**Assignment - 1**

1. How would you define Machine Learning?

Machine Learning means using mathematical algorithms, and building up a relationship between dependent and independent feature by optimizing some important parameters, which are coefficients and intercept, in such a way that when we can be able to predict the value of the dependent variable, it has very minimum or optimum error or residual.

1. Can you name 4 types of problems where it shines?

1) Spam detection in email,

2) Diagnosing Cancer,

3) Detecting Fraudulent credit card transactions,

4) Automatically driving vehicles

5) Future Sales Prediction

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

Labelled training set refer to that data in which variables is marked with label, or class name. Example: Result is a target variable name and its labels are Pass and Fail.

When the model is trained on labelled training data set then the data is marked as the labelled based on the training so that when we give the test data without labels for prediction then the model could be able to predict the labels with the minimum errors.

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

In supervised learning, target variables are given. and un supervised task target variable is not given. The two most important supervised tasks are:

1) Regression- Target variables are numerical in nature.

2) Classification- Target variables are categorical in nature

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

In unsupervised learning, there is no target output column.

The four basic unsupervised tasked are:

1) clustering: Creating a similar group for individual observation.

2) visualization: Visualizing the data into different clusters.

3) Dimensionality reduction: If there is a huge number of feature then there are chances of overfitting due to curse of dimensionality. So, we try to reduce the number of features into few components. By using PCA etc

4) Association rule-based learning

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

Reinforced Learning model is a model in which robot can learn from response, retrain and optimize itself.

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

Clustering algorithms can be used the divide the customer into different groups.

Clustering Algorithm - DBSCAN, and K means.

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

Spam detection is a supervised machine learning task because in this we provide the labels to target feature such as (spam or no spam).

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

It refers to the learning that we can get from the internet. Online learning system is a learning system in which the machine learns as data is given in small streams continuously.

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

Out-of-core learning system is used when a dataset is too large to fit into a computer's memory. So we uses online learning system to feed data in small bits. The algorithm loads the data into small parts, runs the training step, then repeats the training process until it is trained on all data.

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

Instance-based learning algorithms use a similarity measure to generalize and make prediction for new cases.

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

Model parameter determines how the model will predict the result; model usually has more than one parameter.

A hyperparameter is a parameter of the learning algorithm, not the model. For example, in Linear regression task our model is parameterized by theta which is a vector of weights. In order to find the best values for theta we have a cost function which is run repeatedly by the gradient descent algorithm. Gradient descent has a hyperparameter called alpha which is the learning rate of the algorithm.

1. What are the criteria that model-based learning algorithms look for? What is the most popularmethod they use to achieve success? What method do they use to make predictions?

The goal of a model-based algorithm is to build a generalize model. To achieve this, model based algorithms find the optimal values for the model's parameters, often called theta. Model-based system learns to minimizing a cost function that measures how bad the system is at making predictions on new data, and add a penalty to model for complexity if regularization technique is performed on the model. To make the prediction, a new instance's features are fed into a hypothesis function which uses the minimized theta found by repeatedly running the cost function.

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

Four main challenges in Machine Learning include

* 1. Overfitting the data- using too complicated model. Maximum fit on training data and fail to produce similar accuracy on test data
  2. Underfitting the data- using a simple model
  3. Lack of data
  4. Nonrepresentative data

1. 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 an overfitting of the model and happens when model perform poorly on test dataset. To overcome overfitting, we use the following techniques:

1) Reduce the complexity of the model by removing features or constraining the parameters dimension reduction technique such as PCA.

2) Get more data.

3) Reducing noise in the data by fixing errors and removing outliers.

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

Test set is a set which we will use to test the model and it is also used to check the performance or accuracy of the model.

1. What is a validation set’s purpose?

In cross validation we generally used validation set. Validation set check the model accuracy on different samples of data and then decide the best possible prediction for the model.

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

The goal of train 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 can rank our models in terms of their accuracy and select the best performing model.

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

Model will not be generalizable to new examples. If we use this data to choose hyperparameters, model will get a chance to "see" the test data which is a unseen data and to develop a bias towards this test data. Therefore, we actually lose the possibility to find out how good our model would actually be on unseen data