**Machine Learning** is the field of study that gives computers the capability to learn without being explicitly programmed.

**Types of machine learning problems**

There are various ways to classify machine learning problems. Here, we discuss the most obvious ones.

1. On basis of the nature of the learning “signal” or “feedback” available to a learning system

Supervised learning: The computer is presented with example inputs and their desired outputs, given by a “teacher”, and the goal is to learn a general rule that maps inputs to outputs. The training process continues until the model achieves a desired level of accuracy on the training data. Some real life examples are:

Image Classification: You train with images/labels. Then in the future you give a new image expecting that the computer will recognize the new object.

Market Prediction/Regression: You train the computer with historical market data and ask the computer to predict the new price in the future.

Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

Y = f(X)

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance.

Supervised learning problems can be further grouped into regression and classification problems.

* **Classification**: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.
* **Regression**: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Some common types of problems built on top of classification and regression include recommendation and time series prediction respectively.

Some popular examples of supervised machine learning algorithms are:

* Linear regression for regression problems.
* Random forest for classification and regression problems.
* Support vector machines for classification problems.
* Techniques of Supervised Machine Learning algorithms include **linear** and **logistic regression**, **multi-class classification**, **Decision Trees** and **support vector machines**. Supervised learning requires that the data used to train the algorithm is already labeled with correct answers. For example, a classification algorithm will learn to identify animals after being trained on a dataset of images that are properly labeled with the species of the animal and some identifying characteristics.  
  Supervised learning problems can be further grouped into **Regression** and **Classification** problems. Both problems have as goal the construction of a succinct model that can predict the value of the dependent attribute from the attribute variables. The difference between the two tasks is the fact that the dependent attribute is numerical for regression and categorical for classification.

A regression problem is when the output variable is a real or continuous value, such as “salary” or “weight”. Many different models can be used, the simplest is the linear regression.

Linear Regression is the simplest model in machine learning.

Model Representation

In this problem we have an input variable - X and one output variable - Y. And we want to build linear relationship between these variables. Here the input variable is called Independent Variable and the output variable is called Dependent Variable. We can define this linear relationship as follows:

Y = \beta\_0 + \beta\_1XY=β

0

​ +β

1

​ X

The \beta\_1β

1

​ is called a scale factor or coefficient and \beta\_0β

0

​ is called bias coefficient. The bias coeffient gives an extra degree of freedom to this model. This equation is similar to the line equation y = mx + by=mx+b with m = \beta\_1m=β

1

​ (Slope) and b = \beta\_0b=β

0

​ (Intercept). So in this Simple Linear Regression model we want to draw a line between X and Y which estimates the relationship between X and Y.

Regression is used to analyse the forecast of marketing. Suppose if company is making profit every month then after some time we are able to predict the value of the profit.

When do we use KNN algorithm?

KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry.

KNN is a typical example of a lazy learner. It is called lazy not because of its apparent simplicity, but because it doesn't learn a discriminative function from the training data but memorizes the training dataset instead.

We can implement a KNN model by following the below steps:

1. Load the data
2. Initialise the value of k
3. For getting the predicted class, iterate from 1 to total number of training data points
   1. Calculate the distance between test data and each row of training data. Here we will use Euclidean distance as our distance metric since it’s the most popular method. The other metrics that can be used are Chebyshev, cosine, etc.
   2. Sort the calculated distances in ascending order based on distance values
   3. Get top k rows from the sorted array
   4. Get the most frequent class of these rows
   5. Return the predicted class

Logistic is categorical means yes or no. linear is continuous means some number

<https://www.kaggle.com/ash316/ml-from-scratch-with-iris> for KNN.

KNN

When prediction is req forunseen data, it search entire dataset for k most instance. It’s algo that classifies the new data based on asimilar measure.

K are the num of neighbours. Biggest use is the recommendation system. Also helps in the searching similar topics in internet as there are thousands of data in internet.

Hand writing detection advance use.

KNN are supervised learning and also a lazy learner because it does not learn a discriminative func from the training data but memorizes the training dataset instead.

Prediction in KNN is relatively expensive. Each time we want to make a prediction, KNN is searching for the nearest neighbours in entire training data set. We use Euclidian distance formula i.e, sqrt of (x-x2)(y-y2).

Logistic Regression;

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.

The main result of a correlation is called the **correlation coefficient** (or "r"). It ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related.

If r is close to 0, it means there is no relationship between the variables. If r is positive, it means that as one variable gets larger the other gets larger. If r is negative it means that as one gets larger, the other gets smaller (often called an "inverse" correlation).

Sometimes csv file has null values, which are later displayed as NaN in Data Frame. Pandas dropna() method allows the user to analyze and drop Rows/Columns with Null values in different ways.

Syntax:

DataFrameName.dropna(axis=0, how='any', thresh=None, subset=None, inplace=False)

Parameters:

axis: axis takes int or string value for rows/columns. Input can be 0 or 1 for Integer and ‘index’ or ‘columns’ for String.

how: how takes string value of two kinds only (‘any’ or ‘all’). ‘any’ drops the row/column if ANY value is Null and ‘all’ drops only if ALL values are null.

thresh: thresh takes integer value which tells minimum amount of na values to drop.

subset: It’s an array which limits the dropping process to passed rows/columns through list.

inplace: It is a boolean which makes the changes in data frame itself if True.

Link <https://www.kaggle.com/manish2104/predictive-model-90-accuracy-titanic>

ml-wjfltW8sBtU for logistic regression problem. Model name is shwetatitanic.