Recent Trends in Artificial Intelligence And Machine Learning for Smart Computing Age

STUDY MATERIAL

Once the stuff of science fiction novels and futuristic movies, Artificial Intelligence (AI) is now very real to us. From business applications to everyday life, we are almost unaware that many of us interact with AI every day. Nearly 60 percent (57.9) of organizations with Big Data solutions are using AI in some way; it's predicted that AI and machine learning will impact all segments of our daily lives by 2025 with huge implications for industries ranging from transport and logistics to healthcare, home maintenance, and customer service.

1. **What is the philosophy behind Artificial Intelligence?**

With the exploding capabilities of the computer system, it led to the question, "Can a machine think and behave like humans do? Can machines have the same intelligent mechanism humans?" This curiosity led to the development of AI. AI was rooted in the philosophy that machines can behave in a similar intelligent manner, with the add-on benefits of sophisticated automation.

2. What are some advantages of Artificial Intelligence?

a. Low error rate compared to humans, if coded properly. It is capable of very high precision, accuracy, and speed.
b. Can replace humans in repetitive and tedious tasks, thus saving on time and resources.
c. Its ability to predict a human is very significant, as in applications like acting as "assistants".
d. Unlike humans, AI can think logically without emotions, thus making rational decisions with near-zero mistakes
e. It has the ability to assess people intuitively, which is used in health sector applications.
f. Can organise and manage records in an intuitive manner.
g. It has applications in daily lives and across multiple scenarios, like browser search engines or banking fraud detection.
h. As AI is not affected by hostile environments, it can perform dangerous and risk-ridden tasks as in mining, and exploration in space.
i. Can interact with humans to provide entertainment as avatars or robots. For instance, video games.
j. Can do repetitive jobs without any breaks

3. Name some common uses and applications for AI?

This question tests your understanding of the AI field and how well you have grasped the far-reaching applications of AI.

When going for an interview, it is expected that you know about the company and its business. So if possible, highlight uses that are relevant to the interviewing company. It can earn your brownie points!

Applications and/or use cases:

- object detection and classification in navigation
- image recognition and tagging
- predictive maintenance
- data processing
- automation of manual tasks
- data-driven reporting
- natural language processing

- chatbots
- sentiment analysis
- sales prediction
- self-driving cars
- facial expression recognition
- gaming
- speech recognition

4. Why is image recognition a key function of AI?

AI mimics human. As humans are visual, AI is designed to imitate human brains. Teaching machines to recognize and categorise images helps machines to learn and become intuitive. With more and more images processed, AI becomes highly proficient in recognising and processing those images, whether objects, people, places, writing or photographs. The image recognition function of AI is most important today, as it finds widespread applications in daily lives — in security systems, driverless cars, navigation, search engines, robots in logistics, or medical imaging,

5. How is Game Theory related to AI?

Game theory is a framework of strategic hypothetical situations among competing players. AI uses game theory to evaluate potential actions of opponents where actions have a cost and some value. For instance, when writing an 'agent software' to bid for auctions, the agent has to be intelligent to understand the game theory and strategy working behind it.

6. What are the Common Uses and Applications of AI?

Your answer here should show that you recognize the far-reaching and practical applications of AI. Still, your answer is up to you because of your understanding of the AI field is what the interviewer is trying to ascertain. If possible, mention those uses most relevant to the potential employer. Possibilities include contract analysis, object detection, and classification for avoidance and navigation, image recognition, content distribution, predictive maintenance, data processing, automation of manual tasks, or data-driven reporting.

7. What are Intelligent Agents, and How are They Used in AI?

Intelligent agents are autonomous entities that use sensors to know what is going on, and then use actuators to perform their tasks or goals. They can be simple or complex and can be programmed to learn to accomplish their jobs better.

8. What is Tensorflow, and What is It Used For?

TensorFlow is an open-source software library initially developed by the Google Brain Team for use in machine learning and neural networks research. It is used for data-flow programming. TensorFlow makes it much easier to build certain AI features into applications, including natural language processing and speech recognition.

9. What is Machine Learning, and How Does It Relate to AI?

Machine learning is a subset of AI. The idea is that machines will "learn" and get better at tasks over time rather than having humans continually having to input parameters. Machine learning is a practical application of AI.

10. What are Neural Networks, and How Do They Relate to AI?

Neural networks are a class of machine learning algorithms. The neuron part of the neural is the computational component, and the network part is how the neurons are connected. Neural networks pass data among themselves, gathering more and more meaning as the data moves along. Because the networks are interconnected, more complex data can be processed more efficiently.

11. What is Deep Learning, and How Does It Relate to AI?

Deep learning is a subset of machine learning. It refers to using multi-layered neural networks to process data in increasingly sophisticated ways, enabling the software to train itself to perform tasks like speech and image recognition through exposure to these vast amounts of data for continual improvement in the ability to recognize and process information. Layers of neural networks stacked on top of each for use in deep learning are called deep neural networks.

12. What is a Bayesian Network, and How Does It Relate to AI?

A Bayesian network is a graphical model for probabilistic relationships among a set of variables. It mimics the human brain in processing variables.

13. What is Supervised Versus Unsupervised Learning?

This is one of the next important AI questions. Supervised learning is a machine learning process in which outputs are fed back into a computer for the software to learn from, for more accurate results the next time. With supervised learning, the "machine" receives initial training to start. In contrast, unsupervised learning means a computer will learn without initial training to base its knowledge.

14. What are some common misunderstandings about AI?

Since the beginning of the development of artificial intelligence, there have been a number of misunderstandings regarding it. The following are examples of some of these common misunderstandings:

- ***AI Does Not Need Humans***

AI's initial misunderstanding is that it can function without the assistance of humans. But in practice, each AI-based system still relies on humans, and they will continue to do so for the foreseeable future. Human-gathered data is needed to get insight into the information.

- ***AI is Harmful to Humanity***

As long as AI isn't able to outperform humans, it isn't a threat to our survival. It is impossible for a strong technology to be destructive if it is handled properly.

- ***AI Has Attained Its Pinnacle***

However, there is still a significant distance between us and the most advanced level of AI. Getting to the peak of the ridge will be an extremely difficult and lengthy trek.

- ***AI Will Overtake Your Job***

One of the most common misconceptions is that artificial intelligence will eliminate most of the employment, yet the technology is really creating more opportunities for new professions.

- ***AI is a Novel Technological Advance***

This technology was originally conceived for the first time in the year 1840 via an English newspaper, despite the fact that some individuals believe that it is a new kind of technology.

15. What Role Does Computer Vision Play in AI?

Artificial intelligence (AI) is broken down into a number of subfields, one of which is known as computer vision. Computer vision is the process of teaching computers to understand and collect data from the visual environment, such as graphics. Therefore, AI technology is used by computer vision in order to address complicated challenges such as image analysis, object identification, and other similar issues.

16. How Does the Strong AI Differ From the Weak AI?

- ***Strong AI***

The goal of strong artificial intelligence is to create actual intelligence artificially, which refers to an intellect created by humans that possesses feelings, consciousness, and emotions comparable to those of humans. The idea of creating AI entities with perceiving, analyzing, and decision-making skills comparable to those of humans is still only an assumption at this point.

- ***Weak AI***

The present phase of artificial intelligence research is known as "weak AI," and it is concerned with the construction of expert systems and robots that can assist people and solve challenging real-world issues. Weak artificial intelligence systems like Alexa and Siri are examples.

17. How Can Artificial Intelligence Be Used to Identify Fraud?

This is one of the next important AI questions. It is possible to use artificial intelligence in fraud detection utilizing various machine learning techniques (e.g., supervised and unsupervised). Machine learning's rule-based algorithms may be used to identify and stop fraudulent transactions. Machine learning is used to identify fraud in the following ways:

- ***Extracting Data***

Data extraction is the initial stage. Web scraping technologies and surveys are used to collect data. The sort of model we want to build dictates the type of data we gather. Personal information, transactions, and shopping may all be found here.

- ***Data Cleaning***

This stage eliminates any information that was deemed unnecessary or duplicated. Because of the data's inherent unreliability, incorrect predictions may be made.

- ***Data Analysis and Exploration***

This is a critical step in determining the relationship between various predictor variables.

- ***Building Models***

The very last thing that has to be done is to construct the model by applying various machine learning algorithms to it, and this will depend on the requirements of the company.

18. Why Do We Utilize an Inference Engine in AI?

AI's inference engine extracts valuable learning from its knowledge base by following a set of predefined logical rules. For the most part, it operates in two distinct modes:

- ***Backward Chaining***

It starts with the end aim and then works backward to figure out the evidence that points in that direction.

- ***Forward Chaining***

It begins with facts that are already known and then claims new facts.

19. What are Different Platforms for Artificial Intelligence (AI) Development?

Some of the best Artificial Intelligence Platforms are Google AI Platform, Microsoft Azure, TensorFlow, Infosys Nia, Rainbird, Wipro HOLMES, Premonition, Dialogflow, Ayasdi, Meya, MindMeld, KAI, Wit, Vital A.I, Receptiviti, Lumiata, Watson Studio, and Infrrd.

20. What are the Programming Languages Used for Artificial Intelligence?

Prolog (generic core, modules) is an early 1970s logic programming language that is particularly well suited for artificial intelligence applications. Python is presently the most popular language. Others:

- R.
- Julia.
- Java and
- C++.

Python is the most popular AI programming language; it's one of the trendiest languages out there, and it's also simple to learn! Python is a high-level, interpreted programming language with dynamic semantics. For quick development, it is particularly appealing because of its high-level data structures and dynamic typing.

21. What is the Future of Artificial Intelligence?

Machine learning and natural language processing are projected to advance further in the artificial intelligence future (AI), resulting in the creation of more complex and autonomously AI systems. These systems may be used in a wide range of applications, such as autonomous vehicles, personal assistants, and intelligent robots. Additionally, AI is expected to play a significant role in areas such as healthcare, finance, and manufacturing. However, as AI becomes more advanced and integrated into society, it is also important to consider the ethical and societal implications of this technology and to ensure that it is developed and used responsibly.

22. What is the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning?

Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) are related but distinct fields of study.

Artificial Intelligence (AI) is a broad field that encompasses a variety of techniques and approaches for creating intelligent systems that can perform tasks that typically require human intelligence, such as understanding natural language, recognizing speech, and making decisions.

Machine Learning (ML) is a subset of AI that involves the development of algorithms and statistical models that enable systems to improve their performance over time by learning from data. Machine learning algorithms can be categorized into three main categories: supervised learning, unsupervised learning, and reinforcement learning.

Deep Learning (DL) is a subset of ML that involves the use of neural networks, a type of model inspired by the structure and function of the human brain, to learn from data. Deep learning techniques are particularly well-suited for tasks such as image and speech recognition, and are often used in natural language processing and computer vision applications.

In summary, AI is the broad field of creating intelligent systems, ML is a subset of AI that uses algorithms to learn from data and make predictions, and DL is a subset of ML that uses neural networks to learn from data.

23. How are Artificial Intelligence and Machine Learning Related?

This is one of the most basic, yet most important AI questions. Artificial Intelligence (AI) is a broad field that encompasses a variety of techniques and approaches for creating intelligent systems that can perform tasks that typically require human intelligence, such as recognizing speech, understanding natural language, and making decisions.

Machine Learning (ML), on the other hand, is a specific approach to achieving AI. It involves the development of algorithms and statistical models that enable systems to improve their performance over time by learning from data. Machine learning algorithms can be categorized into three main categories: supervised learning, unsupervised learning, and reinforcement learning.

24. What are Different Types of Machine Learning?

Supervised learning: An example of supervised learning would be when a model was trained on a labeled dataset, with the best outputs provided for each input. The model then uses this labeled dataset to make predictions on new, unseen data. Eg: linear regression, and support vector machines.

Unsupervised learning: In this type of machine learning, the model is not provided with labeled data, and instead must find patterns or structure in the input data on its own. Examples of unsupervised learning algorithms include k-means clustering, hierarchical clustering, and Principal Component Analysis (PCA).

Reinforcement learning: In this type of machine learning, the model learns by interacting with its environment and receiving feedback in the form of rewards or penalties. The model's goal is to learn a policy that maximizes the cumulative reward. Examples of reinforcement learning algorithms include Q-learning and SARSA.

Semi-supervised learning: a mixture of supervised and unsupervised learning, it uses a small amount of labeled data and a large amount of unlabelled data to learn the patterns.

Self-supervised learning: a type of machine learning where the model learns from a pre-defined task using unlabelled data.

25. What is Reinforcement Learning, and How Does It Work?

Reinforcement learning (RL) is a type of machine learning where an agent learns to make decisions in an environment by interacting with it and receiving feedback in the form of rewards or penalties. To maximize its cumulative reward over time, the agent must learn a policy that maps environmental states to actions.

26. Explain Markov's Decision Process.

A mathematical framework called the Markov Decision Process (MDP) is used to describe decision-making in circumstances where the result is partially determined by chance and partially controlled by the decision-maker. MDPs are widely used in the field of reinforcement learning as they provide a way to model an agent's decision-making problem.

An MDP is defined by a set of states, a set of actions, a transition function that defines the probability of going from one state to another, a reward function that defines the immediate reward for being in a particular state and taking a particular action, and a discount factor that determines the importance of future rewards.

27. What is Overfitting?

This is one of the next important AI questions. Overfitting in AI occurs when a machine learning model becomes too complex and starts to fit the training data too closely, to the point where it

memorizes the training data rather than learning the underlying patterns and relationships. This means that the model performs very well on the training data, but poorly on new, unseen data.

Overfitting can occur in any machine learning algorithm, and it can happen when the model is too complex relative to the amount and quality of training data available. In some cases, the model may even start to fit the noise in the data, rather than the underlying patterns. This can result in poor performance and accuracy when the model is used for prediction or classification tasks on new data.

To prevent overfitting, it is important to use techniques like regularization, cross-validation, and early stopping during the training process. These techniques can help to prevent the model from becoming too complex and help to ensure that it generalizes well to new, unseen data.

28. What are the Techniques Used to Avoid Overfitting?

Cross-validation: This is a technique where the data is split into multiple subsets, and the model is trained and tested on different subsets. This helps to prevent the model from memorizing the training data and generalizing poorly to new data.

Regularization: This is a technique where a penalty term is added to the model's objective function, which discourages the model from assigning too much importance to any single feature. This helps to prevent the model from fitting to noise in the training data.

Early stopping: This is a technique where the training process is stopped before the model's performance on the training data starts to decrease, this is useful when the model is trained with multiple iterations.

Ensemble methods: This is a technique where multiple models are trained, and their predictions are combined to create a final prediction. This helps to reduce the variance and increase the robustness of the model.

Pruning: This is a technique where the complexity of the model is reduced by removing unimportant features or nodes.

Dropout: This is a technique where a random subset of the neurons is dropped out of the network during training, this prevents the network from relying too much on any one neuron.

Bayesian approaches: This is a technique where prior information is incorporated into the model's parameters.

29. What is Natural Language Processing?

Natural Language Processing (NLP) is a field of artificial intelligence and computer science that focuses on the interaction between computers and humans in natural language. NLP involves using techniques from computer science, linguistics, and mathematics to process and analyze human language.

30. What is the Difference Between Natural Language Processing and Text Mining?

Natural Language Processing (NLP) and Text Mining are related fields that focus on the analysis and understanding of human language, but they have some key differences.

NLP is a branch of artificial intelligence that focuses on the interaction between computers and humans in natural language. It involves using techniques from computer science, linguistics, and mathematics to process and analyze human language. NLP tasks include speech recognition, natural language understanding, natural language generation, machine translation, and sentiment analysis.

Text Mining, on the other hand, is a broader field that involves the use of NLP techniques to extract valuable information from unstructured text data. Text Mining often used in business, social science, and information science. It includes tasks such as information retrieval, text classification, text clustering, text summarization, and entity recognition.

In summary, NLP is a field of AI that deals with the interactions of computers and human languages, while Text Mining is a broader field that deals with the extraction of insights and knowledge from unstructured text data using NLP techniques.

31. Explain SVM and why it is called as maximum margin classifier.

Support Vector Machine (SVM) is a supervised machine learning algorithm, used for both, classification and regression. It sorts the data into one of two categories, and outputs a map of the sorted data with the margins between the two data points as far apart as possible.

It is known as maximum margin classifier because in a binary classification dataset, it places the decision boundary such that the distance between the two clusters is maximised. The SVM aims to find a separating hyperplane between positive and negative instances. It establishes the largest margin to avoid overfitting.

32. Differentiate between Precision and Recall.

Precision and Recall are model evaluation metrics that measure Relevance of results. They are used in pattern recognition, information retrieval and binary classification.

a) Precision means the percentage of the results that are relevant. On the other hand, Recall refers to the percentage of total relevant results that have been retrieved over the total amount of relevant instances.

For example, in a text search on a set of documents, Precision is the fraction of retrieved documents that are relevant to the query: However, Recall is the fraction of the relevant documents that are successfully retrieved.

b) Precision attempts to answer the following question:

What proportion of positive identifications was actually correct?

Recall attempts to answer the following question:

What proportion of actual positives was identified correctly?

c) Precision and Recall are opposite of each other, i.e. increasing one of them reduces the other and vice versa.

d) In simple terms, high precision means that an algorithm returned substantially more relevant results than irrelevant ones, while high recall means that an algorithm returned most of the relevant results

e) Precision measure the quality or relevancy of results of the model. Recall measures the quantity of positives or relevant result returned by the model.

33. How do you choose an algorithm for a classification problem?

There is no one solution fits all. Several factors go into the choice of a machine learning algorithm. The choice depends upon the level of accuracy required, and the size of the training set. Here is a sample answer.

The method followed would be:

a) Define the problem.

Herein, the output of the model is a class, as it is a classification problem.

b) Identify available algorithms from linear and non-linear classifiers

- Logistic Regression.
- Linear Discriminant Analysis
- k-Nearest Neighbours.
- Classification and Regression Trees.
- Naive Bayes.
- Support Vector Machines.

c) Implement all of them.

Next, set up a machine learning pipeline that compares the performance of each algorithm on the dataset, using a set of evaluation criteria or chosen metrics. The best performing one would be selected. Depending upon the results, either it would be run once or in intervals when new data is added.

d) Improve results using various optimisation methods

By using cross-validation (like k-fold) and hyperparameter tuning or Ensembling (bagging, boosting, etc); each algorithm would be tuned to optimise performance, if time is not a constraint. Otherwise, manually select the hyperparameters.

34. If your model suffers from low bias and high variance, which algorithm would you use to tackle it? Why?

The error of a model can either be of bias and/or variance. Very low bias but high variance indicates overfitting, as well as complexity of the model. By averaging these out, we can reduce the variance and increase the bias.

a) A bagging algorithm can handle the high variance. The dataset is randomly subsampled *mm* times and the model trained using each subsample. Then the models are averaged by averaging out the predictions of each mode.

b) By using the k-nearest neighbour algorithm, the trade-off between bias and variance can be achieved. The value of k is increased to increase the number of neighbours that contribute to the prediction, and this in turn increases the bias of the model.

c) By using the support vector machine algorithm, the trade-off can be achieved by increasing the C parameter that influences the number of violations of the margin allowed in the training data, and this in turn increases the bias but decreases the variance.

35. What is Fuzzy Logic?

In the real world, we often encounter situations when we cannot determine whether the state is true or false. In such cases, fuzzy logic provides a valuable reasoning method that closely reflects human reasoning. The approach considers the inaccuracies and uncertainties of any situation which humans engage in to consider possibilities. Thus, fuzzy logic is based on "degrees of truth" rather than the usual "true or false" (1 or 0) Boolean logic on which the modern computer is based. As a subset of AI, it encodes human learning for artificial processing and is represented as IF-THEN rules.

36. What are some methods of reducing dimensionality?

Popular Techniques include:

i) Missing Values Ratio — When data columns contain too many missing values, then removing those columns by setting a threshold for missing values.

ii) Low Variance Filter — When a data column has constant values, its variance would be 0 and such variables will not explain the variation in target variables.

iii) High Correlation Filter — When data columns are interdependent on each other and contain similar information it adds to the redundancy of the model. Highly correlated columns are thus identified using correlation coefficients.

iv) Random Forest — To tackle issues of missing values, outliers and most significant variables, the feature selection method is used to find most informative subset of features.

v) Backward Feature Elimination — Eliminates features that do not add value to the model, one at a time by checking the error rate after each elimination, till the maximum error rate tolerable is reached. The smallest number of features is then defined.

vi) Forward Feature Construction — Find the most significant features that improve the performance of the model, and adding them one at a time for model improvement.

vii) Principle Component Analysis (PCA) — Existing set of variables is transformed into a new set of variables, which is a linear combination of original variables.

viii) Factor Analysis — The variables are modelled as linear combinations of the potential factors, plus "error" terms. It is based on the assumption that there exists several unobserved latent variables that account for the correlations among observed variables.

ix) t-Distributed Stochastic Neighbour Embedding (t-SNE) — It considers the probability that pairs of data points in the high-dimensional space are related, and chooses low-dimensional embeddings that produce a similar distribution.

x) ISOMAP — Uses a matrix of pair-wise distances between all points and computes a position for each point. Then ISOMAP uses classic multi-dimensional scaling (MDS) to compute reduced-dimensional positions of the points

37. What is stratified cross-validation and when is it used?

Where there is a large imbalance in the response variables, the cross-validation technique is used to rearrange data between training and validation sets, so that the distribution in each fold has a good representation of the whole dataset. It forces each fold to have at least m instances of each class.

Stratified cross-validation is used in the following events:

1. When the dataset is small and has multiple categories, which creates an imbalance.
2. When the dataset has different distributions and validation is required to prevent a generalisation problem.

38. What is an imbalanced dataset? Can you list some ways to deal with it?

An imbalanced dataset is one where the distribution of data in the target categories is not uniform. For example, in email classification problem, there will typically be more spam mails than ham (relevant) mails. The class imbalance may be as high as 70–95 % for the spam mail class, and 5–30 % % for the relevant mails. This disproportionate distribution of two classes of data is an imbalanced dataset.

Using an imbalanced dataset affects the performance and accuracy training model and needs modification.

Good ways to deal with imbalanced datasets should focus on correcting the imbalance, when there is no option to use another algorithm. Some ways are:

- Oversampling of minority class when the data is insufficient.
- Under sampling of majority class when there is a good quantity of data to work with.
- Collecting more data and adding the data in the lighter category to control the imbalance.
- Cluster-based oversampling so that all classes are of the same size and clusters of the same class have equal number of instances.
- Generate synthetic samples by randomly sampling the attributes from instances in the minority class and adding to the dataset.
- Resampling with different ratios between the rare and abundant class
- Using appropriate metrics to deal with the imbalance. For instance, precision, confusion, recall, and F-score, to ensure better accuracy of the model.

- Modifying existing classification algorithms and designing own models that work best with imbalanced datasets.

39. What is Ensemble learning?

Ensemble learning is a method that combines multiple machine learning models to create more powerful models.

There are many reasons for a model to be different. Few reasons are:

- Different Population
- Different Hypothesis
- Different modeling techniques

When working with the model's training and testing data, we will experience an error. This error might be bias, variance, and irreducible error.

Now the model should always have a balance between bias and variance, which we call a bias-variance trade-off.

This ensemble learning is a way to perform this trade-off.

There are many ensemble techniques available but when aggregating multiple models there are two general methods:

- Bagging, a native method: take the training set and generate new training sets off of it.
- Boosting, a more elegant method: similar to bagging, boosting is used to optimize the best weighting scheme for a training set.

40. What is a Random Forest? How does it work?

Random forest is a versatile machine learning method capable of performing both regression and classification tasks.

Like bagging and boosting, random forest works by combining a set of other tree models. Random forest builds a tree from a random sample of the columns in the test data.

Here are the steps how a random forest creates the trees:

- Take a sample size from the training data.
- Begin with a single node.
- Run the following algorithm, from the start node:
    - If the number of observations is less than node size then stop.
    - Select random variables.
    - Find the variable that does the "best" job of splitting the observations.
    - Split the observations into two nodes.
    - Call step `a` on each of these nodes.

41. What is Collaborative Filtering? And Content-Based Filtering?

Collaborative filtering is a proven technique for personalized content recommendations. Collaborative filtering is a type of recommendation system that predicts new content by matching the interests of the individual user with the preferences of many users.

Content-based recommender systems are focused only on the preferences of the user. New recommendations are made to the user from similar content according to the user's previous choices.

42. What is Clustering?

Clustering is the process of grouping a set of objects into a number of groups. Objects should be similar to one another within the same cluster and dissimilar to those in other clusters.

A few types of clustering are:

- Hierarchical clustering
- K means clustering
- Density-based clustering
- Fuzzy clustering, etc.

43. How can you select K for K-means Clustering?

There are two kinds of methods that include direct methods and statistical testing methods:

- Direct methods: It contains elbow and silhouette
- Statistical testing methods: It has gap statistics.

The silhouette is the most frequently used while determining the optimal value of k.

44. What are Recommender Systems?

A recommendation engine is a system used to predict users' interests and recommend products that are quite likely interesting for them.

Data required for recommender systems stems from explicit user ratings after watching a film or listening to a song, from implicit search engine queries and purchase histories, or from other knowledge about the users/items themselves.

45. How do check the Normality of a dataset?

Visually, we can use plots. A few of the normality checks are as follows:

- Shapiro-Wilk Test
- Anderson-Darling Test
- Martinez-Iglewicz Test
- Kolmogorov-Smirnov Test
- D'Agostino Skewness Test

46. Can logistic regression use for more than 2 classes?

No, by default logistic regression is a binary classifier, so it cannot be applied to more than 2 classes. However, it can be extended for solving multi-class classification problems **(multinomial logistic regression)**

47. Explain Correlation and Covariance?

Correlation is used for measuring and also for estimating the quantitative relationship between two variables. Correlation measures how strongly two variables are related. Examples like, income and expenditure, demand and supply, etc.

Covariance is a simple way to measure the correlation between two variables. The problem with covariance is that they are hard to compare without normalization.

48. What is P-value?

P-values are used to make a decision about a hypothesis test. P-value is the minimum significant level at which you can reject the null hypothesis. The lower the p-value, the more likely you reject the null hypothesis.

49. What are Parametric and Non-Parametric Models?

Parametric models will have limited parameters and to predict new data, you only need to know the parameter of the model.

Non-Parametric models have no limits in taking a number of parameters, allowing for more flexibility and to predict new data. You need to know the state of the data and model parameters.

50. What evaluation approaches would you work to gauge the effectiveness of a machine learning model?

Answer: You would first split the dataset into training and test sets, or perhaps use cross-validation techniques to further segment the dataset into composite sets of training and test sets within the data. You should then implement a choice selection of performance metrics: here is a fairly comprehensive list. You could use measures such as the F1 score, the accuracy, and the confusion matrix. What's important here is to demonstrate that you understand the nuances of how a model is measured and how to choose the right performance measures for the right situations.

51. Density-Based Clustering Algorithms

Density-Based **Clustering** refers to unsupervised learning methods that identify distinctive groups/clusters in the data, based on the idea that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a base algorithm for density-based clustering. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers.

The DBSCAN algorithm uses two parameters:

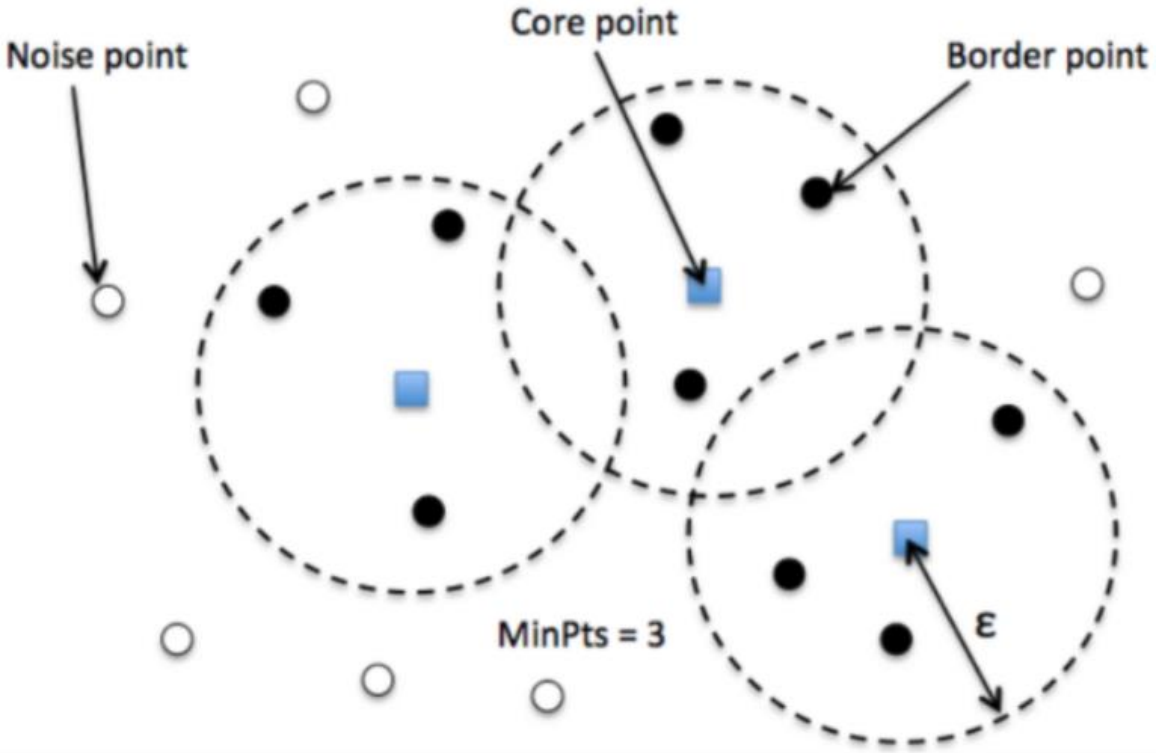
- **minPts:** The minimum number of points (a threshold) clustered together for a region to be considered dense.
- **eps (ε):** A distance measure that will be used to locate the points in the neighborhood of any point.

These parameters can be understood if we explore two concepts called Density Reachability and Density Connectivity.

Reachability in terms of density establishes a point to be reachable from another if it lies within a particular distance (eps) from it.

Connectivity, on the other hand, involves a transitivity based chaining-approach to determine whether points are located in a particular cluster. For example, p and q points could be connected if p->r->s->t->q, where a->b means b is in the neighborhood of a.

There are three types of points after the DBSCAN clustering is complete:



- Core — This is a point that has at least m points within distance n from itself.
- Border — This is a point that has at least one Core point at a distance n.
- Noise — This is a point that is neither a Core nor a Border. And it has less than m points within distance n from itself.

Algorithmic steps for DBSCAN clustering

The algorithm proceeds by arbitrarily picking up a point in the dataset (until all points have been visited). If there are at least 'minPoint' points within a radius of 'ε' to the point then we consider all these points to be part of the same cluster. The clusters are then expanded by recursively repeating the neighborhood calculation for each neighboring point

52. What are the different types of terminologies involved in Logistic Regression?

Here are some common terms involved in logistic regression:

- **Independent variables:** The input characteristics or predictor factors applied to the dependent variable's predictions.
- **Dependent variable:** The target variable in a logistic regression model, which we are trying to predict.
- **Logistic function:** The formula used to represent how the independent and dependent variables relate to one another. The logistic function transforms the input variables into a probability value between 0 and 1, which represents the likelihood of the dependent variable being 1 or 0.
- **Odds:** It is the ratio of something occurring to something not occurring. it is different from probability as the probability is the ratio of something occurring to everything that could possibly occur.
- **Log-odds:** The log-odds, also known as the logit function, is the natural logarithm of the odds. In logistic regression, the log odds of the dependent variable are modeled as a linear combination of the independent variables and the intercept.
- **Coefficient:** The logistic regression model's estimated parameters, show how the independent and dependent variables relate to one another.
- **Intercept:** A constant term in the logistic regression model, which represents the log odds when all independent variables are equal to zero.
- **Maximum likelihood estimation:** The method used to estimate the coefficients of the logistic regression model, which maximizes the likelihood of observing the data given the model.

53. What is an SVM?

Support vector machines are a set of supervised learning methods used for classification, regression, and outliers detection. All of these are common tasks in machine learning.

You can use them to detect cancerous cells based on millions of images or you can use them to predict future driving routes with a well-fitted regression model.

There are specific types of SVMs you can use for particular machine learning problems, like support vector regression (SVR) which is an extension of support vector classification (SVC).

The main thing to keep in mind here is that these are just math equations tuned to give you the most accurate answer possible as quickly as possible.

SVMs are different from other classification algorithms because of the way they choose the decision boundary that maximizes the distance from the nearest data points of all the classes. The decision boundary created by SVMs is called the maximum margin classifier or the maximum margin hyper plane.

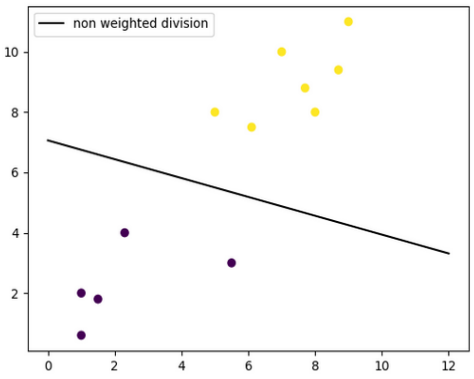
54. How an SVM works

A simple linear SVM classifier works by making a straight line between two classes. That means all of the data points on one side of the line will represent a category and the data points on the other side of the line will be put into a different category. This means there can be an infinite number of lines to choose from.

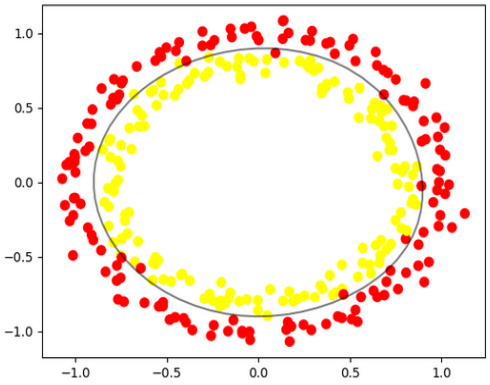
What makes the linear SVM algorithm better than some of the other algorithms, like k-nearest neighbors, is that it chooses the best line to classify your data points. It chooses the line that separates the data and is the furthest away from the closet data points as possible.

A 2-D example helps to make sense of all the machine learning jargon. Basically you have some data points on a grid. You're trying to separate these data points by the category they should fit in, but you don't want to have any data in the wrong category. That means you're trying to find the line between the two closest points that keeps the other data points separated.

So the two closest data points give you the support vectors you'll use to find that line. That line is called the decision boundary.



linear SVM



non-linear SVM using RBF kernel

55. Why SVMs are used in machine learning

SVMs are used in applications like handwriting recognition, intrusion detection, face detection, email classification, gene classification, and in web pages. This is one of the reasons we use SVMs in machine learning. It can handle both classification and regression on linear and non-linear data.

Another reason we use SVMs is because they can find complex relationships between your data without you needing to do a lot of transformations on your own. It's a great option when you are working with smaller datasets that have tens to hundreds of thousands of features. They typically find more accurate results when compared to other algorithms because of their ability to handle small, complex datasets.

Here are some of the pros and cons for using SVMs.

Pros

- Effective on datasets with multiple features, like financial or medical data.
- Effective in cases where number of features is greater than the number of data points.
- Uses a subset of training points in the decision function called support vectors which makes it memory efficient.
- Different kernel functions can be specified for the decision function. You can use common kernels, but it's also possible to specify custom kernels.

Cons

- If the number of features is a lot bigger than the number of data points, avoiding over-fitting when choosing kernel functions and regularization term is crucial.
- SVMs don't directly provide probability estimates. Those are calculated using an expensive five-fold cross-validation.
- Works best on small sample sets because of its high training time.

Since SVMs can use any number of kernels, it's important that you know about a few of them.