**HLT WK 6**

Select one or more choices from the list of common Machine Learning Algorithms, do some investigations, do some investigations, and write me a short summary. I am looking for the following:

* Is it Supervised/Unsupervised/Reinforcement learning?
* What does the algorithm do?
* In which situations will it be most useful?
* (Optional) Can you find any examples of where this algorithm has been used?

**Linear Regression**

Linear Regression is a machine learning algorithm based on *supervised* learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable.

Linear Regression is of two types: Simple and Multiple. Simple Linear Regression is where only one independent variable is present and the model must find the linear relationship of it with the dependent variable

Whereas, In Multiple Linear Regression there are more than one independent variables for the model to find the relationship.

Medical researchers often use linear regression **to understand the relationship between drug dosage and blood pressure of patients**. For example, researchers might administer various dosages of a certain drug to patients and observe how their blood pressure responds.

**Logistic Regression**

Logistic regression is basically a *supervised* classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for a given set of features(or inputs), X.

Contrary to popular belief, logistic regression is a regression model. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as “1”. Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function**.**

Unlike linear regression models, which are used to predict a continuous outcome variable, logistic regression models are **mostly used to predict a dichotomous categorical outcome**, LRAs are frequently used in business analysis applications. An application may use logistic analysis to determine consumer behaviour. For instance whether or not a customer will church in 30 days.

**Decision Tree (Supervised Model)**Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

**SVM (Support Vector Machine)**

In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis

**Naïve Bayes (Supervised model)**

Naive Bayes classifiers are a family of simple "[probabilistic classifiers](https://en.wikipedia.org/wiki/Probabilistic_classification)" based on applying [Bayes' theorem](https://en.wikipedia.org/wiki/Bayes%27_theorem) with strong (naïve) [independence](https://en.wikipedia.org/wiki/Statistical_independence) assumptions between the features (see [Bayes classifier](https://en.wikipedia.org/wiki/Bayes_classifier)). They are among the simplest [Bayesian network](https://en.wikipedia.org/wiki/Bayesian_network) models, but coupled with [kernel density estimation](https://en.wikipedia.org/wiki/Kernel_density_estimation), they can achieve higher accuracy levels.

**KNN(K-Nearest Neighbours) – Supervised model**

In statistics, the k-nearest neighbours algorithm (k-NN) is a non-parametric classification method first developed by Evelyn Fix and Joseph Hodges in 1951, and later expanded by Thomas Cover. It is used for classification and regression. In both cases, the input consists of the k closest training examples in a data set. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbour.

In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbours.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data. Example: Suppose, **we have an image of a creature that looks similar to cat and dog**, but we want to know either it is a cat or dog.

**K- Means** (unsupervised)

Kmeans algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The way kmeans algorithm works is as follows:

Specify number of clusters K.

Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.

Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn’t changing.

Compute the sum of the squared distance between data points and all centroids.

Assign each data point to the closest cluster (centroid).

Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

The approach kmeans follows to solve the problem is called Expectation-Maximization. The E-step is assigning the data points to the closest cluster. The M-step is computing the centroid of each cluster. Below is a breakdown of how we can solve it mathematically (feel free to skip it). K means is often used in marketing in order to group users according to multiple characteristics, such as location, purchasing behaviour, age, and gender. It can also be used in scientific research; for example, to find population clusters within DNA data.

**Random Forest**

Random forest is a *Supervised* Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

Examples: A student named X wants to choose a course after his 10+2, and he is confused about the choice of course based on his skill set. So, he decides to consult various people like his cousins, teachers, parents, degree students, and working people. He asks them varied questions like why he should choose, job opportunities with that course, course fee, etc. Finally, after consulting various people about the course he decides to take the course suggested by most of the people.