estimates the value function with a greedy policy obtained from the last policy improvement.
4. Q-learning is an off-policy learner *i.e.*, it learns the value of the optimal policy independently of the agent's actions.
5. On the other hand, an on-policy learner learns the value of the policy being carried out by the agent, including the exploration steps and it will find a policy that is optimal, taking into account the exploration inherent in the policy.

Que 5.15. Describe Q-learning algorithm process.

Answer

Step 1 : Initialize the Q-table : First the Q-table has to be built. There are n columns, where n = number of actions. There are m rows, where m = number of states.

In our example n = Go left, Go right, Go up and Go down and m = Start, Idle, Correct path, Wrong path and End. First, lets initialize the value at 0.

Step 2 : Choose an action.

Step 3 : Perform an action : The combination of steps 2 and 3 is performed for an undefined amount of time. These steps run until the time training is stopped, or when the training loop stopped as defined in the code.

a. First, an action (a) in the state (s) is chosen based on the Q-table. Note that, when the episode initially starts, every Q-value should be 0.

b. Then, update the Q-values for being at the start and moving right using the Bellman equation.

Step 4 : Measure reward : Now we have taken an action and observed an outcome and reward.

Step 5 : Evaluate : We need to update the function Q(s, a)

This process is repeated again and again until the learning is stopped. In this way the Q table is been updated and the value function Q is maximized. Here the Q returns the expected future reward of that action at that state.

PART-4

Introduction to Deep Q Learning.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.16. Describe deep Q-learning.

Answer

1. In deep Q-learning, we use a neural network to approximate the Q-value function.
2. The state is given as the input and the Q-value of all possible actions is generated as the output.
3. The comparison between Q-learning and deep Q-learning is illustrated below :

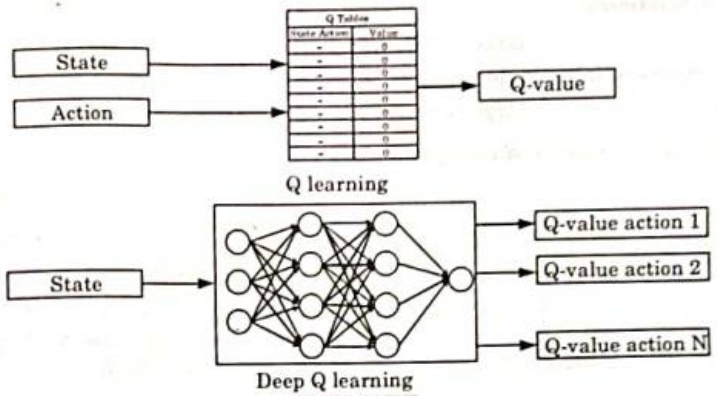


Fig. 5.16.1.

4. On a higher level, Deep Q learning works as such :
   i. Gather and store samples in a replay buffer with current policy.
   ii. Random sample batches of experiences from the replay buffer.
   iii. Use the sampled experiences to update the Q network.
   iv. Repeat 1-3.

Que 5.17. What are the steps involved in deep Q-learning network ?

Answer

Steps involved in reinforcement learning using deep Q-learning networks :

1. All the past experience is stored by the user in memory.
2. The next action is determined by the maximum output of the Q-network.
3. The loss function here is mean squared error of the predicted Q-value and the target Q-value – Q*. This is basically a regression problem.
4. However, we do not know the target or actual value here as we are dealing with a reinforcement learning problem. Going back to the Q-value update equation derived from the Bellman equation, we have :

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)]$$

Que 5.18. Write pseudocode for deep Q-learning.

Answer

Start with $Q_0(s, a)$ for all s, a.

Get initial state s

For $k = 1, 2, \ldots$ till convergence

Sample action a, get next state s'

If s' is terminal :

$$\text{target} = R(s, a, s')$$

Sample new initial state s'

else $\text{target} = R(s, a, s') + \gamma \max Q_k(s', a')$

$$\theta_{k+1} \leftarrow \theta_k - \alpha \nabla_\theta E_{s' \sim P(s'|s,a)}[(Q_\theta(s, a) - \text{target}(s'))^2]|_{\theta = \theta_k}$$

$s \leftarrow s'$

PART-5

Genetic Algorithm, Introduction, Components, GA Cycle of Reproduction, Crossover, Mutation, Genetic Programming, Models of Evolution and Learning, Application.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.19. Write short note on Genetic algorithm.

Answer

1. Genetic algorithms are computerized search and optimization algorithm based on mechanics of natural genetics and natural selection.
2. These algorithms mimic the principle of natural genetics and natural selection to construct search and optimization procedure.
3. Genetic algorithms convert the design space into genetic space. Design space is a set of feasible solutions.
4. Genetic algorithms work with a coding of variables.
5. The advantage of working with a coding of variables space is that coding discretizes the search space even though the function may be continuous.
6. Search space is the space for all possible feasible solutions of particular problem.
7. Following are the benefits of Genetic algorithm :
   a. They are robust.
   b. They provide optimization over large space state.
   c. They do not break on slight change in input or presence of noise.
8. Following are the application of Genetic algorithm :
   a. Recurrent neural network

   b. Mutation testing
   c. Code breaking
   d. Filtering and signal processing
   e. Learning fuzzy rule base

Que 5.20. Write procedure of Genetic algorithm with advantages and disadvantages.

Answer

Procedure of Genetic algorithm :

1. Generate a set of individuals as the initial population.
2. Use genetic operators such as selection or cross over.
3. Apply mutation or digital reverse if necessary.
4. Evaluate the fitness function of the new population.
5. Use the fitness function for determining the best individuals and replace predefined members from the original population.
6. Iterate steps 2–5 and terminate when some predefined population threshold is met.

Advantages of genetic algorithm :

1. Genetic algorithms can be executed in parallel. Hence, genetic algorithms are faster.
2. It is useful for solving optimization problems.

Disadvantages of Genetic algorithm :

1. Identification of the fitness function is difficult as it depends on the problem.
2. The selection of suitable genetic operators is difficult.

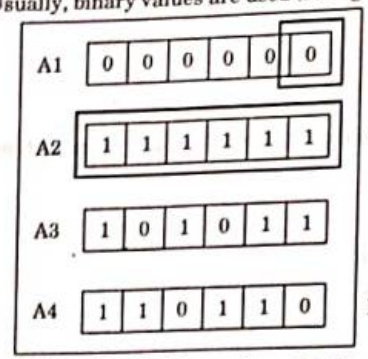
Que 5.21. Explain different phases of genetic algorithm.

Answer

Different phases of genetic algorithm are :

1. **Initial population :**
   a. The process begins with a set of individuals which is called a population.
   b. Each individual is a solution to the problem we want to solve.
   c. An individual is characterized by a set of parameters (variables) known as genes.
   d. Genes are joined into a string to form a chromosome (solution).
   e. In a genetic algorithm, the set of genes of an individual is represented using a string.

f. Usually, binary values are used (string of 1s and 0s).



2. FA (Factor Analysis) fitness function :

a. The fitness function determines how fit an individual is (the ability of all individual to compete with other individual).

b. It gives a fitness score to each individual.

c. The probability that an individual will be selected for reproduction is based on its fitness score.

3. Selection :

a. The idea of selection phase is to select the fittest individuals and let them pass their genes to the next generation.

b. Two pairs of individuals (parents) are selected based on their fitness scores.

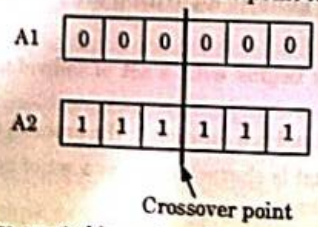
c. Individuals with high fitness have more chance to be selected for reproduction.

4. Crossover :

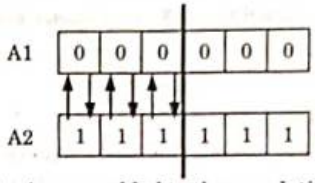
a. Crossover is the most significant phase in a genetic algorithm.

b. For each pair of parents to be mated, a crossover point is chosen at random from within the genes.

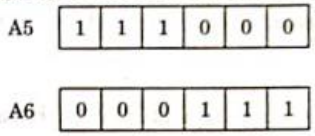
c. For example, consider the crossover point to be 3 as shown :



d. Offspring are created by exchanging the genes of parents among themselves until the crossover point is reached.



e. The new offspring are added to the population.

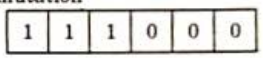


5. Mutation :

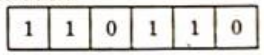
a. When new offspring formed, some of their genes can be subjected to a mutation with a low random probability.

b. This implies that some of the bits in the bit string can be flipped.

Before mutation



After mutation



c. Mutation occurs to maintain diversity within the population and prevent premature convergence.

6. Termination :

a. The algorithm terminates if the population has converged (does not produce offspring which are significantly different from the previous generation).

b. Then it is said that the genetic algorithm has provided a set of solutions to our problem.

Que 5.22. Draw a flowchart of GA and explain the working principle.

Answer

Genetic algorithm : Refer Q. 1.24, Page 1–23L, Unit-1.

Working principle :

1. To illustrate the working principle of GA, we consider unconstrained optimization problem.
2. Let us consider the following maximization problem :

maximize $f(X)$

$X_i^{(L)} \le X_i \le X_i^{(U)}$ for $i = 1, 2 \ldots N$,

3. If we want to minimize $f(X)$, for $f(X) > 0$, then we can write the objective function as :

$$\text{maximize } \frac{1}{1+f(X)}$$

4. If $f(X) < 0$ instead of minimizing $f(X)$, maximize $[-f(X)]$. Hence, both maximization and minimization problems can be handled by GA.

Que 5.23. Write short notes on procedures of GA.

Answer

1. **Start :** Generate random population of n chromosomes.
2. **Fitness :** Evaluate the fitness $f(x)$ of each chromosome x in the population.
3. **New population :** Create a new population by repeating following steps until the new population is complete.
   a. **Selection :** Select two parent chromosomes from a population according to their fitness.
   b. **Crossover :** With a crossover probability crossover the parents to form new offspring (children). If no crossover was performed, offspring is the exact copy of parents.
   c. **Mutation :** With a mutation probability mutate new offspring at each locus (position in chromosome).
   d. **Accepting :** Place new offspring in the new population.
4. **Replace :** Use new generated population for a further run of the algorithm.
5. **Test :** If the end condition is satisfied, stop, and return the best solution in current population.
6. Go to step 2

Que 5.24. What are the benefits of using GA ? What are its limitations ?

Answer

Benefits of using GA :

1. It is easy to understand.
2. It is modular and separate from application.
3. It supports multi-objective optimization.
4. It is good for noisy environment.

Limitations of genetic algorithm are :

1. The problem of identifying fitness function.

2. Definition of representation for the problem.
3. Premature convergence occurs.
4. The problem of choosing the various parameters like the size of the population, mutation rate, crossover rate, the selection method and its strength.
5. Cannot use gradients.
6. Cannot easily incorporate problem specific information.
7. Not good at identifying local optima.
8. No effective terminator.
9. Not effective for smooth unimodal functions.
10. Needs to be coupled with a local search technique.

Que 5.25. Write short notes of genetic representations.

Answer

1. Genetic representation is a way of representing solutions/individuals in evolutionary computation methods.
2. Genetic representation can encode appearance, behavior, physical qualities of individuals.
3. All the individuals of a population are represented by using binary encoding, permutational encoding, encoding by tree.
4. Genetic algorithms use linear binary representations. The most standard method of representation is an array of bits.
5. These genetic representations are convenient because parts of individual are easily aligned due to their fixed size which makes simple crossover operation.

Que 5.26. Give the detail of genetic representation (Encoding).

OR

Explain different types of encoding in genetic algorithm.

Answer

Genetic representations :

1. **Encoding :**
   a. Encoding is a process of representing individual genes.
   b. The process can be performed using bits, numbers, trees, arrays, lists or any other objects.
   c. The encoding depends mainly on solving the problem.
2. **Binary encoding :**
   a. Binary encoding is the most commonly used method of genetic representation because GA uses this type of encoding.

b. In binary encoding, every chromosome is a string of bits, 0 or 1.

Chromosome A	10110010110010101100101
Chromosome B	11111110000011000001111

c. Binary encoding gives many possible chromosomes.

3. **Octal or Hexadecimal encoding :**

a. The encoding is done using octal or hexadecimal numbers.

Chromosome	Octal	Hexadecimal
Chromosome A	54545345	B2CAE5
Chromosome B	77406037	FE0C1F

4. **Permutation encoding (real number encoding) :**

a. Permutation encoding can be used in ordering problems, such as Travelling Salesman Problem (TSP).

b. In permutation encoding, every chromosome is a string of numbers, which represents number in a sequence.

Chromosome A	1 5 3 2 6 4 7 9 8
Chromosome B	8 5 6 7 2 3 1 4 9

5. **Value encoding :**

a. Direct value encoding can be used in problems, where some complicated values, such as real numbers, are used.

b. In value encoding, every chromosome is a string of some values.

c. Values can be anything connected to problem, real numbers or chars to some complicated objects.

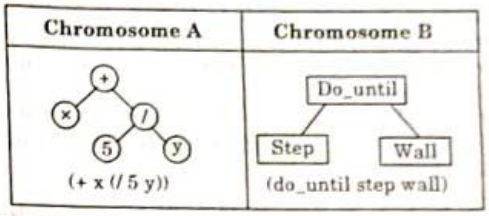
Chromosome A	1.2324 5.3243 0.4556 2.3293 2.4545
Chromosome B	ABDJEIFJDHDIERJFDLDFLFEGT
Chromosome C	(back), (back), (right), (forward), (left)

6. **Tree encoding :**

a. Tree encoding is used for evolving programs or expressions, for genetic programming.

b. In tree encoding, every chromosome is a tree of some objects, such as functions or commands in programming language.

c. Programming language LISP is often used to this, because programs in it are represented in this form and can be easily parsed as a tree, so the cross-over and mutation can be done relatively easily.



Que 5.27. Explain different methods of selection in genetic algorithm in order to select a population for next generation.

Answer

The various methods of selecting chromosomes for parents to cross over are :

a. **Roulette-wheel selection :**

i. Roulette-wheel selection is the proportionate reproductive method where a string is selected from the mating pool with a probability proportional to the fitness.

ii. Thus, ith string in the population is selected with a probability proportional to F_i where F_i is the fitness value for that string.

iii. Since the population size is usually kept fixed in Genetic Algorithm, the sum of the probabilities of each string being selected for the mating pool must be one.

iv. The probability of the ith selected string is

$$p_i = \frac{F_i}{\sum_{j=1}^{n} F_j}$$

where 'n' is the population size.

v. The average fitness is

$$\bar{F} = \sum_{j=1}^{n} F_j / n \qquad \ldots(5.27.1)$$

b. **Boltzmann selection :**

i. Boltzmann selection uses the concept of simulated annealing.

ii. Simulated annealing is a method of functional minimization or maximization.

iii. This method simulates the process of slow cooling of molten metal to achieve the minimum function value in a minimization problem.

iv. The cooling phenomenon is simulated by controlling a temperature so that a system in thermal equilibrium at a temperature T has its energy distributed probabilistically according to

$$P(E) = \exp\left(-\frac{E}{kT}\right) \qquad \ldots(5.27.2)$$

where 'k' is Boltzmann constant.

v. This expression suggests that a system at a high temperature has almost uniform probability of being at any energy state, but at a low temperature it has a small probability of being at a high energy state.

vi. Therefore, by controlling the temperature T and assuming search process follows Boltzmann probability distribution, the convergence of the algorithm is controlled.

c. Tournament selection :

i. GA uses a strategy to select the individuals from population and insert them into a mating pool.

ii. A selection strategy in GA is a process that favours the selection of better individuals in the population for the mating pool.

iii. There are two important issues in the evolution process of genetic search.

1. **Population diversity :** Population diversity means that the genes from the already discovered good individuals are exploited.
2. **Selective pressure :** Selective pressure is the degree to which the better individuals are favoured.

iv. The higher the selective pressure the better individuals are favoured.

d. Rank selection :

i. Rank selection first ranks the population and takes every chromosome, receives fitness from the ranking.

ii. The worst will have fitness 1, the next 2, ..., and the best will have fitness N (N is the number of chromosomes in the population).

iii. The method can lead to slow convergence because the best chromosome does not differ so much from the other.

e. Steady-state selection :

i. The main idea of the selection is that bigger part of chromosome should survive to next generation.

ii. GA works in the following way :

1. In every generation a few chromosomes are selected for creating new off springs.
2. Then, some chromosomes are removed and new offspring is placed in that place.
3. The rest of population survives a new generation.

Que 5.28. Differentiate between Roulette-wheel based on fitness and Roulette-wheel based on rank with suitable example.

Answer

Difference :

S. No.	Roulette-wheel based on fitness	Roulette-wheel based on rank
1.	Population is selected with a probability that is directly proportional to their fitness values.	Probability of a population being selected is based on its fitness rank.
2.	It computes selection probabilities according to their fitness values but do not sort the individual in the population.	It first sort individuals in the population according to their fitness and then computes selection probabilities according to their ranks rather than fitness values.
3.	It gives a chance to all the individuals in the population to be selected.	It selects the individuals with highest rank in the population.
4.	Diversity in the population is preserved.	Diversity in the population is not preserved.

Example :

1. Imagine a Roulette-wheel where all chromosomes in the population are placed, each chromosome has its place accordingly to its fitness function :

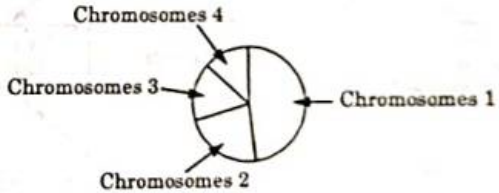


Fig. 5.28.1. Roulette-wheel selection.

2. When the wheel is spun, the wheel will finally stop and pointer attached to it will points to the one of chromosomes with bigger fitness value.
3. **The different between roulette-wheel selection based on fitness and rank is shown in Fig. 5.28.1 and Fig. 5.28.3.**

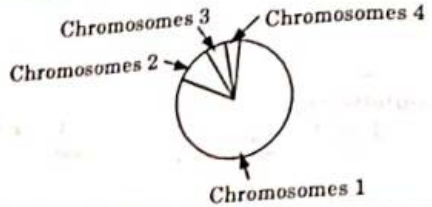


Fig. 5.28.2. Situation before ranking (graph of fitnesses).

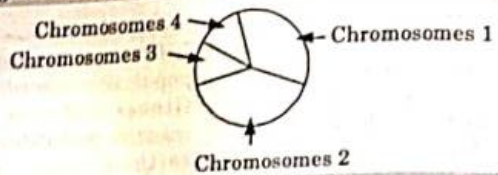


Fig. 5.28.3. Situation after ranking (graph of order numbers).

Que 5.29. Draw genetics cycle for genetic algorithm.

Answer

Generational cycle of GA :

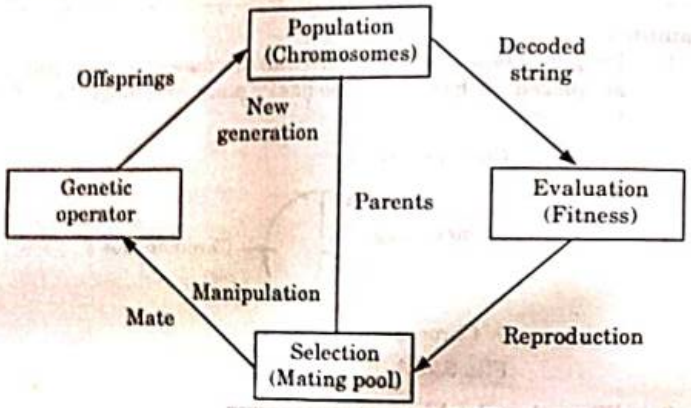


Fig. 5.29.1. The GA cycle.

Components of generational cycle in GA :

1. **Population (Chromosomes) :** A population is collection of individuals. A population consists of a number of individuals being tested, the phenotype parameters defining the individuals and some information about search space.
2. **Evaluation (Fitness) :** A fitness function is a particular type of objective function that quantifies the optimality of a solution (*i.e.*, a chromosome)

in a genetic algorithm so that particular chromosome may be ranked against all the other chromosomes.

3. **Selection :** During each successive generation, a proportion of the existing population is selected to breed a new generation. Individual solutions are selected through a fitness-based process.
4. **Generic operator :** A genetic operator is an operator used in genetic algorithm to guide the algorithm towards a solution to a given problem.

Que 5.30. Why mutation is done in genetic algorithm ? Explain types of mutation.

Answer

Mutation is done in genetic algorithm because :

1. It maintains genetic diversity from one generation of a population of genetic algorithm chromosomes to the next.
2. GA can give better solution of the problem by using mutation.

Types of mutation :

1. **Bit string mutation :** The mutation of bit strings occurs through bit flips at random positions.

   Example : 1 0 1 0 0 1 0

   ↓

   1 0 1 0 1 1 0

   The probability of a mutation of a bit is $1/l$, where l is the length of the binary vector. Thus, a mutation rate of 1 per mutation and individual selected for mutation is reached.
2. **Flip bit :** This mutation operator takes the chosen genome and inverts the bits (*i.e.*, if the genome bit is 1, it is changed to 0 and vice versa).
3. **Boundary :** This mutation operator replaces the genome with either lower or upper bound randomly. This can be used for integer and float genes.
4. **Non-uniform :** The probability that amount of mutation will go to 0 with the next generation is increased by using non-uniform mutation operator. It keeps the population from stagnating in the early stages of the evolution.
5. **Uniform :** This operator replaces the value of the chosen gene with a uniform random value selected between the user-specified upper and lower bounds for that gene.
6. **Gaussian :** This operator adds a unit Gaussian distributed random value to the chosen gene. If it falls outside of the user-specified lower or upper bounds for that gene, the new gene value is clipped.

Que 5.31. **What is the main function of crossover operation in genetic algorithm ?**

Answer

1. Crossover is the basic operator of genetic algorithm. Performance of genetic algorithm depends on crossover operator.
2. Type of crossover operator used for a problem depends on the type of encoding used.
3. The basic principle of crossover process is to exchange genetic material of two parents beyond the crossover points.

Function of crossover operation/operator in genetic algorithm :

1. The main function of crossover operator is to introduce diversity in the population.
2. Specific crossover made for a specific problem can improve performance of the genetic algorithm.
3. Crossover combines parental solutions to form offspring with a hope to produce better solutions.
4. Crossover operators are critical in ensuring good mixing of building blocks.
5. Crossover is used to maintain balance between exploitation and exploration. The exploitation and exploration techniques are responsible for the performance of genetic algorithms. Exploitation means to use the already existing information to find out the better solution and exploration is to investigate new and unknown solution in exploration space.

Que 5.32. **Discuss the different applications of genetic algorithms.**

Answer

Application of GA :

1. **Optimization :** Genetic Algorithms are most commonly used in optimization problems wherein we have to maximize or minimize a given objective function value under a given set of constraints.
2. **Economics :** GAs are also used to characterize various economic models like the cobweb model, game theory equilibrium resolution, asset pricing, etc.
3. **Neural networks :** GAs are also used to train neural networks, particularly recurrent neural networks.
4. **Parallelization :** GAs also have very good parallel capabilities, and prove to be very effective means in solving certain problems, and also provide a good area for research.

5. **Image processing :** GAs are used for various digital image processing (DIP) tasks as well like dense pixel matching.
6. **Machine learning :** Genetics based machine learning (GBML) is a nice area in machine learning.
7. **Robot trajectory generation :** GAs have been used to plan the path which a robot arm takes by moving from one point to another.

Que 5.33. **Explain optimization of travelling salesman problem using genetic algorithm and give a suitable example too.**

Answer

1. The TSP consist a number of cities, where each pair of cities has a corresponding distance.

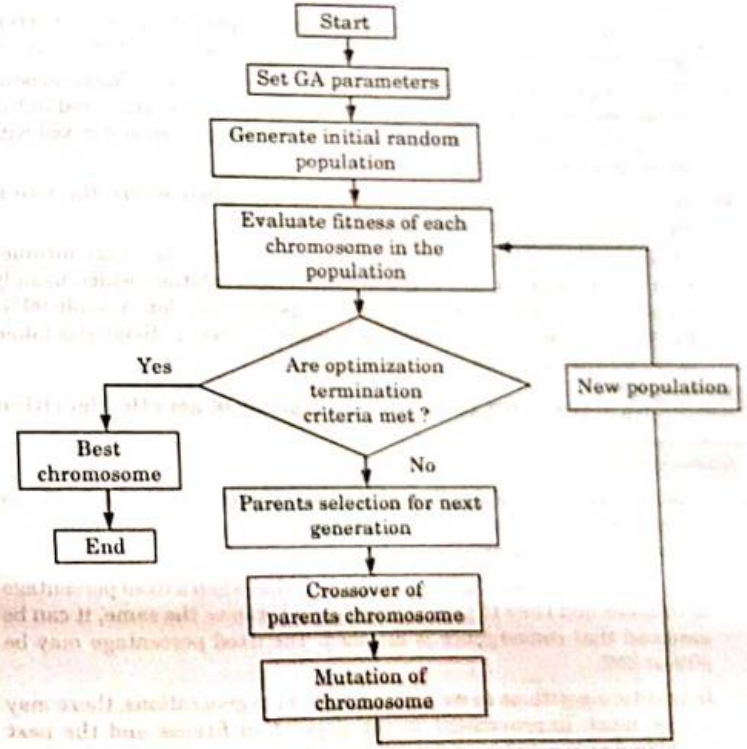


Fig. 5.33.1. Genetic algorithm procedure for TSP.

2. The aim is to visit all the cities such that the total distance travelled will be minimized.
3. A solution, and therefore a chromosome which represents that solution to the TSP, can be given as an order, that is, a path, of the cities.
4. The procedure for solving TSP can be viewed as a process flow given in Fig. 5.33.1.
5. The GA process starts by supplying important information such as location of the city, maximum number of generations, population size, probability of crossover and probability of mutation.
6. An initial random population of chromosomes is generated and the fitness of each chromosome is evaluated.
7. The population is then transformed into a new population (the next generation) using three genetic operators : selection, crossover and mutation.
8. The selection operator is used to choose two parents from the current generation in order to create a new child by crossover and/or mutation.
9. The new generation contains a higher proportion of the characteristics possessed by the good members of the previous generation and in this way good characteristics are spread over the population and mixed with other good characteristics.
10. After each generation, a new set of chromosomes where the size is equal to the initial population size is evolved.
11. This transformation process from one generation to the next continues until the population converges to the optimal solution, which usually occurs when a certain percentage of the population (for example 90 %) has the same optimal chromosome in which the best individual is taken as the optimal solution.

Que 5.34. Write short notes on convergence of genetic algorithm

Answer

1. A genetic algorithm is usually said to converge when there is no significant improvement in the values of fitness of the population from one generation to the next.
2. One criterion for convergence may be such that when a fixed percentage of columns and rows in population matrix becomes the same, it can be assumed that convergence is attained. The fixed percentage may be 80% or 85%.
3. In genetic algorithms as we proceed with more generations, there may not be much improvement in the population fitness and the best individual may not change for subsequent populations.
4. As the generation progresses, the population gets filled with more fit individuals with only slight deviation from the fitness of best individuals

so far found, and the average fitness comes very close to the fitness of the best individuals.

5. The convergence criteria can be explained from schema point of view.
6. A schema is a similarity template describing a subset of strings with similarities at certain positions. A schema represents a subset of all possible strings that have the same bits at certain string positions.
7. Since schema represents a robust of strings, we can associate a fitness value with a schema, *i.e.*, the average fitness of the schema.
8. One can visualize GA's search for the optimal strings as a simultaneous competition among schema increases the number of their instances in the population.

1 UNIT

Introduction (2 Marks Questions)

1.1. Define machine learning.

Ans: Machine learning is an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.

1.2. What are the different types of machine learning algorithm ?

Ans: Different types of machine learning algorithm are :

1. Supervised machine learning algorithm
2. Unsupervised machine learning algorithm
3. Semi-supervised machine learning algorithm
4. Reinforcement machine learning algorithm

1.3. What are the applications of machine learning ?

Ans: Applications of machine learning are :

1. Image recognition
2. Speech recognition
3. Medical diagnosis
4. Statistical arbitrage
5. Learning association

1.4. What are the advantages of machine learning ?

Ans: Advantages of machine learning :

1. Easily identifies trends and patterns.
2. No human intervention is needed.
3. Continuous improvement.
4. Handling multi-dimensional and multi-variety data.

1.5. What are the disadvantages of machine learning ?

Ans: Disadvantages of machine learning :

1. Data acquisition
2. Time and resources
3. Interpretation of results
4. High error-susceptibility

1.6. What is the role of machine learning in human life ?

Ans: Role of machine learning in human life :

1. Learning
2. Reasoning
3. Problem solving
4. Language understanding

1.7. What are the components of machine learning system ?

Ans: Components of machine learning system are :

1. Sensing
2. Segmentation
3. Feature extraction
4. Classification
5. Post processing

1.8. What are the classes of problem in machine learning ?

Ans: Classes of problem in machine learning are :

1. Classification
2. Regression
3. Clustering
4. Rule extraction

1.9. What are the issues related with machine learning ?

Ans: Issues related with machine learning are :

1. Data quality
2. Transparency
3. Traceability
4. Reproduction of results

1.10. Define supervised learning.

Ans: Supervised learning is also known as associative learning, in which the network is trained by providing it with input and matching output patterns.

1.11. Define unsupervised learning ?

Ans: Unsupervised learning is also known as self-organization, in which an output unit is trained to respond to clusters of pattern within the input.

1.12. Define well defined learning problem.

Ans: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

1.13. What are the features of learning problems ?

Ans: Features of learning problems are :

1. The class of tasks (T).
2. The measure of performance to be improved (P).
3. The source of experience (E).

1.14. Define decision tree learning.

Ans: Decision tree learning is the predictive modeling approaches used in statistics, data mining and machine learning. It uses a decision tree to go from observations about an item to conclusions about the item's target values.

1.15. What is decision tree ?

Ans: A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs and utility.

1.16. What are the types of decision tree ?

Ans: There are two types of decision tree :

1. Classification tree
2. Regression tree

1.17. Define classification tree and regression tree.

Ans: Classification tree : A classification tree is an algorithm where the target variable is fixed. This algorithm is used to identify the class within which a target variable would fall.

Regression tree : A regression tree is an algorithm where the target variable is not fixed and this algorithm is used to predict its value.

1.18. Name different decision tree algorithm.

Ans: Different decision tree algorithms are :

1. ID3
2. C4.5
3. CART

1.19. What are the issues related with the decision tree ?

Ans: Issues related with decision tree are :

1. Missing data
2. Multi-valued attribute
3. Continuous and integer valued input attributes
4. Continuous-valued output attributes

1.20. What are the attribute selection measures used in decision tree ?

Ans: Attribute selection measures used in decision tree are :

1. Entropy
2. Information gain
3. Gain ratio

2 UNIT

Regression (2 Marks Questions)

2.1. Define the term regression.

Ans. Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable and a series of other variables (known as independent variables).

2.2. What are the types of regression ?

Ans. **Following are the types of regression :**

1. Linear regression
2. Logistic regression

2.3. Define logistic regression.

Ans. Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

2.4. What are the types of logistic regression ?

Ans. **Following are the types of logistic regression :**

1. Binary or Binomial logistic regression
2. Multinomial logistic regression
3. Ordinal logistic regression

2.5. Define Bayesian decision theory.

Ans. Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification. This approach is based on quantifying the tradeoffs between various classification decisions using probability and costs that accompany such decisions.

2.6. Define Bayesian belief network.

Ans. Bayesian belief networks specify joint conditional probability distributions. They are also known as Belief Networks, Bayesian Networks, or Probabilistic Networks.

2.7. Define EM algorithm.

Ans. The Expectation-Maximization (EM) algorithm is an iterative way to find maximum-likelihood estimates for model parameters when the data is incomplete or has some missing data points or has some hidden variables.

2.8. What are the usages of EM algorithm ?

Ans. **Usages of EM algorithm are :**

1. It can be used to fill the missing data in a sample.
2. It can be used as the basis of unsupervised learning of clusters.
3. It can be used for the purpose of estimating the parameters of Hidden Markov Model (HMM).
4. It can be used for discovering the values of latent variables.

2.9. What are the advantages of EM algorithm ?

Ans. **Advantages of EM algorithm are :**

1. It is always guaranteed that likelihood will increase with each iteration.
2. The E-step and M-step are easy implementation.
3. Solutions to the M-steps exist in the closed form.

2.10. What are the disadvantages of EM algorithm ?

Ans. **Disadvantages of EM algorithm are :**

1. It has slow convergence.
2. It makes convergence to the local optima only.
3. It requires both the probabilities, forward and backward (numerical optimization requires only forward probability).

2.11. Define support vector machine.

Ans. A support vector machine is a supervised machine learning algorithm that looks at data and sorts, analyzes data for classification and regression analysis.

2.12. What are the types of support vector machine ?

Ans. Types of support vector machine are :

1. Linear support vector machine
2. Non-linear support vector machine

2.13. What are the applications of SVM ?

Ans. Applications of SVM :

1. Text and hypertext classification
2. Image classification
3. Recognizing handwritten characters
4. Biological sciences, including protein classification

☺☺☺

Decision Tree Learning (2 Marks Questions)

3.1. What is instance-based learning ?

Ans. Instance-Based Learning (IBL) is an extension of nearest neighbour or KNN classification algorithms that do not maintain a set of abstraction of model created from the instances.

3.2. What are the advantages of KNN algorithm ?

Ans. Advantages of KNN algorithm are :

1. No training period.
2. Since the KNN algorithm requires no training before making predictions, new data can be added seamlessly which will not impact the accuracy of the algorithm.
3. KNN is easy to implement.

3.3. What are the disadvantages of KNN algorithm ?

Ans. Disadvantages of KNN algorithm are :

1. It is does not work well with large dataset.
2. It is does not work well with high dimensions.
3. It need feature scaling.
4. It is sensitive to noisy data, missing values and outliers.

3.4. Define locally weighted regression.

Ans. Locally Weighted Regression (LWR) is a memory-based method that performs a regression around a point of interest using training data that are local to that point.

3.5. Define radial basis function.

Ans. A Radial Basis Function (RBF) is a function that assigns a real value to each input from its domain (it is a real-value function), and the value produced by the RBF is always an absolute value *i.e.*, it is a measure of distance and cannot be negative.

3.6. Define case-based learning algorithms.

Ans. Case-based learning algorithms contain as input a sequence of training cases and as output a concept description, which can be

used to generate predictions of goal feature values for subsequently presented cases.

3.7. What are the disadvantages of CBL (Case-Based Learning) ?

Ans. Disadvantage of case-based learning algorithm :

1. They are computationally expensive because they save and compute similarities to all training cases.
2. They are intolerant of noise and irrelevant features.
3. They are sensitive to the choice of the algorithm's similarity function.
4. There is no simple way they can process symbolic valued feature values.

3.8. What are the functions of CBL ?

Ans. Functions of case-based learning algorithm are :

1. Pre-processor
2. Similarity
3. Prediction
4. Memory updating

3.9. What are the processing stages of CBL ?

Ans. Case-based learning algorithm processing stages are :

1. Case retrieval
2. Case adaptation
3. Solution evaluation
4. Case-base updating

3.10. What are the benefits of CBL as lazy problem solving method ?

Ans. The benefits of CBL as a lazy Problem solving method are :

1. Ease of knowledge elicitation.
2. Absence of problem-solving bias.
3. Incremental learning.
4. Suitability for complex and not-fully formalized solution spaces.
5. Suitability for sequential problem solving.
6. Ease of explanation.
7. Ease of maintenance.

3.11. What are the applications of CBL ?

Ans. Applications of CBL :

1. Interpretation
2. Classification
3. Design
4. Planning
5. Advising

3.12. What are the advantages of instance-based learning ?

Ans. Advantages of instance-based learning :

1. Learning is trivial
2. Works efficiently
3. Noise resistant
4. Rich representation, arbitrary decision surfaces
5. Easy to understand

3.13. What are the disadvantages of instance-based learning ?

Ans. Disadvantages of instance-based learning :

1. Need lots of data.
2. Computational cost is high.
3. Restricted to $x \in R^n$.
4. Implicit weights of attributes (need normalization).
5. Need large space for storage *i.e.*, require large memory.
6. Expensive application time.

4 UNIT

Artificial Neural Network (2 Marks Questions)

4.1. What are neurons ?

Ans: A neuron is a small cell that receives electro-chemical signals from its various sources and in return responds by transmitting electrical impulses to other neurons.

4.2. What is artificial neural network ?

Ans: Artificial neural network are computational algorithm that intended to simulate the behaviour of biological systems composed of neurons.

4.3. Give the difference between supervised and unsupervised learning in artificial neural network.

Ans:

S. No.	Supervised learning	Unsupervised learning
1.	It uses known and labeled data as input.	It uses unknown data as input.
2.	It uses offline analysis.	It uses real time analysis of data.
3.	Number of classes is known.	Number of classes is not known.
4.	Accurate and reliable results.	Moderate accurate and reliable results.

4.4. Define activation function.

Ans: An activation function is the basic element in neural model. It is used for limiting the amplitude of the output of a neuron. It is also called squashing function.

4.5. Give types of activation function.

Ans: **Types of activation function :**

1. Signum function

4 UNIT

Artificial Neural Network (2 Marks Questions)

4.1. What are neurons ?

Ans. A neuron is a small cell that receives electro-chemical signals from its various sources and in return responds by transmitting electrical impulses to other neurons.

4.2. What is artificial neural network ?

Ans. Artificial neural network are computational algorithm that intended to simulate the behaviour of biological systems composed of neurons.

4.3. Give the difference between supervised and unsupervised learning in artificial neural network.

Ans.

S. No.	Supervised learning	Unsupervised learning
1.	It uses known and labeled data as input.	It uses unknown data as input.
2.	It uses offline analysis.	It uses real time analysis of data.
3.	Number of classes is known.	Number of classes is not known.
4.	Accurate and reliable results.	Moderate accurate and reliable results.

4.4. Define activation function.

Ans. An activation function is the basic element in neural model. It is used for limiting the amplitude of the output of a neuron. It is also called squashing function.

4.5. Give types of activation function.

Ans. **Types of activation function :**

1. Signum function

2. Sigmoidal function
3. Identity function
4. Binary step function
5. Bipolar step function

4.6. Give advantages of neural network.

Ans. **Advantages of neural network :**

1. A neural network can perform tasks that a linear program cannot.
2. It can be implemented in any application.
3. A neural network learns and does not need to be reprogrammed.

4.7. What are disadvantages of neural network (NN) ?

Ans. **Disadvantages of neural network :**

1. The neural network needs training to operate.
2. It requires high processing time for large NN.

4.8. List the various types of soft computing techniques and mention some application areas for neural network.

Ans. **Types of soft computing techniques :**

1. Fuzzy logic control
2. Neural network
3. Genetic algorithms
4. Support vector machine

Application areas for neural network :

1. Speech recognition
2. Character recognition
3. Signature verification application
4. Human face recognition

4.9. Draw a biological NN and explain the parts.

Ans.

1. Biological neural networks are made up of real biological neurons that are connected in the peripheral nervous system.
2. In general a biological neural network is composed of a group of chemically connected or functionally associated neurons.

Fig. 4.9.1.

A biological neural network has three major parts :

1. **Soma or cell body :** It contains the cell's nucleus and other vital components called organelles which perform specialized tasks.
2. **A set of dendrites :** It forms a tree like structure that spread out from the cell. The neuron receives its input electrical signal along these set of dendrites.
3. **Axon :** It is tabular extension from the cell (Soma) that carries an electrical signal away from Soma to another neuron for processing.

4.10. What is single layer feed forward network ?

Ans. Single layer feed forward network is the simplest form of a layered network where an input layer of source nodes that projects onto an output layer of neurons, but not vice versa.

4.11. Write different applications of neural networks (NN).

Ans. **Applications of NN are :**

1. Image recognition
2. Data mining
3. Machine translation
4. Spell checking
5. Stock and sport bet prediction
6. Statistical modeling

4.12. Draw an artificial neural network.

Ans. An Artificial Neuron Network (ANN) is a computational model based on the structure and functions of biological neural networks.

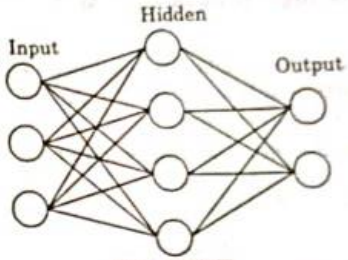


Fig. 4.12.1.

4.13. What do you mean by neural network architecture ?

Ans. Neural network architecture refers to the arrangement of neurons into layers and the connection patterns between layers, activation functions, and learning methods. The neural network model and the architecture of a neural network determine how a network transforms its input into an output.

4.14. What are the types of neuron connection ?

Ans. **Following are the types of neuron connection :**

1. Single-layer feed forward network
2. Multilayer feed forward network
3. Single node with its own feedback
4. Single-layer recurrent network
5. Multilayer recurrent network

4.15. What is gradient descent ?

Ans. Gradient descent is an optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient.

4.16. What are the types of gradient descent ?

Ans. **Types of gradient descent are :**

1. Batch gradient descent
2. Stochastic gradient descent
3. Mini-batch gradient descent

4.17. What is self organizing map (SOM) ?

Ans.

1. Self Organizing Map (SOM) provides a data visualization technique which helps to understand high dimensional data by reducing the dimensions of data to a map.
2. SOM also represents clustering concept by grouping similar data together.

Reinforcement Learning (2 Marks Questions)

5.1. Define genetic algorithm.

Ans. Genetic algorithms are computerized search and optimization algorithm based on mechanics of natural genetics and natural selection. These algorithms mimic the principle of natural genetics and natural selection to construct search and optimization procedure.

5.2. Give the benefits of genetic algorithm.

Ans. **Benefits of genetic algorithm are :**

1. They are Robust.
2. They provide optimization over large space state.
3. They do not break on slight change in input or presence of noise.

5.3. What are the applications of genetic algorithm ?

Ans. **Following are the applications of genetic algorithms :**

1. Recurrent neural network
2. Mutation testing
3. Code breaking
4. Filtering and signal processing
5. Learning fuzzy rule base

5.4. What are the disadvantages of genetic algorithm ?

Ans. **Disadvantages of genetic algorithm :**

1. Identification of the fitness function is difficult as it depends on the problem.
2. The selection of suitable genetic operators is difficult.

5.5. Define genetic programming.

Ans. Genetic Programming (GP) is a type of Evolutionary Algorithm (EA), a subset of machine learning. EAs are used to discover solution to problems that human do not know how to solve.

5.6. What are the advantages of genetic programming ?

Ans. Advantages of genetic programming are :

1. In GP, the number of possible programs that can be constructed by the algorithm is immense.
2. Although GP uses machine code which helps in providing result very fast but if any of the high level language is used which needs to be compile, and can generate errors and can make our program slow.
3. There is a high probability that even a very small variation has a disastrous effect on fitness of the solution generated.

5.7. What are the disadvantages of genetic programming ?

Ans. Disadvantages of genetic programming are :

1. It does impose any fixed length of solution, so the maximum length can be extended up to hardware limits.
2. In genetic programming it is not necessary for an individual to have maximum knowledge of the problem and their solutions.

5.8. What are different types of genetic programming ?

Ans. Different types of genetic programming are :

1. Tree-based genetic programming
2. Stack-based genetic programming
3. Linear genetic programming
4. Grammatical evolution
5. Cartesian Genetic Programming (CGP)
6. Genetic Improvement Programming (GIP)

5.9. What are the functions of learning in evolution ?

Ans. Function of learning in evolution :

1. It allows individuals to adapt changes in the environment that occur in the life span of an individual or across few generations.
2. It allows evolution to use information extracted from the environment thereby channeling evolutionary search.
3. It can help and guide evolution.

5.10. What are the disadvantages of learning in evolution ?

Ans. Disadvantages of learning in evolution are :

1. A delay in the ability to acquire fitness.
2. Increased unreliability.

5.11. Define learnable evolution model.

Ans. Learnable Evolution Model (LEM) is a non-Darwinian methodology for evolutionary computation that employs machine learning to guide the generation of new individuals (candidate problem solutions).

5.12. What are different phases of genetic algorithm ?

Ans. Different phases of genetic algorithm are :

1. Initial population
2. FA (Factor Analysis) fitness function
3. Selection
4. Crossover
5. Mutation
6. Termination

5.13. Define sequential covering algorithm.

Ans. Sequential covering algorithm is a general procedure that repeatedly learns a single rule to create a decision list (or set) that covers the entire dataset rule by rule.

5.14. Define Beam search.

Ans. Beam search is a heuristic search algorithm that explores a graph by expanding the most promising node in a limited set.

5.15. What are the properties of heuristic search ?

Ans. Properties of heuristic search are :

1. Admissibility condition
2. Completeness condition
3. Dominance properties
4. Optimality property

5.16. What are different types of reinforcement learning ?

Ans: **Different types of reinforcement learning are :**

1. Positive reinforcement learning
2. Negative reinforcement learning

5.17. What are the elements of reinforcement learning ?

Ans: **Elements of reinforcement learning are :**

1. Policy (π)
2. Reward function (r)
3. Value function (V)
4. Transition model (M)

5.18. Define *Q*-learning.

Ans: Reinforcement learning is the problem faced by an agent that must learn behaviour through trial-and-error interactions with a dynamic environment, *Q*-learning is model-free reinforcement learning, and it is typically easier to implement.

5.19. Define positive and negative reinforcement learning.

Ans: **Positive reinforcement learning :** Positive reinforcement learning is defined as when an event, occurs due to a particular behaviour such as, increases the strength and the frequency of the behaviour.

Negative reinforcement learning : Negative reinforcement is defined as strengthening of a behaviour because a negative condition is stopped or avoided.

☺☺☺

B. Tech.
(SEM. V) ODD SEMESTER THEORY EXAMINATION, 2020-21
MACHINE LEARNING TECHNIQUES

Time : 3 Hours **Max. Marks : 100**

Note : Attempt all Sections. If require any missing data; then choose suitably.

SECTION-A

1. Attempt **all** questions in brief. **(2 × 10 = 20)**

a. Explain the concept of machine learning.

Ans: Refer Q. 1.1, Page SQ–1L, Unit-1, Two Marks Questions.

b. Compare ANN and Bayesian networks.

Ans:

S. No.	ANN	BNN
1.	It stands for Artificial Neural Network.	It stands for Biological Neural Network.
2.	Processing speed is fast as compared to Biological Neural Network.	They are slow in processing information.
3.	Allocation for Storage to a new process is strictly irreplaceable as the old location is saved for the previous process.	Allocation for storage to a new process is easy as it is added just by adjusting the interconnection strengths.

c. What is the difference between linear and logistics regression ?

Ans: Refer Q. 2.5, Page 2–4L, Unit-2.

d. Discuss support vectors in SVM.

Ans: Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help us to build SVM.

e. Discuss overfitting and underfitting situation in decision tree learning.

Ans. **Overfitting :** A statistical model is said to be overfitted, when we train it with a lot of data. When a model gets trained with so much of data, it starts learning from the noise and inaccurate data entries in our data set.

Underfitting : A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data. Underfitting destroys the accuracy of our machine learning model.

f. What is the task of the E-step of the EM-algorithm ?

Ans. Refer Q. 2.14, Page 2–16L, Unit-2.

g. Define the learning classifiers.

Ans. Learning classifier are a paradigm of rule-based machine learning methods that combine a discovery component with a learning component.

h. What is the difference between machine learning and deep learning ?

Ans.

S. No.	Machine learning	Deep learning
1.	Machine learning is a superset of deep learning.	Deep learning is a subset of machine learning.
2.	The data represented in machine learning uses structured data.	The data representation is used in deep learning uses neural networks.
3.	Machine learning is an evolution of AI.	Deep learning is an evolution to machine learning.

i. What objective function do regression trees minimize ?

Ans. The objective function that regression tree minimize are :

1. Output variable variance in validation data, taken one terminal node at a time.
2. Product of the cost complexity factor and the number of terminal nodes.

j. What is the difference between Q-learning and deep Q-learning ?

Ans. **Q-learning :** Refer Q. 5.14, Page 5–12L, Unit-5.
Deep Q-learning : Refer Q. 5.16, Page 5–13L, Unit-5.

SECTION-B

2. Attempt any three of the following : (3 × 10 = 30)

a. Apply KNN for following dataset and predict class of test example (A1 = 3, A2 = 7). Assume K = 3

A1	A2	Class
7	7	True
7	4	True
3	4	False
1	4	True
5	3	False
6	3	True

Ans. **Step 1 :** Calculate the distance between the query instance and all training samples.
Coordinates of query instance is (3, 7)

A1	A2	Square distance to query instance (3, 7)
7	7	$(7-3)^2 + (7-7)^2 = 16$
7	4	$(7-3)^2 + (4-7)^2 = 25$
3	4	$(3-3)^2 + (4-7)^2 = 9$
1	4	$(1-3)^2 + (4-7)^2 = 13$
5	3	$(5-3)^2 + (3-7)^2 = 20$
6	3	$(6-3)^2 + (3-7)^2 = 25$

Step 2 : Sort the distance and determine nearest neighbours based on k th minimum distance.

A1	A2	Rank minimum distance	It is included in 3-nearest neighbour
7	7	16 = 3	Yes
7	4	25 = 5	No
3	4	9 = 1	Yes
1	4	13 = 2	No
5	3	20 = 4	No
6	3	25 = 6	No

Step 3 : Gather the category y of the nearest neighbour.

A1	A2	Rank minimum distance	It is included in 3-nearest neighbour	Y = category of nearess neighbour
7	7	16 = 3	Yse	Ture
7	4	25 = 5	No	—
3	4	9 = 1	Yes	True
1	4	13 = 2	Yes	True
5	3	20 = 4	No	True
6	3	25 = 6	No	True

Step 4 : Using majority of the category of nearest neighbour as prediction value of query instant

$A_1 = 3$ and $A_2 = 7$ is included in Ture category.

b. Describe the Kohonen Self-Organizing Maps and its algorithm.

Ans. **Kohonen self-organizing maps :** Refer Q. 4.21, Page 4–20L, Unit-4.

Algorithm : Refer Q. 4.22, Page 4–21L, Unit-4.

c. Explain the various learning models for reinforcement-learning.

Ans. Refer Q. 5.11, Page 5–11L, Unit-5.

d. Explain the role of genetic algorithm. Discuss the various phases considered in genetic algorithm.

Ans. **Genetic algorithm :** Refer Q. 5.19, Page 5–15L, Unit-5.

Phases of genetic algorithm : Refer Q. 5.21, Page 5–16L, Unit-5.

e. Describe BPN algorithm in ANN along with a suitable example.

Ans. Refer Q. 4.17, Page 4–18L, Unit-4.

SECTION-C

3. Attempt any one part of the following : (1 × 10 = 10)

a. Why SVM is an example of a large margin classifier ? Discuss the different kernels functions used in SVM.

Ans. **SVM is an example of a large margin classifier :**

1. SVM is a type of classifier which classifies positive and negative examples, here black and white data points

2. As shown in the image, the largest margin is found in order to avoid overfitting *i.e.*, the optimal hyperplane is at the maximum distance from the positive and negative examples(Equal distant from the boundary lines).
3. To satisfy this constraint, and also to classify the data points accurately, the margin is maximised, that is why this is called the large margin classifier.

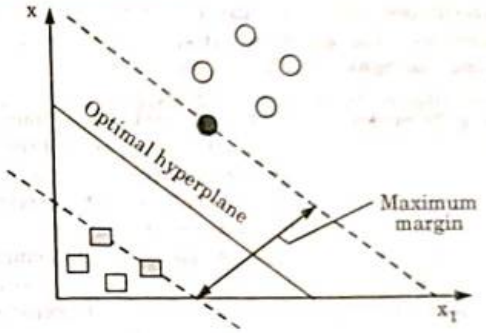


Fig. 1.

Different kernal function used in SVM :

1. **Sigmoid kernel :** This function is equivalent to a two-layer, perceptron model of neural network, which is used as activation function for artificial neurons.
2. **Gaussian kernel :** It is used to perform transformation, when there is no prior knowledge about data.
3. **Polynomial kernel :** It represents the similarity of vectors in training set of data in a feature space over polynomials of the original variables used in kernel.

b. Explain the relevance of CBR. How CADET tool employs CBR ?

Ans.

1. When a new case arrises to classify, a Case-based Reasoner(CBR) will first check if an identical training case exists.
2. If one is found, then the accompanying solution to that case is returned. If no identical case is found, then the CBR will search for training cases having components that are similar to those of the new case.
3. These training cases may be considered as neighbours of the new case. If cases are represented as graphs, this involves searching for subgraphs that are similar to subgraphs within the new case.
4. The CBR tries to combine the solutions of the neighbouring training cases to propose a solution for the new case.
5. If compatibilities arise with the individual solutions, then backtracking to search for other solutions may be necessary.

6. The CBR may employ background knowledge and problem-solving strategies to propose a feasible solution.

CADET tools employes CBR :

1. CADET is a Case-based Design Tool.
2. CADET is a system that aids conceptual design of electro-mechanical devices and is based on the paradigm of Case-based Reasoning.

4. Attempt any **one** part of the following : **(1 × 10 = 10)**

a. Discuss the applications, properties, issues, and disadvantages of SVM.

Ans. Following are the application of SVM :

1. **Face detection :** SVM classify parts of the image as a face and non-face and create a square boundary around the face.
2. **Text and hypertext categorization :**
   a. SVM allow text and hypertext categorization for both inductive and transductive models.
   b. They use training data to classify documents into different categories.
   c. It categorizes on the basis of the score generated and then compares with the threshold value.
3. **Classification of images :**
   a. Use of SVMs provides better search accuracy for image classification.
   b. It provides better accuracy in comparison to the traditional query-based searching techniques.
4. **Bioinformatics :**
   a. It includes protein classification and cancer classification.
   b. We use SVM for identifying the classification of genes, patients on the basis of genes and other biological problems.
5. **Protein fold and remote homology detection :** We use SVM algorithms for protein remote homology detection.
6. **Handwriting recognition :** We use SVM to recognize handwritten characters used widely.

Properties : Refer Q. 2.25, Page 2–23L, Unit-2.

Issues, disadvantages : Refer Q. 2.24, Page 2–22L, Unit-2.

b. Explain the Confusion Matrix with respect to machine learning algorithms.

Ans.

1. A confusion matrix is a technique for summarizing the performance of a classification algorithm.
2. Classification accuracy alone can be misleading if we have an unequal number of observations in each class or if we have more than two classes in our dataset.
3. Calculating a confusion matrix can give us a better idea of what our classification model is getting right and what types of errors it is making.

4. The general idea is to count the number of times instances of class A are classified as class B.
5. Machine learning algorithm are techniques used for estimating the target function (f) to predict the output variable (Y) given input variable (X).

5. Attempt any **one** part of the following : **(1 × 10 = 10)**

a. Illustrate the operation of the ID3 training example. Consider information gain as attribute measure.

PlayTennis : training examples

Data	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Ans. Here,dataset is of binary classes(yes and no), where 9 out of 14 are "yes" and 5 out of 14 are "no". Complete entropy of dataset is :

$$H(S) = - p(yes) * log2(p(yes)) - p(no) * log2(p(no))$$
$$= - (9/14) * log2(9/14) - (5/14) * log2(5/14)$$
$$= - (- 0.41) - (- 0.53)$$
$$= 0.94$$

Step1 : First Attribute – Outlook

Categorical values – sunny, overcast and rain

H(Outlook=sunny) = – (2/5)*log(2/5) – (3/5)*log(3/5) = 0.971

H(Outlook=rain) = – (3/5)*log(3/5) – (2/5)*log(2/5) = 0.971

H(Outlook=overcast) = – (4/4)*log(4/4) – 0 = 0

Average Entropy Information for Outlook –

I(Outlook) = p(sunny) * H(Outlook=sunny) + p(rain) * H(Outlook=rain) + p(overcast) * H(Outlook=overcast)

= (5/14)*0.971 + (5/14)*0.971 + (4/14)*0

= 0.693

Information Gain = H(S) – I(Outlook)

= 0.94 – 0.693

= 0.247

Step2 : Second Attribute – Temperature

Categorical values – hot, mild, cool
H(Temperature=hot) = – (2/4)*log(2/4) – (2/4)*log(2/4) = 1
H(Temperature=cool) = – (3/4)*log(3/4)-(1/4)*log(1/4) = 0.811
H(Temperature=mild) = – (4/6)*log(4/6)–(2/6)*log(2/6) = 0.9179
Average Entropy Information for Temperature
I(Temperature) = p(hot)*H(Temperature=hot) + p(mild)*H(Temperature=mild) + p(cool)*H(Temperature=cool)
= (4/14)*1 + (6/14)*0.9179 + (4/14)*0.811
= 0.9108
Information Gain = H(S) – I(Temperature)
= 0.94 – 0.9108
= 0.0292

Step3 : Third Attribute – Humidity

Categorical values – high, normal
H(Humidity=high) = – (3/7)*log(3/7) – (4/7)*log(4/7) = 0.983
H(Humidity=normal) = – (6/7)*log(6/7) – (1/7)*log(1/7) = 0.591
Average Entropy Information for Humidity –
I(Humidity) = p(high)*H(Humidity=high) + p(normal)*H(Humidity=normal)
= (7/14)*0.983 + (7/14)*0.591
= 0.787
Information Gain = H(S) – I(Humidity)
= 0.94 – 0.787
= 0.153

Step4 : Fourth Attribute – Wind

Categorical values – weak, strong
H(Wind=weak) = – (6/8)*log(6/8) – (2/8)*log(2/8) = 0.811
H(Wind=strong) = – (3/6)*log(3/6) – (3/6)*log(3/6) = 1
Average Entropy Information for Wind –
I(Wind) = p(weak)*H(Wind=weak) + p(strong)*H(Wind=strong)
= (8/14)*0.811 + (6/14)*1
= 0.892
Information Gain = H(S) – I(Wind)
= 0.94 – 0.892
= 0.048

Step5 : Now, finding the best attribute for splitting the data with Outlook=Sunny values{ Dataset rows = [1, 2, 8, 9, 11]}.

Complete entropy of Sunny is –
H(S) = – p(yes) * log2(p(yes)) – p(no) * log2(p(no))
= – (2/5) * log2(2/5) – (3/5) * log2(3/5)
= 0.971

Step6 : First Attribute – Temperature

Categorical values – hot, mild, cool
H(Sunny, Temperature=hot) = – 0 – (2/2)*log(2/2) = 0
H(Sunny, Temperature=cool) = – (1)*log(1) – 0 = 0
H(Sunny, Temperature=mild) = – (1/2)*log(1/2) – (1/2)*log(1/2)
= 1

Average Entropy Information for Temperature –
I(Sunny, Temperature) = p(Sunny, hot)*H(Sunny, Temperature=hot) + p(Sunny, mild)*H(Sunny, Temperature=mild) + p(Sunny, cool)*H(Sunny, Temperature=cool)
= (2/5)*0 + (1/5)*0 + (2/5)*1
= 0.4
Information Gain = H(Sunny) – I(Sunny, Temperature)
= 0.971 – 0.4
= 0.571

Step7 : Second Attribute – Humidity

Categorical values - high, normal
H(Sunny, Humidity=high) = – 0 – (3/3)*log(3/3) = 0
H(Sunny, Humidity=normal) = – (2/2)*log(2/2) – 0 = 0
Average Entropy Information for Humidity –
I(Sunny, Humidity) = p(Sunny, high)*H(Sunny, Humidity=high) + p(Sunny, normal)*H(Sunny, Humidity=normal)
= (3/5)*0 + (2/5)*0
= 0
Information Gain = H(Sunny) – I(Sunny, Humidity)
= 0.971 – 0
= 0.971

Step8 : Third Attribute – Wind

Categorical values - weak, strong
H(Sunny, Wind=weak) = –(1/3)*log(1/3) – (2/3)*log(2/3) = 0.918
H(Sunny, Wind=strong) = – (1/2)*log(1/2) – (1/2)*log(1/2) = 1
Average Entropy Information for Wind
I(Sunny, Wind) = p(Sunny, weak)*H(Sunny, Wind=weak) + p(Sunny, strong)*H(Sunny, Wind=strong)
= (3/5)*0.918 + (2/5)*1
= 0.9508
Information Gain = H(Sunny) – I(Sunny, Wind)
= 0.971 – 0.9508
= 0.0202

Step9 : Here, when Outlook = Sunny and Humidity = High, it is a pure class of category "no". And When Outlook = Sunny and Humidity = Normal, it is again a pure class of category "yes". Therefore, we do not need to do further calculations.

Step10 : Now, finding the best attribute for splitting the data with Outlook=Sunny values{ Dataset rows = [4, 5, 6, 10, 14]}.

Complete entropy of Rain is –
H(S) = – p(yes) * log2(p(yes)) – p(no) * log2(p(no))
= – (3/5) * log(3/5) - (2/5) * log(2/5)
= 0.91

Step11 : First Attribute – Temperature

Categorical values – mild, cool
H(Rain, Temperature=cool) = – (1/2)*log(1/2) – (1/2)*log(1/2) = 1

H(Rain, Temperature=mild) = – (2/3)*log(2/3) – (1/3)*log(1/3) = 0.918

Average Entropy Information for Temperature

I(Rain, Temperature) = p(Rain, mild)*H(Rain, Temperature=mild) + p(Rain, cool)*H(Rain, Temperature=cool)

= (2/5)*1 + (3/5)*0.918

= 0.9508

Information Gain = H(Rain) – I(Rain, Temperature)

= 0.971 – 0.9508

= 0.0202

Step12: Second Attribute – Wind

Categorical values – weak, strong

H(Wind=weak) = – (3/3)*log(3/3)-0 = 0

H(Wind=strong) = 0 – (2/2)*log(2/2) = 0

Average Entropy Information for Wind -

I(Wind) = p(Rain, weak)*H(Rain, Wind=weak) + p(Rain, strong)*H(Rain, Wind=strong)

= (3/5)*0 + (2/5)*0

= 0

Information Gain = H(Rain) – I(Rain, Wind)

= 0.971 – 0

= 0.971

Step13: Here, the attribute with maximum information gain is Wind. So, the decision tree built so far – maximum information gain is Wind.

Here, when Outlook = Rain and Wind = Strong, it is a pure class of category "no". And When Outlook = Rain and Wind = Weak, it is again a pure class of category "yes".

b. Describe Markov decision process in reinforcement learning.

Ans. Refer Q. 5.11, Page 5–11L, Unit-5.

6. Attempt any one part of the following : **(1 × 10 = 10)**

a. What is instance-based learning ? How locally weighted regression is different from radial basis function networks ?

Ans. **Instance-based learning :** Refer Q. 3.14, Page 3–13L, Unit-3.
Locally weighted regression : Refer Q. 3.14, Page 3–13L, Unit-3.
Radial basis function networks : Refer Q. 3.21, Page 3–18L, Unit-3.

b. How is Bayes theorem used in machine learning ? How naive Bayes algorithm is different from Bayes theorem ?

Ans. Refer Q. 2.6, Page 2–5L, Unit-2.

7. Attempt any one part of the following : **(1 × 10 = 10)**

a. Compare regression, classification and clustering in machine learning along with suitable real life applications ?

Ans. **Regression :** Refer Q. 2.1, Page 2–2L, Unit-2.
Classification and clustering : Refer Q. 1.16, Page 1–17L, Unit-1.

b. Given below is an input matrix named I, kernel matrix, calculate the Convoluted matrix C using stride = 1 also apply max pooling on C.

Input Matrix 1

1	0	0	1	1	0	1
0	0	1	1	1	0	1
1	1	1	0	1	0	1
1	1	0	1	0	0	0
1	0	1	0	1	1	0
0	1	1	0	0	1	1
0	1	1	1	0	1	1

Kernel Matrix

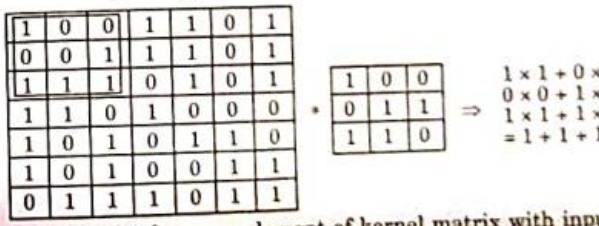
1	0	0
0	1	1
1	1	0

Ans. Here $n = 7, f = 3$

Output = $n - f + 1$

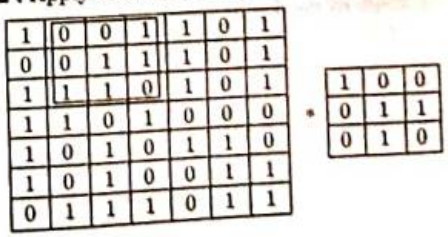
$= 7 - 3 + 1 = 4 + 1 = 5$

Step 1 : Applying kernel matrix on input matrix as :



multiply every element of kernel matrix with input matrix and add

Step 2 : Apply the same using stride = 1 i.e.,



Step 3 : Similarly follow the steps using stride = 1 column wise we will get the output as

4	3	2	3	2
3	1	3	2	2
2	3	2	1	2
2	3	1	2	2
3	2	2	1	4

B. Tech.
(SEM. V) ODD SEMESTER THEORY EXAMINATION, 2021-22
MACHINE LEARNING TECHNIQUES

Time : 3 Hours **Max. Marks : 100**

Note : 1. Attempt **all** Sections. If require any missing data; then choose suitably.

SECTION-A

1. Attempt **all** questions in brief. **(2 × 10 = 20)**

a. What is a "Well -posed Learning "problem ? Explain with an example.

Ans. A Well-posed learning problem is a problem whose solution exists, is unique and the solution depends on data and is not sensitive to small changes in data.

The three important components of a well-posed learning problem are Task, Performance Measure and Experience.

For examples, Learning to play Checkers :

A computer might improve its performance as an ability to win at the class of tasks that are about playing checkers. The performance keeps improving through experience by playing against itself.

To simplify,

T -> Play the checkers game.

P -> Percentage of games won against the opponent.

E -> Playing practice games against itself.

b. What is Occam's razor in ML ?

Ans. Occam's razor suggests that in machine learning, we should prefer simpler models with fewer coefficients over complex models. It is a heuristic that suggests more complex hypotheses make more assumptions that, in turn, will make them too narrow and not generalize well.

c. What is the role of Inductive Bias in ANN ?

Ans. Inductive biases play an important role in the ability of ANN to generalize to the unseen data. The inductive bias (also known as learning bias) of a learning algorithm is the set of assumptions that the learner uses to predict outputs of given inputs that it has not encountered.

d. What is gradient descent delta rule ?

Ans. **Gradient descent :** Refer Q. 4.15, Page SQ–14L, Unit-4, Two Marks Question.

Delta rule : The delta rule is a straight-forward application of gradient descent. It converges toward a best-fit approximation to the target concept if the training instances are not linearly separable.

e. What is Paired *t*-Tests in Hypothesis evaluation ?

Ans. Paired *t*-test is a statistical technique that is used to compare two population means in the case of two samples that are correlated.

f. How do you find the confidence interval for a hypothesis test ?

Ans. Confidence intervals use data from a sample to estimate a population parameter. Hypothesis tests use data from a sample to test a specified hypothesis. Confidence intervals go hand-in-hand with the hypothesis tests. The conclusion drawn from confidence interval is usually the same as the conclusion drawn from hypothesis test. In other words, if the 95% confidence interval contains the hypothesized parameter, then a hypothesis test at the 0.05 α level will almost always fail to reject the null hypothesis. If the 95% confidence interval does not contain the hypothesize parameter, then a hypothesis test at the 0.05 α level will almost always reject the null hypothesis.

g. What is sample complexity of a Learning Problem ?

Ans. Sample complexity is the number of training-samples that we need to supply to the algorithm, so that the function returned by the algorithm is within an arbitrarily small error of the best possible function, with probability arbitrarily close to 1.

h. Differentiate between Lazy and Eager Learning.

Ans.

S. No.	Lazy learning	Eager learning
1.	Lazy learning methods simply store the data and generalizing beyond these data is postponed until an explicit request is made.	Eager learning methods construct general, explicit description of the target function based on the provided training examples.

i. What is the problem of crowding in GA ?

Ans.

1. Crowding is a technique used in genetic algorithms to preserve population diversity by pairing each offspring with a similar individual in the current population (pairing phase) and deciding which of the two will survive (replacement phase).

2. The replacement phase of crowding is usually carried out through deterministic or probabilistic crowding, which have the limitations (problem) that they apply the same selective pressure regardless of the problem being solved and the stage of genetic algorithm search.

j. Comparison of purely analytical and purely inductive learning.

Ans.

S. No.	Purely analytical learning	Purely inductive learning
1.	Inductive learning information is obtained through observation.	Analytical learning information is obtained by explaining and analyzing these observations.
2.	Inductive learning mechanisms are required in order to learn in situations where prior knowledge is incomplete or incorrect.	Analytical learning mechanisms are required in order to scale up to learning complex concepts, and to handle situations in which available training data is limited.

SECTION-B

2. Attempt any three of the following : $(3 \times 10 = 30)$

a. Design the Final design of checkers learning program.

Ans.

A. Steps for designing checkers learning program are :

1. Choosing the Training Experience
2. Choosing the Target Function
3. Choosing a Representation for the Target Function
4. Choosing a Function Approximation Algorithm
5. The Final Design

B. Design of the final design of checkers learning program :

1. The performance System :

i. To solve the given performance task by using the learned target function(s).

ii. It takes a new board as input and outputs a trace of the game it played against itself.

2. The Critic :

i. To take as input the history or trace of the game and produce as output a set of training examples of the target function.

3. The Generalizer :

i. To take as input the training examples and produce an output hypothesis that is its estimate of the target function.

ii. It generalizes from the specific training examples, hypothesizing a general function that covers examples and other cases beyond the training examples.
iii. Good generalization to new cases is crucial.

4. **The Experiment Generator :**
   i. Takes the current hypothesis (currently learned function) as input and outputs a new problem (an initial board state) for the performance system to explore.
   ii. Its role is to pick new practice problems that will maximize the learning rate of the overall system.

b. What is Maximum Likelihood and Least Squared Error Hypothesis ?

Ans.

A. Maximum Likelihood Hypothesis :

1. A common modeling problem involves how to estimate a joint probability distribution for a dataset.
2. This problem is made more challenging as sample (X) drawn from the population is small and has noise.
3. One solution to probability density estimation is referred to as Maximum Likelihood Estimation.
4. Maximum Likelihood Estimation involves treating the problem as an optimization or search problem, where we seek a set of parameters that results in the best fit for the joint probability of the data sample (X).
5. First, it involves defining a parameter called theta that defines both the choice of the probability density function and the parameters of that distribution. It is stated formally as :

   P(X ; theta)

6. This resulting conditional probability is referred to as the likelihood of observing the data given the model parameters and written using the notation L() to denote the likelihood function. For example :

   L(X ; theta)

7. The objective of Maximum Likelihood Estimation is to find the set of parameters (theta) that maximize the likelihood function, e.g., result in the largest likelihood value.

   maximize L(X ; theta)

B. Least Squared Error Hypothesis :

1. The least-square method is a technique commonly used in Regression Analysis.
2. It is a mathematical method used to find the best fit line that represents the relationship between an independent and dependent variable.
3. Many learning approaches such as neural network learning, linear regression, etc.,, try to learn a continuous-valued target function.

4. Under certain assumptions any learning algorithm that minimizes the squared error between the output hypothesis predictions and the training data will output a Maximum Likelihood Hypothesis.
5. The significance of this result is that it provides a Bayesian justification (under certain assumptions) for curve fitting methods that attempt to minimize the sum of squared errors over the training data.

c. What problem does the EM algorithm solve ?

Ans. Following are the problem that EM algorithm solve :

1. The EM algorithm is used for obtaining maximum likelihood estimates of parameters when some of the data is missing.
2. The EM algorithm can also be applied when there is latent, i.e. unobserved, data which was never intended to be observed in the first place.
3. In Bayesian statistics the EM algorithm is used to obtain the mode of the posterior marginal distributions of parameters.

d. Highlight the importance of Case Based Learning.

Ans. Refer Q. 3.26, Page 3-23L, Unit-3.

e. Write short notes on Learning First Order Rules.

Ans.

1. Propositional logic allows the expression of individual propositions and their truth-functional combination.
2. First order logic allows the expression of propositions and their truth functional combination, but it also allows us to represent propositions as assertions of predicates about individuals or sets of individuals.
3. Inference rules permit conclusions to be drawn about sets/ individuals.
4. First order logic is much more expressive than propositional logic- *i.e.*, it allows a finer-grain of specification and reasoning when representing knowledge.
5. First order rule learners can generalise over relational concepts (which propositional learners cannot).
6. All expressions in first-order logic are composed of the following attributes :
   i. Constants - *e.g.* Tyler, 23, *a*
   ii. Variables - *e.g.* *A*, *B*, *C*
   iii. Predicate symbols - *e.g.* male, father (True or False values only)
   iv. Function symbols - *e.g.* age (can take on any constant as a value)
   v. Connectives - *e.g.* $\wedge, \vee, \neg, \rightarrow, \leftarrow$
   vi. Quantifiers - *e.g.* $\forall, \exists$
   vii. **Term :** It can be defined as any constant, variable or function applied to any term. *E.g.* age(bob).

viii. **Literal :** It can be defined as any predicate or negated predicate applied to any terms, e.g., female(sue), father(X, Y).

SECTION-C

3. Attempt any one part of the following : (1 × 10 = 10)

a. Explain the "Concept Learning" task giving an example.

Ans.

1. Concept Learning can be seen as a problem of searching through a predefined space of potential hypotheses for the hypothesis that best fits the training examples.
2. The hypothesis space has a general-to-specific ordering of hypotheses, and the search can be efficiently organized by taking advantage of a naturally occurring structure over the hypothesis space.
3. Concept Learning can be understood from following example :

A concept learning task – Enjoy sport training example :

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	YES
2	Sunny	Warm	High	Strong	Warm	Same	YES
3	Rainy	Cold	High	Strong	Warm	Change	NO
4	Sunny	Warm	High	Strong	Warm	Change	YES

Attributes

Concept

4. A set of example days and each is described by six attributes.
5. The task is to learn to predict the value of EnjoySport for arbitrary day, based on the values of its attribute values.
6. Each hypothesis consists of a conjunction of constraints on the instance attributes.
7. Each hypothesis will be a vector of six constraints, specifying the values of the six attributes
   – (Sky, AirTemp, Humidity, Wind, Water, and Forecast).
8. Each attribute will be :
   ? – indicating any value is acceptable for the attribute (don't care)
   single value – specifying a single required value (e.g. Warm) (specific)
   0 – indicating no value is acceptable for the attribute (no value)

b. Find the maximally general hypothesis and maximally specific hypothesis for the training examples given in the table using the candidate elimination algorithm.

Given Training Example :

Sky	Temp	Humidity	Widd	Water	Forecast	Sport
Sunny	warm	Normal	Strong	warm	same	Yes
Sunny	warm	High	Strong	warm	same	Yes
Rainy	cold	High	Strong	warm	change	No
Sunny	warm	High	Strong	cool	change	Yes

Ans. **Step 1 :**

Initialize G & S as most General and specific hypothesis.

G = {'?', '?','?','?', '?','?'}

S = {'ϕ','ϕ','ϕ','ϕ','ϕ','ϕ'}

Step 2 :

for each +ve example : make a specific hypothesis more general.

s = {'ϕ','ϕ','ϕ','ϕ','ϕ','ϕ'}

Take the most specific hypothesis as your 1st positive instance.

h = {'sunny', 'warm', 'Normal', 'Strong', 'warm', 'same'}

General hypothesis will remain same : G = {'?', '?','?','?', '?','?'}

Step 3 :

Compare with another positive instance for each attribute.

if (attribute value = hypothesis value) do nothing.

else

replace the hypothesis value with more general constraint '?'.

Since instance 2 is also positive so we will compare with it. In instance 2 attribute humidity is changing so we will generalize that attribute.

S = {'sunny', 'warm','?', 'Strong', 'warm', 'same'}

General hypothesis will remain same: G = {'?', '?','?','?', '?','?'}

Step 4 :

Instance 3 is negative so for each –ve example make general hypothesis more specific.

We will make the general hypothesis more specific by comparing all the attributes of the negative instance with the positive instance if attribute found different to create a dedicated set for the attribute.

G = {<'sunny', '?','?','?', '?','?'> , <'?', 'warm','?','?', '?','?'> , <'?', '?','Normal','?', '?','?'> , < '?', '?','?','?', '?','same'>}

Specific hypothesis will be same : S = {'sunny', 'warm','?', 'Strong', 'warm', 'same'}

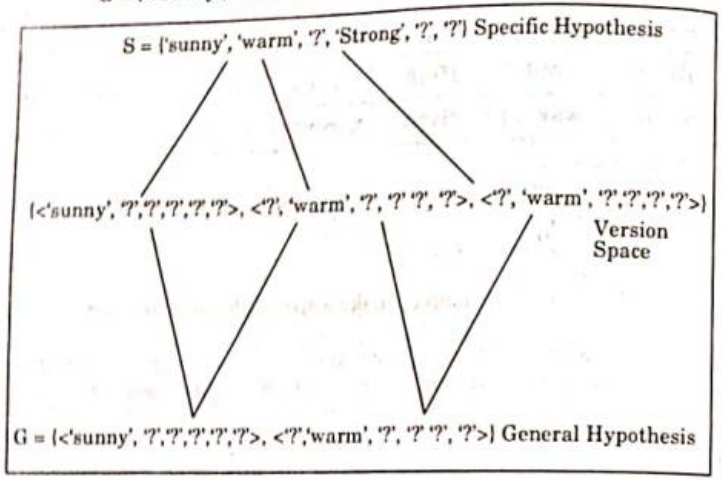
Step 5 :

Instance 4 is positive so repeat step 3 :

S = {'sunny', 'warm', '?', 'Strong', '?', '?'}

Discard the general hypothesis set which is contradicting with a resultant specific hypothesis. Here humidity and forecast attribute is contradicting.

G = {<'sunny', '?','?','?', '?','?'> , <'?', 'warm','?','?', '?','?'> }
∴ Maximally specific and general hypothesis are :
S = {'sunny', 'warm','?', 'Strong', '?', '?'}
G = {<'sunny', '?','?','?', '?','?'> , <'?', 'warm','?','?', '?','?'> }



4. Attempt any one part of the following : (1 × 10 = 10)

a. Comment on the Algorithmic convergence & Generalization property of ANN.

Ans. Algorithmic convergence of ANN :

1. Convergence generally refers to the values of a process that have a tendency in behavior over time.
2. It is a useful idea when working with optimization algorithms.
3. Optimization refers to a type of problem that requires finding a set of inputs that result in the maximum or minimum value from an objective function.
4. Optimization is an iterative process that produces a sequence of candidate solutions until ultimately arriving upon a final solution at the end of the process.
5. This behavior or dynamics of the optimization algorithm arriving on a stable-point final solution is referred to as convergence.
6. The convergence of optimization algorithms, is an important concept in machine learning for those algorithms that fit (learn) on a training dataset via an iterative optimization algorithm, such as artificial neural networks.

Generalization property of ANN :

1. The performance of Artificial Neural Networks (ANN) is mostly dependent upon its generalization capability.
2. Generalization of the ANN is ability to handle unseen data.
3. The generalization capability of the network is mostly determined by system complexity and training of the network.
4. Poor generalization is observed when the network is over-trained or system complexity (or degree of freedom) is relatively more than the training data.
5. A smaller network which can fit the data will have the k good generalization ability.
6. Network parameter pruning is one of the promising methods to reduce the degree of freedom of a network and hence improve its generalization.

b. Discuss the following issues in Decision Tree Learning :

1. **Overfitting the data**
2. **Guarding against bad attribute choices**
3. **Handling continuous valued attributes**
4. **Handling missing attribute values**
5. **Handling attributes with differing costs**

Ans.

A. Overfitting the data :

1. A machine learning algorithm is said to be overfitted, when we train it with a lot of data.
2. When a model gets trained with so much of data, it starts learning from the noise and inaccurate data entries in our data set.
3. Then the model does not categorize the data correctly, because of too much of details and noise.
4. The causes of overfitting are the non-parametric and non-linear methods.
5. A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.

B. Guarding against bad attribute choices :

1. The primary challenge in the decision tree implementation is to identify which attributes do we need to consider as the root node at each level.
2. Handling this is known as attribute selection.
3. As an example consider the attribute Date, which has a very large number of possible values.

4. What is wrong with the attribute Date ? Simply put, it has so many possible values that it is bound to separate the training examples into very small subsets.
5. Because of this, it will have a very high information gain relative to the training examples.
6. To guard against bad attribute choice we use gain ratio.
7. The gain ratio measure penalizes attributes such as Date by incorporating a term, called split information that is sensitive to how broadly and uniformly the attribute splits the data.

C. Handling continuous valued attributes :

1. In ID3 algorithm the attributes tested in the decision nodes of the tree must be discrete valued.
2. The above restriction can easily be removed so that continuous-valued decision attributes can be incorporated into the learned tree.
3. For an attribute A that is continuous-valued, the algorithm can dynamically create a new boolean attribute A, that is true if $A < c$ and false otherwise.
4. The only catch is how to select the best value for the threshold c.

D. Handling missing attribute values :

1. In certain cases, the available data may be missing values for some attributes.
2. In such cases, it is common to estimate the missing attribute value based on other examples for which this attribute has a known value.
3. One strategy for dealing with the missing attribute value is to assign it the value that is most common among training examples at node n.
4. A second, more complex procedure is to assign a probability to each of the possible values of A.

E. Handling attributes with differing costs :

1. In some learning tasks the instance attributes may have associated costs.
2. These attributes vary significantly in their costs.
3. In such tasks, we would prefer decision trees that use low-cost attributes where possible, relying on high-cost attributes only when needed to produce reliable classifications.
4. ID3 algorithm can be modified to consider attribute costs by introducing a cost term into the attribute selection measure.

5. Attempt any **one** part of the following : (1 × 10 = 10)

a. How is Naive Bayesian Classifier different from Bayesian Classifier ?

Ans.

1. The difference between Bayes theorem and Naive Bayes is that Naive Bayes assumes conditional independence where Bayes theorem does not.
2. The Naive Bayes classifier is an approximation to the Bayes classifier, in which they assume that the features are conditionally independent given the class instead of modeling their full conditional distribution given the class.
3. A Bayes classifier is best interpreted as a decision rule.
4. Suppose we seek to estimate the class of an observation given a vector of features. Denote the class C and the vector of features $(F_1, F_2, ..., F_k)$.
5. Given a probability model underlying the data (that is, given the joint distribution of $(C, F_1, F_2, ..., F_k)$), the Bayes classification function chooses a class by maximizing the probability of the class given the observed features :

   $\text{argmax}_c P(C = c \mid F_1 = f_1, ..., F_k = f_k)$
6. Although the Bayes classifier seems appealing, in practice the quantity $P(C = c \mid F_1 = f_1, ..., F_k = f_k)$ is very difficult to compute.
7. We can make it a bit easier by applying Bayes theorem and ignoring the resulting denominator, which is a constant.
8. Then we have the slightly better

   $P(C = c)\, P(F_1 = f_1, ..., F_k = f_k \mid C = c)$

   but this is often still intractable : lots of observations are required to estimate these conditional distributions, and this gets worse as k increases.
9. As a consequence, an approximation is used : we pretend

$$P(F_1 = f_1, ..., F_k = f_k \mid C = c)$$

$$\approx \prod_{i=1}^{k} P(F_i = f_i \mid C = c)$$

10. This is a pretty naive approximation, but in practice it works surprisingly well. Substituting this into the Bayes classifier yields the naive Bayes classifier :

    $\text{argmax}_c P(C = c) \prod_{i=1}^{k} P(F_i = f_i \mid C = c)$.

b. Explain the role of Central Limit Theorem Approach for deriving Confidence Interval.

Ans:

1. An interval can be calculated, within which we would expect the population mean to lie. This gives us our degree of accuracy.
2. However, as we can never be 100% confident of anything, we need to put a confidence level on to this interval.
3. We start by considering a confidence interval for the population mean.
4. Confidence intervals give you a way of quantifying how much variation will appear in repeated measurements and statistical calculations.
5. They use the central limit theorem to quantify how much confidence you can place in any of your measurements or statistical conclusions from samples.
6. The central limit theorem states that the sampling distribution of the mean approaches a normal distribution, as the sample size increases.
7. This fact holds especially true for sample sizes over 30.
8. Therefore, as a sample size increases, the sample mean and standard deviation will be closer in value to the population mean μ and standard deviation σ.
9. Thus, the central limit theorem helps in finding out how much confidence you can place in repeated measurements and statistical calculations.

6. Attempt any one part of the following : (1 × 10 = 10)

a. Write short notes on Probably Approximately Correct (PAC) learning model.

Ans:

1. Probably approximately correct (PAC) learning is a framework for mathematical analysis of machine learning.
2. It was proposed in 1984 by Leslie Valiant of the Harvard University.
3. The PAC model belongs to that class of learning models which is characterized by learning from examples.
4. In this framework, the learner receives samples and must select a generalization function (called the hypothesis) from a certain class of possible functions.
5. The goal is that, with high probability the selected function will have low generalization error.
6. The learner must be able to learn the concept given any arbitrary approximation ratio, probability of success, or distribution of the samples.

7. The model was later extended to treat noise (misclassified samples).
8. An important innovation of the PAC framework is the introduction of computational complexity theory concepts to machine learning.
9. In particular, the learner is expected to find efficient functions and the learner itself must implement an efficient procedure.

b. Discuss various Mistake Bound Model of Learning.

Ans: Following are the various Mistake Bound Model of Learning :

A. Mistake Bound Model: Find-S

1. Instances drawn at random from X according to distribution D.
2. Learner must classify each instance before receiving correct classification from teacher.
3. Consider Find-S when H = conjunction of boolean literals.
   i. Initialize h to the most specific hypothesis $|1 \wedge \neg|1 \wedge |2 \wedge \neg|2 \ldots |n \wedge \neg|n$.
   ii. For each positive training instance x, remove from h any literal that is not satisfied by x.
   iii. Output hypothesis h.
4. How many mistakes before converging to correct h ? n +1.
5. The first hypothesis eliminates n terms, each subsequent mistake will eliminate at least one more term.

B. Mistake Bound Model: Halving Algorithm

1. Consider the Halving Algorithm
   i. Learn concept using version space Candidate-Elimination algorithm.
   ii. Classify new instances by majority vote of version space members.
2. How many mistakes before converging to correct h ? log 2 $|H|$.
3. Every mistake eliminates at least half of the hypothesis from the version space (majority vote).
4. The version space starts at $|H|$.

C. Optimal Mistake Bounds :

1. Let C be an arbitrary nonempty concept class.
2. The optimal mistake bound for C, denoted Opt(C), is the minimum over all possible learning algorithms A of $M_A(C)$.

$$\text{Opt}(C) = \min_{A \text{ elearning algorithms}} M_A(C)$$

3. This definition states that Opt(C) is the number of mistakes made for the hardest target concept in C, using the hardest training sequence, by the best algorithm.
4. Now, for any concept class C, there is an interesting relationship among the optimal mistake bound for C, the bound of the HALVING algorithm, and the VC dimension of C, namely

$$VC(C) \leq \text{Opt}(C) \leq M_{\text{Halving}}(C) \leq \log_2(|C|)$$

7. Attempt any one part of the following : (1 × 10 = 10)

a. What is the significance of Learn-one Rule Algorithm ?

Ans.

1. This method is used in the sequential learning algorithm for learning the rules.
2. It returns a single rule that covers at least some examples.
3. However, what makes it really powerful is its ability to create relations among the attributes given, hence covering a larger hypothesis space.
4. The Learn-One-Rule algorithm follows a greedy searching paradigm where it searches for the rules with high accuracy but its coverage is very low.
5. It classifies all the positive examples for a particular instance.
6. It involves a PERFORMANCE method that calculates the performance of each candidate hypothesis, (*i.e.*, how well the hypothesis matches the given set of examples in the training data.

   Performance(NewRule, h) :

   h-examples = the set of rules that match h

   return (h-examples)
7. It starts with the most general rule precondition, then greedily adds the variable that most improves performance measured over the training examples.

b. Describe a prototypical genetic algorithm along with various operations possible in it.

Ans.

Table 1 :

GA(Fitness, Fitness_threshold, p, r, m)

Fitness : A function that assigns an evaluation score, given a hypothesis.

Fitness_threshold : A threshold specifying the termination criterion.

p : The number of hypotheses to be included in the population.

r : The fraction of the population to be replaced by Crossover at each step.

m : The mutation rate.

- **Initialize population :** $P \leftarrow$ Generate p hypotheses at random
- **Evaluate :** For each h in P, compute Fitness(h)
- While [max Fitness(h)] < Fitness_threshold do
  h

  Create a new generation, P_s :

1. **Select :** Probabilistically select $(1 - r)p$ members of P to add to P_s. The probability $Pr(h_i)$ of selecting hypothesis h_i from P is given by

$$Pr(h_i) = \frac{\text{Fitness}(h_i)}{\sum_{j=1}^{p} \text{Fitness}(h_j)}$$

2. **Crossover :** Probabilistically select $\frac{r \cdot p}{2}$ pairs of hypotheses from P, according to $Pr(h_i)$ given above. For each pair, (h_1, h_2), produce two offspring by applying the Crossover operator. Add all offspring to P_s.
3. **Mutate :** Choose m percent of the members of P_s with uniform probability. For each, invert one randomly selected bit in its representation.
4. **Update :** $P \leftarrow P_s$.
5. **Evaluate :** for each h in P, compute Fitness(h)

- Return the hypothesis from P that has the highest fitness.

1. A prototypical genetic algorithm is described in Table 1.
2. The inputs to this algorithm include the fitness function for ranking candidate hypotheses, a threshold defining an acceptable level of fitness for terminating the algorithm, the size of the population to be maintained, and parameters that determine how successor populations are to be generated: the fraction of the population to be replaced at each generation and the mutation rate.
3. In this algorithm each iteration through the main loop produces a new generation of hypotheses based on the current population.
4. First, a certain number of hypotheses from the current population are selected for inclusion in the next generation.
5. These are selected probabilistically, where the probability of selecting hypothesis h_i is given by

$$Pr(h_i) = \frac{\text{Fitness}(h_i)}{\sum_{j=1}^{p} \text{Fitness}(h_j)}$$

6. Thus, the probability that a hypothesis will be selected is proportional to its own fitness and is inversely proportional to the fitness of the other competing hypotheses in the current population.
7. Once these members of the current generation have been selected for inclusion in the next generation population, additional members are generated using a crossover operation.
8. After new members have been created by this crossover operation, the new generation population now contains the desired number of members.
9. At this point, a certain fraction m of these members are chosen at random, and random mutations all performed to alter these members.
10. This GA algorithm thus performs a randomized, parallel beam search for hypotheses that perform well according to the fitness function.

☺☺☺

B. Tech.
(SEM. V) ODD SEMESTER THEORY EXAMINATION, 2022-23
MACHINE LEARNING TECHNIQUES

Time : 3 Hours **Max. Marks : 100**

Note : 1. Attempt **all** sections. If require any missing data; then choose suitably.

Section-A

1. Attempt **all** questions in brief. (2 × 10 = 20)

a. Discuss model representation of artificial neuron.

Ans. Refer Q. 4.12, Page SQ–13L, Unit-4, Two Marks Questions.

b. Explain general-to-specific ordering hypothesis in concept learning.

Ans. The general-to-specific ordering hypothesis is a concept learning hypothesis that suggests that people tend to learn concepts by first acquiring general, abstract concepts and then gradually refining them to become more specific and concrete.

c. Discuss support vectors in SVM.

Ans. In SVM, support vectors are the data points that lie closest to the decision boundary or hyperplane that separates the different classes in the data. These support vectors are used to define the hyperplane and to make predictions on new data.

d. Compare artificial intelligence and machine learning.

Ans.

S. No.	Artificial intelligence	Machine learning
1.	Artificial intelligence is a technology which enables a machine to simulate human behavior.	Machine learning is a subset of AI which allows a machine to automatically learn from past data without programming explicitly.
2.	The goal of AI is to make a smart computer system like humans to solve complex problems.	The goal of ML is to allow machines to learn from data so that they can give accurate output.

e. Discuss reinforcement learning.

Ans. Refer Q. 5.1, Page 5–2L, Unit-5.

f. Illustrate the advantages of instance-based learning techniques over other machine learning techniques.

Ans. Refer Q. 3.17, Page 3–15L, Unit-3.

g. Differentiate between gradient descent and stochastic gradient descent.

Ans.

S. No.	Gradient Descent	Stochastic Gradient Descent
1.	Suitable for small datasets.	Suitable for large datasets.
2.	Fixed learning rate at each iteration.	Dynamic learning rate that decreases over time.
3.	High computational cost.	Low computational cost.

h. Compare ANN and Bayesian network.

Ans.

S. No.		Artificial Neural Network (ANN)	Bayesian Network
1.	Architecture	Non-linear interconnected network of nodes and layers.	Directed Acyclic Graph (DAG).
2.	Learning	Supervised, unsupervised or reinforcement learning.	Supervised, unsupervised or semi-supervised learning.
3.	Probabilistic	Not inherently probabilistic.	Inherently probabilistic.

i. Illustrate Markov decision model.

Ans. The graphical representation of the Markov decision model is as follows :

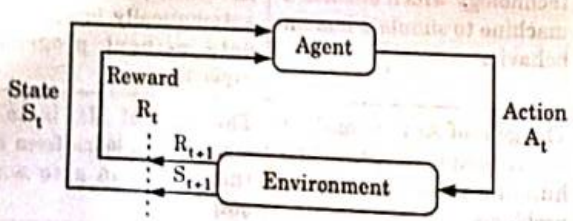


Fig. 1.

j. Differentiate between Q-learning and deep learning.

Ans.

S. No.		Q-Learning	Deep Learning
1.	Definition	Reinforcement learning technique for estimating Q-values.	Subset of machine learning that uses artificial neural networks.
2.	Algorithm	Model-free, iterative approach.	Model-based or model-free, depending on the architecture.
3.	Goal	To learn an optimal policy based on reward signals.	To perform supervised or unsupervised learning on complex data.

Section-B

2. Attempt any **three** of the following : $(10 \times 3 = 30)$

a. Explain supervised and unsupervised learning techniques.

Ans. Refer Q. 1.3, Page 1–4L, Unit-1.

b. Discuss linear regression and logistic regression in detail.

Ans. **Linear regression :** Refer Q. 2.2, Page 2–2L, Unit-2.
Logistics regression : Refer Q. 2.3, Page 2–3L, Unit-2.

c. Describe the following concepts in decision tree in detail :
i. Avoiding overfitting in derision tree.
ii. Incorporating continuous valued attributes.

Ans.

i. Avoiding overfitting in decision tree : Overfitting is a common problem in decision trees, which occurs when the tree is too complex and tailored to the training data, leading to poor generalization on new, unseen data. There are several techniques to avoid overfitting in decision trees :

1. **Pre-pruning :**
   a. The pre-pruning technique refers to the early stopping of the growth of the decision tree.
   b. The pre-pruning technique involves tuning the hyperparameters of the decision tree model prior to the training pipeline.
   c. The hyperparameters of the decision tree including max_depth, min_samples_leaf, min_samples_split can be tuned to early stop the growth of the tree and prevent the model from overfitting.

2. **Post-pruning :**
   a. The Post-pruning technique allows the decision tree model to grow to its full depth, then removes the tree branches to prevent the model from overfitting.
   b. Cost Complexity Pruning (CCP) is one type of post-pruning technique.
   c. In case of cost complexity pruning, the ccp_alpha can be tuned to get the best fit model.

ii. Incorporating continuous valued attributes : Decision trees are designed to work with discrete or categorical data, but many real-world problems involve continuous-valued attributes, such as age, height, or income. There are several techniques to incorporate continuous-valued attributes into decision trees :

1. **Discretization :**
   i. Convert continuous-valued attributes into discrete intervals or categories, such as age groups or income ranges, and treat them as nominal attributes.
   ii. This approach can simplify the tree and reduce the number of splits, but may lose some information and lead to less accurate models.
2. **Binary splitting :**
   i. Perform a binary split on each continuous-valued attribute by selecting a threshold value that maximizes the information gain or other splitting criterion.
   ii. This approach can preserve more information than discretization, but may lead to overfitting if the threshold values are too specific or not representative of the data.

d. Explain various types of activation functions with examples.

Ans. Types of activation function are :

1. **Linear function :**
   a. Linear function has the equation similar to as of a straight line *i.e.*, $y = x$.
   b. No matter how many layers we have, if all are linear in nature, the final activation function of last layer is nothing but just a linear function of the input of first layer.
   c. For example, calculation of price of a house is a regression problem. House price may have any big/small value, so we can apply linear activation at output layer.
2. **Sigmoid function :**
   a. It is a function which is plotted as '*S*' shaped graph.
   b. Equation : $A = 1/(1 + e - x)$.
   c. For example, calculation of multiple value.
3. **Tanh function :**
   a. The activation that works almost always better than sigmoid function is Tanh function also knows as Tangent Hyperbolic function.

   b. It's actually mathematically shifted version of the sigmoid function. Both are similar and can be derived from each other.
   c. For example, calculation for finding mean for hidden layer of neural network.
4. **RELU function :**
   a. It stands for rectified linear unit. It is the most widely used activation function. Chiefly implemented in hidden layers of neural network.
   b. Equation : $A(x) = \max(0, x)$. It gives an output x if x is positive and 0 otherwise.
   c. For example, calculate activation function for convolutional neural networks.

e. Illustrate the process of *Q*-learing and discuss the following terms :

i. *Q*-values or action value

ii. Rewards and episode

iii. Temporal difference or TD update.

Ans. Q-learning process : Refer Q. 5.14 and Q. 5.15 Page 5–12L, Unit-5.

i. Q-values or action value :

1. The Q-values represent the expected utility of taking an action in a given state, and are stored in a Q-table.
2. The Q-value for a state-action pair (s, a) is updated based on the observed reward and the maximum Q-value of the next state :

$$Q(s, a) = Q(s, a) + \alpha [r + \gamma \max Q(s', a') - Q(s, a)]$$

where, α is the learning rate that controls the weight of the new observation,
γ is the discount factor that balances immediate and future rewards,
r is the observed reward, and
a' is the optimal action in the next state.

ii. Rewards and episode :

1. Rewards are feedback signals that the agent receives from the environment based on the actions it takes in a given state.
2. The goal of the agent is to maximize the cumulative reward over multiple steps or episodes.
3. An episode is a sequence of state-action-reward tuples that starts from the initial state and ends in a terminal state or a maximum number of steps.

iii. Temporal difference or TD update :

1. The Temporal Difference (TD) update is a key concept in Q-learning that computes the error between the observed reward and the predicted Q-value, and updates the Q-value accordingly.
2. The TD error is defined as :

$$d = r + \gamma \max Q(s', a') - Q(s, a)$$

where, r is the observed reward,
γ is the discount factor that balances immediate and future rewards,
a' is the optimal action in the next state, and
s' is the next state.

3. The TD update equation updates the Q-value for the state-action pair based on the TD error and the learning rate :
$Q(s, a) = Q(s, a) + a\delta$
where, a is the learning rate that controls the weight of the TD error.
4. The TD update is the basis for the iterative Q-value update in Q-learning.

Section-C

3. Attempt any **one** part of the following : **(10 × 1 = 10)**

a. Illustrate the various areas in which you can apply machine learning.

Ans. Refer Q. 1.10, Page 1–11L, Unit-1.

b. Compare regression, classification and clustering in machine learning along with suitable real life applications.

Ans.

S. No.	Regression	Classification	Clustering
1.	Supervised learning technique.	Supervised learning technique.	Unsupervised learning technique.
2.	Output variable-continuous.	Output variable-discrete.	Output variable-not given.
3.	It uses linear regression algorithm.	It uses logistic regression algorithm.	It uses k-means algorithm.
4.	It aims to forecast.	It compute the category of data.	It group similar item cluster.
5.	For example, predict stock market price.	For example, classify emails as spam or non-spam.	For example, find all transaction which are fraudulent in nature.

4. Attempt any **one** part of the following : **(10 × 1 = 10)**

a. Discuss the role of Bayes theorem in machine learning. How naive Bayes algorithm is different from Bayes theorem ?

Ans. **Role of Bayes theorem in machine learning :**

1. In machine learning, Bayes theorem is often used to estimate the probability of a class given some input features in a classification task.
2. This is known as Bayesian classification or probabilistic classification.
3. The goal is to find the class that maximizes the posterior probability given the input features.
4. Bayes theorem provides a framework for combining prior knowledge about the classes with observed data to make predictions.
5. Bayes theorem is also used in Bayesian regression.
6. Bayesian regression combines a prior distribution over the model parameters with the likelihood of the data to compute a posterior distribution over the parameters.
7. This allows us to estimate the uncertainty of the predictions and make more informed decisions based on the available evidence.

Difference : The main difference between Naive Bayes algorithm and Bayes theorem is that Naive Bayes algorithm assumes that the features are conditionally independent given the class, while Bayes theorem makes no assumption about the relationship between the features.

b. Explain hyperplane (decision boundary) in SVM. Categorize various popular kernels associated with SVM.

Ans. **Hyperplane :** Refer Q. 2.23, Page 2–21L, Unit-2.
Kernel associated with SVM :

1. **Polynomial kernel :** Refer Q. 2.21, Page 2–20L, Unit-2.
2. **Gaussian kernel :** Refer Q. 2.22, Page 2–21L, Unit-2.

5. Attempt any **one** part of the following : **(10 × 1 = 10)**

a. Demonstrate K-Nearest neighbors-algorithm for classification with the help of an example.

Ans. **KNN algorithm :** Refer Q. 3.18, Page 3–16L, Unit-3.
Example : Refer Q. 2(a), Page SP–3L, Solved Paper (2020-21).

b. Explain instance-based learning. Compare locally weighted regression and radial basis function networks.

Ans. **Instance-based learning :** Refer Q. 3.13, Page 3–12L, Unit-3.

S. No.		Locally Weighted Regression (LWR)	Radial Basis Function Networks (RBFN)
1.	Purpose	Non-parametric regression method.	Feed forward neural network.
2.	Training	Instance-based learning method.	Supervised learning method.

3.	Number of parameters	No fixed set of parameters.	Fixed set of parameters determined by the number of basis functions and their width.
4.	Model complexity	Low complexity.	Higher complexity.
5.	Applications	Used in robotics, control, and signal processing.	Used in function approximation, classification, and clustering tasks.

6. Attempt any **one** part of the following : **(10 × 1 = 10)**

a. Explain the different layers used in convolutional neural network with suitable examples.

Ans. Following are the different layers used in a typical CNN :

1. **Input layer :** This layer represents the input image or data. It is usually a rectangular grid of pixels or values that represent the features of the image.

   For example, an input layer of a CNN for recognizing handwritten digits may be a 28x28 grayscale image of a digit.

2. **Convolutional layer :** This layer applies a set of filters or kernels to the input image to extract features. Each filter slides over the image and performs a dot product between its weights and the pixels in the corresponding receptive field. The result is a set of feature maps that highlight different aspects of the input image. **For example,** a convolutional layer in a CNN for recognizing faces may extract features such as edges, corners, and textures.

3. **ReLU layer :** This layer applies the rectified linear unit (ReLU) activation function to the output of the convolutional layer. ReLU sets all negative values to zero and keeps positive values unchanged. This introduces non-linearity into the network and helps to speed up the training process.

   For example, a ReLU layer in a CNN for recognizing animals may enhance the edges and contours of the animals in the feature maps.

4. **Pooling layer :** This layer reduces the dimensionality of the feature maps by down-sampling them. It applies a pooling function, such as max pooling or average pooling, to a small window of pixels in the feature map and outputs the maximum or average value. This reduces the size of the feature maps and helps to improve the robustness of the network to small variations in the input.

   For example, a pooling layer in a CNN for recognizing objects may down-sample the feature maps and preserve the most salient features.

5. **Fully connected layer :** This layer takes the flattened output of the previous layers and applies a set of weights to compute a score for each possible class. It is similar to the dense layer in a traditional neural network.

   For example, a fully connected layer in a CNN for recognizing flowers may take the flattened output of the previous layers and compute a score for each possible flower species.

6. **Softmax layer :** This layer applies the softmax function to the scores computed by the previous layer to produce a probability distribution over the possible classes. The output of the softmax layer represents the predicted class of the input image.

   For example, a softmax layer in a CNN for recognizing fruits may take the scores computed by the previous layer and produce a probability distribution over the possible fruit types.

b. Illustrate backpropagation algorithm by assuming the training rules for output unit weights and hidden unit weights.

Ans. Following is an illustration of the backpropagation algorithm :

1. **Initialize the weights :** Set the weights for each neuron in the network to small random values.

2. **Forward propagation :** Feed the input data into the network and calculate the output of each neuron by applying the activation function.

3. **Calculate the error :** Compute the error between the predicted output and the true output for each training example using a cost function, such as mean squared error.

4. **Backward propagation for the output layer :** Update the weights of the output layer neurons using the chain rule of calculus to calculate the partial derivatives of the cost function with respect to the output weights. The update rule for the output weights is :

   delta_w_out = learning_rate * error *

   derivative_of_activation_function(output) * hidden_output

   w_out = w_out - delta_w_out

   where, 'delta_w_out' is the change in the weight,

   'learning_rate' is a hyperparameter that controls the size of the weight update,

   'error' is the difference between the predicted output and the true output,

   'derivative_of_activation_function' is the derivative of the activation function used for the output layer, and

   'hidden_output' is the output of the hidden layer neurons.

5. **Backward propagation for the hidden layer :** Update the weights of the hidden layer neurons using the chain rule of calculus to calculate the partial derivatives of the cost function with respect to the hidden weights. The update rule for the hidden weights is :

   delta_w_hidden = learning_rate *

   derivative_of_activation_function(hidden_output) * error * w_out * derivative_of_activation_function(hidden_input) * input

   w_hidden = w_hidden - delta_w_hidden

6. Repeat steps 2-5 for a fixed number of epochs or until the error converges.

7. Attempt any one part of the following : (10 × 1 = 10)

a. Explain various types of reinforcement learning techniques with suitable examples.

Ans. Refer Q. 5.10, Page 5–10L, Unit-5.

b. How to identify the reproduction cycle of genetic algorithm ? Explain with suitable example.

Ans. **Reproduction cycle of genetic algorithm :**

1. The reproduction cycle of a genetic algorithm refers to the process of creating a new generation of candidate solutions from the previous generation using genetic operators.

2. The reproduction cycle typically consists of the following steps :

   i. **Selection :** Select a subset of the fittest individuals from the previous generation to become parents for the next generation.

   ii. **Crossover :** Create new individuals by combining the genetic material of the selected parents through a crossover operation.

   iii. **Mutation :** Introduce random changes into the genetic material of the new individuals through a mutation operation.

   iv. **Replacement :** Replace the least fit individuals from the previous generation with the new individuals to form the next generation.

3. The reproduction cycle is repeated for a fixed number of generations or until a termination criterion is met, such as reaching a maximum fitness or a maximum number of generations.

Example :

1. Consider a genetic algorithm for optimizing the parameters of a machine learning model.

2. The genetic algorithm starts with a randomly initialized population of candidate solutions, each of which corresponds to a different set of parameters for the model.

3. In each generation, the fitness of each individual is evaluated.

4. The fittest individuals are selected as parents for the next generation.

5. The new individuals are created by combining their genetic material through crossover and introducing random changes through mutation.
6. The new individuals are then evaluated for fitness and used to form the next generation.
7. This process is repeated for a fixed number of generations or until a termination criterion is met.
7. By following this reproduction cycle, the genetic algorithm can gradually improve the performance of the machine learning model.

☺☺☺

B. Tech.
(SEM. V) ODD SEMESTER THEORY EXAMINATION, 2023-24
MACHINE LEARNING TECHNIQUES

Time : 3 Hours **Max. Marks : 100**

Note : 1. Attempt **all** sections. If require any missing data; then choose suitably.

Section-A

1. Attempt **all** questions in brief. (2 × 10 = 20)

a. Discuss the important objectives of machine learning.

Ans. The key objectives of machine learning include :
1. Enhancing decision-making by enabling predictive analytics,
2. Improving efficiency through automation of tasks,
3. Uncovering patterns and insights from large datasets,
4. Facilitating adaptive learning in systems, and
5. Personalizing user experiences.

b. Discuss overfitting and underfitting situation in decision tree learning.

Ans. Refer Q. 1(e), Page SP–2L, Solved Paper (2020-21).

c. Discuss support vectors in SVM.

Ans. Support vectors in SVM are the critical data points that lie closest to the decision boundary. They define the margin of the classifier and are essential in constructing the optimal hyperplane.

d. What is gradient descent delta rule ?

Ans. The gradient descent delta rule, also known as the delta rule, is an optimization algorithm used in training neural networks. It involves adjusting the weights of the network to minimize the error between predicted and actual outputs.

e. Explain case-based learning.

Ans. Case-based learning is an educational approach where learners analyze real-life scenarios, or "cases," to apply theoretical knowledge in practical contexts.

f. For which problem decision tree is best suitable.

Ans. Decision trees are best suited for classification and regression problems where the data has clear, hierarchical decision rules.

g. Define the term ANN and CNN.

Ans. **ANN :** Refer Q. 1.13, Page 1–14L, Unit-1.
CNN : Refer Q. 4.28, Page 4–24L, Unit-4.

h. Differentiate between Lazy and Eager Learning.

Ans. Refer Q. 1(h), Page SP–14L, Solved Paper (2021-22).

i. Comparison of purely analytical and purely inductive learning.

Ans. Refer Q. 1(j), Page SP–15L, Solved Paper (2021-22).

j. Define the term Offspring, Chromosome and Genes used in GA.

Ans. **Offspring :** An offspring is a new solution generated by combining parent solutions through crossover and mutation.
Chromosomes : A chromosome represents a potential solution encoded as a string of genes.
Genes : Genes are individual units within a chromosome, representing specific parameters or traits that influence the solution's performance.

Section-B

2. Attempt any **three** of the following : **(10 × 3 = 30)**

a. Compare supervised and unsupervised learning techniques with examples.

Ans.

S. No.	Aspect	Supervised learning	Unsupervised learning
1.	Definition	Uses labeled data to train models.	Uses unlabeled data to find hidden patterns.
2.	Objective	Predict outcomes based on input-output pairs.	Discover structure and relationships in data.
3.	Output	Known labels or values.	Groupings or feature representations.
4.	Complexity	Typically higher.	Typically lower.
5.	Examples	Classification (*e.g.*, spam detection), regression (*e.g.*, house price prediction).	Clustering (*e.g.*, customer segmentation), dimensionality reduction (*e.g.*, PCA).

b. Explain maximum likelihood and least squared error Hypothesis with example.

Ans. Refer Q. 2(b), Page SP–16L, Solved Paper (2021-22).
Example of maximum likelihood hypothesis : Given data points from a normal distribution, MLE will compute the mean (μ) and standard deviation (σ) that maximize the likelihood function $L(\mu, \sigma) = \Pi\, P(x\_i \mid \mu, \sigma)$.
Example of least squared error hypothesis : In linear regression, given data points (x_i, y_i), the least squares method finds the line $y = mx + b$ that minimizes $\Sigma(y\_i - (mx\_i + b))^2$, providing the best fit line through the data points.

c. Compare and contrast Information Gain, Gain Ratio, and Gini Index in detail.

Ans.

S. No.	Criteria	Information Gain	Gain Ratio	Gini Index
1.	Definition	Measures reduction in entropy or impurity post-split.	Adjusts Information Gain considering split information.	Measures dataset impurity via class probability.
2.	Purpose	Identifies most informative attribute for dataset split.	Reduces bias towards attributes with many values.	Evaluates dataset impurity for decision tree split.
3.	Bias	Biased towards attributes with many distinct values.	Reduces bias towards attributes with many values.	Unbiased measure of dataset impurity.
4.	Splitting metric	Primary metric for decision tree splitting splitting (e.g., ID3, C4.5).	Used in decision trees (e.g., C4.5) to address bias.	Widely used in decision trees (e.g., CART).
5.	Com-putation	Entropy of parent and children nodes.	Information Gain and split information.	Probabilities of each class in the dataset.
6.	Com-plexity	Simpler computation compared to others.	Slightly more complex than Information Gain.	Similar complexity to Information Gain.

d. Explain the different layers used in convolutional neural network with suitable examples.

Ans. Convolutional Neural Networks (CNNs) consist of following key layers :

1. **Convolutional layer :** Applies convolutional filters to input data, extracting features like edges or textures.

   Example : In an image of a cat, initial filters might detect edges, while deeper layers identify shapes and patterns unique to cats.

2. **Activation layer :** Introduces non-linearity using functions like ReLU (Rectified Linear Unit).

   Example : ReLU replaces negative pixel values with zero, allowing the network to learn complex patterns.

3. **Pooling layer :** Reduces spatial dimensions, keeping essential information while decreasing computation.

   Example : Max pooling selects the highest value from a region of the feature map, retaining prominent features and reducing size.

4. **Fully connected layer :** Connects neurons from the previous layer to every neuron in the next layer, facilitating high-level reasoning.

   Example : In the final layers of an image classifier, these neurons might recognize entire objects like "cat" or "dog" from the features extracted by earlier layers.

5. **Output layer :** Produces the final prediction, often using a softmax function for classification.

   Example : In a digit recognition task, the output layer assigns probabilities to each digit (0 – 9).

e. Discuss the applications of reinforcement learning. In which problems reinforcement learning is used ?

Ans. Refer Q. 5.12, Page 5–11L, Unit-5.

Section-C

3. Attempt any **one** part of the following : **(10 × 1 = 10)**

a. Compare regression, classification and clustering in machine learning along with suitable real life applications.

Ans.

S. No.	Criteria	Regression	Classification	Clustering
1.	Definition	Predicts continuous numerical values.	Assigns class labels or categories.	Groups similar data points.
2.	Output	Continuous values	Discrete classes	Cluster memberships.
3.	Algorithms	Linear regression, Random forest regression.	Logistic regression, Decision trees.	K-means, Hierarchical clustering.
4.	Input features	Numerical features.	Categorical or numerical features.	Numerical features.
5.	Real life application	Predictive analytics, forecasting.	Medical diagnosis, fraud detection.	Customer segmentation, anomaly detection.

b. Explain the "Concept Learning" task giving an example.

Ans. Refer Q. 3(a), Page SP–18L, Solved Paper (2021-22).

4. Attempt any **one** part of the following : **(10 × 1 = 10)**

a. Explain hyperplane (decision boundary) in SVM. Categorize various popular kernels associated with SVM.

Ans. **Hyperplane in SVM :** Refer Q. 2.23, Page 2–21L, Unit-2.

Some popular kernels associated with SVM :

1. **Linear kernel :**
   a. Linear kernel is the simplest and most commonly used kernel function.
   b. It defines the dot product between the input vectors in the original feature space.
   c. The linear kernel can be defined as :
      $K(x, y) = x.y$
2. **Polynomial Kernel :** Refer Q 2.21, Page 2-20L, Unit-2.
3. **Gaussian Kernel :** Refer Q 2.22, Page 2-21L, Unit-2.

b. Differentiate between Naive Bayes classifier and Bayesian belief networks. Give an application of Bayesian belief networks.

Ans:

S. No.	Aspect	Naive Bayes Classifier	Bayesian Belief Networks
1.	Basis	Utilizes Bayes theorem for classification.	Constructs a graphical model showing variable dependencies.
2.	Structure	Typically represented as a single probabilistic model.	Exhibits a graph structure illustrating variable relations.
3.	Complexity	Simple model, easy to understand and implement.	More complex, requires learning network structure.
4.	Use Cases	Well-suited for text classification, spam filtering, etc.	Effective in decision support, medical diagnosis, etc.
5.	Learning	Fast to train with small datasets.	Slower learning, especially with large datasets.

Application of Bayesian belief network :

1. **Medical diagnosis :** BBNs assist in diagnosing diseases by modeling relationships between symptoms, aiding in accurate treatment decisions.
2. **Fault diagnosis :** BBNs pinpoint faults in systems like manufacturing and networks by modeling dependencies between components and symptoms.

5. Attempt any **one** part of the following : **(10 × 1 = 10)**

a. Discuss decision tree and explain its working in detail.

Ans: **Decision tree :** A decision tree is a machine learning model used for classification and regression tasks. It operates by recursively splitting the data into subsets based on feature values, creating a tree-like structure of decisions.

Working : Refer Q. 3.3, Page 3–3L, Unit-3.

b. Demonstrate K-nearest neighbors algorithm for classification with the help of an example.

Ans: **K-NN algorithm :** Refer Q. 3.18, Page 3–16L, Unit-3.

Example : Let's classify a new point based on a dataset of fruit features (weight and color) with known labels (Apple or Orange).

Dataset :

Weight	Color (1 = Red, 0 = Orange)	Label
150	1	Apple
170	0	Orange
140	1	Apple
130	0	Orange
160	1	Apple

New point : Weight = 155, Color = 0

1. **Choose k :** Assume $k = 3$.
2. **Calculate distance :** Compute Euclidean distances from the new point to all training points :
   a. Distance to (150, 1) : $\sqrt{((155 - 150)^2 + (0 - 1)^2)} = \sqrt{(25 + 1)} = \sqrt{26} \approx 5.10$
   b. Distance to (170, 0) : $\sqrt{((155 - 170)^2 + (0 - 0)^2)} = \sqrt{(225)} = 15$
   c. Distance to (140, 1) : $\sqrt{((155 - 140)^2 + (0 - 1)^2)} = \sqrt{(225 + 1)} = \sqrt{226} \approx 15.03$
   d. Distance to (130, 0) : $\sqrt{((155 - 130)^2 + (0 - 0)^2)} = \sqrt{(625)} = 25$
   e. Distance to (160, 1) : $\sqrt{((155 - 160)^2 + (0 - 1)^2)} = \sqrt{(25 + 1)} = \sqrt{26} \approx 5.10$
3. **Find nearest neighbors :** The three closest points are :
   a. (150, 1) Apple
   b. (160, 1) Apple
   c. (170, 0) Orange
4. **Majority vote :** Among the 3 nearest neighbors, 2 are Apples and 1 is Orange. Therefore, the new point is classified as Apple.

6. Attempt any **one** part of the following : **(10 × 1 = 10)**

a. Illustrate back propagation algorithm by assuming the training rules for output unit weights and hidden unit weights.

Ans: Refer Q. 6(b), Page SP–37L, Solved Paper (2022-23).

b. Write short notes on probably approximately correct (PAC) learning.

Ans: Refer Q. 6(a), Page SP–24L, Solved Paper (2021-22).

7. Attempt any **one** part of the following : **(10 × 1 = 10)**

a. Explain Q-learning with its key terms, key feature and elements. Discuss its applications used in real life.

Ans: **Q-learning :** Refer Q. 5.14, Page 5–12L, Unit-5.

Key terms :

1. **States :** The state, S, represents the current position of an agent in an environment.

2. **Action :** The action, A, is the step taken by the agent when it is in a particular state.
3. **Rewards :** For every action, the agent will get a positive or negative reward.
4. **Episodes :** When an agent ends up in a terminating state and can't take a new action.
5. **Q-values :** Used to determine how good an action, A, taken at a particular state, S, is $Q(A, S)$.
6. **Temporal difference :** A formula used to find the Q-value by using the value of current state and action and previous state and action.

Key features and elements :

1. **Exploration vs. exploitation :** Q-learning balances exploration (trying new actions) and exploitation (choosing actions based on learned Q-values) to find the optimal policy.
2. **Q-value update rule :** The Q-value of a state-action pair is updated iteratively based on the observed rewards and the estimated value of the next state.
3. **Bellman equation :** Q-learning is based on the Bellman equation, which expresses the optimal Q-value of a state-action pair as the sum of the immediate reward and the discounted value of the next state's optimal Q-value.
4. **Greedy policy :** After learning the Q-values, the agent selects actions greedily by choosing the action with the highest Q-value in each state.
5. **Learning rate (α) :** A parameter that determines the rate at which the Q-values are updated. It balances new information with previously learned values.

Applications in real life :

1. **Robotics :** Q-learning is used to teach robots how to navigate in unknown environments, optimize their paths, and make decisions to achieve specific goals.
2. **Game playing :** Q-learning has been applied to develop game-playing agents that learn optimal strategies in games like chess, and video games.
3. **Resource management :** In telecommunications and network management, Q-learning can be used to optimize resource allocation, routing decisions, and network performance.
4. **Finance :** Q-learning techniques are applied in algorithmic trading to learn optimal trading strategies based on market data and maximize profits while minimizing risks.

b. Define the term genetic algorithm. Discuss the working of genetic algorithm with the help of flowchart.

Ans: Genetic algorithm : Refer Q. 1.24, Page 1–23L, Unit-1.
Working : Refer Q. 5.22, Page 5–18L, Unit-5.

☺☺☺

Quantum Series
ENGINEERING

QUANTUM Series

Related titles in Quantum Series

For Semester - 5
(Computer Science & Engineering / Information Technology)

- Database Management System
- Web Technology
- Design and Analysis of Algorithm
- Artificial Intelligence

Departmental Elective-I

- Data Analytics
- Computer Graphics
- Object Oriented System Design with C++

Departmental Elective-II

- Machine Learning Techniques
- Application of Soft Computing
- Image Processing
- Data Warehousing & Data Mining

Commom Non Credit Course (NC)

- Constitution of India
- Essence of Indian Traditional Knowledge

A comprehensive book to get the big picture without spending hours over lengthy text books.

Quantum Series is the complete one-stop solution for engineering student looking for ...mple yet effective guidance system for ... engineering subject. Based on the ...ds of students and catering to the ...irements of the syllabi, this series ...quely addresses the way in which ...ncepts are tested through university examinations. The easy to comprehend question answer form adhered to by the books in this series is suitable and recommended for student. The students are able to effortlessly grasp the concepts and ideas discussed in their course books with the help of this series. The solved question papers of previous years act as a additional advantage for students to comprehend the paper pattern, and thus anticipate and prepare for examinations accordingly.

The coherent manner in which the books in this series present new ideas and concepts to students makes this series play an essential role in the preparation for university examinations. The detailed and comprehensive discussions, easy to understand examples, objective questions and ample exercises, all aid the students to understand everything in an all-inclusive manner.

- **Topic-wise coverage in Question-Answer form.**
- **Clears course fundamentals.**
- **Includes solved University Questions.**
- **The perfect assistance for scoring good marks.**
- **Good for brush up before exams.**
- **Ideal for self-study.**

Quantum Publications®
(A Unit of Quantum Page Pvt. Ltd.)
Plot No. 59/2/7, Site-4, Industrial Area, Sahibabad,
Ghaziabad, 201010, (U.P.) Phone: 0120-4160479
E-mail: pagequantum@gmail.com Web: www.quantumpage.co.in
Find us on: facebook.com/quantumseriesofficial