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

**Machine learning** is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

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

*1.* Predictive analytics and intelligent decision-making

*2.* Cybersecurity and threat intelligence

3. internet of things (IoT) and smart cities

4. Traffic prediction and transportation

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

**Labelled training set** :

As the name suggests, labeled data (aka annotated data) is when you put **meaningful labels**, add tags, or assign classes to the raw data that you've collected. What is a label in machine learning? Let’s say you are building an image recognition system and have already collected several thousand photographs. Labels would be telling the AI that the photos contain a ‘person’, a ‘tree’, a ‘car’, and so on.

Labeled data makes the training process much more efficient and simple. The idea behind labeling data is to teach the AI to recognize patterns according to the task or target. This way, after the training process, the input of new unlabeled data will lead to predictable labels.

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

* **Classification**uses an algorithm to accurately assign test data into specific categories. It recognizes specific entities within the dataset and attempts to draw some conclusions on how those entities should be labeled or defined. Common classification algorithms are linear classifiers, support vector machines (SVM), decision trees, k-nearest neighbor, and random forest, which are described in more detail below.
* **Regression** is used to understand the relationship between dependent and independent variables. It is commonly used to make projections, such as for sales revenue for a given business. [Linear regression](https://www.ibm.com/topics/linear-regression), [logistical regression](https://www.ibm.com/topics/logistic-regression), and polynomial regression are popular regression algorithms.

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

**Unsupervised machine learning** is the process of inferring underlying hidden patterns from historical data. Within such an approach, a machine learning model tries to find any similarities, differences, patterns, and structure in data by itself. No prior human intervention is needed.

**Examples :** *data exploration,*

*customer segmentation,*

*recommender systems,*

*target marketing campaigns, and*

*data preparation and visualization, etc.*

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

There are several types of Machine Learning algorithms that can be used to allow a robot to walk in various unknown terrains, including:

*Reinforcement learning, Neural networks, Probabilistic methods.*

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

Here are a few popular algorithms for customer segmentation:

* *K-means Clustering*
* *Hierarichal clustering*
* *DBSCAN*
* *Decision trees*

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

**Supervised learning problem.**

In supervised learning, the algorithm learns from labeled training data, where each example is associated with a class label (in this case, spam or not spam). The algorithm then generalizes from the labeled examples to make predictions on new, unseen data.

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

An online learning system, also known as incremental learning or streaming learning, is a machine learning approach where the model learns from data that arrives in a sequential and online manner. Unlike batch learning, where the model is trained on a fixed dataset, online learning allows the model to update and adapt to new data as it becomes available, without retraining on the entire dataset.

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

**Out-of-core learning**, also known as "out-of-memory learning" or "disk-based learning," is a machine learning approach designed to handle large datasets that do not fit into the available memory (RAM) of a computing system. It involves processing and learning from data that is stored on disk or in a distributed file system, rather than loading the entire dataset into memory.

In contrast, **core learning** refers to traditional machine learning approaches where the entire dataset is assumed to be available in memory during training and prediction. These approaches assume that the dataset can fit into the available RAM, allowing for faster and more efficient processing.

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

**k-nearest neighbors algorithm**

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

**Model parameters** are the internal variables or weights that the learning algorithm adjusts during the training process to optimize the model's performance on the training data.

**Hyperparameters**, on the other hand, are settings or configuration choices made by the practitioner or data scientist before training the model. They are not learned from the data but are set before the learning process begins.

It's important to note that hyperparameters need to be tuned and optimized to achieve the best model performance. This is typically done through techniques such as grid search, random search, or more advanced optimization methods like Bayesian optimization.

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?

Model-based learning algorithms typically aim to achieve success by identifying and constructing a model that accurately represents the underlying patterns and relationships in the training data. The criteria that these algorithms commonly consider include:

**Fit to the Training Data, Generalization, Simplicity**.

The most popular method used by model-based learning algorithms to achieve success is through the process of optimization, typically using techniques like gradient descent or variants thereof.

The specific method for making predictions depends on the type of model being used. For example:

* In linear regression, the model predicts the output by computing a weighted sum of the input features and adding a bias term.
* In logistic regression, the model applies a sigmoid function to the linear combination of features to predict the probability of belonging to a specific class.
* In decision trees, the model traverses a tree structure based on the input features to reach a leaf node that corresponds to the predicted output value.

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

*Data Quality and Quantity*

*Overfitting and Generalization*

*Feature Selection and Engineering*

*Model Interpretability and Explainability.*

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?

When a model performs well on the training data but fails to generalize to new situations, it indicates an issue of **overfitting**. Overfitting occurs when the model becomes too complex and starts memorizing the training data rather than learning the underlying patterns. Here are three different options to address this problem:

*Regularization,*

*Cross valiation,*

*Feature selection and Dimensionality Reduction.*

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

A **test set** refers to a portion of labeled data that is set aside and not used during the training process of a machine learning model. It is a separate dataset from the training data and is used to evaluate the performance and generalization ability of the trained model on unseen instances.

We need test set to serve :

*Performance Evaluation*

*Model Selection*

*Preventing overfitting.*

17.What is a validation set's purpose?

*Hyperparameter Tuning, Model selection, Early stopping.*

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

It referring to a combination of a training set and a development/validation set.

In machine learning, the typical workflow involves splitting the available labeled data into three main subsets: the training set, the validation set (sometimes called the development set), and the test set. The purpose of this split is to train, fine-tune, and evaluate machine learning models.

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

Using the test set to tune hyperparameters can lead to several issues and undermine the integrity of model evaluation. Here are some potential problems that can arise:

1. Overfitting to the Test Set: By repeatedly evaluating and adjusting the model's hyperparameters based on the test set performance, you effectively "leak" information about the test set into the model. This can lead to overfitting, where the model becomes overly optimized for the specific characteristics of the test set, but may not generalize well to new, unseen data.
2. Optimistic Performance Estimates: When hyperparameters are tuned using the test set, the model's performance on that particular test set may no longer be a reliable indicator of its generalization ability. The reported performance metrics might be overly optimistic and not reflect the true performance on unseen data.
3. Lack of Generalization: Hyperparameters tuned specifically to the test set may result in a model that performs well on that particular test set but fails to generalize to other unseen datasets. The goal of hyperparameter tuning is to find configurations that improve generalization, and using the test set directly for tuning undermines this objective.
4. Limited Evaluation Options: If the test set is used for hyperparameter tuning, you lose the ability to independently assess the model's performance on completely unseen data. This hinders your ability to validate the model's ability to generalize and makes it difficult to compare the performance of different models objectively.