**Machine Learning – 1**

**Lokashwar Nallagatla**   
  
  
What does one mean by the term "machine learning"?

Machine learning is a field of artificial intelligence that enables computers to learn from data and improve their performance over time without explicit programming. It involves developing algorithms that recognize patterns in data, make predictions, and take actions based on those insights. These systems improve their predictions or actions through experience, learning from past data.

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

Machine learning excels in tasks that involve large, complex datasets or require real-time adaptation. Examples include image recognition, natural language processing, recommendation systems, and fraud detection. These problems often involve finding patterns or making predictions that are difficult to define with explicit rules.

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

A labelled training set consists of input data that is paired with the correct output or label. For example, in an image recognition task, each image is labelled with its correct classification, such as "cat" or "dog." The model learns to associate features of the data with the corresponding labels, which helps it predict labels for new, unseen data.

What are the two most important tasks that are supervised?

The two most important supervised learning tasks are classification and regression. Classification involves predicting a discrete label, such as identifying whether an email is spam or not, while regression predicts continuous values, like forecasting house prices based on various features.

Can you think of four examples of unsupervised tasks?

Unsupervised learning is useful for tasks where the goal is to discover hidden patterns or structures in data. Examples include clustering (grouping similar data points), dimensionality reduction (reducing the number of features), anomaly detection (identifying outliers in data), and association rule mining (finding relationships between variables).

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

A reinforcement learning model would be best for teaching a robot to navigate unfamiliar terrains. The robot learns by interacting with the environment, receiving feedback in the form of rewards or penalties for its actions, and gradually improving its decision-making process over time.

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

To divide customers into groups, a clustering algorithm like K-means would be effective. This unsupervised method identifies natural groupings in customer data based on similarities in their behaviour or characteristics, such as purchase history or demographic information.

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

Spam detection is a supervised learning problem because it requires a labelled dataset, where emails are marked as either spam or not spam. The model learns the patterns that distinguish these two classes and uses that knowledge to classify new emails.

What is the concept of an online learning system?

An online learning system updates its model continuously as new data becomes available, rather than waiting to process all data at once. This is particularly useful in situations where data is constantly changing or arriving in real-time, such as in stock market predictions or personalized recommendations.

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

Out-of-core learning refers to algorithms that can handle data too large to fit into memory by processing it in smaller chunks. Core learning, on the other hand, assumes that the entire dataset can be loaded into memory at once. Out-of-core learning is necessary for big data applications, while core learning is used for smaller datasets.

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

Instance-based learning algorithms, such as k-nearest neighbours, make predictions by comparing new data points to the most similar instances in the training set. These algorithms rely on a similarity measure, like Euclidean distance, to find the nearest neighbours and predict the outcome based on those neighbours.

What’s the difference between a model parameter and a hyperparameter in a learning algorithm?

Model parameters are learned from the training data during the training process, such as the weights in a neural network. Hyperparameters are set before training begins and control the learning process, such as the learning rate or the number of trees in a random forest. Hyperparameters are typically tuned for optimal performance.

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 aim to find the best model that minimizes error or maximizes a certain objective function. The most popular method they use is optimization, such as gradient descent. To make predictions, they apply the trained model to new data and output the predicted value or label based on learned patterns.

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

Some of the biggest challenges in machine learning include overfitting (where the model fits the training data too closely), underfitting (where the model is too simple to capture data patterns), data quality issues (such as noise or missing values), and model interpretability (making complex models understandable to humans).

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?

This is a sign of overfitting. To address it, you could collect more training data to help the model generalize, apply regularization techniques to penalize overly complex models, or simplify the model by reducing its features or parameters to prevent it from memorizing the training data.

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

A test set is a portion of the dataset that is used to evaluate the model’s performance after training. It helps assess how well the model generalizes to unseen data and provides an unbiased estimate of its real-world effectiveness. The test set should not be used during training to avoid overfitting.

What is a validation set’s purpose?

A validation set is used to tune model hyperparameters and make decisions about model selection during training. It serves as an intermediate step between training and testing, helping to evaluate different model configurations before final testing to avoid overfitting to the training data.

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

A train-dev kit refers to splitting the dataset into training, development (validation), and test sets. This is crucial when you need to train, tune, and test the model in a way that ensures it is not overfitting. The development set is used for tuning hyperparameters, while the test set evaluates final performance.

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

Using the test set to tune hyperparameters can lead to overfitting, as the model may memorize the test data, resulting in a misleadingly high-performance estimate. The test set should remain untouched during training and hyperparameter tuning to provide an unbiased evaluation of the model's generalization ability.