Course 2, Machine Learning

# Section 1 : (Revise from Handwritten Notes)

* Machine Learning is essentially a set of machine learning algorithms that learn patterns within a data to predict certain features. Two types, Supervised and Unsupervised
* The algorithm aims to correct learnable parameters such that the error incurred is minimized.
* The error for the entire models prediction is done via cost function such as MSE, and for a singular data point it’s called a loss function. We can visualize the cost function to get a better understanding of the model’s performance for different parameters.
* Gradient Descent is an algorithm that we use to achieve local minimum for the cost function. We calculate the derivate of our loss function with respect to each parameters to get steepness as well as direction of each slope, We then take a step in the direction that would lead to minimized error as determined by the learning rate.

# Section 2 :

* Regression is a type of supervised learning task that predicts continuous values i.e numeric. Huber loss function can also be used in place of MSE, MAE as the error doesn’t grow exponentially after a certain point but rather grows linearly.
* EDA is mandatory to get an understanding of the data we will be working with.
* Data Splitting is done to divide the data into training and testing sets. Training set is used to actually train the model and testing set is used to see how accurate our model is.
* There are multiple types of regression models, simple/linear regression comprises of only two parameters, and multiple linear regression comprises of more than 2 features this improves the model performance.

# Section 3 :

* Classification is a type of supervised learning where we predict classes/categories i.e discrete values. Binary Cross Entropy Loss Function is used as the Loss Function.
* Logistic regression is a type of classification model that predicts the probability of an outcome to belong to a certain class.
* Data Processing is needed to transform the data into a suitable form to be used in training.
* Evaluation of our model can be done via the following parameters such as False Positive, False Negative, Precision, Recall, F1 Score. (Revise from notes what each term means)
* Decision Tree is a supervised learning model which is primarily used for classification but can also be adapted to be used for regression task. It primarily works on the concept of iterative splitting of data, where gini index and entropy determines the split in such a way that the split represents the largest distribution/segmentation of data. The leaf nodes provide the prediction. (Revise the ppt of Decision Tree)
* Random Forrest is a type of ensemble learning where we make use of multiple decision trees and make use of voting/aggregation to get the output.

# Section 4 :

* Clustering is a type of unsupervised learning that groups similar data points together. K-means is a popular clustering algorithm.(revise ML written notes)
* This is primarily used for unlabeled data where a target variable equivalent does not exist, in addition we make use of PCA i.e principal component analysis to compress the multidimensional nature of unlabeled data.
* We can visualize them easily via plotting each datapoint to spot the different clusters where each cluster represents a similar set of data points.