**Q1. What does one mean by the term “machine learning”?**

**Answer:** Machine learning is a subset of AI. which is broadly defined as the capability of a machine to imitate intelligent human behavior. Artificial intelligence systems are used to perform complex tasks in a way that is similar to how humans solve problems.

**Q2. Can you think of 4 distinct types of issues where it shines?**

**Answer:**

* Predictive Analytics,
* Personalization and Recommendation,
* cancer diagnosis,
* automatically driving vehicles.

**Q3. What is a labeled training set, and how does it work?**

**Answer:** A labeled training set is a dataset used in supervised machine learning where each input data point is paired with a corresponding output (or label). It serves as the foundational data for teaching a machine learning model to make predictions or classifications by providing it with examples of the correct answers.

**Q4. What are the two most important tasks that are supervised?**

**Answer:** The two most important tasks in supervised learning are classification and regression. These tasks are widely used across various domains and serve different purposes depending on the type of data and the problem at hand.

**Q5. Can you think of four examples of unsupervised tasks?**

**Answer:**

1. Clustering (K-Means, DBSCAN): Customer segmentation
2. Dimensionality Reduction (PCA, t-SNE, Autoencoders): Data visualization
3. Anomaly Detection (Isolation Forest, One-Class SVM): Fraud detection
4. Association Rule Mining (Apriori, FP-Growth): Market basket analysis

**Q6. State the machine learning model that would be best to make a robot walk through various**

**unfamiliar terrains?**

**Answer:** The best machine learning model for enabling a robot to walk through various unfamiliar terrains is a Reinforcement Learning (RL) model. Specifically, a Deep Reinforcement Learning (DRL) approach, such as Deep Q-Networks (DQN) or Proximal Policy Optimization (PPO), is well-suited for this task.

**Q7. Which algorithm will you use to divide your customers into different groups?**

**Answer:** Clustering algorithm would be the best choice.

**Q8. Will you consider the problem of spam detection to be a supervised or unsupervised learning** **problem?**

**Answer:** The problem of spam detection is a supervised learning problem.

**q9.** **What is the concept of an online learning system?**

**Answer:** An online learning system a virtual environment that allows students to learn from an institution or other students over the internet.

**Q10. What is out-of-core learning, and how does it differ from core learning?**

**Answer:** Out-of-core learning refers to the ability to train machine learning models on datasets that are too large to fit into a computer's main memory (RAM). It processes the data in chunks or batches, loading and using only a portion of the dataset in memory at a time, rather than loading the entire dataset.

**Q11. What kind of learning algorithm makes predictions using a similarity measure?**

**Answer:** The type of learning algorithm that makes predictions based on a similarity measure is typically a instance-based learning algorithm (or lazy learning algorithm). These algorithms do not explicitly learn a model during training but instead rely on comparing new inputs to stored instances from the training data.

**Example:**

1. k-Nearest Neighbors (k-NN)

2. Support Vector Machines (SVM) with RBF Kernel

3. Case-based reasoning (cbr)

**Q12. What the difference between a model parameter and a hyperparameter in a learning algorithm?**

**Answer:**

|  |  |  |
| --- | --- | --- |
| **Aspect** | **Model parameter** | **Hyperparameter** |
| **Definition** | Internal values learned from the data | External values set before training |
| **Learning process** | Automatically learned and adjusted during training | Set manually before training begins |
| **Examples** | Weights, biases, split points, thresholds | Learning rate, number of layers, batch size |
| **Tuning** | Not tuned manually; optimized during training | Tuning required through techniques like grid search |
| **Dependence on data** | Depends directly on the training data | Does not depend directly on data |

**Q13. 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:**

Criteria for Success:

* Accuracy, generalization, efficiency, model complexity, and flexibility.

Popular Method for Achieving Success:

* Optimization of Parameters, primarily through methods like gradient descent, Stochastic Gradient Descent and regularization to prevent overfitting.

Prediction Method:

* The model generates predictions by applying the learned parameters (weights, biases) to new data, based on the chosen model type (e.g., linear combination, decision tree traversal, neural network forward pass).

**Q14. Can you name four of the most important Machine Learning challenges?**

**Answer:**

1. Data Quality and Quantity
2. Overfitting and Underfitting
3. Model Interpretability and Explainability
4. Scalability and Computational Resources

**Q15. 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:**

Regularization: Add penalties to control model complexity.

Cross-Validation: Use different subsets of data to evaluate generalization.

Data Augmentation: Increase data diversity to help the model generalize better.

**Q16. What exactly is a test set, and why would you need one?**

**Answer:** A test set is a subset of the dataset that is used to evaluate the performance of a machine learning model after it has been trained. Unlike the training set, which is used to fit the model, the test set is kept separate and is only used after training to assess how well the model generalizes to new, unseen data.

**Q17. What is a validation set’s purpose?**

**Answer:** A validation set is a subset of the dataset used to evaluate and tune the model during the training process. It is separate from both the training set (used for training the model) and the test set (used for final evaluation). The validation set helps to monitor the model's performance during training, adjust hyperparameters, and avoid overfitting.

**Q18. What precisely is the train-dev kit, when will you need it, how do you put it to use?**

**Answer:** A Train-Dev Kit (also referred to as a training-development kit) is a combination of tools, resources, and techniques used in machine learning to split the dataset into different parts for training, development, and validation purposes. Specifically, it helps to separate the dataset into training, development (dev), and test sets in a structured and efficient way.

The dev set (short for development set) is typically a subset used to tune hyperparameters, monitor model performance during training, and prevent overfitting. It's used similarly to a validation set, but it might also be used to fine-tune the model after training with the training set, ensuring the model generalizes well.

**Q19. What could go wrong if you use the test set to tune hyperparameters?**

**Answer:** Using the test set for hyperparameter tuning leads to overfitting and biased performance estimates.

The test set should only be used for final evaluation, after all decisions regarding training and hyperparameter tuning have been made.

Use a validation set (or cross-validation) for tuning hyperparameters to avoid contaminating the test set.