1. What does one mean by the term "machine learning"?

Ans:

"Machine learning" refers to a field of artificial intelligence (AI) that focuses on developing algorithms and models that enable computers to learn and make predictions or decisions without being explicitly programmed. It involves using statistical techniques to enable computers to analyze and interpret complex patterns and data, and then make informed decisions or predictions based on that analysis.

2.Can you think of 4 distinct types of issues where it shines?

Ans:

a. Image Recognition: Machine learning algorithms can be trained to accurately identify and classify objects within images, enabling applications such as facial recognition, object detection, and autonomous driving.

b. Natural Language Processing: Machine learning techniques can be applied to analyze and understand human language, enabling tasks such as sentiment analysis, language translation, and chatbot interactions.

c. Fraud Detection: Machine learning models can learn patterns of fraudulent behavior by analyzing large volumes of data, helping to identify and prevent fraudulent activities in various domains, such as finance, insurance, and e-commerce.

d. Recommendation Systems: Machine learning algorithms can analyze user preferences and behaviour to provide personalized recommendations, such as movie recommendations on streaming platforms, product recommendations on e-commerce websites, or content recommendations on social media.

3.What is a labeled training set, and how does it work?

Ans:

A labeled training set is a dataset used in supervised learning where each data instance (or example) is accompanied by a corresponding label or target value. The labels represent the desired output or outcome for a given input. During training, the machine learning algorithm learns from the labeled training set by adjusting its internal parameters to minimize the difference between its predicted outputs and the actual labels. This process allows the algorithm to generalize from the training data and make predictions on unseen data based on what it has learned.

4.What are the two most important tasks that are supervised?

Ans:

The two most important supervised tasks are:

a. Classification: In classification, the goal is to assign predefined categories or labels to input data instances. For example, classifying emails as spam or non-spam, or classifying images as different types of objects.

b. Regression: In regression, the goal is to predict a continuous numerical value based on input variables. For example, predicting housing prices based on features like size, location, and number of rooms.

5.Can you think of four examples of unsupervised tasks?

Ans:

Four examples of unsupervised tasks are:

a. Clustering: Grouping similar data together based on their inherent patterns or similarities. Examples include customer segmentation, document clustering, or image segmentation.

b. Dimensionality Reduction: Reducing the number of input variables while preserving important information. Techniques like Principal Component Analysis (PCA) or t-SNE can be used for visualizations or feature extraction.

c. Anomaly Detection: Identifying rare or abnormal instances in a dataset. This can be applied to detect fraudulent transactions, network intrusions, or equipment failures.

d. Recommender system

6.State the machine learning model that would be best to make a robot walk through various unfamiliar terrains?

Ans: Reinforcement Learning mode

7.Which algorithm will you use to divide your customers into different groups?

Ans:

To divide customers into different groups, a clustering algorithm such as K-means or hierarchical clustering can be used. These algorithms analyze patterns in the data and group customers based on similarities in their attributes or behaviours.

8.Will you consider the problem of spam detection to be a supervised or unsupervised learning problem?

Ans: Supervised

9.What is the concept of an online learning system?

Ans:

An online learning system refers to a machine learning system that can continuously update and adapt its model in real-time as new data becomes available. Unlike batch learning, where the model is trained on a fixed dataset and then applied to new data, online learning systems can learn incrementally from individual data instances or small batches of data. These systems are particularly useful in dynamic environments where data distribution or patterns may change over time. Online learning allows the model to adapt and respond quickly to new information, making it suitable for tasks like online recommendation systems, fraud detection.

10.What is out-of-core learning, and how does it differ from core learning?

Ans:

Out-of-core learning is a technique used when dealing with large datasets that cannot fit into the memory of a single machine. In out-of-core learning, the data is divided into smaller chunks or batches, and the learning algorithm processes these batches sequentially, updating the model parameters as it goes. This approach allows the model to learn from data that is stored on disk enabling the training of models on datasets that exceed the memory capacity of a single machine. In contrast, core learning refers to traditional machine learning approaches where the entire dataset is loaded into memory for training.

11.What kind of learning algorithm makes predictions using a similarity measure?

Ans:

These algorithms do not explicitly build a model during training but instead store the training data in memory. When making predictions on new instances, the algorithm measures the similarity between the new data and the stored training data and uses this similarity to make predictions. Examples include k-nearest neighbors (KNN) and locality-sensitive hashing (LSH).

12.What's the difference between a model parameter and a hyperparameter in a learning algorithm?

Ans:

In a learning algorithm, model parameters are the internal variables that are learned from the training data and are used to make predictions. These parameters capture the patterns and relationships present in the data. On the other hand, hyperparameters are external to the model and are set by the user before the learning process begins. Hyperparameters control the behavior and performance of the learning algorithm, such as the learning rate, regularization strength, or the number of hidden units in a neural network. They are typically tuned or optimized through techniques like cross-validation to find the best configuration for the given problem.

13.What are the criteria that model-based learning algorithms look for? What is the most popular method they use to achieve success? What method do they use to make predictions?

Ans:

Model-based learning algorithms look for criteria such as accuracy, generalization ability, simplicity, and computational efficiency. The most popular method used by model-based learning algorithms is the minimization of a loss function, which measures the difference between the predicted outputs and the true labels in the training data. This minimization is often achieved through optimization algorithms like gradient descent. Once the model is trained, it can make predictions by applying the learned model parameters to new unseen data.

14.Can you name four of the most important Machine Learning challenges?

Ans:

a. Overfitting: This occurs when a model performs well on the training data but fails to generalize to new, unseen data. Overfitting can lead to poor performance and lack of generalization ability.

b. Data scarcity: When the amount of available training data is limited, it becomes challenging to build accurate and robust models. Insufficient data may result in models that are prone to high variance or biased representations.

c. Feature engineering: Selecting and transforming the relevant features from the raw data is a critical challenge. It requires domain knowledge, creativity, and understanding of the problem to identify the most informative and discriminative features.

d. Interpretability and explainability: As machine learning models become more complex, understanding and interpreting their decisions become challenging

15.What happens if the model performs well on the training data but fails to generalize the results to new situations? Can you think of three different options?

Ans:

If a model performs well on the training data but fails to generalize to new situations, it means that the model has overfit the training data. This can happen due to various reasons such as the model being too complex or the training data not being representative of the real-world scenarios. Here are three different options to address this issue:

a. Regularization: Apply regularization techniques such as L1 or L2 regularization to the model. This helps in reducing overfitting by adding a penalty term to the loss function, discouraging the model from relying too heavily on specific features or parameters.

b. Feature selection: Review and select the most relevant features for training the model. Removing irrelevant or noisy features can help improve the model's ability to generalize.

c. Increase training data: Collect more diverse and representative training data. A larger and more varied dataset can provide the model with a better understanding of different scenarios, enabling it to generalize better.

16.What exactly is a test set, and why would you need one?

Ans:

A test set is a portion of the data that is not used during model training or hyperparameter tuning. It is used to assess the performance of the trained model on unseen data. The purpose of a test set is to provide an unbiased evaluation of the model's generalization ability. By evaluating the model on the test set, we can estimate how well the model is likely to perform in real-world scenarios.

17.What is a validation set's purpose?

Ans:

The purpose of a validation set is to fine-tune the hyperparameters of a model. During the training process, the model is evaluated on the validation set using different hyperparameter configurations. The performance on the validation set helps in selecting the best hyperparameter values that yield the optimal performance of the model. The validation set serves as an intermediate step between the training set and the final evaluation on the test set.

18.What precisely is the train-dev kit, when will you need it, how do you put it to use?

Ans:

The train-dev kit, or training development kit, refers to a subset of the training data that is set aside for model development and evaluation. It is used to iteratively refine the model during the development process. The train-dev kit helps in assessing the model's performance on a separate dataset, ensuring that it is not overfitting to the training data. It is typically used when the model is being developed and refined, and it helps in making informed decisions about model improvements and adjustments.

19.What could go wrong if you use the test set to tune hyperparameters?

Ans:

There will be data leakage purpose of test is evaluate the tuned model and check how well it is performing on unseen data i.e. its ability to generalise. If we use test data to tune the model we may get a very good test accuracy but it will perform poorly on unseen data.