**ML assignment 1**

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

*Ans. Machine learning is a subfield of artificial intelligence that involves developing algorithms and statistical models that enable computers to learn from and make predictions or decisions based on data. Machine learning algorithms are designed to automatically improve their performance on a specific task by learning from examples and feedback.*

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

*Ans. 4 different types where ML shines are:*

* *Predictive Analysis (ex. Customer Churn, Fraudulent Transactions etc.)*
* *Recommender System (ex. Streaming services)*
* *Autonomous Vehicles (ex. Self-driving cars)*
* *Image and speech recognition*

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

*Ans. A labelled training set is a dataset that contains input data and their output labels and the ML algorithm learns the output for the inputs given and gives output for unseen data with good predictions and accuracy.*

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

*Ans. Two most important tasks that are supervised are:*

* *Regression*
* *Classification*

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

*Ans. four examples of unsupervised tasks are:*

* *Clustering*
* *Dimensionality Reduction*
* *Anomaly Detection*
* *Association Rule Mining*

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

**unfamiliar terrains?**

*Ans. Reinforcement Learning because Reinforcement Learning (RL) is a type of machine learning that enables an agent (such as a robot) to learn by interacting with an environment and receiving feedback in the form of rewards or penalties based on its actions.*

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

*Ans. Clustering because the algorithm identifies patterns and similarities in the data to group similar customers together.*

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

**problem?**

*Ans. Supervised Machine Learning because labelled data is needed to be classified whether the mail is “Spam” or “Not Spam”. Clustering algorithms group data points based on their similarities, but without a priori knowledge of what is "Spam" and what is "Not Spam", it is not possible to assign cluster labels to the email messages.*

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

*Ans. online learning is a kind of flexible approach to machine learning that is where the data is constantly changing or evolving with time.*

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

*Ans. Out-of-core learning allows models to be trained on datasets that are too large to fit into memory, by loading only a small subset of the data into memory at a time and performing computations on that subset.* *In out-of-core learning, the dataset is divided into smaller chunks, and the model is trained on each chunk sequentially. As each chunk is processed, the results are stored in a disk-based storage system, which can be accessed later to make predictions or to update the model. This allows the model to learn from large datasets without requiring all the data to be loaded into memory at once.*

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

*Ans. Nearest Neighbour algorithms, where the model is trained on a set of labelled data points, and then makes predictions based on the similarity between the new input data point and the training data points.*

**12.What is the difference between a model parameter and a hyperparameter in a learning**

**algorithm?**

*Ans. A model parameter is a variable that is learned from the training data and is used by the model to make predictions on new data. A hyperparameter, on the other hand, is a variable that is set before training begins and is used to control the learning process itself.*

**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. The criteria that model-based learning algorithms look for:*

* *Accuracy*
* *Generalization*
* *Simplicity*
* *Interpretability*

*The most popular method that model-based learning algorithms use to achieve success is to fit a mathematical function to the training data using optimization techniques such as gradient descent. This involves minimizing a loss function that measures the difference between the predicted outputs and the true outputs on the training data.*

*The method that algorithms use to make predictions depends on the specific model being used. For example, linear regression models make predictions by computing a weighted sum of the input features, while decision tree models make predictions by traversing a tree structure based on the input features.*

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

*Ans. The most important Machine Learning challenges:*

* *Data Quality and Quantity*
* *Overfitting and Underfitting*
* *Interpretability*
* *Generalization*

**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. This is case of overfitting where the data learns the complex data and performs well on training data but fails on test data.*

*Options to avoid this includes:*

* *Regularisation*
* *Cross-Validation*
* *Increase quantity of data*

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

*Ans. A test set is a set of data that is held out from the training data and used to evaluate the performance of a trained model. The purpose of a test set is to assess how well the model generalizes to new, unseen data.*

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

*Ans. validation set is a set of data that is used during training to evaluate the performance of a model and to tune its hyperparameters. The purpose of a validation set is to help improve the generalization performance of the model. The validation set is typically used to evaluate the model after each epoch of training. This allows us to monitor the model's performance and detect overfitting early on.*

**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, also known as the development set or hold-out set, is a subset of the training data that is used to evaluate the performance of a machine learning model during development. The train-dev kit is typically used when the model is being iteratively developed and fine-tuned. The purpose of the train-dev kit is to simulate the performance of the model on new, unseen data, without using the actual test set.*

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

*Ans. Risk of Overfitting increases. When you tune hyperparameters using the test set, you effectively leak information from the test set into the model. The test set should only be used once, at the end of model development, to get an unbiased estimate of the model's performance on new, unseen data. If you use the test set to tune hyperparameters, you are effectively using it multiple times, and the performance estimate is no longer unbiased.*