**MACHINE LEARNING THEORY ASSIGNMENT 1**

Q1. Machine learning is the concept that a computer program can learn and adapt to new data without human intervention. Machine learning is a field of artificial intelligence (AI) that keeps a computer's built-in algorithms current regardless of changes in the worldwide economy.

Q2. Four main challenges in Machine Learning include overfitting the data (using a model too complicated), underfitting the data (using a simple model), lacking in data and nonrepresentative data.

Q3. You split up the data containing known response variable values into two pieces. The training set is used to train the algorithm, and then you use the trained model on the test set to predict the response variable values that are already known.

Q4. The two most common supervised learning tasks are regression and classification. In a regression problem we our prediction is a scalar value.

Q5. Common unsupervised tasks include clustering, visualization, dimensionality reduction, and association rule learning.

Q6. The best Machine Learning algorithm to allow a robot to walk in unknown terrain is Reinforced Learning, where the robot can learn from response of the terrain to optimize itself.

Q7. The k-means clustering algorithm K-means clustering is a machine learning algorithm that arranges unlabeled data points around a specific number of clusters.

Q8. Spam detection is a supervised learning problem because the labels are known (spam or no spam).

Q9. Online machine learning is a method of machine learning in which data becomes available in a sequential order and is used to update the best predictor for future data at each step, as opposed to batch learning techniques which generate the best predictor by learning on the entire training data set.

Q10. Out-of-core learning refers to a set of algorithms working with data that cannot fit into the memory of a single computer, but that can easily fit into some data storage such as a local hard disk or web repository.

Q11. Learning algorithm that relies on a similarity measure to make predictions is instance-based algorithm.

Q12. Model parameters are estimated based on the data during model training and model hyperparameters are set manually and are used in processes to help estimate model parameters. Model hyperparameters are often referred to as parameters because they are the parts of the machine learning that must be set manually and tuned.

Q13. The goal for a model-based algorithm is to be able to generalize to new examples. To do this, model based algorithms search for optimal values for the model's parameters, often called theta . This searching, or "learning", is what machine learning is all about.

Q14. Four main challenges in Machine Learning include overfitting the data (using a model too complicated), underfitting the data (using a simple model), lacking in data and nonrepresentative data.

Q15. Representative training data is also key: If your training data doesn't reflect the actual datasets your model will encounter, you may end up with a model that won't perform once you've reached testing or production. Another issue that can occur during training is overfitting and underfitting.

Q16. The test set is a separate set of data used to test the model after completing the training. It provides an unbiased final model performance metric in terms of accuracy, precision, etc

Q17. A validation set is a set of data used to train artificial intelligence (AI) with the goal of finding and optimizing the best model to solve a given problem. Validation sets are also known as dev sets. A supervised AI is trained on a corpus of training data.

Q18. The goal of dev-set is to rank the models in term of their accuracy and helps us decide which model to proceed further with. Using Dev set we rank all our models in terms of their accuracy and pick the best performing model.

Q19. If you use the same set as training and validation set, you have the risk that you choose (hyper-)parameters that model well this set - however, you cannot say that you measure its performance on unseen data /its generalization performance.