CS3491 ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

COURSE OBJECTIVES:

The main objectives of this course are to:

- Study about uninformed and Heuristic search techniques.
- Learn techniques for reasoning under uncertainty
- Introduce Machine Learning and supervised learning algorithms
- Study about ensembling and unsupervised learning algorithms
- Learn the basics of deep learning using neural networks

UNIT I PROBLEM SOLVING

Introduction to AI - AI Applications - Problem solving agents – search algorithms – uninformed search strategies – Heuristic search strategies – Local search and optimization problems – adversarial search – constraint satisfaction problems (CSP)

UNIT II PROBABILISTIC REASONING

Acting under uncertainty – Bayesian inference – naïve bayes models. Probabilistic reasoning – Bayesian networks – exact inference in BN – approximate inference in BN – causal networks.

UNIT III SUPERVISED LEARNING

Introduction to machine learning – Linear Regression Models: Least squares, single & multiple variables, Bayesian linear regression, gradient descent, Linear Classification Models: Discriminant function – Probabilistic discriminative model - Logistic regression, Probabilistic generative model – Naive Bayes, Maximum margin classifier – Support vector machine, Decision Tree, Random forests

UNIT IV ENSEMBLE TECHNIQUES AND UNSUPERVISED LEARNING

Combining multiple learners: Model combination schemes, Voting, Ensemble Learning - bagging, boosting, stacking, Unsupervised learning: K-means, Instance Based Learning: KNN, Gaussian mixture models and Expectation maximization

UNIT V NEURAL NETWORKS

Perceptron - Multilayer perceptron, activation functions, network training – gradient descent optimization – stochastic gradient descent, error backpropagation, from shallow networks to deep networks –Unit saturation (aka the vanishing gradient problem) – ReLU, hyperparameter tuning, batch normalization, regularization, dropout.

PRACTICAL EXERCISES:

1. Implementation of Uninformed search algorithms (BFS, DFS)
2. Implementation of Informed search algorithms (A*, memory-bounded A*)
3. Implement naïve Bayes models
4. Implement Bayesian Networks
5. Build Regression models
6. Build decision trees and random forests
7. Build SVM models
8. Implement ensembling techniques
9. Implement clustering algorithms

10. Implement EM for Bayesian networks
11. Build simple NN models
12. Build deep learning NN models

COURSE OUTCOMES:

At the end of this course, the students will be able to:

CO1: Use appropriate search algorithms for problem solving
CO2: Apply reasoning under uncertainty
CO3: Build supervised learning models
CO4: Build ensembling and unsupervised models
CO5: Build deep learning neural network models

TEXT BOOKS:

1. Stuart Russell and Peter Norvig, "Artificial Intelligence – A Modern Approach", Fourth Edition, Pearson Education, 2021.
2. Ethem Alpaydin, "Introduction to Machine Learning", MIT Press, Fourth Edition, 2020.

REFERENCES:

1. Dan W. Patterson, "Introduction to Artificial Intelligence and Expert Systems", Pearson Education,2007
2. Kevin Night, Elaine Rich, and Nair B., "Artificial Intelligence", McGraw Hill, 2008
3. Patrick H. Winston, "Artificial Intelligence", Third Edition, Pearson Education, 2006
4. Deepak Khemani, "Artificial Intelligence", Tata McGraw Hill Education, 2013 (http://nptel.ac.in/)
5. Christopher M. Bishop, "Pattern Recognition and Machine Learning", Springer, 2006.
6. Tom Mitchell, "Machine Learning", McGraw Hill, 3rd Edition,1997.
7. Charu C. Aggarwal, "Data Classification Algorithms and Applications", CRC Press, 2014
8. Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar, "Foundations of Machine Learning", MIT Press, 2012.
9. Ian Goodfellow, Yoshua Bengio, Aaron Courville, "Deep Learning", MIT Press, 2016

UNIT I PROBLEM SOLVING

Introduction to AI - AI Applications - Problem solving agents – search algorithms – uninformed search strategies – Heuristic search strategies – Local search and optimization problems – adversarial search – constraint satisfaction problems (CSP)

Part A

1. **What is Artificial Intelligence?**

   Artificial Intelligence is the study of how to make computers do things which at the moment people do better.

2. **What is an agent?**

   An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators.

3. **What are the different types of agents?**

   A human agent has eyes, ears, and other organs for sensors and hands, legs, mouth, and other body parts for actuators.

   A robotic agent might have cameras and infrared range finders for sensors and various motors for actuators.

   A software agent receives keystrokes, file contents, and network packets as sensory inputs and acts on the environment by displaying on the screen, writing files, and sending network packets.

   Generic agent - A general structure of an agent who interacts with the environment.

4. **Define rational agent.**

   For each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has. A rational agent should be autonomous.

5. **List down the characteristics of intelligent agent.**

   Internal characteristics are

   Learning/reasoning: an agent has the ability to learn from previous experience and to successively adapt its own behaviour to the environment.

   Reactivity: an agent must be capable of reacting appropriately to influences or information from its environment.

   Autonomy: an agent must have both control over its actions and internal states. The degree of the agent's autonomy can be specified. There may need intervention from the user only for important decisions.

   Goal-oriented: an agent has well-defined goals and gradually influence its environment and so achieve its own goals.

   External characteristics are

Communication: an agent often requires an interaction with its environment to fulfil its tasks, such as human, other agents, and arbitrary information sources.

Cooperation: cooperation of several agents permits faster and better solutions for complex tasks that exceed the capabilities of a single agent.

Mobility: an agent may navigate within electronic communication networks.

Character: like human, an agent may demonstrate an external behaviour with many human characters as possible.

6. What are various applications of AI? or What can AI do today?

- Robotic vehicles
- Speech recognition
- Autonomous planning and scheduling
- Game playing
- Spam fighting
- Logistics planning
- Robotics
- Machine Translation

7. Are reflex actions (such as flinching from a hot stove) rational? Are they intelligent?

Reflex actions can be considered rational. If the body is performing the action, then it can be argued that reflex actions are rational because of evolutionary adaptation. Flinching from a hot stove is a normal reaction, because the body wants to keep itself out of danger and getting away from something hot is a way to do that.

Reflex actions are also intelligent. Intelligence suggests that there is reasoning and logic involved in the action itself.

8. Is AI a science, or is it engineering? Or neither or both? Explain.

AI is both science and engineering. Observing and experimenting, which are at the core of any science, allows us to study artificial intelligence. From what we learn by observation and experimentation, we are able to engineer new systems that encompass what we learn and that may even be capable of learning themselves.

9. What are the various agent programs in intelligent systems?

Simple reflex agents
Model-based reflex agents
Goal-based agents
Utility-based agents

10. Define the problem solving agent.

A Problem solving agent is a goal-based agent. It decides what to do by finding sequence of actions that lead to desirable states. The agent can adopt a goal and aim at satisfying it. Goal formulation is the first step in problem solving.

11. Define the terms goal formulation and problem formulation.

Goal formulation based on the current situation and the agent's performance measure is the first step in problem solving. The agent's task is to find out which sequence of actions will get to a goal state.

Problem formulation is the process of deciding what actions and states to consider given a goal.

12. List the steps involved in simple problem solving agent.

(i) Goal formulation
(ii) Problem formulation
(iii) Search
(iv) Search Algorithm
(v) Execution phase

13. What are the components of well-defined problems? (or)
What are the four components to define a problem? Define them?

The four components to define a problem are,

1) Initial state – it is the state in which agent starts in.

2) A description of possible actions – it is the description of possible actions which are available to the agent.

3) The goal test – it is the test that determines whether a given state is goal (final) state.

4) A path cost function – it is the function that assigns a numeric cost (value) to each path.

The problem-solving agent is expected to choose a cost function that reflects its own performance measure.

14. Differentiate toy problems and real world problems?

A toy problem is intended to illustrate various problem solving methods. It can be easily used by different researchers to compare the performance of algorithms. A real world problem is one whose solutions people actually care about.

15. Give example for real world end toy problems.

Real world problem examples:

- Airline travel problem.
- Touring problem.
- Traveling salesman problem
- VLSI Layout problem
- Robot navigation

- Automatic Assembly
- Internet searching

Toy problem Examples:

- 8 – Queen problem
- 8 – Puzzle problem
- Vacuum world problem

16. How will you measure the problem-solving performance?

We can evaluate an algorithm's performance in four ways:

Completeness: Is the algorithm guaranteed to find a solution when there is one?

Optimality: Does the strategy find the optimal solution?

Time complexity: How long does it take to find a solution?

Space complexity: How much memory is needed to perform the search?

17. What is the application of BFS?

It is simple search strategy, which is complete i.e. it surely gives solution if solution exists. If the depth of search tree is small then BFS is the best choice. It is useful in tree as well as in graph search.

18. State on which basis search algorithms are chosen?

Search algorithms are chosen depending on two components.

1) How is the state space – That is, state space is tree structured or graph? Critical factor for state space is what is branching factor and depth level of that tree or graph.

2) What is the performance of the search strategy? A complete, optimal search strategy with better time and space requirement is critical factor in performance of search strategy.

19. Evaluate performance of problem-solving method based on depth-first search algorithm?

DFS algorithm performance measurement is done with four ways –

1) Completeness – It is complete (guarantees solution)

2) Optimality – it is not optimal.

3) Time complexity – It's time complexity is O (b).

4) Space complexity – its space complexity is O (b d+1).

20. List some of the uninformed search techniques.

The uninformed search strategies are those that do not take into account the location of the goal. That is these algorithms ignore where they are going until they find a goal and report success. The various uninformed search strategies are

- Breadth-first search
- Uniform-cost search
- Depth-first search

- Depth-limited search
- Iterative deepening depth-first search
- Bidirectional search

21. What is the power of heuristic search? (or) Why does one go for heuristics search?

Heuristic search uses problem specific knowledge while searching in state space. This helps to improve average search performance. They use evaluation functions which denote relative desirability (goodness) of a expanding node set. This makes the search more efficient and faster. One should go for heuristic search because it has power to solve large, hard problems in affordable times.

22. What are the advantages of heuristic function?

Heuristics function ranks alternative paths in various search algorithms, at each branching step, based on the available information, so that a better path is chosen. The main advantage of heuristic function is that it guides for which state to explore now, while searching. It makes use of problem specific knowledge like constraints to check the goodness of a state to be explained. This drastically reduces the required searching time.

23. State the reason when hill climbing often gets stuck?

Local maxima are the state where hill climbing algorithm is sure to get struck. Local maxima are the peak that is higher than each of its neighbour states, but lower than the global maximum. So we have missed the better state here. All the search procedure turns out to be wasted here. It is like a dead end.

24. When a heuristic function h is said to be admissible? Give an admissible heuristic function for TSP?

Admissible heuristic function is that function which never over estimates the cost to reach the goal state. It means that h(n) gives true cost to reach the goal state 'n'. The admissible heuristic for TSP is

a. Minimum spanning tree.

b. Minimum assignment problem

25. What do you mean by local maxima with respect to search technique?

Local maximum is the peak that is higher than each of its neighbour states, but lowers than the global maximum i.e. a local maximum is a tiny hill on the surface whose peak is not as high as the main peak (which is a optimal solution). Hill climbing fails to find optimum solution when it encounters local maxima. Any small move, from here also makes things worse (temporarily). At local maxima all the search procedure turns out to be wasted here. It is like a dead end.

26. How can we avoid ridge and plateau in hill climbing?

Ridge and plateau in hill climbing can be avoided using methods like backtracking, making big jumps. Backtracking and making big jumps help to avoid plateau, whereas, application of multiple rules helps to avoid the problem of ridges.

27. Differentiate Blind Search and Heuristic Search.

Parameters	Blind search	Heuristic search
Known as	It is also known Uninformed Search	It is also known Informed Search
Using Knowledge	It doesn't use knowledge for the searching process.	It uses knowledge for the searching process.
Performance	It finds solution slow as compared to an informed search.	It finds a solution more quickly.
Completion	It is always complete.	It may or may not be complete.
Cost Factor	Cost is high.	Cost is low.
Time	It consumes moderate time because of slow searching.	It consumes less time because of quick searching.
Direction	No suggestion is given regarding the solution in it.	There is a direction given about the solution.
Implementation	It is lengthier while implemented.	It is less lengthy while implemented.
Efficiency	It is comparatively less efficient as incurred cost is more and the speed of finding the Breadth-First solution is slow.	It is more efficient as efficiency takes into account cost and performance. The incurred cost is less and speed of finding solutions is quick.
Computational requirements	Comparatively higher computational requirements.	Computational requirements are lessened.
Size of search problems	Solving a massive search task is challenging.	Having a wide scope in terms of handling large search problems.
Examples of Algorithms	Example a) Breadth first search b) Uniform cost search c) Depth first Search d) Depth limited search e) Iterative deepening search f) Bi – Directional Search	Example a) Best first search b) Greedy search c) A* search d) AO* Search e)Hill Climbing Algorithm

28. What is CSP?

CSP are problems whose state and goal test conform to a standard structure and very simple representation. CSPs are defined using set of variables and a set of constraints on those variables. The variables have some allowed values from specified domain. For example – Graph coloring problem.

29. How can minimax also be extended for game of chance?

In a game of chance, we can add extra level of chance nodes in game search tree. These nodes have successors which are the outcomes of random element. The minimax algorithm uses probability P attached with chance node di based on this value. Successor function S(N,di) give moves from position N for outcome di

Part B

1. **Enumerate Classical "Water jug Problem". Describe the state space for this problem and also give the solution.**
2. **How to define a problem as state space search? Discuss it with the help of an example**
3. **Solve the given problem. Describe the operators involved in it.**
   Consider a Water Jug Problem : You are given two jugs, a 4-gallon one and a 3-gallon one. Neither has any measuring markers on it. There is a pump that can be used to fill the jugs with water. How can you get exactly 2 gallons of water into the 4-gallon jug ? Explicit Assumptions: A jug can be filled from the pump, water can be poured out of a jug onto the ground, water can be poured from one jug to another and that there are no other measuring devices available.
4. **Define the following problems. What types of control strategy is used in the following problem.**
   i.The Tower of Hanoi
   ii.Crypto-arithmetic
   iii.The Missionaries and cannibals problems
   iv.8-puzzle problem
5. **Discuss uninformed search methods with examples.**
6. **Give an example of a problem for which breadth first search would work better than depth first search.**
7. **Explain the algorithm for steepest hill climbing**
8. **Explain the A* search and give the proof of optimality of A***
9. **Explain AO* algorithm with a suitable example. State the limitations in the algorithm?**
10. **Explain the nature of heuristics with example. What is the effect of heuristics accuracy?**
11. **Explain the various types of hill climbing search techniques.**
12. **Discuss about constraint satisfaction problem with a algorithm for solving a crypt arithmetic Problem.**
13. **Solve the following Crypt arithmetic problem using constraints satisfaction search procedure.**

```
  CROSS
+ROADS
-----------
 DANGER
----------------
```

14. **Explain alpha-beta pruning algorithm and the Minmax game playing algorithm with example?**
15. **Solve the given problem. Describe the operators involved in it.**
    Consider a water jug problem: You are given two jugs, a 4-gallon one and a 3-gallon one. Neither have any measuring Markers on it. There is a pump that can be used to fill the jug with water. How can you get exactly 2 gallons of water into the 4 gallon jug? Explicit Assumptions: A jug can be filled from the pump, water can be poured out of a jug onto the ground, water can be poured from one jug to another and that there are no other measuring devices available.

UNIT II PROBABILISTIC REASONING

Acting under uncertainty – Bayesian inference – naïve bayes models. Probabilistic reasoning – Bayesian networks – exact inference in BN – approximate inference in BN – causal networks.

Part A

1. **Why does uncertainty arise?**
   Agents almost never have access to the whole truth about their environment. Uncertainty arises because of both laziness and ignorance. It is inescapable in complex, nondeterministic, or partially observable environments
   - Agents cannot find a categorical answer.
   - Uncertainty can also arise because of incompleteness, incorrectness in agents understanding of properties of environment.
2. **Differentiate uncertainty with ignorance.**
   A key condition that differentiates ignorance from uncertainty is the absence of knowledge about the factors that influence the issues

3. **What is the need for probability theory in uncertainty?**
   Probability provides the way of summarizing the uncertainty that comes from our laziness and ignorance. Probability statements do not have quite the same kind of semantics known as evidences.

4. **What is the need for utility theory in uncertainty?**
   Utility theory says that every state has a degree of usefulness, or utility to in agent, and that the agent will prefer states with higher utility.

5. **Define principle of maximum expected utility (MEU)?**
   `The fundamental idea of decision theory is that an agent is rational if and only if it chooses the action that yields the highest expected utility, averaged over all the possible outcomes of the action. This is called the principle of maximum expected utility (MEU).

6. **Mention the needs of probabilistic reasoning in AI.**
   - When there are unpredictable outcomes.
   - When specifications or possibilities of predicates becomes too large to handle.
   - When an unknown error occurs during an experiment.

7. **What does the full joint probability distribution specify?**

The full joint probability distribution specifies the probability of each complete assignment of values to random variables. It is usually too large to create or use in its explicit form, but when it is available it can be used to answer queries simply by adding up entries for the possible worlds corresponding to the query propositions.

8. **State Bayes' Theorem in Artificial Intelligence.**

Bayes' theorem is also known as **Bayes' rule, Bayes' law**, or **Bayesian reasoning**, which determines the probability of an event with uncertain knowledge. It is a way to calculate the value of P(B|A) with the knowledge of P(A|B). Bayes' theorem allows updating the probability prediction of an event by observing new information of the real world.

Example: If cancer corresponds to one's age then by using Bayes' theorem, we can determine the probability of cancer more accurately with the help of age.

P(A/B)=[P(A)*P(B/A)]/P(B)

9. **Given that P(A)=0.3,P(A|B)=0.4 and P(B)=0.5, Compute P(B|A).**

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) * P(B|A)}{P(B)}$$

0.4 = (0.3*P(B/A))/0.5

P(B/A) = 0.66

10. **What is Bayesian Belief Network?**

A Bayesian network is a probabilistic graphical model which represents a set of variables and their conditional dependencies using a directed acyclic graph. It is also called a Bayes network, belief network, decision network, or Bayesian model.

Bayesian networks are probabilistic, because these networks are built from a probability distribution, and also use probability theory for prediction and anomaly detection.

A Bayesian network is a directed graph in which each node is annotated with quantitative probability information. The full specification is as follows:

1. Each node corresponds to a random variable, which may be discrete or continuous.

2. A set of directed links or arrows connects pairs of nodes. If there is an arrow from node X to node Y , X is said to be a parent of Y. The graph has no directed cycles (and hence is a directed acyclic graph, or DAG).

3. Each nodeXi has a conditional probability distribution P(Xi |Parents(Xi)) that quantifies the effect of the parents on the node.

Part B

1. **How to get the exact inference form Bayesian network?**

2. **Explain variable elimination algorithm for answering queries on Bayesian networks?**
3. **Define uncertain knowledge, prior probability and conditional probability. State the Bayes' theorem. How it is useful for decision making under uncertainty? Explain belief networks briefly?**
4. **Explain the method of handling approximate inference in Bayesian networks.**
5. **What is Bayes' rule? Explain how Bayes' rule can be applied to tackle uncertain Knowledge.**
6. **Discuss about Bayesian Theory and Bayesian network.**
7. **Explain how does Bayesian statistics provide reasoning under various kinds of uncertainty?**
8. **How to get the approximate inference from Bayesian network.**
9. **Construct a Bayesian Network and define the necessary CPTs for the given scenario. We have a bag of three biased coins a,b and c with probabilities of coming up heads of 20%, 60% and 80% respectively. One coin is drawn randomly from the bag (with equal likelihood of drawing each of the three coins) and then the coin is flipped three times to generate the outcomes X1, X2 and X3.**
    a. **Draw a Bayesian network corresponding to this setup and define the relevant CPTs.**
    b. **Calculate which coin is most likely to have been drawn if the flips come up HHT**
10. **Consider the following set of propositions**

    Patient has spots

    Patient has measles

    Patient has high fever

    Patient has Rocky mountain spotted fever.

    Patient has previously been inoculated against measles.

    Patient was recently bitten by a tick

    Patient has an allergy.

    a) Create a network that defines the casual connections among these nodes.

    b) Make it a Bayesian network by constructing the necessary conditional probability matrix.

UNIT III SUPERVISED LEARNING

Introduction to machine learning – Linear Regression Models: Least squares, single & multiple variables, Bayesian linear regression, gradient descent, Linear Classification Models: Discriminant function – Probabilistic discriminative model - Logistic regression, Probabilistic generative model – Naive Bayes, Maximum margin classifier – Support vector machine, Decision Tree, Random forests.

PART - A

1. **What is Machine Learning?**

   Machine learning is a branch of computer science which deals with system programming in order to automatically learn and improve with experience. For example: Robots are programed so that they can perform the task based on data they gather from sensors. It automatically learns programs from data.

2. **Mention the difference between Data Mining and Machine learning?**

   Machine learning relates with the study, design and development of the algorithms that give computers the capability to learn without being explicitly programmed. While, data mining can be defined as the process in which the unstructured data tries to extract knowledge or unknown interesting patterns.

3. **What is 'Overfitting' in Machine learning?**

   In machine learning, when a statistical model describes random error or noise instead of underlying relationship 'overfitting' occurs. When a model is excessively complex, overfitting is normally observed, because of having too many parameters with respect to the number of training data types. The model exhibits poor performance which has been overfit.

4. **Why overfitting happens?**

   The possibility of overfitting exists as the criteria used for training the model is not the same as the criteria used to judge the efficacy of a model.

5. **How can you avoid overfitting?**

   By using a lot of data overfitting can be avoided, overfitting happens relatively as you have a small dataset, and you try to learn from it. But if you have a small database and you are forced to come with a model based on that. In such situation, you can use a technique known as cross validation. In this method the dataset splits into two section, testing and training datasets, the testing dataset will only test the model while, in training dataset, the datapoints will come up with the model. In this technique, a model is usually given a dataset of a known data on which training (training data set) is run and a dataset of unknown data against which the model is tested. The idea of cross validation is to define a dataset to "test" the model in the training phase.

6. **What are the five popular algorithms of Machine Learning?**

   - Decision Trees
   - Neural Networks (back propagation)
   - Probabilistic networks
   - Nearest Neighbor
   - Support vector machines

7. What are the different Algorithm techniques in Machine Learning?

The different types of techniques in Machine Learning are

- Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning
- Reinforcement Learning
- Transduction
- Learning to Learn

8. What are the three stages to build the hypotheses or model in machine learning?

- Model building
- Model testing
- Applying the model

9. What is the standard approach to supervised learning?

The standard approach to supervised learning is to split the set of example into the training set and the test.

10. What is 'Training set' and 'Test set'?

In various areas of information science like machine learning, a set of data is used to discover the potentially predictive relationship known as 'Training Set'. Training set is an examples given to the learner, while Test set is used to test the accuracy of the hypotheses generated by the learner, and it is the set of example held back from the learner. Training set are distinct from Test set.

11. What is the difference between artificial learning and machine learning?

Designing and developing algorithms according to the behaviours based on empirical data are known as Machine Learning. While artificial intelligence in addition to machine learning, it also covers other aspects like knowledge representation, natural language processing, planning, robotics etc.

12. What are the advantages of Naive Bayes?

In Naïve Bayes classifier will converge quicker than discriminative models like logistic regression, so you need less training data. The main advantage is that it can't learn interactions between features.

13. What is the main key difference between supervised and unsupervised machine learning?

supervised learning	Unsupervised learning
The supervised learning technique needs labelled data to train the model. For example, to solve a classification problem (a supervised learning task), you need to have label data to train the model and to classify the data into your labelled groups.	Unsupervised learning does not need any labelled dataset. This is the main key difference between supervised learning and unsupervised learning.

14. What is a Linear Regression?

In simple terms, linear regression is adopting a linear approach to modeling the relationship between a dependent variable (scalar response) and one or more independent variables (explanatory variables). In case you have one explanatory variable, you call it a simple linear regression. In case you have more than one independent variable, you refer to the process as multiple linear regressions.

15. What are the disadvantages of the linear regression model?

One of the most significant demerits of the linear model is that it is sensitive and dependent on the outliers. It can affect the overall result. Another notable demerit of the linear model is overfitting. Similarly, underfitting is also a significant disadvantage of the linear model.

16. What is the difference between classification and regression?

Classification is used to produce discrete results; classification is used to classify data into some specific categories. For example, classifying emails into spam and non-spam categories. Whereas, we use regression analysis when we are dealing with continuous data, for example predicting stock prices at a certain point in time.

17. What is the difference between stochastic gradient descent (SGD) and gradient descent (GD)?

Both algorithms are methods for finding a set of parameters that minimize a loss function by evaluating parameters against data and then making adjustments. In standard gradient descent, you'll evaluate all training samples for each set of parameters. This is akin to taking big, slow steps toward the solution. In stochastic gradient descent, you'll evaluate only 1 training sample for the set of parameters before updating them. This is akin to taking small, quick steps toward the solution.

18. What are the different types of least squares?

Least squares problems fall into two categories: linear or ordinary least squares and nonlinear least squares, depending on whether or not the residuals are linear in all unknowns. The linear least-squares problem occurs in statistical regression analysis; it has a closed-form solution.

19. What is the difference between least squares regression and multiple regression?

The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.

20. What is the principle of least squares?

Principle of Least Squares" states that the most probable values of a system of unknown quantities upon which observations have been made, are obtained by making the sum of the squares of the errors a minimum.

21. What are some advantages to using Bayesian linear regression?

Doing Bayesian regression is not an algorithm but a different approach to statistical inference. The major advantage is that, by this Bayesian processing, you recover the whole range of inferential solutions, rather than a point estimate and a confidence interval as in classical regression.

22. What Is Bayesian Linear Regression?

In Bayesian linear regression, the mean of one parameter is characterized by a weighted sum of other variables. This type of conditional modeling aims to determine the prior distribution of the regressors as well as other variables describing the allocation of the regress and eventually permits the out-of-sample forecasting of the regress and conditional on observations of the regression coefficients.

23. What are the advantages of Bayesian Regression?

- Extremely efficient when the dataset is tiny.
- Particularly well-suited for online learning as opposed to batch learning, when we know the complete dataset before we begin training the model. This is so that Bayesian Regression can be used without having to save data.
- The Bayesian technique has been successfully applied and is quite strong mathematically. Therefore, using this requires no additional prior knowledge of the dataset.

24. What are the disadvantages of Bayesian Regression?

- The model's inference process can take some time.
- The Bayesian strategy is not worthwhile if there is a lot of data accessible for our dataset, and the regular probability approach does the task more effectively.

25. What are types of classification models?

- Logistic Regression
- Naive Bayes
- K-Nearest Neighbors
- Decision Tree
- Support Vector Machines

26. Why is random forest better than SVM?

Random Forest is intrinsically suited for multiclass problems, while SVM is intrinsically two-class. For multiclass problem you will need to reduce it into multiple binary classification problems. Random Forest works well with a mixture of numerical and categorical features.

27. Which is better linear regression or random forest?

Multiple linear regression is often used for prediction in neuroscience. Random forest regression is an alternative form of regression. It does not make the assumptions of linear regression. We show that linear regression can be superior to random forest regression.

28. Which is better linear or tree based models?

If there is a high non-linearity & complex relationship between dependent & independent variables, a tree model will outperform a classical regression method. If you need to build a model which is easy to explain to people, a decision tree model will always do better than a linear model.

29. Is linear discriminant analysis classification or regression?

Linear Discriminant Analysis is a simple and effective method for classification. Because it is simple and so well understood, there are many extensions and variations to the method.

30. What is probabilistic discriminative model?

Discriminative models are a class of supervised machine learning models which make predictions by estimating conditional probability P(y|x). In order to use a generative model, more unknowns should be solved: one has to estimate probability of each class and probability of observation given class.

31. What is SVM?

It is a supervised learning algorithm used both for classification and regression problems. A type of discriminative modelling, support vector machine (SVM) creates a decision boundary to segregate n-dimensional space into classes. The best decision boundary is called a hyperplane created by choosing the extreme points called the support vectors.

32. What is Decision tree?

A type of supervised machine learning model where data is continuously split according to certain parameters. It has two main entities–decision nodes and leaves. While leaves are the final outcomes or decisions, nodes are the points where data is split.

33. What is Random forest?

It is a flexible and easy-to-use machine learning algorithm that gives great results without even using hyper-parameter tuning. Because of its simplicity and diversity, it is one of the most used algorithms for both classification and regression tasks.

34. What is Decision Tree Classification?

A decision tree builds classification (or regression) models as a tree structure, with datasets broken up into ever-smaller subsets while developing the decision tree, literally in a tree-like way with branches and nodes. Decision trees can handle both categorical and numerical data.

35. What Is Pruning in Decision Trees, and How Is It Done?

Pruning is a technique in machine learning that reduces the size of decision trees. It reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting.

Pruning can occur in:

- Top-down fashion. It will traverse nodes and trim subtrees starting at the root
- Bottom-up fashion. It will begin at the leaf nodes

There is a popular pruning algorithm called reduced error pruning, in which:

- Starting at the leaves, each node is replaced with its most popular class

- If the prediction accuracy is not affected, the change is kept
- There is an advantage of simplicity and speed

36. Do you think 50 small decision trees are better than a large one? Why?

Yes. Because a random forest is an ensemble method that takes many weak decision trees to make a strong learner. Random forests are more accurate, more robust, and less prone to overfitting.

37. You've built a random forest model with 10000 trees. You got delighted after getting training error as 0.00. But, the validation error is 34.23. What is going on? Haven't you trained your model perfectly?

The model has overfitted. Training error 0.00 means the classifier has mimicked the training data patterns to an extent, that they are not available in the unseen data. Hence, when this classifier was run on an unseen sample, it couldn't find those patterns and returned predictions with higher error. In a random forest, it happens when we use a larger number of trees than necessary. Hence, to avoid this situation, we should tune the number of trees using cross-validation.

38. When would you use random forests vs SVM and why?

There are a couple of reasons why a random forest is a better choice of the model than asupport vector machine:

- Random forests allow you to determine the feature importance. SVM's can't do this.
- Random forests are much quicker and simpler to build than an SVM.
- For multi-class classification problems, SVMs require a one-vs-rest method, which is less scalable and more memory intensive.

Part – B

1. **Assume a disease so rare that it is seen in only one person out of every million. Assume also that we have a test that is effective in that if a person has the disease, there is a 99 percent chance that the test result will be positive; however, the test is not perfect, and there is a one in a thousand chance that the test result will be positive on a healthy person. Assume that a new patient arrives and the test result is positive. What is the probability that the patient has the disease?**
2. **Explain Naïve Bayes Classifier with an Example.**
3. **Explain SVM Algorithm in Detail.**
4. **Explain Decision Tree Classification.**
5. **Explain the principle of the gradient descent algorithm. Accompany your explanation with a diagram. Explain the use of all the terms and constants that you introduce and comment on the range of values that they can take.**
6. **Explain the following**
   a) Linear regression
   b) Logistic Regression

UNIT IV ENSEMBLE TECHNIQUES AND UNSUPERVISED LEARNING

Combining multiple learners: Model combination schemes, Voting, Ensemble Learning - bagging, boosting, stacking, Unsupervised learning: K-means, Instance Based Learning: KNN, Gaussian mixture models and Expectation maximization

PART - A

1. **What is bagging and boosting in ensemble learning?**

Bagging is a way to decrease the variance in the prediction by generating additional data for training from dataset using combinations with repetitions to produce multi-sets of the original data. Boosting is an iterative technique which adjusts the weight of an observation based on the last classification.

2. **What is stacking in ensemble learning?**

Stacking is one of the most popular ensemble machine learning techniques used to predict multiple nodes to build a new model and improve model performance. Stacking enables us to train multiple models to solve similar problems, and based on their combined output, it builds a new model with improved performance.

3. **Which are the three types of ensemble learning?**

The three main classes of ensemble learning methods are bagging, stacking, and boosting, and it is important to both have a detailed understanding of each method and to consider them on your predictive modeling project.

4. **Why ensemble methods are used?**

There are two main reasons to use an ensemble over a single model, and they are related; they are: Performance: An ensemble can make better predictions and achieve better performance than any single contributing model. Robustness: An ensemble reduces the spread or dispersion of the predictions and model performance.

5. **What is a voting classifier?**

A voting classifier is a machine learning estimator that trains various base models or estimators and predicts on the basis of aggregating the findings of each base estimator. The aggregating criteria can be combined decision of voting for each estimator output

6. **What type of classifiers are used in weighted voting method?**

The performance-weighted-voting model integrates five classifiers including logistic regression, SVM, random forest, XGBoost and neural networks. We first used cross-validation to get the predicted results for the five classifiers.

7. **What is difference between K means and Gaussian mixture?**

K-Means is a simple and fast clustering method, but it may not truly capture heterogeneity inherent in Cloud workloads. Gaussian Mixture Models can discover complex patterns and group them into cohesive, homogeneous components that are close representatives of real patterns within the data set.

8. **What are Gaussian mixture models How is expectation maximization used in it?**

Expectation maximization provides an iterative solution to maximum likelihood estimation with latent variables. Gaussian mixture models are an approach

to density estimation where the parameters of the distributions are fit using the expectation-maximization algorithm.

9. What is k-means unsupervised learning?

K-Means clustering is an unsupervised learning algorithm. There is no labeled data for this clustering, unlike in supervised learning. K-Means performs the division of objects into clusters that share similarities and are dissimilar to the objects belonging to another cluster. The term 'K' is a number.

10. What is the difference between K-means and KNN?

KNN is a supervised learning algorithm mainly used for classification problems, whereas K-Means (aka K-means clustering) is an unsupervised learning algorithm. K in K-Means refers to the number of clusters, whereas K in KNN is the number of nearest neighbors (based on the chosen distance metric).

11. What is expectation maximization algorithm used for?

The EM algorithm is used to find (local) maximum likelihood parameters of a statistical model in cases where the equations cannot be solved directly. Typically these models involve latent variables in addition to unknown parameters and known data observations.

12. What is the advantage of Gaussian process?

Gaussian processes are a powerful algorithm for both regression and classification. Their greatest practical advantage is that they can give a reliable estimate of their own uncertainty.

13. What are examples of unsupervised learning?

Some examples of unsupervised learning algorithms include K-Means Clustering, Principal Component Analysis and Hierarchical Clustering.

14. How do you implement expectation maximization algorithm?

The two steps of the EM algorithm are:

E-step: perform probabilistic assignments of each data point to some class based on the current hypothesis h for the distributional class parameters;

M-step: update the hypothesis h for the distributional class parameters based on the new data assignments.

15. What is the principle of maximum likelihood?

The principle of maximum likelihood is a method of obtaining the optimum values of the parameters that define a model. And while doing so, you increase the likelihood of your model reaching the “true” model.

Part – B

1. **Explain briefly about unsupervised learning structure?**
2. **Explain various learning techniques involved in unsupervised learning?**
3. **What is Gaussian process? And explain in detail of Gaussian parameter estimates with suitable examples.**
4. **Explain the concepts of clustering approaches. How it differ from classification.**
5. **List the applications of clustering and identify advantages and disadvantages of clustering algorithm.**
6. **Explain about EM algorithm.**

7. **List non-parametric techniques and Explain K-nearest neighbour estimation.**

UNIT V NEURAL NETWORKS

Perceptron - Multilayer perceptron, activation functions, network training – gradient descent optimization – stochastic gradient descent, error backpropagation, from shallow networks to deep networks –Unit saturation (aka the vanishing gradient problem) – ReLU, hyperparameter tuning, batch normalization, regularization, dropout.

1. **What is perceptron and its types?**

A Perceptron is an Artificial Neuron. It is the simplest possible Neural Network. Neural Networks are the building blocks of Machine Learning.

2. **Which activation function is used in multilayer perceptron?**

Image result for Perceptron - Multilayer perceptron, activation functions

The Sigmoid Activation Function: Activation in Multilayer Perceptron Neural Networks.

3. **What are the activation functions of MLP?**

In MLP and CNN neural network models, ReLU is the default activation function for hidden layers. In RNN neural network models, we use the sigmoid or tanh function for hidden layers. The tanh function has better performance. Only the identity activation function is considered linear.

4. **Does MLP have activation function?**

Multilayer perceptrons (MLP) has been proven to be very successful in many applications including classification. The activation function is the source of the MLP power. Careful selection of the activation function has a huge impact on the network performance.

5. **What is the difference between a perceptron and a MLP?**

The Perceptron was only capable of handling linearly separable data hence the multi-layer perception was introduced to overcome this limitation. An MLP is a neural network capable of handling both linearly separable and non-linearly separable data.

6. **What are the types of activation function?**

Popular types of activation functions and when to use them

- Binary Step Function
- Linear Function
- Sigmoid
- Tanh
- ReLU
- Leaky ReLU
- Parameterised ReLU
- Exponential Linear Unit

7. **What is MLP and how does it work?**

A multilayer perceptron (MLP) is a feedforward artificial neural network that generates a set of outputs from a set of inputs. An MLP is characterized by several layers of input nodes connected as a directed graph between the input and output layers. MLP uses backpropogation for training the network.

8. Why do you require Multilayer Perceptron?

MLPs are useful in research for their ability to solve problems stochastically, which often allows approximate solutions for extremely complex problems like fitness approximation.

9. What are the advantages of Multilayer Perceptron?

Advantages of Multi-Layer Perceptron:

A multi-layered perceptron model can be used to solve complex non-linear problems.

It works well with both small and large input data.

It helps us to obtain quick predictions after the training.

It helps to obtain the same accuracy ratio with large as well as small data.

10. What do you mean by activation function?

An activation function is a function used in artificial neural networks which outputs a small value for small inputs, and a larger value if its inputs exceed a threshold. If the inputs are large enough, the activation function "fires", otherwise it does nothing.

11. What are the limitations of perceptron?

Perceptron networks have several limitations. First, the output values of a perceptron can take on only one of two values (0 or 1) because of the hard-limit transfer function. Second, perceptrons can only classify linearly separable sets of vectors.

12. How many layers are there in perceptron?

This is known as a two-layer perceptron. It consists of two layers of neurons. The first layer is known as hidden layer, and the second layer, known as the output layer, consists of a single neuron.

13. is stochastic gradient descent same as gradient descent?

Compared to Gradient Descent, Stochastic Gradient Descent is much faster, and more suitable to large-scale datasets. But since the gradient it's not computed for the entire dataset, and only for one random point on each iteration, the updates have a higher variance.

14. How is stochastic gradient descent used as an optimization technique?

Stochastic gradient descent is an optimization algorithm often used in machine learning applications to find the model parameters that correspond to the best fit between predicted and actual outputs. It's an inexact but powerful technique. Stochastic gradient descent is widely used in machine learning applications.

15. Does stochastic gradient descent lead to faster training?

Gradient Descent is the most common optimization algorithm and the foundation of how we train an ML model. But it can be really slow for large datasets. That's why we use a variant of this algorithm known as Stochastic Gradient Descent to make our model learn a lot faster.

16. What is stochastic gradient descent and why is it used in the training of neural networks?

Stochastic Gradient Descent is an optimization algorithm that can be used to train neural network models. The Stochastic Gradient Descent algorithm requires gradients to be calculated for each variable in the model so that new values for the variables can be calculated.

17. What are the three main types gradient descent algorithm?

There are three types of gradient descent learning algorithms: batch gradient descent, stochastic gradient descent and mini-batch gradient descent.

18. What are the disadvantages of stochastic gradient descent?

SGD is much faster but the convergence path of SGD is noisier than that of original gradient descent. This is because in each step it is not calculating the actual gradient but an approximation. So we see a lot of fluctuations in the cost.

19. How do you solve the vanishing gradient problem within a deep neural network?

The vanishing gradient problem is caused by the derivative of the activation function used to create the neural network. The simplest solution to the problem is to replace the activation function of the network. Instead of sigmoid, use an activation function such as ReLU

20. What is the problem with ReLU?

Key among the limitations of ReLU is the case where large weight updates can mean that the summed input to the activation function is always negative, regardless of the input to the network. This means that a node with this problem will forever output an activation value of 0.0. This is referred to as a "dying ReLU"

21. Why is ReLU used in deep learning?

The ReLU function is another non-linear activation function that has gained popularity in the deep learning domain. ReLU stands for Rectified Linear Unit. The main advantage of using the ReLU function over other activation functions is that it does not activate all the neurons at the same time.

22. Why is ReLU better than Softmax?

As per our business requirement, we can choose our required activation function. Generally , we use ReLU in hidden layer to avoid vanishing gradient problem and better computation performance , and Softmax function use in last output layer .

Part – B

1. **Draw the architecture of a single layer perceptron (SLP) and explain its operation. Mention its advantages and disadvantages.**
2. **Draw the architecture of a Multilayer perceptron (MLP) and explain its operation. Mention its advantages and disadvantages.**
3. **Explain the stochastic optimization methods for weight determination.**
4. **Describe back propagation and features of back propagation.**
5. **Write the flowchart of error back-propagation training algorithm.**
6. **Develop a Back propagation algorithm for Multilayer Feed forward neural network consisting of one input layer, one hidden layer and output layer from first principles.**
7. **List the factors that affect the performance of multilayer feed-forward neural network.**
8. **Difference between a Shallow Net & Deep Learning Net.**
9. **How do you tune hyperparameters for better neural network performance? Explain in detail.**