**DSM Machine Learning1\_Assignment 20**

Q1. What are the three stages to build the hypotheses or model in machine learning?

Ans: a) Model building b) Model testing c) Applying the model

In a machine learning problem where the input is denoted by x and the output is y. In order to do machine learning, there should exist a relationship (pattern) between the input and output values. Let’s say that this the function y=f(x)y=f(x), this known as the **target function.** However, f(.) is unknown function to us.  so, machine learning algorithms try to guess a “hypothesis'' function h(x) that approximates the unknown f(.), the set of all possible hypotheses is known as the Hypothesis set H(.), the goal is the learning process is to find the final hypothesis that best approximates the unknown target function. Different machine learning models have different hypothesis sets. We have to build the model based on the available data set, we have to train and test the model. Then apply on live dataset to predict the results.

Q2. What is the standard approach to supervised learning?

Ans: The standard approach to supervised learning is to split the set of examples into the training set and the test. The most widely used learning algorithms are:

* Support Vector Machines.
* linear regression.
* logistic regression.
* naive Bayes.
* linear discriminant analysis.
* decision trees.
* k-nearest neighbor algorithm.
* Neural Networks (Multilayer perceptron)

Q3. What is Training set and Test set?

Ans: In various areas of information science like machine learning, a set of data is used to discover the potentially predictive relationship known as ‘Training Set’. Training set is an example given to the learner, while Test set is used to test the accuracy of the hypotheses generated by the learner, and it is the set of examples held back from the learner. Training set are distinct from Test set.

Q4. What is the general principle of an ensemble method and what is bagging and

boosting in ensemble method?

Ans: The general principle of an ensemble method is to combine the predictions of several models built with a given learning algorithm in order to improve robustness over a single model.

Bagging is a method in ensemble for improving unstable estimation or classification schemes. Bagging both can reduce errors by reducing the variance term.

Boosting method are used sequentially to reduce the bias of the combined model. Boosting can reduce errors by reducing the variance term.

Q5. How can you avoid overfitting?

Ans: In Machine learning we have to train over model over available data with us. When we do this activity, we get 99% accuracy, but when we use the same model to predict the live(unseen) data, we found that the accuracy of predicted result is just 50%. **Our model doesn’t generalize well from our training data to unseen data.** This is known as overfitting, and it’s a common problem in machine learning and data science. In fact, overfitting occurs in the real world all the time. A key challenge with overfitting, and with machine learning in general, is that we can’t know how well our model will perform on new data until we actually test it. To address this, we can split our initial dataset into separate training and test subsets. A) We can do the cross-validation. This is a powerful preventative measure against overfitting. We use our initial training data to generate multiple mini train-test splits. Use these splits to tune our model. In standard k-fold cross-validation, we partition the data into k subsets, called folds. Then, we iteratively train the algorithm on k-1 folds while using the remaining fold as the test set (called the “holdout fold”). B) We can **Train with more data.** It won’t work every time, but training with more data can help algorithms detect the signal better. C) **Remove features** Some algorithms have built-in feature selection. For those that don’t, you can manually improve their generalizability by removing irrelevant input features. D) **Early stopping -**When we’re iteratively, we can measure how well each iteration of the model performs. E) **Regularization -** refers to a broad range of techniques for artificially forcing your model to be simpler.