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

Ans: Machine learning refers to the field of study and practice that focuses on developing computer systems capable of learning from data and improving their performance without being explicitly programmed. It involves designing algorithms and models that can automatically learn and make predictions or take actions based on patterns and insights discovered in data.

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

Ans: Four distinct types of issues where machine learning shines include:

a) **Image Classification:** Machine learning algorithms can be trained to accurately classify images into different categories, such as identifying objects or recognizing handwritten digits. For example, the MNIST dataset is a popular benchmark for image classification tasks.

b) **Natural Language Processing (NLP):** Machine learning techniques have been highly successful in tasks such as sentiment analysis, text classification, machine translation, and speech recognition. Companies like Google and Amazon utilize machine learning to improve their voice assistants' understanding and response capabilities.

c) **Fraud Detection**: Machine learning models can analyze patterns in large volumes of data to identify fraudulent activities or transactions. For instance, credit card companies use machine learning to detect fraudulent transactions by learning from historical data and spotting anomalies.

d) **Recommendation Systems:** Machine learning algorithms can analyze user preferences and behaviours to provide personalized recommendations. Examples include movie or product recommendations on platforms like Netflix and Amazon.

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

Ans: A labelled training set is a dataset used to train a machine learning model. It consists of input data samples (features) and their corresponding output or target values (labels). The labels are provided by human experts or obtained through manual annotation. During training, the model learns the underlying patterns and relationships between the features and labels, enabling it to make predictions on new, unseen data.

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

The two most important tasks in supervised learning are:

a) **Classification**: This task involves predicting a discrete class or category for a given input. For example, classifying emails as spam or non-spam, predicting whether a customer will churn or not, or determining whether an image contains a cat or a dog.

b) **Regression**: This task involves predicting a continuous value or quantity. For instance, predicting house prices based on various features like size, location, and number of rooms.

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

Four examples of unsupervised learning tasks are:

a) **Clustering:** Grouping similar data points together based on their inherent patterns or similarities. For example, clustering customer data to identify market segments or clustering news articles into topics.

b) **Dimensionality Reduction:** Reducing the number of input features while preserving the essential information. Techniques like Principal Component Analysis (PCA) can be used to achieve this. It finds applications in data visualization and feature extraction.

c) **Anomaly Detection:** Identifying unusual or outlier data points that do not conform to the expected patterns. This can be useful for fraud detection or detecting faults in systems.

d) **Association Rule Learning:** Finding interesting relationships or associations among items in a dataset. For example, identifying products frequently bought together in a retail setting.

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

Ans: The machine learning model that would be best to make a robot walk through various unfamiliar terrains is a reinforcement learning model. Reinforcement learning involves training an agent (in this case, the robot) to interact with an environment, learn from the feedback (rewards or penalties) received based on its actions, and optimize its behaviour over time. By using reinforcement learning, the robot can learn to explore and navigate different terrains through trial and error, adjusting its actions to maximize rewards and minimize penalties.

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

Ans: The algorithm used to divide customers into different groups would depend on the specific requirements and characteristics of the customer data. One popular algorithm for clustering analysis is the k-means algorithm. It partitions the data into k clusters based on their similarity and assigns each customer to the most appropriate cluster. Other clustering algorithms like hierarchical clustering or density-based clustering can also be used based on the nature of the customer data and the desired outcomes.

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

Ans: The problem of spam detection is typically considered a supervised learning problem. In this case, a labeled training set is used, consisting of a large number of emails or text messages labeled as either spam or not spam. By training a supervised learning model on this labeled data, it can learn to classify new, unseen messages as spam or non-spam based on the patterns and characteristics observed in the training set.

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

Ans: An online learning system is a machine learning system that continuously learns and updates its model in real-time as new data becomes available. It processes data in small batches or one data point at a time, updating the model's parameters incrementally as new information arrives. This allows the system to adapt to changes in the data distribution or concept drift over time. Online learning is useful when dealing with streaming data or when the model needs to quickly adapt to new information.

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

Ans: Out-of-core learning refers to a technique used when the dataset is too large to fit into the memory (RAM) of a single machine. In out-of-core learning, the data is read and processed in smaller **manageable chunks (also called batches)** from disk rather than loading the entire dataset at once. This technique enables training machine learning models on large-scale datasets without requiring excessive memory resources. It differs from in-core learning, where the entire dataset can fit into memory, allowing for faster processing.

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

Ans: A learning algorithm that makes predictions using a similarity measure is known as an instance-based learning algorithm. These algorithms compare new instances or examples to the instances in the training set and make predictions based on their similarity or proximity. One popular instance-based learning algorithm is k-nearest neighbours (KNN), which classifies a new data point based on the class labels of its k nearest neighbours in the feature space.

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

Ans: In a learning algorithm, model parameters are the internal variables that the model learns from the training data. They represent the coefficients or weights assigned to different features or attributes in the model. Hyperparameters, on the other hand, are settings or configurations that are set before the learning process begins and are not directly learned from the data. They control the behaviour of the learning algorithm and influence how the model learns and generalizes. Examples of hyperparameters include learning rate, regularization strength, and the number of hidden layers in a neural network.

**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, simplicity, and interpretability. They aim to find a model that fits the training data well while avoiding overfitting. The most popular method used by model-based learning algorithms to achieve success is regularization. Regularization techniques such as L1 or L2 regularization help prevent overfitting by adding a penalty term to the model's objective function, favouring simpler models with smaller coefficients. To make predictions, model-based learning algorithms use the learned model and apply it to new input data to generate predictions or estimates.

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

Ans: Four important machine learning challenges are:

**a) Data Quality and Quantity**: Obtaining sufficient and high-quality data for training machine learning models is often a challenge. Data collection, cleaning, and annotation can be time-consuming and expensive. Additionally, imbalanced or biased datasets can impact the model's performance.

**b) Overfitting and Underfitting**: Balancing the model's ability to capture the underlying patterns in the data without overfitting or underfitting is crucial. Overfitting occurs when the model becomes too complex and fits the training data too closely, resulting in poor generalization to new data. Underfitting happens when the model is too simple and fails to capture the underlying patterns, leading to low accuracy.

**c) Interpretability and Explainability**: Some machine learning models, such as deep neural networks, can be complex and act as black boxes, making it challenging to understand and interpret their decision-making process. Explainability is important in domains where transparency, accountability, and regulatory compliance are required.

**d) Scalability and Efficiency**: Scaling machine learning algorithms to handle large datasets or real-time applications can be a significant challenge. Training complex models on big data can require substantial computational resources and time. Designing efficient algorithms and leveraging distributed computing frameworks can help address scalability issues.

**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 could indicate the following:

**a) Overfitting**: The model may have memorized the training data too well, capturing noise or irrelevant patterns that do not generalize to new data. To address this, techniques like regularization or increasing the size of the training set can be applied.

**b) Insufficient Representativeness**: The training data may not adequately represent the true distribution of the problem. Collecting more diverse and representative data or applying data augmentation techniques can help improve generalization.

**c) Concept Drift**: The underlying patterns in the data may change over time, rendering the model's learned knowledge outdated. Monitoring and updating the model periodically or employing online learning techniques can address concept drift.

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

Ans: A test set is a portion of the dataset that is held out and not used during model training. It is used to evaluate the performance of the trained model on unseen data. By measuring the model's performance on the test set, such as calculating accuracy or error metrics, we can assess how well the model generalizes to new instances. The test set provides an unbiased estimate of the model's performance and helps in comparing different models or tuning hyperparameters.

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

Ans: The purpose of a validation set is to fine-tune the model's hyperparameters and assess its performance during the training process. While the training set is used to optimize the model's parameters, the validation set helps in selecting the best hyperparameters, such as learning rate, regularization strength, or the number of hidden units. By evaluating different hyperparameter configurations on the validation set, we can choose the hyperparameters that yield the best performance before finally testing the model on the unseen 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 development set) is a subset of the training data that is used to tune the model's performance and assess its progress during development. It is separate from the validation and test sets and can be used to iteratively refine the model's architecture, feature engineering, or hyperparameter settings. The train-dev kit helps identify issues like underfitting, overfitting, or convergence problems early in the development process, providing insights for improving the model before evaluation on the final test set.

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

Ans: If the test set is used to tune hyperparameters, it can lead to over-optimization and biased performance estimation. When hyperparameters are tuned based on the test set, the model implicitly learns from this data, compromising its ability to generalize to new, unseen data. As a result, the model's performance on the test set may be overly optimistic and not representative of its true performance. To avoid this, it is important to have a separate validation set for hyperparameter tuning and reserve the test set for final unbiased evaluation.