Assignment 01 Solutions

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

Ans: Machine learning Popularly known as ML is a branch of Artificial Intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning refers to the development of algorithms and models that enable computers to learn and improve from data without being explicitly programmed. It involves training computers to recognize patterns, make predictions, and take actions based on the knowledge gained from the data.

Machine learning has a wide range of applications across industries, including image and speech recognition, natural language processing, recommendation systems, autonomous vehicles, financial analysis, healthcare, and many others.

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

Ans: The following are some of the issues where Machine Learning can be used:

`Image and Object Recognition:` Machine learning excels at tasks like image classification and object recognition. It can be used to build models that accurately identify and classify objects, people, or scenes in images or videos. This is valuable in applications such as autonomous vehicles, surveillance systems, medical imaging, and facial recognition.

`Natural Language Processing:` Machine learning is effective in natural language processing tasks, including sentiment analysis, language translation, chatbots, and voice assistants. It enables computers to understand, generate, and respond to human language, making it useful in customer support, language translation services, and information retrieval.

`Fraud Detection and Anomaly Detection:` Machine learning is well-suited for detecting fraudulent activities and anomalies in large datasets. By analyzing patterns and identifying deviations from expected behavior, machine learning models can effectively detect fraudulent transactions, cybersecurity threats, network intrusions, or unusual patterns in financial transactions.

`Recommendation Systems:` Machine learning plays a crucial role in building recommendation systems that suggest personalized recommendations to users. By analyzing user behavior, preferences, and historical data, machine learning algorithms can recommend products, movies, music, or articles that are likely to be of interest to individual users. This is commonly seen in platforms like e-commerce, streaming services, and content aggregators.

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

Ans: You split up the data containing known response variable values into two pieces. The training set is used to train the algorithm, and then we use the trained model on the test set to predict the response variable values that are already known. Clearly, it is better to have labeled data than unlabeled data since we can get much more information from them.

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

Ans: The two most common supervised learning tasks are **Regression** and **Classification**.

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

Ans: Four examples of unsupervised learning tasks are: `1.Clustering` , `2.Dimensionality reduction` , `3.Anomaly detection` , `4.Association rule learning.`

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

Ans: To make a robot walk through various unfamiliar terrains, the machine learning model that would be best suited is a combination of Reinforcement Learning (RL) and Deep Learning, specifically using Deep `Reinforcement Learning` algorithms.

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

Ans: The Best Algorithm to Segment Customers into different groups is either **Supervised Learning** (if the groups have known labels) or **Unsupervised Learning** (if there are no group labels).

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

Ans: Spam detection is a Supervised Machine Learning problem because the labels are known (spam or no spam).

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

Ans: Online learning system is a learning system in which the machine learns continously, as data is given in small streams continuously.

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

Ans: Out-of-core learning system is a system that can handle data that cannot fit into our computer memory. It uses online learning system to feed data in small bits.

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

Ans: Learning algorithm that relies on a similarity measure to make predictions is **Instance Based Algorithm**. The Machine Learning systems which are categorized as instance-based learning are the systems that learn the training examples by heart and then generalizes to new instances based on some similarity measure. It is called instance-based because it builds the hypotheses from the training instances.

It is also known as memory-based learning or lazy-learning. The time complexity of this algorithm depends upon the size of training data. The worst-case time complexity of this algorithm is O (n), where n is the number of training instances.

For example, If we were to create a spam filter with an instance-based learning algorithm, instead of just flagging emails that are already marked as spam emails, our spam filter would be programmed to also flag emails that are very similar to them. This requires a measure of resemblance between two emails. A similarity measure between two emails could be the same sender or the repetitive use of the same

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

Ans: Model parameter determines how a model will predict given a new instance. Model usually has more than one parameter (i.e. slope of a linear model). Hyperparameter is a parameter for the learning algorithm, not of a model. 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.

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: Model based learning algorithm search for the optimal value of parameters in a model that will give the best results for the new instances. We often use a cost function or similar to determine what the parameter value has to be in order to minimize the function. The model makes prediction by using the value of the new instance and the parameters in its function. 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. Model-based system learn by minimizing a cost function that measures how bad the system is at making predicitons on new data, plus a penalty for model complexity if the model is regularized. To make a prediction, a new instance's features are fed into a hypothesis function which uses the minimized theta found by repeatedly running the cost function.

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

Ans: Four main challenges in Machine Learning include the following:

1. Overfitting the Data (using a model too complicated)
2. Underfitting the data (using a simple model)
3. Lacking in Data
4. Non Representative Data.

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: If the model performs poorly to new instances, then it has overfitted on the training data. To solve this, we can do any of the following three:

- Get more data
- Implement a simpler model
- Eliminate outliers or noise from the existing data set.

17.What is a validation set's purpose ?

Ans: Validation set is a set used to compare between different training models. Let's say we have a linear model and we want to perform some hyperparameter tuning to reduce the generalization error. One way to do this 100 different models with 100 different hyperparameter values using the training set and finding the generalization error with the test set. You find the best hyperparameter value gives you 5% generalization error.

So you launch the model into production and find you're seeing 15% generalization error. This isn't going as expected. What happened?

The problem is that for each iteration of hyperparameter tuning, you measured the generalization error then updated the model using the same test set. In other words, your produced the best generalization error for the test set. The test set no longer represents cases the model hasn't seen before.

A common solution to this problem is to have a second holdout set called the validation set. You train multiple models with various hyperparameters using the training set, you select the model and hyperparameters that perform best on the validation set, and when you are happy about your model you run a single final test against the test set to get an estimate of the generalization error.

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

Ans: Cross-validation is a tool to compare models without needing a separate validation set. It is preferred over validation set because we can save from breaking of part of the training set to create a validation set, as having more data is valuable regardless.

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

Ans: If you tune hyperparameters using the test sets, then it may not perform well on the out-of-sample data because the model is tuned just for that specific set. Our model will not be generalizable to new examples.

In [ ]:

```
1
```