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

**ANSWER:** Machine learning refers to teaching computers to learn and make decisions on their own, without being explicitly programmed for each specific task. Instead of giving step-by-step instructions, we provide computers with lots of data and let them figure out patterns and rules by themselves. It's like teaching a computer to recognize cats by showing it many pictures of cats. Once the computer learns, it can identify cats in new pictures it has never seen before. Machine learning helps computers become smarter and more useful in solving different kinds of problems.

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

**ANSWER:** Here are four distinct types of issues where machine learning shines:

Image Recognition: Machine learning excels in tasks related to image recognition. It can be used to build models that can accurately identify objects, people, or specific features within images. For example, it can be used in facial recognition systems, autonomous vehicles, or even medical imaging to detect diseases.

Natural Language Processing: Machine learning is highly effective in processing and understanding human language. It can be used for tasks such as sentiment analysis, language translation, chatbots, and voice recognition. Natural language processing models can learn to understand and generate human-like text, enabling applications like virtual assistants and language translation services.

Fraud Detection: Machine learning is valuable in detecting fraudulent activities or anomalies in large volumes of data. By learning patterns from historical data, machine learning models can identify suspicious transactions, fraudulent behaviour, or potential cybersecurity threats. This is particularly useful in financial institutions, e-commerce platforms, and online payment systems.

Personalized Recommendations: Machine learning is widely employed in recommendation systems. By analyzing user behaviour and preferences, it can suggest personalized content, products, or services to individuals. For instance, streaming platforms use machine learning algorithms to recommend movies or TV shows based on a user's viewing history, ratings, and similar users' preferences.

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

**ANSWER:** A labeled training set is a collection of data used in supervised machine learning. It consists of input data (features) and corresponding output labels or target values. The labels provide information about the correct or desired output for each input data point. The purpose of a labeled training set is to teach a machine learning model to recognize patterns and make accurate predictions or classifications when presented with new, unseen data.

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

**ANSWER:** Classification and regression, are fundamental and widely applied in supervised machine learning.

Classification: Classification is a task where the goal is to predict the category or class label of a given input data point. For example, classifying emails as spam or non-spam, recognizing handwritten digits, or determining whether an image contains a cat or a dog.

Regression: Regression involves predicting a continuous numerical value or a numeric quantity based on input features. Regression can be used for various purposes such as predicting house prices based on features like location, size, and number of rooms, forecasting sales based on historical data and market factors etc.

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

**ANSWER:** Here are four examples of unsupervised learning tasks:

Clustering : Clustering is a task where the goal is to identify inherent patterns or groupings in a dataset without any prior knowledge of the class labels. The algorithm groups similar data points together based on their features or attributes. Common clustering algorithms include k-means clustering, hierarchical clustering, and DBSCAN. Clustering can be used for customer segmentation, image segmentation, document clustering, and anomaly detection.

Dimensionality Reduction: Dimensionality reduction is the process of reducing the number of input variables or features while preserving the most important information. It helps in dealing with high-dimensional datasets, reducing computational complexity, and visualizing data. It finds applications in data visualization, feature extraction, and noise reduction.

Association Rule Mining: Association rule mining aims to discover interesting relationships or associations among items in large datasets. It identifies common co-occurrence patterns or dependencies between items. This task is commonly used in market basket analysis, where it helps uncover relationships between products frequently purchased together.

Anomaly Detection: Anomaly detection involves identifying unusual or abnormal observations in a dataset that deviate from the expected patterns or behaviours. It is used to detect outliers or anomalies that could indicate fraudulent activities, system failures, or anomalies in manufacturing processes. Unsupervised anomaly detection techniques include clustering-based methods, statistical approaches, and auto encoders.

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

**ANSWER:** To make a robot walk through various unfamiliar terrains, a machine learning model that is commonly used is a Reinforcement Learning (RL) model, particularly in the context of robotics. Reinforcement Learning involves training an agent to make sequential decisions in an environment to maximize a cumulative reward signal.

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

**ANSWER:** To divide customers into different groups, a commonly used algorithm is k-means clustering. K-means clustering is an unsupervised learning algorithm that aims to partition a dataset into k distinct clusters based on similarity or distance measures.

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

**ANSWER:** The problem of spam detection is typically considered a supervised learning problem. In supervised learning, the model is trained on labeled data, where each data point is associated with a class label (spam or non-spam in this case). The goal is to learn a mapping between input features (such as email content, sender information, etc.) and the corresponding class labels.

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

**ANSWER:** The concept of an online learning system, also known as incremental learning or streaming learning, refers to a machine learning paradigm where the model learns and adapts continuously as new data arrives in a sequential manner. Unlike batch learning, where the model is trained on a fixed dataset and requires retraining from scratch when new data is introduced, online learning enables models to update their knowledge incrementally and dynamically.

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

**ANSWER:** Out-of-core learning, also known as disk-based learning or external memory learning, is a technique used in machine learning to handle datasets that are too large to fit into the available memory (RAM) of a single machine. It is a specific approach within online learning that addresses the challenge of processing and learning from massive amounts of data.

In core learning, the entire dataset is loaded into memory, and the learning algorithm operates directly on the data stored in RAM. This approach is feasible when the dataset can fit comfortably in memory. In contrast, out-of-core learning deals with datasets that exceed the memory capacity of the machine. Instead of loading the entire dataset into memory, out-of-core learning processes the data in smaller, manageable chunks or batches that can fit into memory. The algorithm iteratively reads and processes these batches, updating the model's parameters incrementally.

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

**ANSWER:** The learning algorithm that makes predictions using a similarity measure is called k-nearest neighbors (k-NN). k-NN is a non-parametric supervised learning algorithm that can be used for both classification and regression tasks.

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

**ANSWER:** Model parameters are learned from the data and determine the internal representation of the model, while hyperparameters are set by the practitioner and affect the learning process itself. Selecting appropriate hyperparameter values is crucial to ensure good model performance, and it often requires experimentation and tuning to find the best settings for a specific learning 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?**

**ANSWER:** Model-based learning algorithms typically look for patterns, relationships, or statistical structures within the data to build a model that can generalize and make predictions on unseen data. The primary criteria they aim to achieve are:

Accuracy: Model-based learning algorithms strive to build models that can accurately predict or estimate the target variable or make correct classifications. The goal is to minimize the error or maximize the accuracy of predictions on both the training data and unseen data.

Generalization: Model-based algorithms aim to create models that can generalize well to unseen data. They seek to capture underlying patterns or relationships in the training data that can be applied to new, unseen instances, avoiding overfitting to the training data.

The most popular method used by model-based learning algorithms to achieve success is to learn and optimize the model's parameters based on the available training data.

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

**ANSWER:** four of the most important Machine Learning challenges are:

* Data quality and quantity: Obtaining high-quality and sufficient data for training accurate models.
* Overfitting and generalization: Addressing the issue of models fitting too closely to the training data and failing to generalize well to unseen data.
* Feature engineering and selection: Identifying and selecting relevant features that capture the underlying patterns in the data effectively.
* Computational complexity and scalability: Dealing with the computational demands and scalability issues when handling large datasets or complex models.

**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?**

**ANSWER:** When a model performs well on the training data but fails to generalize to new situations, it is a clear indication of overfitting. Overfitting occurs when the model learns the training data too closely, capturing noise or irrelevant patterns, and thus fails to generalize well to unseen data.

By applying regularization, utilizing cross-validation techniques, and increasing the training data, it is possible to mitigate overfitting and improve the model's generalization performance.

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

**ANSWER:** A test set refers to a subset of data that is independent of the training process and is used to evaluate the performance and generalization ability of a trained machine learning model. It is distinct from the training set and contains data that the model has not seen during training.

The purpose of a test set is to provide an unbiased estimate of how well the trained model performs on unseen data. By evaluating the model on a separate test set, you can assess its ability to generalize and make predictions accurately on new, unseen instances. It helps measure the model's performance metrics, such as accuracy, precision, recall, F1-score, or mean squared error, among others.

**17. What is a validation set's purpose?**

**ANSWER:** A validation set, sometimes referred to as a development set or holdout set, is a subset of data that is used during the training process of a machine learning model to tune hyperparameters and assess the model's performance before final evaluation on the test set or unseen data. By using a validation set, you can iteratively refine the model, adjust hyperparameters, and select the best model configuration before the final evaluation on the test set or deployment on unseen data. This process helps ensure that the model generalizes well and performs optimally on new instances.

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

**ANSWER:** If you use the test set to tune hyperparameters, it can lead to overfitting to the test set, resulting in biased and unreliable performance estimates. This is because the hyperparameters are selected based on the performance on the test set, making the model specifically optimized for that data. When the model is evaluated on new, unseen data, its performance may significantly degrade, leading to poor generalization and disappointing results. It is essential to have a separate validation set for hyperparameter tuning to ensure unbiased evaluation and better model performance on unseen data.