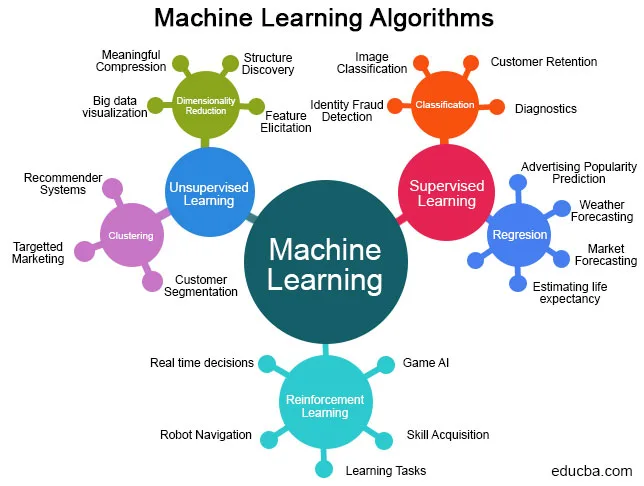
Machine Learning Notes

**Introduction to Machine Learning**

Machine Learning Algorithms are defined as the algorithms that are used for training the models, in machine learning it is divide into three different types, i.e., Supervised Learning( in this dataset are labeled and Regression and Classification techniques are used), Unsupervised Learning(in this dataset are not labeled and techniques like Dimensionality reduction and Clustering are used) and Reinforcement Learning(algorithm in which model learn from its every action) for the development of machine learning solution for applications such as Customer Retention, Image Classification, Skill Acquisition, Customer Segmentation, Game AI, Weather forecasting, Market Forecasting, Diagnostics, etc.



**Categories/Types of Machine Learning Algorithms**

The field of Machine Learning Algorithms could be categorized into –

* **Supervised Learning** – In Supervised Learning, the data set is labeled, i.e., for every feature or independent variable, there is a corresponding target data which we would use to train the model.
* **UN-Supervised Learning** – Unlike in Supervised Learning, the data set is not labeled in this case. Thus clustering technique is used to group the data based on its similarity among the data points in the same group.
* **Reinforcement Learning** – A special type of Machine Learning where the model learns from each action taken. The model is rewarded for any correct decision made and penalized for any wrong decision, which allows it to learn the patterns and make better accurate decisions on unknown data.

**Division of Machine Learning Algorithms**

The problems in Machine Learning Algorithms could be divided into –

* **Regression** – There is a continuous relationship between the dependent and the independent variables. The target variable is numeric in nature, while the independent variables could be numeric or categorical.
* **Classification** – The most common problem statement you would find in the real world is classifying a data point into some binary, multinomial, or ordinal class. The target variable has only two outcomes (Yes/No, 0/1, True/False). In the Multinomial Classification problem, there are multiple classes in the target variable (Apple/ Orange/Mango, and so on). In the Ordinal classification problem, the target variable is ordered (e.g., the grade of students).

To solve this kind of problem, programmers and scientists have developed some programs *or algorithms that could be used on the data* to make predictions. **These algorithms could be divided into linear and non-linear or tree-based algorithms.** Linear algorithms like Linear Regression, Logistic Regression are generally used when there is a linear relationship between the feature and the target variable, whereas the data exhibits non-linear patterns, the tree-based methods such as Decision Tree, Random Forest, Gradient Boosting, etc., are preferred.

**Algorithms**

There are numerous Machine Learning algorithms in the market currently, and it’s only going to increase considering the amount of research done in this field. Linear and Logistic Regression are generally the first algorithms we will learn as to start understanding and growing in this field of machine learning and on path towards becoming a data scientist.

1. **Linear Regression**

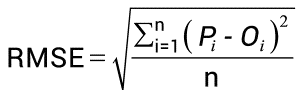
As the name suggests, this algorithm could be used in cases where the target variable, which is continuous in nature, is linearly dependent on the dependent variables. It is represented by –

y = a\*x + b + e, where y is the target variable we are trying to predict, a is the intercept, and b is the slope, x is our dependent variable used to make the prediction. This is a Simple Linear Regression as there is only one independent variable. In the case of Multiple Linear Regression, the equation would have been –

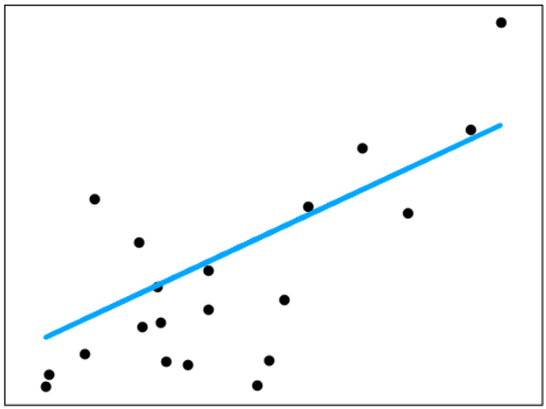
y = a1\*x1 + a2\*x2 + …… + a(n)\*x(n) + b + e

Here, e is the error term, and a1, a2.. a (n) are the coefficient of the independent variables.

A metric is used to evaluate the model’s performance, which could be Root Mean Square Error, which is the square root of the mean of the sum of the difference between the actual and the predicted values.



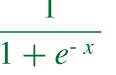
The goal of Linear Regression is to find the best fit line which would minimize the difference between the actual and the predicted data points.



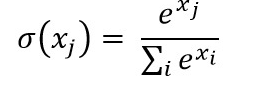
**Problem Solving Time**

1. **Logistic Regression**

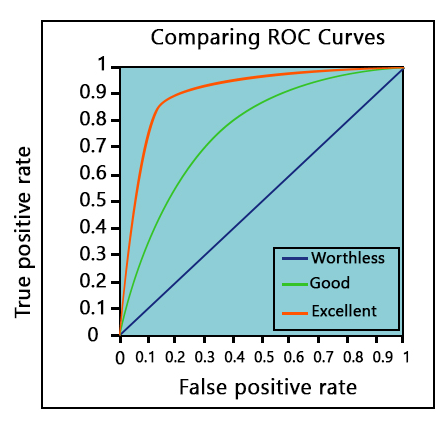
In terms of maintaining a linear relationship, it is the same as Linear Regression. However, unlike in Linear Regression, the target variable in Logistic Regression is categorical, i.e., binary, multinomial or ordinal in nature. Moreover, the choice of the activation function is important in Logistic Regression as for binary classification problems, the log of odds in favor, i.e., the sigmoid function, is used.



In the case of a multi-class problem, the softmax function is preferred as a sigmoid function takes a lot of computation time.



The metric used to evaluate a classification problem is generally Accuracy or the ROC curve. The more the area under the ROC, the better is the model. For example, a random graph would have an AUC of 0.5. The value of 1 indicates the most accuracy, whereas 0 indicates the least accuracy.



**5 sessions completed**

Problem Solving Time

1. K-Nearest Neighbors

The k-nearest neighbors (KNN) algorithm is a simple,

easy-to-implement supervised machine learning algorithm that can

be used to *solve both classification and regression problems*.

6th class was on December 11, 2021 --- KNN theory

<https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761>

7th class was on December 18, 2021 --- KNN practicals

8th class was on December 25, 2021 --- Confusion Matrix and SVM theory

1. Support Vector Machine (SVM)  
     
   <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>  
     
   <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>
2. Decision Tree
3. K-Means Clustering
4. Random Forest

**References**

<https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761>

<https://www.mathsisfun.com/algebra/distance-2-points.html>

<https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc>