QUANTUM SERIES

For
B.Tech Students of Fourth 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

By

Kanika Dhama

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

PUBLISHED BY : **Apram Singh**
Quantum Publications®
(A Unit of Quantum Page Pvt. Ltd.)
Plot No. 59/2/7, Site - 4, Industrial Area,
Sahibabad, Ghaziabad-201 010

Phone : 0120 - 4160479
Email : pagequantum@gmail.com **Website: www.quantumpage.co.in**
Delhi Office : 1/6590, East Rohtas Nagar, Shahdara, Delhi-110032

Machine Learning (CS/IT/OE-II : Sem-8)
1st Edition : *2019-20*

Price: Rs. 80/- only

Printed at : Giriraj Printers, Delhi.

CONTENTS

RCS080 / ROE083 : MACHINE LEARNING

1 UNIT

Introduction

CONTENTS

PART-1

Introduction, Well Defined Learning Problems, Designing a Learning System, Issues in Machine Learning

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.1. 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 1.2. Describe different types of machine learning algorithm.

Answer

Different types of machine learning algorithms are :

1. **Supervised machine learning algorithms :**
   a. Supervised learning is defined when the model is getting trained on a labelled dataset.
   b. Labelled dataset have both input and output parameters.
   c. In this type of learning, both training and validation datasets are labelled.
2. **Unsupervised machine learning algorithms :**
   a. Unsupervised machine learning is used when the information is neither classified nor labelled.
   b. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabelled data.

c. The system does not figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabelled data.

3. **Semi-supervised machine learning algorithms :**
   a. Semi-supervised machine learning algorithm fall in between supervised and unsupervised learning, since they use both labelled and unlabelled data for training.
   b. The systems that use this method are able to improve learning accuracy.
   c. Semi-supervised learning is chosen when labelled data requires skilled and relevant resources in order to train / learn from it.
4. **Reinforcement machine learning algorithms :**
   a. Reinforcement machine learning algorithm is a learning method that interacts with environment by producing actions and discovers errors or rewards.
   b. Trial, error search and delayed reward are the most relevant characteristics of reinforcement learning.
   c. This method allows machines and software agents to automatically determine the ideal behaviour within a specific context in order to maximize performance.
   d. Simple reward feedback is required for the agent to learn which action is best.

Que 1.3. 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 1.4. What are the applications of machine learning ?

Answer

Applications of machine learning are :

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 1.5. **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 Flipcart, 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 1.6. **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.
   d. Similar techniques have potential applications in many signal-interpretation problems.
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.
   c. Similar techniques have possible applications in many sensor based control problems.
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.
   c. It now plays at a level competitive with the human world champion.
   d. Similar techniques have applications in many practical problems where large search spaces must be examined efficiently.

Que 1.8. What is learning ? Explain the important components of learning.

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 :

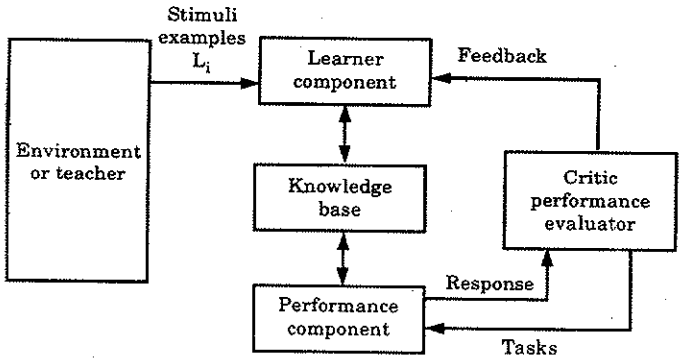


Fig. 1.8.1. General learning model.

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 1.9. Differentiate between artificial intelligence and machine learning.

Answer

S. No.	Artificial Intelligence (AI)	Machine Learning (ML)
1.	AI is human intelligence demonstrated by machines to perform simple to complex tasks.	It provides machines the ability to learn and understand without being explicitly programmed.
2.	The idea behind AI is to program machines to carry out tasks in human ways or smart ways.	The idea behind ML is to teach computers to think and understand like humans.
3.	It is based on characteristics of *human intelligence*.	It is based on the system of probability.
4.	It is used in healthcare, finance, transportation, marketing, media, education, etc.	It is used for optical character recognition, web security, imitation learning, etc.

Que 1.10. 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.

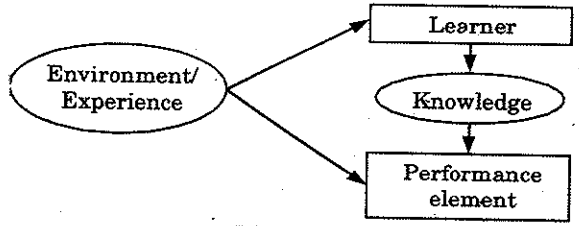


Fig. 1.10.1.

Que 1.11. How we split data in machine learning ?

Answer

Data is splitted in three ways in machine learning :

1. **Training data :**
   a. The part of data we use to train our model.
   b. This is the data which our model actually sees (both input and output) and learn from.
2. **Validation data :**
   a. The part of data which is used to do a frequent evaluation of model, fit on training dataset along with improving involved hyperparameters (initially set parameters before the model begins learning).
   b. This data plays its part when the model is actually training.
3. **Testing data :**
   a. Once our model is completely trained, testing data provides the unbiased evaluation.

b. When we feed in the inputs of testing data, our model will predict some values without seeing actual output.

c. After prediction, we evaluate our model by comparing it with actual output present in the testing data.

d. This is how we evaluate and see how much our model has learned from the experiences feed in as training data, set at the time of training.

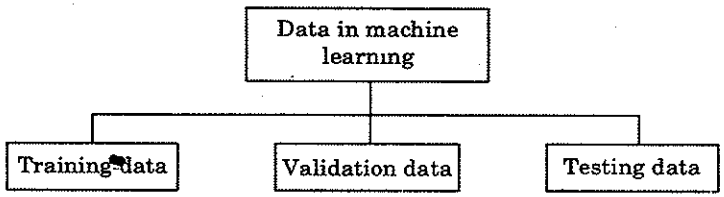


Fig. 1.11.1.

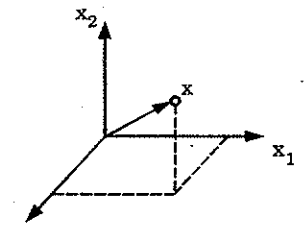
Que 1.12. Describe the terminologies used in machine learning.

Answer

Terminologies used in machine learning are :

1. **Features :** A set of variables that carry discriminating and characterizing information about the objects under consideration.
2. **Feature vector :** A collection of d features, ordered in meaningful way into a d-dimensional column vector that represents the signature of the object to be identified.
3. **Feature space :** Feature space is d-dimensional space in which the feature vectors lie. A d-dimensional vector in a d-dimensional space constitutes a point in that space.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \begin{matrix} \text{feature 1} \\ \text{feature 2} \\ \\ \text{feature 4} \end{matrix}$$



Feature space (3D)

Fig. 1.12.1.

4. **Class :** The category to which a given object belongs, denoted by ω.
5. **Decision boundary :** A boundary in the d-dimensional feature space that separates patterns of different classes from each other.
6. **Classifier :** An algorithm which adjusts its parameters to find the correct decision boundaries through a learning algorithm using a training dataset Such that a cost function is minimized.
7. **Error :** Incorrect labelling of the data by the classifier.
8. **Training performance :** The ability/performance of the classifier in correctly identifying the classes of the training data, which it has already seen. It is not a good indicator of the generalization performance.
9. **Generalization (Test performance) :** Generalization is the ability/performance of the classifier in identifying the classes of previously unseen data.

Que 1.13. Explain the components of machine learning system.

Answer

Components of machine learning system are :

1. **Sensing :**
   a. It uses transducer such as camera or microphone for input.
   b. PR (Pattern Recognition) system depends on the bandwidth, resolution, sensitivity, distortion, etc., of the transducer.
2. **Segmentation :** Patterns should be well separated and should not overlap.
3. **Feature extraction :**
   a. It is used for distinguishing features.
   b. This process extracts invariant features with respect to translation, rotation and scale.
4. **Classification :**
   a. It use a feature vector provide by a feature extractor to assign the object to a category.
   b. It is not always possible to determine the values of all the features.
5. **Post processing :**

a. Post processor uses the output of the classifier to decide on the recommended action.

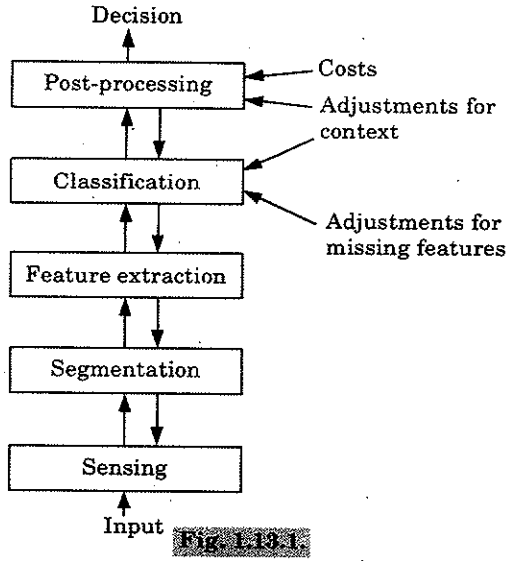


Fig. 1.13.1.

Que 1.14. 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 1.15. 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.

PART-2

The Concept Learning Task, General-to-Specific Ordering of Hypothesis, Find-S, List-Then-Eliminate Algorithm, Candidate-Elimination Algorithm, Inductive Bias

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.16. Write short note on concept learning task.

Answer

1. The task learning, in contrast to learning from observations, can be described by being given a set of training data $\{(\bar{A}_1, C'_1), (\bar{A}_2, C'_2), ..., (\bar{A}_n, C'_n)\}$, where the $\bar{A}_i - [\bar{A}_{i_1}, A_{i_2}, ..., A_{i_1}]^T$ with $A_{i1} \in A$ represent the observable part of the data (here denoted as vector of attributes in the common formalism) and the C'_i represent a valuation of this data.
2. If a functional relationship between the $\bar{A}_i$ and C'_i values is to be discovered, this task is either called regression (in the statistics domain) or supervised learning (in the machine learning domain).
3. The more special case where the C' values are restricted to some finite set C' is called classification or concept learning in computational learning theory.
4. The classical approach to concept learning is concerned with learning concept descriptions for predefined classes C_i of entities from E.
5. A concept is regarded as a function mapping attribute values $\bar{A}_i$ of discrete attributes to a Boolean value indicating concept membership.
6. In this case, the set of entities E is defined by the outer product over the range of the considered attributes in A.
7. Concepts are described as hypotheses, *i.e.*, the conjunction of restrictions on allowed attribute values like allowing just one specific, a set of or any value for an attribute.
8. The task of classical concept learning consists of finding a hypothesis for each class C_i that matches the training data.
9. This task can be performed as a directed search in hypotheses space by exploiting a pre-existing ordering relation, called general to specific ordering of hypotheses.

10. A hypotheses thereby is more general than another if its set of allowed instances is a superset to the set of instances belonging to the other hypothesis.

Que 1.17. Define the term concept and concept learning. How can we represent a concept ?

Answer

Concept : Concept is Boolean-valued function defined over a large set of objects or events.

Concept learning : Concept learning is defined as inferring a Boolean-valued function from training examples of input and output of the function.

Concept learning can be represented using :

1. **Instance x :** Instance x is a collection of attributes (Sky, AirTemp, Humidity, etc.)
2. **Target function c :** Enjoysport : $X \rightarrow \{0, 1\}$
3. **Hypothesis h :** Hypothesis h is a conjunction of constraints on the attributes. A constraint can be :

   a specific value (for example water = warm)

   do not care (for example water = ?)

   no value allowed (for example water = ϕ)
4. **Training example d :** An instance x_i paired with the target function c, $<x_i, c(x_i)>$

   $c(x_i) = 0$ negative example

   $c(x_i) = 1$ positive example

Que 1.18. Explain the working of find-S algorithm with flow chart.

Answer

Working of find-S algorithm :

1. The process starts with initializing 'h' with the most specific hypothesis, generally, it is the first positive example in the data set.
2. We check for each positive example. If the example is negative, we will move on to the next example but if it is a positive example we will consider it for the next step.
3. We will check if each attribute in the example is equal to the hypothesis value.
4. If the value matches, then no changes are made.
5. If the value does not match, the value is changed to '?'.

6. We do this until we reach the last positive example in the data set.

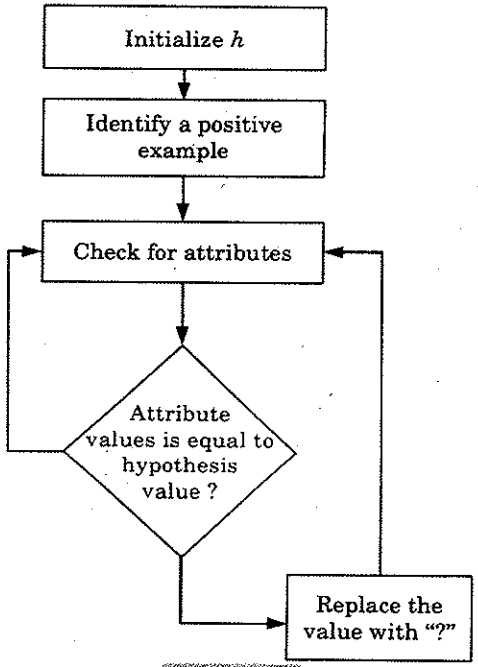


Fig. 1.18.1.

Que 1.19. Write find-S algorithm, with its advantages and disadvantages.

Answer

Find-S algorithm :

1. Initialize h to the most specific hypothesis in H.
2. For each positive training instance x.

   For each attribute constraint a_i in h

   IF the constraint a_i in h is satisfied by x THEN

   do nothing

   ELSE

   replace a_i in h by next more general constraint satisfied by x
3. Output hypothesis h.

Advantages of Find-S algorithm :

1. If the correct target concept is contained in H and the training data are correct, the Find-S algorithm can guarantee that the output is the most specific hypothesis in H that is consistent with the positive examples.

Disadvantages of Find-S algorithm :

1. There is no way to determine if the hypothesis is consistent throughout the data.
2. Inconsistent training sets can mislead the Find-S algorithm, since it ignores the negative examples.
3. Find-S algorithm does not provide a backtracking technique to determine the best possible changes that could be done to improve the resulting hypothesis.
4. The convergence of the learning process is poor, and convergence to the correct objective function cannot be guaranteed.
5. The robustness to noise is weak, and, for a number of special assumptions, the algorithm becomes powerless.

Que 1.20. What is version space ? Explain list-then-eliminate algorithm.

Answer

Version space :

1. A hypothesis h is consistent with a set of training examples D of target concept c if and only if $h(x) = c(x)$ for each training example in D.

   Consistent $(h, D) \equiv (\forall \ < x, c(x) > \in D)\ h(x) = c(x)$
2. The version space, $VS_{H,D}$, with respect to hypothesis space H and training examples D, is the subset of hypotheses from H consistent with all training examples in D.

$$VS_{H,D} \equiv \{h \in H \mid \text{Consistent } (h, D)\}$$

List-then-eliminate algorithm :

1. List-Then-Eliminate algorithm initializes the version space to contain all hypotheses in H, then eliminates hypothesis that are inconsistent, from training example.
2. The version space of candidate hypotheses thus shrinks as more examples are observed, until one hypothesis remain that is consistent with all the observed examples.

   a. Presumably this is the desired target concept.

   b. If insufficient data is available to narrow the version space to a single hypothesis, then the algorithm can output the entire set of

hypothesis consistent with the observed data.

3. List-Then-Eliminate algorithm can be applied whenever the hypothesis space H is finite.
   a. It has many advantages, including the fact that it is guaranteed to output all hypotheses consistent with the training data.
   b. It requires exhaustively enumerating all hypotheses in H - an unrealistic requirement for all but the most trivial hypothesis spaces.

Que 1.21. Explain candidate elimination algorithm with its procedure.

Answer

1. The Candidate-Elimination algorithm computes the version space containing all hypotheses from H that are consistent with an observed sequence of training examples.
2. It begins by initializing the version space to the set of all hypotheses in H, that is, by initializing the G boundary set to contain the most general hypothesis in H

$$G_0 \leftarrow \{<?,?,?,?,?,?>\}$$

and initializing the S boundary set to contain the most specific hypothesis

$$S_0 \leftarrow \{<0, 0, 0, 0, 0, 0>\}$$

3. These two boundary sets delimit the entire hypothesis space, because every other hypothesis in H is both more general than S_0 and more specific than G_0.
4. As each training example is considered, the S and G boundary sets are generalized and specialized, respectively, to eliminate from the version space any hypotheses found inconsistent with the new training example.
5. After all examples have been processed, the computed version space contains all the hypotheses consistent with these examples and hypotheses.

Algorithm :

1. Initialize G to the set of maximally general hypotheses in H.
2. Initialize S to the set of maximally specific hypotheses in H.
3. For each training example d, do
   a. If d is a positive example

      Remove from G any hypothesis that does not included

      For each hypothesis s in S that does not include d

      Remove s from S

   b. Add to S all minimal generalizations h of s such that

      h includes d, and

   c. Some member of G is more general than h

      Remove from S any hypothesis that is more general than another hypothesis in S.
4. For each training example d, do

   If d is a negative example

   Remove from S any hypothesis that does included

   For each hypothesis g in G that does include d

   Remove g from G
5. Add to G all minimal generalizations h of g such that
   a. h does not include d and
   b. Some member of S is more specific than h
6. Remove from G any hypothesis that is less general than another hypothesis in G.
7. If G or S ever becomes empty, data not consistent (with H).

Que 1.22. 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 :

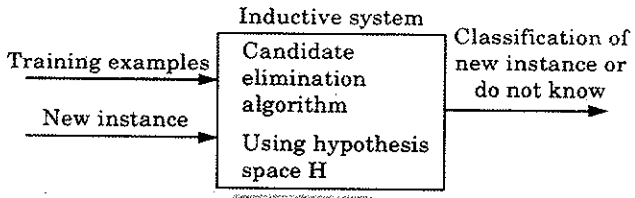


Fig. 1.22.1.

Que 1.23. 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 1.24. What are the learning algorithm 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 1.25. Differentiate between supervised and unsupervised learning.

Answer

S. No.	Supervised learning	Unsupervised 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.	Unsupervised learning is also known as self-organization, in which an output unit is trained to respond to clusters of pattern within the input.
2.	Supervised training requires the pairing of each input vector with a target vector representing the desired output.	Unsupervised training is employed in self-organizing neural networks.
3.	During the training session, an input vector is applied to the network, and it results in an output vector. This response is then compared with the target response.	During training, the neural network receives input patterns and organizes these patterns into categories. When new input pattern is applied, the neural network provides an output response indicating the class to which the input patterns belong.
4.	If the actual response differs from the target response, the network will generate an error signal.	If a class cannot be found for the input pattern, a new class is generated.
5.	The error minimization in this kind of training requires a supervisor or teacher. These input-output pairs can be provided by an external teacher, or by the system which contains neural network.	Unsupervised training does not require a teacher, it requires certain guidelines to form groups. Grouping can be done based on colour, shape, and any other property of the object.
6.	Supervised training methods are used to perform non-linear mapping in pattern classification networks, pattern association networks and multi-layer neural networks.	Unsupervised learning is useful for data compression and clustering.
7.	Supervised learning generates a global model and a local model.	In this, a system is supposed to discover statistically salient features of the input population.

☺☺☺

2 UNIT

Decision Tree Learning

CONTENTS

PART-1

Decision Tree Learning : Decision Tree Learning Algorithm, Inductive Bias, Issues in Decision Tree Learning.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 2.1. 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 2.1.1, illustrate the basic flow of decision tree for decision making with labels (Rain(Yes), Rain(No)).

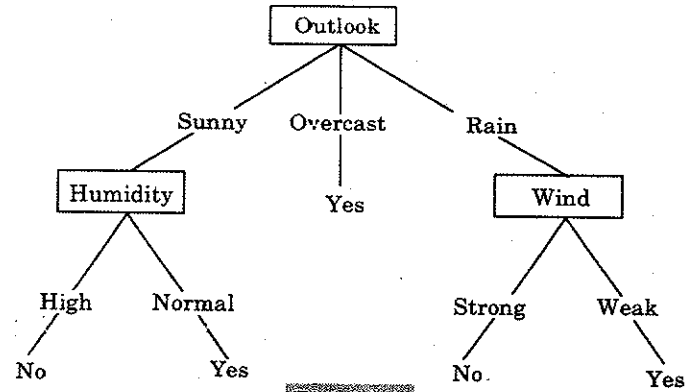


Fig. 2.1.1.

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 2.2. 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 2.3. Write short note on Gini impurity and Gini impurity index.

Answer

1. Gini impurity is the probability of incorrectly classifying a randomly chosen element in the dataset if it were randomly labeled according to the class distribution in the dataset.
2. The Gini impurity index measures the impurity of an input feature with respect to the classes.
3. Gini impurity index reaches its minimum (zero) when all attributes in the node fall into a single information class.
4. The Gini index associated with attribute $X = [x_1, x_2, ..., x_r]$ for node t is denoted by $I_G(t_{X(x_i)})$ and is expressed as

$$I_G(t_{X(x_i)}) = 1 - \sum_{j=1}^{m} f(t_{X(x_i)}, j)^2$$

where $f(t_{X(x_i)}, j)$ is the proportion of samples with the value x_i belonging to class j at node t as defined in equation.

5. The decision tree splitting criterion is based on choosing the attribute with the lowest Gini impurity index of the split.
6. Let a node t be split into r children, n_i be the number of records at child i and N_t be the total number of samples at node t.
7. The Gini impurity index of the split at node t for attribute X is then computed by

$$\text{Gini}(t, X) = \left(\frac{n_1}{N_t}\right) I_G(t_{X(x_1)}) + \left(\frac{n_2}{N_t}\right) I_G(t_{X(x_2)}) \ldots + \left(\frac{n_r}{N_t}\right) I_G(t_{X(x_r)})$$

Que 2.4. 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 2.5. How to avoid overfitting in decision tree model ?

Answer

1. Overfitting is the phenomenon in which the learning system tightly fits the given training data so that it would be inaccurate in predicting the outcomes of the untrained data.
2. In decision trees, overfitting occurs when the tree is designed to perfectly fit all samples in the training data set.

3. To avoid decision tree from overfitting, we remove the branches that make use of features having low importance. This method is called as pruning or post-pruning.
4. It reduces the complexity of tree, and hence improves predictive accuracy by the reduction of overfitting.
5. Pruning should reduce the size of a learning tree without reducing predictive accuracy as measured by a cross-validation set.
6. There are two major pruning techniques:
   a. **Minimum error :** The tree is pruned back to the point where the cross-validated error is minimum.
   b. **Smallest tree :** The tree is pruned back slightly further than the minimum error. Pruning creates a decision tree with cross-validation error within 1 standard error of the minimum error. The smaller tree is more intelligible at the cost of a small increase in error.
7. Another method to prevent over-fitting is to try and stop the tree-building process early, before it produces leaves with very small samples. This heuristic is known as early stopping or pre-pruning decision trees.
8. At each stage of splitting the tree, we check the cross-validation error. If the error does not decrease significantly enough then we stop.
9. Early stopping is a quick fix heuristic. If early stopping is used together with pruning, it can save time.

Que 2.6. 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. 2.6.1.
3. This process is then repeated for the subtree rooted at the new node.
4. The decision tree in Fig. 2.6.1 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 ∧ Humidity = Normal) ∨ (Outlook = Overcast) ∨ (Outlook = Rain ∧ Wind = Weak)

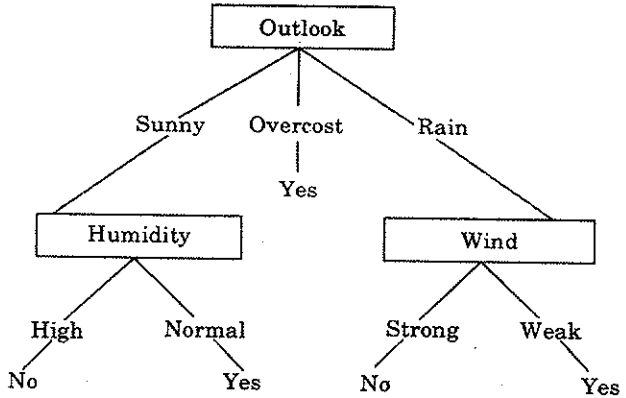


Fig. 2.6.1.

Que 2.7. 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 attribute :**
   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.

Que 2.8. 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.

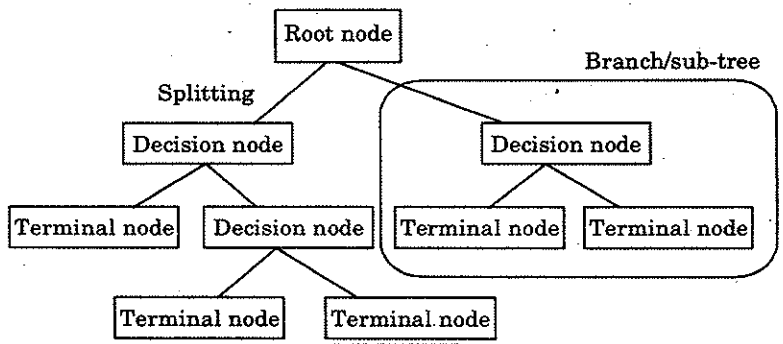


Fig. 2.8.1.

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 2.9. 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 2.10. 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 2.11. 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.

Que 2.12. 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.

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

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 :

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

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.

4. **Gini index :** Refer Q. 2.3, Page 2–3G, Unit-2.

Que 2.13. 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 2.14. 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 2.15. Explain inductive bias with example.

Answer

Refer Q. 1.22, Page 1–20G, Unit-1.

PART-2

Artificial Neural Network : Perceptrons, Gradients Descent and the Delta Rule.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 2.16. 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.
11. 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 2.17. 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 2.18. 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 2.19. 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.

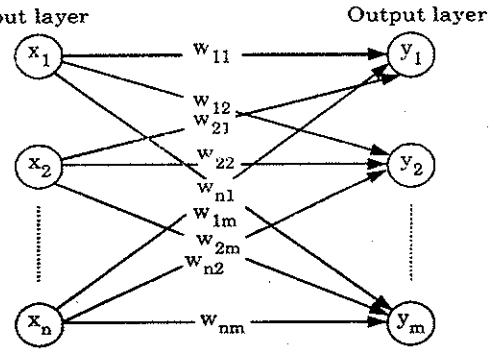
   b. Signature verification technique is a non-vision based technique.
   c. For this application, the first approach is to extract the feature or rather the geometrical feature set representing the signature.
   d. With these feature sets, we have to train the neural networks using an efficient neural 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 2.20. Explain different types of neuron connection with 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.

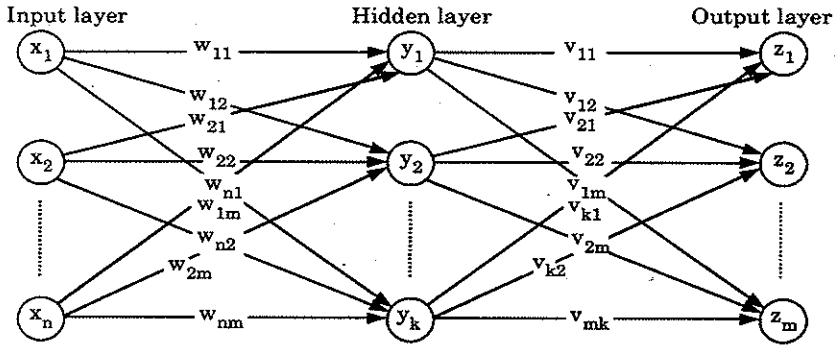


2. Multilayer feed forward network :

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. 2.20.1 shows a single recurrent network having single neuron with feedback to itself.

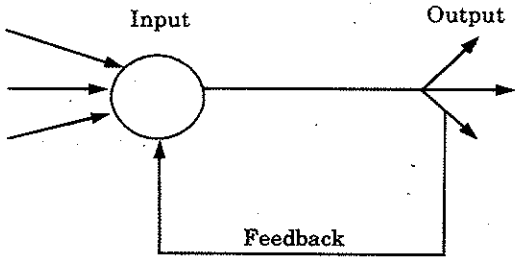


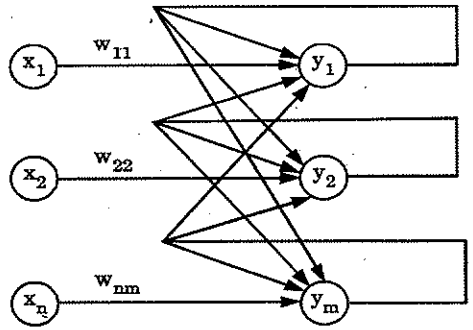
Fig. 2.20.1

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.

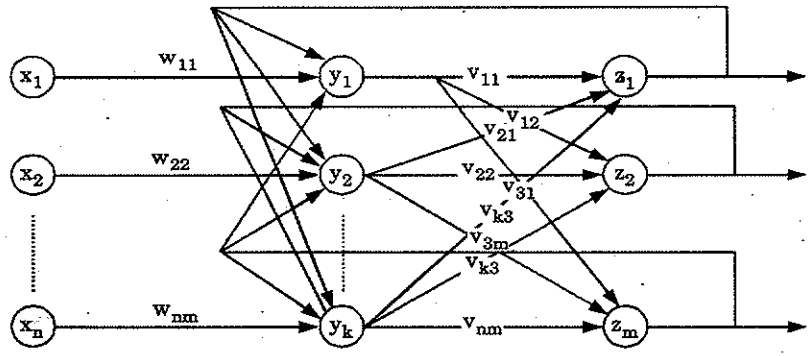


5. Multilayer recurrent network :

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 2.21. 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 2.22. Write short note on gradient descent.

Answer

1. Gradient descent is an optimization technique in machine learning and deep learning and it can be used with all the learning algorithms.
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 is 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 2.23. 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 2.24. 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 2.25. 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 2.26. 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 :

$$O_l = \frac{1}{1 + \exp(- W_l^T x)}$$

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-3

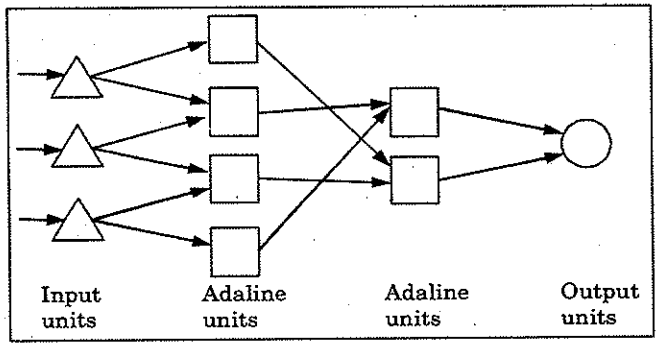
Adaline, Multilayer Network, Derivation of Backpropagation Rule, Backpropagation Algorithm Convergence, Generalization

Long Answer Type and Medium Answer Type Questions

Que 2.27. Explain adaline network with its architecture.

Answer

1. ADALINE is an Adaptive Linear Neuron network with a single linear unit. The Adaline network is trained using the delta rule.
2. It receives input from several units and bias unit.
3. An Adaline model consists of trainable weights. The inputs are of two values (+ 1 or – 1) and the weights have signs (positive or negative).
4. Initially random weights are assigned. The net input calculated is applied to a quantizer transfer function (activation function) that restores the output to + 1 or – 1.
5. The Adaline model compares the actual output with the target output and with the bias units and then adjusts all the weights.



Que 2.28. Explain training and testing algorithm used in adaline network.

Answer

Adaline network training algorithm is as follows :

Step 0 : Weights and bias are to be set to some random values but not zero. Set the learning rate parameter α.

Step 1 : Perform steps 2-6 when stopping condition is false.

Step 2 : Perform steps 3-5 for each bipolar training pair.

Step 3 : Set activations for input units $i = 1$ to n.

Step 4 : Calculate the net input to the output unit.

Step 5 : Update the weight and bias for $i = 1$ to n.

Step 6 : If the highest weight change that occurred during training is smaller than a specified tolerance then stop the training process, else continue. This is the test for the stopping condition of a network.

Adaline networks testing algorithm is as follows :

When the training has been completed, the Adaline can be used to classify input patterns. A step function is used to test the performance of the network.

Step 0 : Initialize the weights. (The weights are obtained from the training algorithm.)

Step 1 : Perform steps 2-4 for each bipolar input vector x.

Step 2 : Set the activations of the input units to x.

Step 3 : Calculate the net input to the output units.

Step 4 : Apply the activation function over the net input calculated.

Que 2.29. 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 backpropagation 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 2.30. 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.

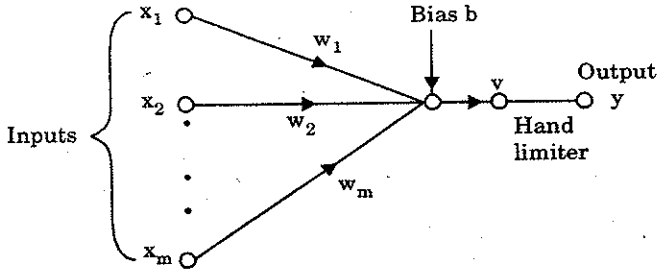


Fig. 2.30.1. Signal flow graph of the perceptron.

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. 2.30.2, 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 :

$$w_1x_1 + w_2x_2 + b = 0$$

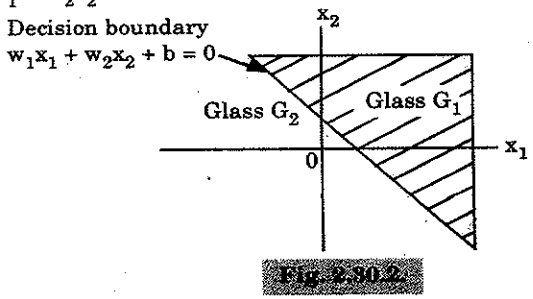


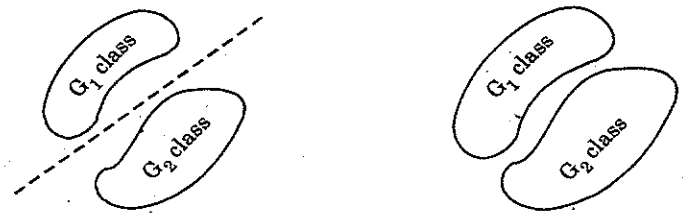
Fig. 2.30.2.

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. 2.30.3.

Que 2.31. 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 to fast learning. For a constant α, the learning algorithm is termed as fixed increment algorithm.

Que 2.32. 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 :

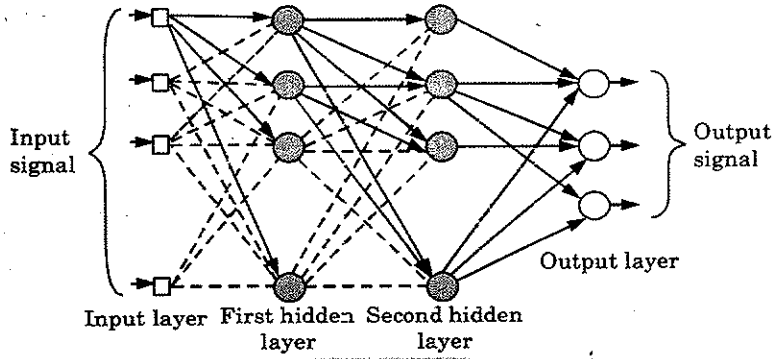


Fig. 2.32.1.

1. Fig. 2.32.1 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.
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 2.33. 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. An formula to select learning coefficient has been :

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

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

$$O = \frac{1}{(1+e^{-\lambda(1+\theta)})} \qquad ...(2.33.1)$$

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. θ in eq. (2.33.1) 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 2.34. 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

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

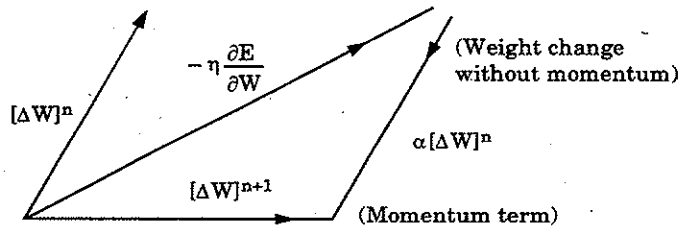


Fig. 2.34.1. Influence of momentum term on weight change

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.

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.

3 UNIT

Evaluating Hypotheses

CONTENTS

PART-1

Evaluating Hypotheses, Estimating Hypotheses Accuracy

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.1. Explain hypotheses with its characteristics and importance.

Answer

Hypothesis (h) :

1. A hypothesis is a function that describes the target in supervised machine learning.
2. The hypothesis depends upon the data and also depends upon the restrictions and bias that we have imposed on the data.
3. A hypothesis is a tentative relationship between two or more variables which direct the research activity.
4. A hypothesis is a testable prediction which is expected to occur. It can be a false or a true statement that is tested in the research to check its authenticity.

Characteristics of hypothesis :

1. Empirically testable
2. Simple and clear
3. Specific and relevant
4. Predictable
5. Manageable

Importance of hypothesis :

1. It gives a direction to the research.
2. It specifies the focus of the researcher.
3. It helps in devising research techniques.
4. It prevents from blind research.
5. It ensures accuracy and precision.
6. It saves resources *i.e.*, time, money and energy.

Que 3.2. Explain different types of hypotheses.

Answer

Different types of hypotheses are :

1. **Simple hypothesis :** A simple hypothesis is a hypothesis that reflects a relationship between two variables *i.e.*, independent and dependent variable.

   Examples :

   i. Higher the unemployment, higher would be the rate of crime in society.

   ii. Lower the use of fertilizers, lower would be agricultural productivity.

   iii. Higher the poverty in a society, higher would be the rate of crimes.

2. **Complex hypothesis :** A complex hypothesis is a hypothesis that reflects relationship among more than two variables.

   Examples :

   i. Higher the poverty, higher the illiteracy in a society, higher will be the rate of crime (three variables - two independent variables and one dependent variable).

   ii. Lower the use of fertilizer, improved seeds and modern equipments, lower would be the agricultural productivity (Four variables - three independent variables and one dependent variable).

   iii. Higher the illiteracy in a society, higher will be poverty and crime rate, (three variables - one independent variable and two dependent variables).

3. **Working hypothesis :**

   i. A hypothesis, that is accepted to put to test and work in a research, is called a working hypothesis.

   ii. It is a hypothesis that is assumed to be suitable to explain certain facts and relationship of phenomena.

   iii. It is supposed that this hypothesis would generate a productive theory and is accepted to put to test for investigation.

   iv. It can be any hypothesis that is processed for work during the research.

4. **Alternative hypothesis :**

   i. If the working hypothesis is proved wrong or rejected, another hypothesis (to replace the working hypothesis) is formulated to be tested to generate the desired results - this is known as an alternate hypothesis.

   ii. It is an alternate assumption (a relationship or an explanation) which is adopted after the working hypothesis fails to generate required theory. Alternative hypothesis is denoted by H.

5. **Null hypothesis :** A null hypothesis is a hypothesis that has no relationship between variables. It negates association between variables.

   Examples :

   i. Poverty has nothing to do with the rate of crime in a society.

   ii. Illiteracy has nothing to do with the rate of unemployment in a society.

   iii. A null hypothesis is made with an intention where the researcher wants to disapprove, reject or nullify the null hypothesis to confirm a relationship between the variables.

   iv. A null hypothesis is made for a reverse strategy - to prove it wrong in order to confirm that there is a relationship between the variables. A null hypothesis is denoted by HQ.

6. **Statistical hypothesis :**

   i. A hypothesis that can be verified statistically is known as a statistical hypothesis.

   ii. It means using quantitative techniques, to generate statistical data, can easily verify it.

   iii. The variables in a statistical hypothesis can be transformed into quantifiable sub-variable to test it statistically.

7. **Logical hypothesis :**

   i. A hypothesis that can be verified logically is known as a logical hypothesis.

   ii. It is a hypothesis expressing a relationship whose interlinks can be joined on the basis of logical explanation.

   iii. It is verified by logical evidence.

   iv. Being verified logically does not necessarily mean that it cannot be verified statistically.

Que 3.3. **Describe the difficulties faced in estimating the accuracy of hypotheses.**

Answer

Difficulties faced in estimating the accuracy of hypotheses :

1. **Bias in the estimate :**

   a. The observed accuracy of the learned hypothesis over the training examples is a poor estimator of its accuracy over future examples.

   b. Because the learned hypothesis was derived from these examples, will provide an optimistically biased estimate of hypothesis accuracy over future examples.

c. To obtain an unbiased estimate of future accuracy, we test the hypothesis on some set of test examples chosen independently of the training examples and the hypothesis.

2. Variance in the estimate :

a. Even if the hypothesis accuracy is measured over an unbiased set of test examples, independent of the training examples, the measured accuracy can still vary from the true accuracy, depending on the makeup of the particular set of test examples.

b. The smaller the set of test examples, the greater the expected variance.

PART-2

Basics of Sampling Theory, Comparing Learning Algorithm.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.4. What are the steps used in the process of sampling ?

Answer

Steps used in the process of sampling are :

1. Identifying the population set.
2. Determination of the size of our sample set.
3. Providing a medium for the basis of selection of samples from the Population medium.
4. Picking out samples from the medium using sampling techniques like simple random, systematic or stratified sampling.
5. Checking whether the formed sample set contains elements (actually matches the different attributes of population set) without large variations in between.
6. Checking for errors or inaccurate estimations in the formed sample set that may or may not have occurred.
7. The set which we get after performing the above steps contributes to the sample set.

Que 3.5. Write short note on sampling frame.

Answer

1. Sampling Frame is the basis of the sample medium.
2. It is a collection of all the sample elements taken into observation.
3. It might even happen that all elements in the sampling frame, did not even take part in the actual statistics.
4. In that case, the elements that took part in the study are called samples and potential elements that could have been in the study but did not take part forms the sampling frame.
5. Thus, sampling frame is the potential list of elements on which we will perform our statistics.
6. Sampling frame is important because it will help in predicting the reaction of the statistics result with the population set.
7. A sampling frame is not just a random set of handpicked elements rather it even consists of identifiers which help to identify each and every element in the set.

Que 3.6. What are different methods of sampling ?

Answer

Different methods of sampling are :

1. Simple Random Sampling (SRS) :

a. Simple random sampling is the elementary form of sampling.

b. In this method, all the elements in populations are first divided into random sets of equal sizes.

c. Random sets have no defining property among themselves, *i.e.*, one set cannot be identified from another set based on some specific identifiers.

d. Thus every element has an equal property of being selected, *i.e.*,

$$P(\text{of getting selected}) = 1/2$$

e. The basic methods for employing SRS are :

i. Choose the population set.

ii. Identify the basis of sampling.

iii. Use of random number/session generators to pick an element from each set.

2. Systematic sampling :

a. Systematic sampling is also known as a type of probability sampling.

b. It is more accurate than SRS and also the standard error formation percentage is very low but not error-free.

c. In this method, first, the population tray elements are arranged based on a specific order or scheme known as being sorted.

d. It can be of any order, which totally depends upon the person performing the statistics.

e. The elements are arranged either in ascending, descending, lexicographically or any other known methods deemed fit by the tester.

f. After being arranged the sample elements are picked on the basis of pre-defined interval set or function.

P(of getting selected) = depends upon the ordered population tray after it has been sorted

g. The basic methods of employing systematic random sampling are :

i. Choosing the population set wisely.

ii. Checking whether systematic sampling will be the efficient method or not.

iii. If yes, then application of sorting method to get an ordered pair of population elements.

iv. Choosing a periodicity to crawl out elements.

3. Stratified sampling :

a. Stratified sampling is a hybrid method concerning both simple random sampling as well as systematic sampling.

b. It is one of the most advanced types of sampling method available, providing accurate result to the tester.

c. In this method, the population tray is divided into sub-segments also known as stratum (singular).

d. Each stratum can have their own unique property. After being divided into different sub-stratum, SRS or systematic sampling can be used to create and pick out samples for performing statistics.

e. The elementary methods for stratified sampling are :

i. Choosing the population tray.

ii. Checking for periodicity or any other features, so that they can be divided into different strata.

iii. Dividing the population tray into sub-sets and sub-groups on the basis of selective property.

iv. Using SRS or systematic sampling of each individual strata to form the sample frame.

v. We can even apply different sampling methods to different sub-sets.

Que 3.7. What are the advantages and disadvantages of different methods of sampling ?

Answer

Advantages of Simple Random Sampling (SRS) :

1. Less exhaustive with respect to time as it is the most elementary form of sampling.
2. Very useful for population set with very less number of elements.
3. SRS can be employed anywhere, anytime even without the use of special random generators.

Disadvantages of Simple Random Sampling (SRS) :

1. Not efficient for large population sets.
2. There are chances of bias and then SRS would not be able to provide a correct result.
3. It does not provide a specific identifier to separate statistically similar samples.

Advantages of systematic sampling :

1. Accuracy is higher than SRS.
2. Standard probability of error is lesser.
3. No problem for bias to creep in during creation of sample frame.

Disadvantages of systematic sampling :

1. Not much efficient when comes to the time wise.
2. Periodicity in population tray elements can lead to absurd results.
3. Systematic sampling can either provide the most accurate result or an impossible one.

Advantages of stratified sampling :

1. It provides results with high accuracy measurements.
2. Different results can be desired just by changing the sampling method.
3. This method also compares different strata when samples are being drawn.

Disadvantages of stratified sampling :

1. Inefficient and expensive when comes to resources as well as money.
2. This method will fail only where homogeneity in elements is present.

Que 3.8. Differentiate between supervised and unsupervised learning.

Answer

Refer Q. 1.25, Page 1–22G, Unit-1.

Que 3.9. **Differentiate between unsupervised and reinforcement learning.**

Answer

Basis	Unsupervised learning	Reinforcement learning
Definition	No external teacher or pre-trained data.	Works on interacting with the environment.
Preference	Assets are depreciable	Liabilities are non-depreciable.
Tasks	Clustering and association.	Exploitation and exploration.
Mapping between input and output	To find the underlying patterns rather than the mapping.	Will get constant feedback from the user by suggesting few news articles and then build a knowledge graph.
Platform	Operated with interactive software or applications.	Supports and works better in AI where human interaction is prevalent.
Algorithms	Many algorithms exist in using this learning.	Neither supervised nor unsupervised algorithms are used.
Integration	Runs on any platform or with any applications.	Runs with any hardware or software devices.

PART-3

Bayesian Learning : Bayes Theorem, Concept Learning, Bayes Optimal Classifier.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.10. **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.
5. For example, there are a number of balls which are red and blue in colour then $\omega = \omega_1$ when the ball is red and $\omega = \omega_2$ when the ball is blue. Because the state of nature is so unpredictable, we consider ω to be a variable that must be described probabilistically.
6. If one ball is red then we can say that the next ball is equally likely to be red or blue.
7. We assume that there is a prior probability $p(\omega_1)$ that the next ball is blue.
8. These prior probabilities reflect the prior knowledge of how likely a ball obtained is red or blue before the ball actually appears.
9. Now after defining the state of nature and prior probabilities, the decision has to be made that a particular ball is present in which class.
10. A decision rule is used to take decision as :

    Decide ω_1 if $p(\omega_1) > p(\omega_2)$, otherwise ω_2 .

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,

$$p(\omega_i \mid x) = \frac{p(x \mid \omega_i)p(\omega_i)}{p(x)} \quad ...(3.10.1)$$

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

$$p(x) = \sum_{i=1}^{2} p(x \mid \omega_i)p(\omega_i) \quad ...(3.10.2)$$

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 ...(3.10.3)
9. In the case of equality the pattern can be assigned to either of the two classes. Using equation (3.10.1), 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. $p(x \mid \omega_1)p(\omega_1) < p(x \mid \omega_2)p(\omega_2)$ ...(3.10.4)
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. (3.10.4) 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. 3.10.1, 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 ($l = 1$).

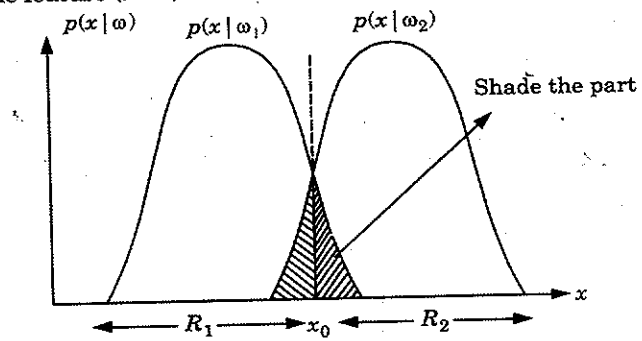


Fig. 3.10.1. Bayesian classifier for the case of two equiprobable classes.

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. 3.10.1, 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.
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. 3.10.1.

Que 3.11. 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. 3.10.1, 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.*,

$$P_e = p(x \in R_2, \omega_1) + p(x \in R_1, \omega_2) \quad ...(3.11.1)$$

6. P_e can be written as,

$$P_e = p(x \in R_2 \mid \omega_1)p(\omega_1) + p(x \in R_1 \mid \omega_2)p(\omega_2)$$
$$= P(\omega_1)\int_{R_2} p(x \mid \omega_1)dx + p(\omega_2)\int_{R_1} p(x \mid \omega_2)dx \quad ...(3.11.2)$$

7. Using the Baye's rule,

$$= P\int_{R_2} p(\omega_1 \mid x)p(x)dx + \int_{R_1} p(\omega_2 \mid x)p(x)dx \quad ...(3.11.3)$$

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

$$R_1 : p(\omega_1 \mid x) > p(\omega_2 \mid x)$$
$$R_2 : p(\omega_2 \mid x) > p(\omega_1 \mid x) \quad ...(3.11.4)$$

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

$$\int_{R_1} p(\omega_1 \mid x)p(x)dx + \int_{R_2} p(\omega_1 \mid x)p(x)dx = 1 \quad ...(3.11.5)$$

10. Combining equation (3.11.3) and (3.11.5), we get,

$$P_e = p(\omega_1) \int_{R_1} (p(\omega_1|x) - p(\omega_2|x))\, p(x)dx \quad \ldots (3.11.6)$$

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 3.12. **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. 3.12.1.

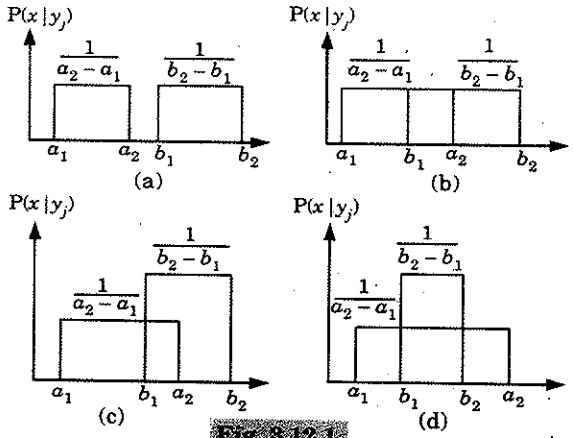


Fig. 3.12.1.

Que 3.13. **What is Baye's theorem ? Explain.**

Answer

1. Let x be a thing in a pattern. In Bayesian term, x is considered "evidence". Let H be some hypothesis such as that x belongs to a specified class C.
2. For classification problem, $p(H|x)$ is determined which is the probability that the hypothesis holds given the observed x.
3. In other words, the probability that x belongs to class C is determined, given that description of x is known.
4. $p(H|x)$ is the posterior probability of H conditioned on x.
5. For example, suppose there are a number of customers described by the attributes age and income, respectively and that x is a 35 years old customer with an income of Rs. 40,000.
6. Suppose that H is the hypothesis that our customer will buy a computer. Then $p(H|x)$ rejects the probability that customer x will buy a computer given that the customer's age and income is known.
7. Similarly $p(x|H)$ is the posterior probability of x conditioned on H.
8. It is the probability that customer x is 35 years old and earns Rs. 40,000 given that we know that the customer will buy computer. $p(x)$ is the prior probability of x.
9. It is the probability that a person from the set of customers is 35 years old and earns Rs. 40,000.
10. Baye's theorem is useful in that it provides a way of calculating the posterior probability $p(H|x)$, from $p(H)$, $p(x|H)$ and $p(x)$

    Baye's theorem is, $p(H|x) = \dfrac{p(x|H)p(H)}{p(x)}$

Que 3.14. **Write short note on Bayes classifier.**

OR

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. In simple terms, 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. 3.14.1(a) depicts a block diagram representation of the Bayes classifier. The important points in this block diagram are twofold :
   a. The data processing in designing the Bayes classifier is confined entirely to the computation of the likelihood ratio $\wedge(x)$.

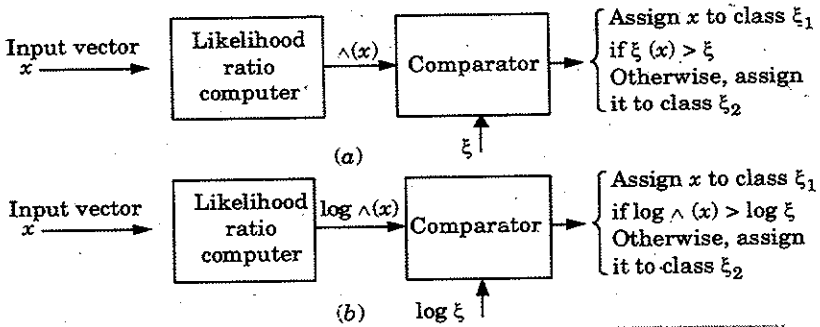


Fig. 3.14.1. Two equivalent implementations of the Bayes classifier : (a) Likelihood ratio test, (b) Log-likelihood ratio test

   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.

Que 3.15. Discuss Bayes classifier using some example in detail.

Answer

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. a. Given data sets with many attributes, the computation of $p(X|C_i)$ will be extremely expensive.
   b. To reduce computation in evaluating $p(X|C_i)$, the assumption of class conditional independence is made.
   c. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the feature.

      Thus, $$p(X|C_i) = \prod_{k=1}^{n} p(x_k|C_i)$$
      $$= p(x_1|C_2) \times p(x_2|C_2)x.... \times p(x_n|C_i)$$

   d. 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.

e. For example, to compute $p(X \mid C_i)$ we consider,

i. If A_k is categorical then $p(x_k \mid C_i)$ 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.

ii. 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 $p(x_k \mid C_i) = g(x_k)$.

f. 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 $p(x_k \mid C_i)$.

g. 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.

h. 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.

i. 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, $p(X \mid C_i)\, p(C_i)$ 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

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

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

Que 3.16. **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.

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

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

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

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

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

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

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.

PART-4

Naive Bayes Classifier, Bayesian Belief Network, EM Algorithm.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.17. 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 conditionally independent of each other, given the class.
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 widely 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.

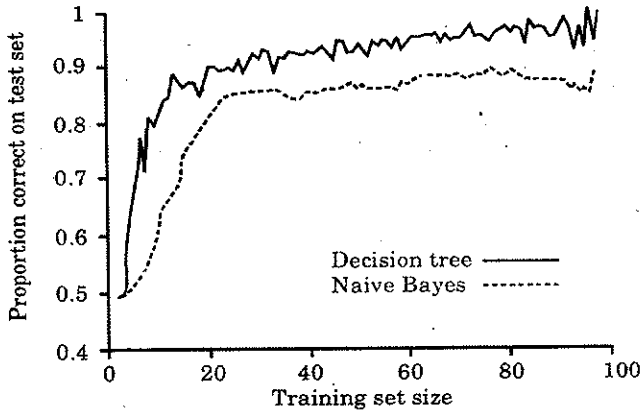


Fig. 3.17.1. The learning curve for Naive Bayes learning.

Que 3.18. **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 3.19. **Describe Bayesian networks. How are the Bayesian networks powerful representation for uncertainty knowledge ?**

Answer

A Bayesian network is a directed acyclic graph in which each node is annotated with quantitative probability information.

The full specification is as follows :

1. A set of random variables makes up the nodes of the network variables may be discrete or continuous.
2. 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.
3. 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.
4. The graph has no directed cycles (and hence is a Directed Acyclic Graph or DAG).

For example :

In Fig. 3.19.1 the weather is independent of the other three variables and Toothache and catch are conditionally independent, given cavity.

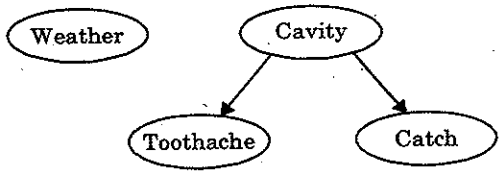


Fig. 3.19.1. A simple bayesian network.

a. 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.

b. Bayesian networks provide a concise way to represent conditional independence relationships in the domain.

c. A Bayesian network specifies a full joint distribution; each joint entry is defined as the product of the corresponding entries in the local conditional distributions.

d. A Bayesian network is often exponentially smaller than the full joint distribution.

e. Hybrid Bayesian networks, includes both discrete and continuous variables, use a variety of canonical distributions.

e. Inference in Bayesian networks means computing the probability distribution of a set of query variables, given a set of evidence variables.

g. Exact inference algorithms, such as variable elimination, evaluate sums of products of conditional probabilities as efficiently as possible.

Bayesian networks are powerful representation for uncertainty knowledge :

1. Two variables A and B are called conditionally independent; if $P(A, B \mid C) = P(A \mid C).\ P(B \mid C)$ or, equivalently, if $P(A \mid B,C) = P(A \mid C)$.
2. Besides the foundational rules of computation for probabilities, the following rules are also true :
   - **a. Bayes Theorem :** $P(A \mid B) = P(B \mid A)\ .\ P(A)\ P(B)$
   - **b. Marginalization :** $P(B) = P(A, B) + P(\neg A, B) = P(B/A)\ :\ P(A) + P(B \mid \neg A)\ .\ P(\neg A)$
   - **c. Conditioning :** $P(A \mid B) = \Sigma_c P(A \mid B, C = c)\ P(C = c \mid B)$
     - i. A variable in Bayesian network is conditionally independent of all non-successor variables. If $X_1, \ldots, X_{n-1}$ are successors of X_n, we have $P(X_n \mid X_1, \ldots, X_{n-1}) = P(X_n \mid \text{Parents}(X_n))$. This condition must be honored during the construction of a network.
     - ii. During construction of a Bayesian network, the variables should be ordered according to causality. First the cause, then the hidden variables, and the diagnosis variables last.
   - **d. Chain rule :** $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} = 1\ P(X_i \mid \text{Parents}(X_i))$

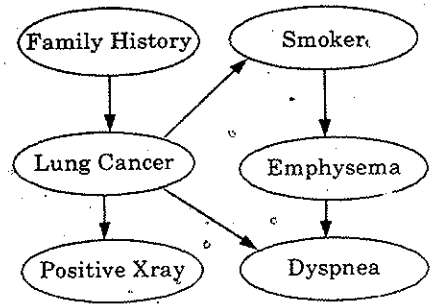
3. Bayesian Network (BN) has been accepted as a powerful tool for common knowledge representation and reasoning of partial beliefs under uncertainty.
4. Bayesian networks utilize knowledge about the independence of variables to simplify the model.
5. One of the most important features of Bayesian networks is the fact that they provide an elegant mathematical structure for modeling complicated relationships among random variables while keeping a relatively simple visualization of these relationships.

Que 3.20. 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 3.21. 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 3.22. 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).

4
UNIT

Computational Learning Theory

CONTENTS

PART-1

Computational Learning Theory, Sample Complexity for Finite Hypothesis Spaces, Sample Complexity for Infinite Hypothesis Space, The Mistake Bound Model of Learning.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.1. Write short note on computational learning theory.

Answer

1. Computational Learning Theory (CLT) is a field of AI used for studying the design of machine learning algorithms to determine what sorts of problems are learnable.
2. The ultimate goals are to understand the theoretical ideas of deep learning programs, what makes them work or not, while improving accuracy and efficiency.
3. This field merges many disciplines, such as probability theory, statistics, programming optimization, information theory, calculus and geometry.
4. Computational learning theory is used to :
   i. Provide a theoretical analysis of learning.
   ii. Show when a learning algorithm can be expected to succeed.
   iii. Show when learning may be impossible.
5. There are three areas comprised by CLT :
   i. Sample complexity : Sample complexity described the examples we need to find in a good hypothesis.
   ii. Computational complexity : Computational complexity defined the computational power we need to find in a good hypothesis.
   iii. Mistake bound : Mistake bound find the mistakes we will make before finding a good hypothesis.

Que 4.2. Describe sample complexity for finite hypothesis space.

Answer

1. The sample complexity of a machine learning algorithm represents the number of training samples that it needs in order to successfully learn a target function.

2. 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.
3. There are two variants of sample complexity :
   a. The weak variant fixes a particular input-output distribution.
   b. The strong variant takes the worst-case sample complexity over all input-output distributions.
3. It characterizes classes of learning problems or specific algorithms in terms of sample complexity, *i.e.*, the number of training examples necessary or sufficient to learn hypotheses of a given accuracy.
4. Complexity of a learning problem depends on :
   a. Size or expressiveness of the hypothesis space.
   b. Accuracy to which target concept must be approximated.
   c. Probability with which the learner must produce a successful hypothesis.
   d. Manner in which training examples are presented, for example, randomly or by query to an oracle.

Que 4.3. Discuss mistake bound model of learning.

Answer

1. An algorithm A learns a class C with mistake bound M *iff*

   $$\text{Mistake}(A, C) \leq M.$$

2. In mistake bound model, learning proceeds in rounds, one by one. Suppose $Y = \{-1, +1\}$.
3. At the beginning of round t, the learning algorithm A has the hypothesis h_t, in round t, we see x_t and predict $h_t(x_t)$.
4. At the end of the round, y_t is revealed and A makes a mistake if $h_t(x_t) \neq y_t$. The algorithm then updates its hypothesis to h_{t+1} and this continues till time T.
5. Suppose the labels were actually produced by some function f in a given concept class C.
6. Then we bound the total number of mistakes the learner commits as :

   $$\text{Mistake}(A, C) := \max_{f \in C, T, x_{1 \ldots T}} \sum_{t=1}^{T} 1[h_t(x_t) \neq f(x_t)]$$

7. Amount of computation A has to do in each round in order to update its hypothesis from h_t to h_{t+1}.
8. Setting this issue aside for a moment, we have a remarkably simple algorithm halving (C) that has a mistake bound of $lg(|C|)$ for any finite concept class C.

9. For a finite set H of hypotheses, define the hypothesis majority (H) as follows :

   $$\text{Majority}(H)(x) := \begin{cases} +1 & |\{h \in H \mid h(x) = +1\}| \geq |H|/2, \\ -1 & \text{otherwise} \end{cases}$$

   Algorithm :
   HALVING (C)
   $C_1 \leftarrow C$
   $h_1 \leftarrow$ majority (C_1)
   for $t = 1$ to T do
   Receive x
   Predict $h_t(x_t)$
   Receive y_t
   $C_{t+1} \leftarrow \{f \in Ct \mid f(x_t) = y_t\}$
   h_{t+1} majority (C_{t+1})
   end for

PART-2

Instance-based Learning, K-Nearest Neighbour Learning, Locally Weighted Regression, Radial Basis Function Network.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.4. 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 4.5. Explain instance-based learning representation.

Answer

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 neighbor 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. When nominal attributes are present, we may use the following procedure.
3. A distance of 0 is assigned if the values are identical, otherwise the distance is 1.
4. 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.
5. 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.

(a) (b) (c)

Fig. 4.5.1.

Que 4.6. 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 describe 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 4.7. 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 4.8. 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.

Que 4.9. 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 4.10. 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 4.11. 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.

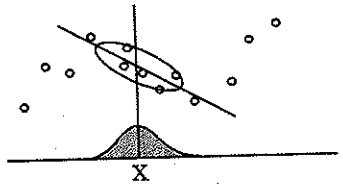


Fig. 4.11.1.

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 :

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

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

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

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

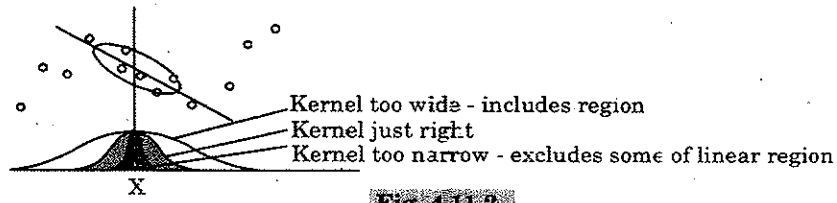


Fig. 4.11.2.

Que 4.12. 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 4.13. 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, $\phi : R^n \rightarrow R$, and is given by

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

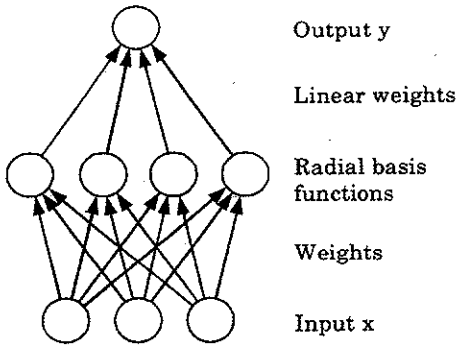


Fig. 4.13.1. Architecture of a radial basis function network. An input vector x is used as input to all radial basis functions, each with different parameters. The output of the network is a linear combination of the outputs from radial basis functions.

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

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

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 optimize the fit between ϕ and the data.

PART-3

Case-Based Learning

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.14. 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 4.15. 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 4.16. 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.

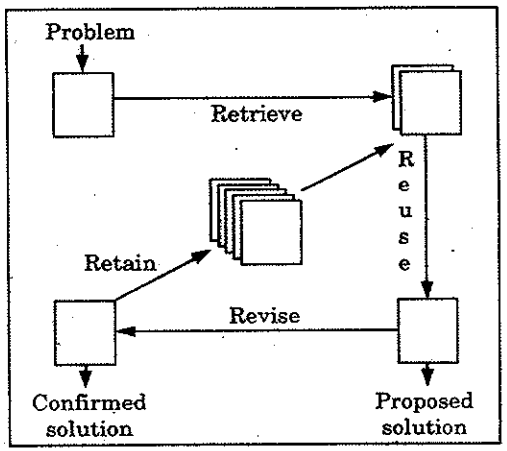


Fig. 4.16.1 The CBL cycle

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 4.17. 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 system 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 4.18. 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 4.19. 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 4.20. 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 4.21 **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).

5 UNIT

Genetic Algorithm

CONTENTS

PART-1

Genetic Algorithm : An Illustrative Example, Hypothesis Space Search, Genetic Programming

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.1. 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 algorithms :
   a. Recurrent neural network
   b. Mutation testing
   c. Code breaking
   d. Filtering and signal processing
   e. Learning fuzzy rule base

Que 5.2. 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 5.3. 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.

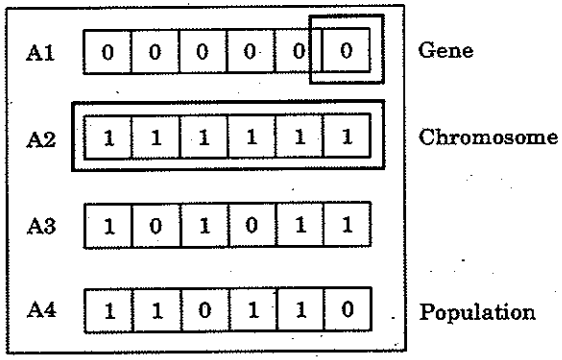
b. Each individual is a solution to the problem we want to solve.

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.

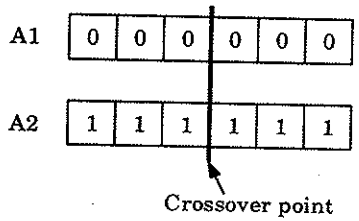
c. Individuals with high fitness have more chance to be selected for reproduction.

4. Crossover :

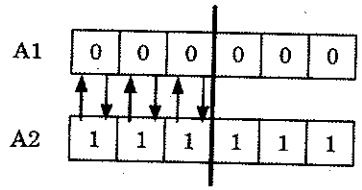
a. Crossover is the most significant phase in a genetic algorithm.

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

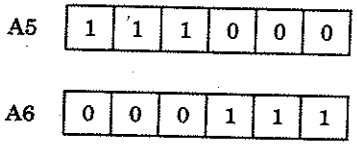
c. For example, consider the crossover point to be 3 as shown :



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

A5	1	1	1	0	0	0

After mutation

A5	1	1	0	1	1	0

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

6. Termination :

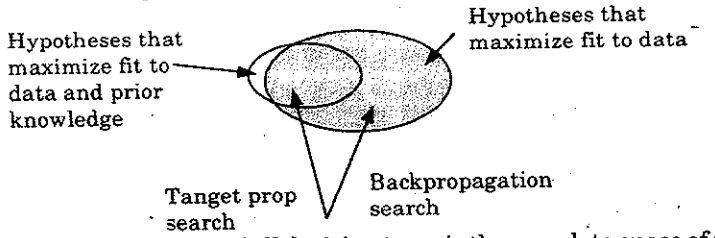
a. The algorithm terminates if the population has converged (does not produce offspring which are significantly different from the previous generation).

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

Que 5.4. Explain briefly hypothesis space search.

Answer

1. The hypothesis space is the set of possible decision trees.
2. ID3 performs a simple to complex, hill-climbing search through hypothesis space for finding locally-optimal solution.
3. ID3 can be characterized as searching a space of hypothesis for one that fits the training examples.



4. ID3 hypothesis space of all decision trees is the complete space of finite discrete-valued functions, relative to the available attributes.
5. It maintains only single current hypothesis *i.e.*, candidate – elimination algorithm.
6. No backtracking in the search is required for converging locally optimal solution.
7. Using training examples at each step resulting search is less sensitive to errors in individual training examples.

Que 5.5. Write short note on Genetic Programming.

Answer

1. Genetic Programming (GP) is a type of Evolutionary Algorithm (EA), *i.e.*, a subset of machine learning.
2. EAs are used to discover solutions to problems that humans do not know how to solve.
3. Free of human preconceptions or biases, the adaptive nature of EAs can generate solutions that are comparable to, and often better than the best human efforts.
4. GP software systems implement an algorithm that uses random mutation, crossover, a fitness function, and multiple generations of evolution to resolve a user-defined task.
5. GP is used to discover a functional relationship between features in data (symbolic regression), to group data into categories (classification), and to assist in the design of electrical circuits, antennae, and quantum algorithms.
6. GP is applied to software engineering through code synthesis, genetic improvement, automatic bug-fixing, and in developing game-playing strategies.

Que 5.6. What are the advantages and disadvantages of Genetic programming ?

Answer

Advantages 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 to their solutions.

Disadvantages of Genetic programming are :

1. In GP, the number of possible programs that can be constructed by the algorithm is immense. This is one of the main reasons why people thought that it would be impossible to find programs that are good solutions to a given problem.
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, it can generate errors and can make the program slow.
3. There is a high probability that even a very small variation has a disastrous effect on fitness of the solution generated.

Que 5.7. Explain different types of Genetic programming.

Answer

Different types of Genetic programming are :

1. **Tree-based Genetic programming :**
   a. In tree-based GP, the computer programs are represented in tree structures that are evaluated recursively to produce the resulting multivariate expressions.
   b. Traditional nomenclature states that a tree node (or just node) is an operator [+, –, *, /] and a terminal node (or leaf) is a variable [a, b, c, d].
2. **Stack-based Genetic programming :**
   a. In stack-based genetic programming, the programs in the evolving population are expressed in a stack-based programming language.
   b. In stack-based genetic programming, programs are composed of instructions that take arguments from data stacks and push results back on data stacks.
   c. A separate stack is provided for each data type, and program code itself can be manipulated on data stacks and subsequently executed.
3. **Linear Genetic Programming (LGP) :**
   a. Linear Genetic Programming (LGP) is a subset of genetic programming where computer programs in a population are represented as a sequence of instructions from imperative programming language or machine language.
4. **Grammatical Evolution (GE) :**
   a. Grammatical Evolution is a new evolutionary computation technique use to find an executable program or program fragment that will achieve a good fitness value for the given objective function.
   b. Grammatical Evolution applies genetic operators to an integer string, subsequently mapped to a program (or similar) through the use of a grammar.
   c. The benefit of GE is that this mapping simplifies the application of search to different programming languages and other structures.
5. **Cartesian Genetic Programming (CGP) :**
   a. CGP is a highly efficient and flexible form of Genetic programming that encodes a graph representation of a computer program.
   b. CGP represents computational structures (mathematical equations, circuits, computer programs etc) as a string of integers.
   c. These integers, known as genes determine the functions of nodes in the graph, the connections between nodes, the connections to inputs and the locations in the graph from where outputs are taken.

6. **Genetic Improvement Programming (GIP) :**
   a. GIP uses replacement software components that maximise achievement of multiple objectives, while retaining the interfaces between the components so-evolved and the surrounding system.
   b. The GISMO project will develop theory, algorithms and techniques for GIP as a way to automatically optimise multiple software engineering objectives such as maximal throughput, fastest response time and most reliable performance, while minimising power consumption, faults, memory use, compiled code size, peak disk usage and disk transfers.

Que 5.8. **Write procedure for cartesian genetic programming.**

Answer

Procedure for cartesian Genetic programming :

1. For all i such that $0 \leq i < 1 + \lambda$ do
2. Randomly generate individual i
3. End for
4. Select the fittest individual, which is promoted as the parent
5. While solution is not found or the generation limit is not reached do
6. For all i such that $0 \leq i < 1 + \lambda$ do
7. Mutate the parent to generate offspring i
8. End for
9. Generate the new parent using the following rules :
   i. If A single offspring has a better fitness than any other member of the population then
   ii. The offspring is chosen as parent
10. Else if one or more offspring have an equal fitness to the parent then
11. Randomly choose one of these as parent
12. Else
13. The parent chromosome remains the same as before
14. End if
15. End while

Que 5.9. **Explain genetic algorithm with steps.**

Answer

Genetic algorithm : Refer Q. 5.1, Page 5–2G, Unit-5.

Algorithm :

GA(Fitness, Fitness_threshold, p, r, m)

Fitness : Fitness function, Fitness_threshold : termination criterion,

p : Number of hypotheses in the population,

r : Fraction to be replaced by crossover,

m : Mutation rate.

1. Initialize population : $P \leftarrow$ Generate p hypotheses at random.
2. Evaluate : For each h in P, compute Fitness(h).
3. While [$\max_h$ Fitness(h)] < Fitness_threshold, Do :
   - i. **Select :** Probabilistically select $(1-r) \cdot p$ members of P to add to Ps.
   - ii. **Crossover :** Probabilistically select $r \cdot p/2$ pairs of hypotheses from P. For each pair $< h_1, h_2 >$ produce two offspring and add to Ps.
   - iii. **Mutate :** Choose m percent of the members of Ps with uniform probability. For each, invert one randomly selected bit.
   - iv. **Update :** $P \leftarrow Ps$.
   - v. **Evaluate :** For each $h \in P$, compute Fitness(h).
4. Return the hypothesis from P that has the highest fitness.

PART-2

Models of Evolution and Learning.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.10. Explain the adaptive functions of learning in evolution.

Answer

Adaptive functions of learning in evolution :

1. **It allows individuals to adapt to changes in the environment that occur in the life span of an individual or across few generations :**
   - a. Learning has the same function attributed to evaluation, adaptation to the environment.
   - b. Learning evolution enables an organism to adapt to changes in the environment that happen too quickly to be tracked by evolution.

2. **It allows evolution to use information extracted from the environment thereby channeling evolutionary search :**
   - a. Whereas onto genetic adaptation can rely on a very rich, although not always explicit, amount of feedback from the environment, evolutionary adaptation relies on a single value which reflects how well an individual coped with its environment.
   - b. This value is the number of offspring in the case of natural evolution and the fitness value in the case of artificial evolution.
   - c. Instead, from the point of view of onto genetic adaptation, individuals continuously receive feedback information from the environment through their sensors during the whole lifetime.
   - d. However, this amount of information encodes only how well an individual is doing in different moments of its life or how it should modify its own behaviour in order to increase its fitness.
   - e. However, onto genetic and phylogenetic adaptation together might be capable of exploiting this information.
   - f. Indeed evolution is able to transform sensory information into self-generated reinforcement signals or teaching patterns.
3. **It can help and guide evolution :**
   - a. Although physical changes of the phenotype cannot be written back into the genotype, learning might indeed affect the evolutionary course in subtle but in effective ways learning accelerates evolution because sub-optimal individuals can reproduce by acquiring during life necessary features for survival.
   - b. Learning allows to produce complex phenotype with short genotypes by extracting some of the information necessary to build the corresponding phenotypes from the environment.
   - c. Moreover learning can allow the maintenance of more genetic diversity.
   - d. Different genes have more chances to be preserved in the population if the individuals who incorporate those genes are able to learn the same fit behaviours.

Que 5.11. What are the disadvantages of learning in evolution ?

Answer

Disadvantages of learning in evolution :

1. **A delay in the ability to acquire fitness :**
   - a. Learning individuals will have a sub-optimal behaviour during the learning phase.
   - b. As a consequence they will collect less fitness than individuals who have the same behaviour genetically specified.

c. The longer the learning period, the more accumulated costs have to be paid.

2. **Increased unreliability :**
   a. Since learned behaviour is determined, atleast partly, by the environment, if a vital behaviour-defining stimulus is non-encountered by a particular individual, then it will suffer as a consequence.
   b. The plasticity of learned behaviours provides the possibility that an individual may simply learn the wrong thing, causing it to incur an incorrect behaviour cost.
   c. Learning thus has a stochastic element that it is not present in instinctive behaviours.
3. **Other costs :**
   a. In natural organisms or in biologically inspired artificial organisms learning might implies additional costs.
   b. If individuals are considered young during the learning period, learning also implies a delayed reproduction time.
   c. Moreover, learning might implies the waste of energy resources for the accomplishment if learning process itself or for parental investment.
   d. Finally, while learning, individuals without a fully formed behaviour may damage themselves.

Que 5.12. Write short note on learnable evolution model.

Answer

1. The 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).
2. Unlike standard, Darwinian-type evolutionary computation methods that use random or semi-random operators for generating new individuals (such as mutations and/or recombination), LEM employs hypothesis generation and instantiation operators.
3. The hypothesis generation operator applies a machine learning program to induce descriptions that distinguish between high-fitness and low-fitness individuals in each consecutive population.
4. Such descriptions delineate the search space that contain the desirable solutions.
5. Subsequently the instantiation operator samples these areas to create new individuals.

6. LEM has been modified from optimization domain to classification domain by augmented LEM with ID3.
7. Learnable evolution model describes :
   a. Lamarckian evolution :
      i. Generation is directly influenced by the experiences of individual organisms during their lifetime.
      ii. Direct influence of genetic makeup on the offspring.
      iii. It is completely contradicted by science.
      iv. Lamarckian processes improve the effectiveness of genetic algorithms.
   b. Baldwin effect :
      i. A species in a changing environment underlies evolutionary pressure that favours individuals with the ability to learn.
      ii. Such individuals perform a small local search to maximize their fitness.
      iii. Additionally, such individuals rely less on genetic code.
      iv. Thus, they support a more diverse gene pool, relying on individual learning to overcome missing traits.
      v. Indirect influence of evolutionary adaption for the entire population.

PART-3

Learning First Order Rules, Sequential Covering Algorithm, General to Specific Beam Search.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.13. Explain the procedure of learn-one rule algorithm.

Answer

Learn-one-rule (Target_attribute, Attributes, Examples, k)

Returns a single rule that covers some of the examples. Conducts a general-to-specific greedy beam search for the best rule, guided by the performance metric.

1. Initialize best_hypothesis to the most general hypothesis $\varnothing$.
2. Initialize candidate_hypotheses to the set (Best_hypothesis).

3. While candidate_hypotheses is not empty. Do :
   a. Generate the next more specific candidate_hypotheses :
      i. All_constraints ← the set of all constraints of the form $(a = v)$, where a is a member of attributes, and v is a value of a that occurs in the current set of examples.
      ii. New_candidate_hypotheses
          for each h in candidate_hypotheses.
              for each c in All_canstraints.
      iii. Create a specialization of h by adding the constraint c.
      iv. Remove from New_candidate_hypotheses any hypotheses that are duplicates, inconsistent, or not maximally specific.
   b. Update best_hypothesis :
      i. For all h in New_candidate_hypotheses do.
      ii. If (performance(h_Examples, Target_attribute)
          > Performance(Best_hypothesis, Examples, Target_attribute))
          Then Best_hypothesis ← h.
   c. Update candidate_hypotheses :
      i. Candidate_hypotheses ← the k best members of New_candidate_hypotheses, according to the performance measure.
      ii. Return a rule of the form :
          "IF Best_hypothesis THEN prediction"
          where prediction is the most frequent value of Target_attribute among those Examples that match Best_hypothesis.
          Performance(h, Examples, Target_attribute).
      iii. h_examples ← the subset of Examples that match h.
4. Return – Entropy(h_example), where entropy is with respect to Target_attribute.

Que 5.14. Write short note on sequential covering algorithm.

Answer

1. Sequential covering 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.
2. Many rule learning algorithms are variants of the sequential covering algorithm.
3. This is the most popular algorithm implementing rule learning.

4. A covering algorithm, in the context of propositional learning systems, is an algorithm that develops a cover for the set of positive examples *i.e.*, a set of hypotheses that account for all the positive examples but none of the negative examples.
5. This is called sequential covering because it learn one rule at a time and repeat this process to gradually cover the full set of positive examples.
6. The most effective approach to Learn-One-Rule is beam search.
7. The characteristic of this algorithm is it requires high accuracy, *i.e.*, the prediction should be correct with high probability.
8. Another is that it is possibly low coverage, means that it does not make prediction for all problems.

Que 5.15. Write procedure of sequential covering algorithm.

Answer

Sequential-convering(Target_attribute, Attributes, Examples, Threshold) :
1. Learned_rules ← { }.
2. Rule ← learn-one-rule(Target_attribute, Attributes, Examples).
3. While performance(Rule, Examples) > Threshold, do :
   i. Learned_rule ← Learned_rules + Rule.
   ii. Examples ← Examples – {examples correctly classified by Rule}.
   iii. Rule ← Learn-one-rule(Target_attribute, Attributes, Examples).
4. Learned_rules ← sort Learned_rules according to performance over Examples.
5. Return Learned_rules.

Que 5.16. Explain general-to-specific beam search in details.

Answer

1. Beam search is a heuristic search algorithm that explores a graph by expanding the most promising node in a limited set.
2. In beam search, only a predetermined number of best partial solutions are kept as candidates. It is a greedy algorithm.
3. Beam search uses breadth-first search to build its search tree.
4. At each level of the tree, it generates all successors of the states at the current level, sorting them in increasing order of heuristic cost.
5. However, it only stores a predetermined number of states at each level (called the beam width). Only those states are expanded next. The greater the beam width, the fewer states are pruned.
6. With an infinite beam width, no states are pruned and beam search is identical to breadth-first search. The beam width bounds the memory required to perform the search.

7. Since a goal state could potentially be pruned, beam search sacrifices completeness (guarantee that an algorithm will terminate with a solution, if one exists).
8. Beam search is not optimal (that is, there is no guarantee that it will find the best solution).
9. In general, beam search returns the first solution found. Beam search for machine translation is a different case : once reaching the configured maximum search depth (*i.e.*, translation length), the algorithm will evaluate the solutions found during search at various depths and return the best one (the one with the highest probability).
10. The beam width can either be fixed or variable. Approach that uses a variable beam width starts with the width at a minimum. If no solution is found, the beam is widened and the procedure is repeated.

Que 5.17. Write the procedure of general-to-specific beam search.

Answer

1. Initialize a set of most general complexes.
2. Evaluate performances of those complexes over the example set.
   a. Count how many positive and negative examples it covers.
   b. Evaluate their performances.
3. Sort complexes according to their performances.
4. If the best complex satisfies some threshold, form the hypothesis and return.
5. Otherwise, pick k best performing complexes for the next generation.
6. Specializing all k complexes in the set to find new set of less general complexes.
7. Go to step 2.

Que 5.18. What are the properties of heuristic search ? Give example of heuristic search.

Answer

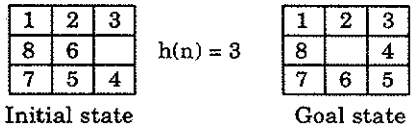
Properties of heuristic search are :

1. **Admissibility condition :** Algorithm A is admissible if it is generated to return as optimal solution.
2. **Completeness condition :** Algorithm A is complete if it always terminates with a solution.
3. **Dominance properties :** Let A_1 and A_2 be admissible algorithms with heuristic estimation function h_1 and h_2 respectively. A_1 is said to be more informed than A_2 whenever $h_1'(n) > h_2'(n)$ for all n, A_1 is said to dominate A_2.

4. **Optimality property :** Algorithm A is optimal over a class of algorithms if A dominates all members of the class.

Example :

8-puzzle ; $h(n)$ = tiles out of place (heuristic function)



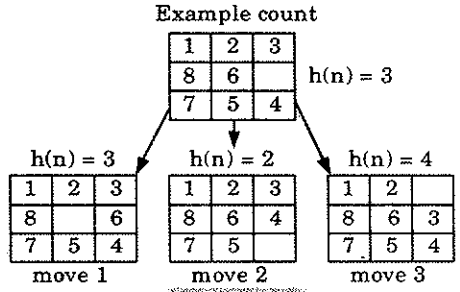


Fig. 5.18.1.

After opting the move 2, we will be in the state as per Fig. 5.18.2.

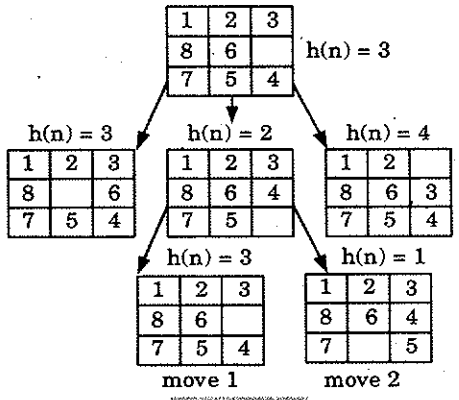


Fig. 5.18.2.

After moving and selecting move 2, we will be in the position as per in Fig. 5.18.3.

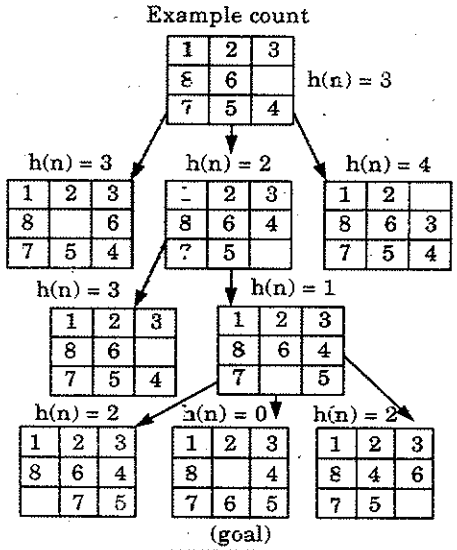


Fig. 5.18.3.

This procedure is continuous process until we reach the goal node as end of the search.

Que 5.19. Explain hill climbing algorithm.

Answer

1. Search methods based on hill climbing get their names from the way the nodes are selected for expansion.
2. At each point in the search path, a successor node that appears to lead most quickly to the top of the hill (the goal) is selected for exploration.
3. This method requires that some information be available with which to evaluate and order the most promising choices.
4. Hill climbing is like depth-first searching where the most promising child is selected for expansion.
5. When the children have been generated, alternative choices are evaluated using some type of heuristic function.
6. Hill climbing can produce substantial savings over blind searches when an informative, reliable function is available to guide the search to a global goal.
7. Here, the generate and test method is augmented by a heuristic function which measures the closeness of the current state to the goal state.

a. Calculate the initial state, if it is a goal state, then return and quit otherwise continue with the initial state as the current state.

b. Follow the loop, until a solution is found or until there are no new operators left to be applied in the current state :

   i. Select an operator that has not yet been applied to the current state and apply it to produce a new state.

   ii. Evaluate the new state.

      1. If it is a goal state, then return and quit.
      2. If it is not a goal state but it is better than the current state, then make it as the current state.
      3. If it is not better than the current state, then continue in the current loop.

PART-4

FOIL, Reinforcement Learning, The Learning Task, Q Learning.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.20. Explain FOIL algorithm with steps.

Answer

FOIL is similar to the propositional rule learning approach except for the following :

1. FOIL accommodates first-order rules and thus needs to accommodate variables in the rule pre-conditions.
2. FOIL uses a special performance measure (FOIL-GAIN) which takes into account the different variable bindings.
3. FOIL seeks only rules that predict when the target literal is true (instead of predicting when it is true or when it is false).
4. FOIL performs a simple hill-climbing search rather than a beam search.
5. The FOIL algorithm is as follows :

```
Input List of examples
Output Rule in first-order predicate logic
FOIL (Examples)
    Let Pos be the positive examples
    Let Pred be the predicate to be learned
```

```
Until Pos is empty do :
    Let Neg be the negative examples
    Set Body to empty
    Call LearnClauseBody
    Add Pred ← Body to the rule
    Remove from Pos all examples which satisfy Body
Procedure LearnClauseBody
  Until Neg is empty do
    Choose a literal L
    Conjoin L to Body
    Remove from Neg examples that do not satisfy L
```

Que 5.21. 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. 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.
10. In many cases humans learn in a very similar way. For example, when a child learns to walk, this usually happens without instruction, rather simply through reinforcement.
11. Successful attempts at working are rewarded by forward progress, and unsuccessful attempts are penalized by often painful falls.

12. Positive and negative reinforcement are also important factors in successful learning in school and in many sports.
13. In many complex domains, reinforcement learning is the only feasible way to train a program to perform at high levels.

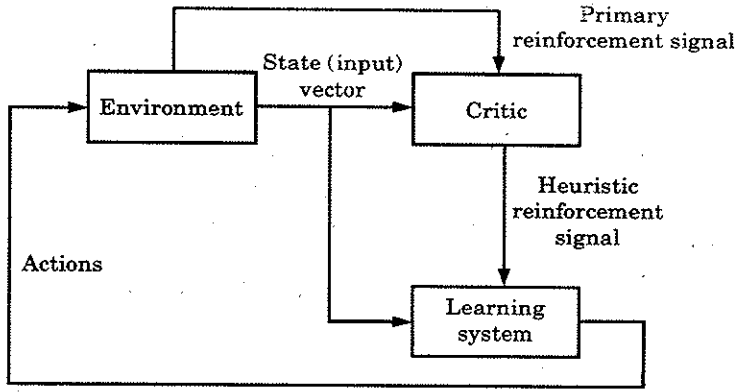


Fig. 5.121.1. Block diagram of reinforcement learning.

Que 5.22. 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 5.23. What is reinforcement learning ? Explain passive reinforcement learning and active reinforcement learning.

Answer

Reinforcement learning : Refer Q. 5.21, Page 5–19G, 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. 5.23.1 shows a policy for the world and the corresponding utilities.
4. In Fig. 5.23.1(*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.

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

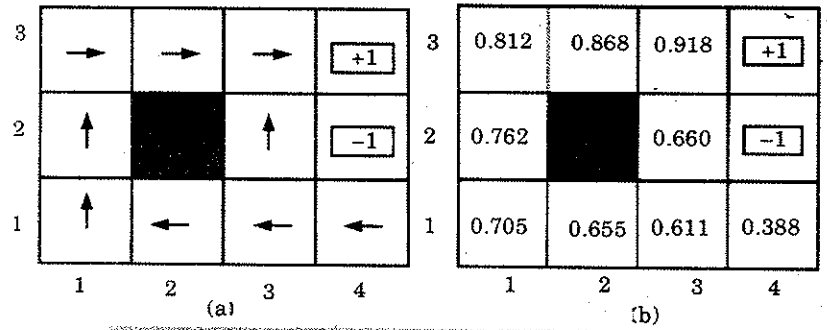


Fig. 5.23.1. (a) A policy π for the 4 × 3 world; (b) The utilities of the states in the 4 × 3 world, given policy π.

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 5.24. 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 5.25. 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 sole objective is to maximise 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) :

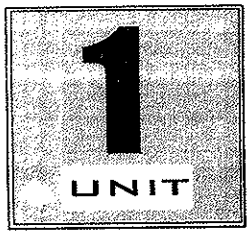
a. It defines the transitions in the environment action a taken in the states, will lead to state s^2.

b. It can be probabilistic.

Que 5.26. Write short note on Q-learning.

Answer

1. 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.
2. Each residential load defined as an agent. Agents should learn how to participate in the electrical market and optimize their cost, simultaneously.
3. A reinforcement learning algorithm so-called a Q-learning algorithm is employed.
4. When an agent i is modeled by a Q-learning algorithm, it keeps in memory a function $Q_i : A_i \rightarrow R$ such that $Q_i(a_i)$ represents it will obtain the expected reward if playing action a_i.
5. It then plays with a high probability the action it believes is going to lead to the highest reward, observes the reward it obtains and uses this observation to update its estimate of Q_i. Suppose that the t^{th} time the game is played, the joint action $(a_1{}^t, \ldots\ldots a_n{}^t)$ represents the actions the different agents have taken.



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. Tracability
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).

2 UNIT

Decision Tree Learning (2 Marks Questions)

2.1. 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.

2.2. 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.

2.3. What are the types of decision tree ?

Ans: **There are two types of decision tree :**

1. Classification tree
2. Regression tree

2.4. Define classification tree and regression tree.

Ans: **Classification :** 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.

2.5. Name the decision-tree algorithm.

Ans: **Decision-tree algorithms are :**

1. ID3
2. C4.5
3. CART

2.6. 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

2.7. 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
4. Gini index

2.8. Define artificial neural network.

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

2.9. What are the advantages of ANN ?

Ans: **Advantages of ANN are :**

1. Easy to use.
2. Alter to unknown condition.
3. It can model difficult function.
4. It can be imposed in any application.

2.10. What are the disadvantages of ANN ?

Ans: **Disadvantages of ANN are :**

1. Hardware dependence.
2. Unexplained functioning of the network.
3. Assurance of proper network structure.
4. The difficulty of showing the problem to the network.
5. The duration of the network is unknown.

2.11. Name different types of neuron connection.

Ans: **Different types of neuron connection are :**

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.

2.12. Define gradient descent.

Ans: Gradient descent is an optimization technique in Machine Learning and Deep Learning and it can be used with most, if not all, of the learning algorithms.

2.13. What are the different types of gradient descent ?

Ans: **Different types of gradient descent are :**

1. Batch gradient descent.
2. Stochastic gradient descent.
3. Mini-batch gradient descent.

2.14. Define ADALINE.

Ans. ADALINE is an Adaptive Linear Neuron network with a single linear unit. The Adaline network is trained using the delta rule. It receives input from several units and bias unit. An Adeline model consists of trainable weights. The inputs are of two values (+ 1 or – 1) and the weights have signs (positive or negative).

2.15. Define backpropagation algorithm.

Ans. Backpropagation algorithm is an algorithm used in the training of feedforward neural networks for supervised learning. Backpropagation efficiently computes the gradient of the loss function with respect to the weights of the network for a single input-output example.

2.16. What is perceptron ?

Ans. The perceptron is the simplest form of a neural network used for classification of patterns said to be linearly separable. It consists of a single neuron with adjustable synaptic weights and bias.

2.17. What is multilayer perceptron ?

Ans. The perceptrons which are arranged in layers are called multilayer perceptron. This model has three layers : an input layer, output layer and hidden layer.

2.18. What are the parameters that affect the backpropagation neural network ?

Ans. **Parameters that affect backpropagation neural networks are :**

1. Momentum factor
2. Learning coefficient
3. Sigmoidal gain
4. Threshold value

2.19. What are the selection parameters used in BPN (Backpropagation Neural Network) ?

Ans. **Selection parameters used in BPN are :**

1. Number of hidden nodes
2. Momentum coefficient α
3. Sigmoidal gain λ
4. Local minima
5. Learning coefficient η

2.20. What are the ANNs used for speech and character recognition ?

Ans. **ANNs used for speech recognition are :**

1. Multilayer network.
2. Multilayer network with recurrent connections.

ANNs used for character recognition are :

1. Multilayer neural network such as backpropagation neural network.

2.21. What are the advantages of mini-Batch gradient descent ?

Ans. **Advantages of mini-Batch gradient descent are :**

1. Easily fits in the memory.
2. It is computationally efficient.
3. It performs vectorization.

3 UNIT

Evaluating Hypotheses (2 Marks Questions)

3.1. Define hypothesis.

Ans: Hypothesis is a function that describes the target in supervised learning. A hypothesis is a tentative relationship between two or more variables which direct the research activity.

3.2. What are characteristics of hypothesis ?

Ans: Characteristics of hypothesis are :

1. Empirically testable
2. Simple and clear
3. Specific and relevant
4. Predictable
5. Manageable

3.3. What is the importance of hypothesis ?

Ans: Importance of hypothesis are :

1. It gives a direction to the research.
2. It specifies the focus of the researcher.
3. It helps in devising research techniques.
4. It prevents from blind research.
5. It ensures accuracy and precision.
6. It saves resources, time, money and energy.

3.4. What are different types of hypotheses ?

Ans: Different types of hypotheses are :

1. Simple hypothesis
2. Complex hypothesis
3. Working hypothesis
4. Alternative hypothesis
5. Null hypothesis
6. Statistical hypothesis
7. Logical hypothesis

3.5. What are the difficulties faced in estimating the accuracy of hypotheses ?

Ans: Difficulties faced in estimating the accuracy of hypotheses are :

1. Bias in the estimate
2. Variance in the estimate

3.6. What are different methods of sampling ?

Ans: Different methods of sampling are :

1. Simple random sampling
2. Systematic sampling
3. Stratified sampling

3.7. 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.

3.8. 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.

3.9. 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.

3.10. What are the usage of EM algorithm ?

Ans: Usage 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.

3.11. What are the advantages of EM algorithm ?

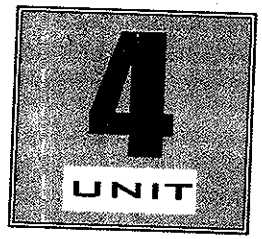
Ans: Advantages of EM algorithm are :

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.

3.12. 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).



Computational Learning Theory (2 Marks Questions)

4.1. Define computational learning theory.

Ans: Computational Learning Theory (CLT) is a field of AI research studying the design of machine learning algorithms to determine what sorts of problems are learnable.

4.2. What are the uses of computational learning theory ?

Ans: Uses of computational learning theory (CLT) are :

i. It provides a theoretical analysis of learning.
ii. It shows when a learning algorithm can be expected to succeed.
iii. It shows when learning may be impossible.

4.3. Define sample complexity.

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.

4.4. What the variants of sample complexity ?

Ans: There are two variants of sample complexity :

1. **Weak variant :** The weak variant fixes a particular input-output distribution.
2. **Strong variant :** The strong variant takes the worst-case sample complexity over all input-output distributions.

4.5. What are the parameters on which complexity of a learning problem depends ?

Ans: Complexity of a learning problem depends on :

1. Size or expressiveness of the hypothesis space.
2. Accuracy to which target concept must be approximated.
3. Probability with which the learner must produce a successful hypothesis.
4. Manner in which training examples are presented, for example, randomly or by query to an oracle.

4.6. Define principal component analysis.

Ans: Principal Component Analysis (PCA) is a statistical procedure that uses an orthogonal transformation which converts a set of correlated variables to a set of uncorrelated variables.

4.7. Define mistake bound model of learning.

Ans: An algorithm A learns a class C with mistake bound M *iff* Mistake $(A, C) \leq M$.

4.8. 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.

4.9. 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.

4.10. 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.

4.11. 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.

4.12. 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.

4.13. Define case-based learning.

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.

4.14. 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.

4.15. What are the functions of CBL ?

Ans: **Functions of case-based learning algorithm are :**

1. Pre-processor
2. Similarity
3. Prediction
4. Memory updating

4.16. What are the processing stage of CBL ?

Ans: **Case-based learning algorithm processing stages are :**

1. Case retrieval
2. Case adaptation
3. Solution evaluation
4. Case-base updating

4.17. 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 formalised solution spaces.
5. Suitability for sequential problem solving.
6. Ease of explanation.
7. Ease of maintenance.

4.18. What are the applications of CBL ?

Ans: **Applications of CBL :**

1. Interpretation
2. Classification
3. Design
4. Planning
5. Advising

4.19. What are the advantages of instance-based learning ?

Ans: **Advantages of instance-based learning :**

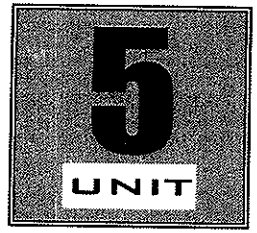
1. Learning is trivial.
2. Works efficiently.

3. Noise resistant.
4. Rich representation, arbitrary decision surfaces.
5. Easy to understand.

4.20. 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.



Genetic Algorithm (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 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 tc implement.

5.19. Define positive and negative reinforcement learning.

Ans: Positive reinforcement learning :

a. 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.

b. In other words, it has a positive effect on the behaviour.

Negative reinforcement learning : Negative reinforcement is defined as strengthening of a behaviour because a negative condition is stopped or avoided.

☺☺☺