MACHINE LEARNING

Name: Devika Srivastava

Registration Number: 18BCE10091

EMAIL.ID: devika.srivastava2018@vitbhopal.ac.in

Topic: Machine Learning

MACHINE LEARNING

Introduction

The term Machine Learning was coined by Arthur Samuel in 1959, an American pioneer in the field of computer gaming and artificial intelligence and stated that “it gives computers the ability to learn without being explicitly programmed”. And in 1997, Tom Mitchell gave a “well-posed” mathematical and relational definition that “A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

Machine Learning is undeniably one of the most influential and powerful technologies in today’s world. More importantly, we are far from seeing its full potential. There’s no doubt, it will continue to be making headlines for the foreseeable future. Machine learning is a tool for turning information into knowledge. In the past 50 years, there has been an explosion of data. This mass of data is useless unless we analyse it and find the patterns hidden within. Machine learning techniques are used to automatically find the valuable underlying patterns within complex data that we would otherwise struggle to discover. The hidden patterns and knowledge about a problem can be used to predict future events and perform all kinds of complex decision making.

From medical diagnoses to fraud detection, machine learning is improving our capability to solve societal problems. Paypal acquired the ML startup Simility to analyze millions of transactions and flag anomalies to prevent money laundering. Machine learning can also predict hypoxaemia (low oxygen levels) during surgery; recognize cardiovascular risk factors like high blood pressure, age, and smoking in retinal images; and identify abnormal growths during colonoscopies.

Terminology

Data

It can be any unprocessed fact, value, text, sound or picture that is not being interpreted and analyzed. Data is the most important part of all Data Analytics, Machine Learning, Artificial Intelligence. Without data, we can’t train any model and all modern research and automation will go vain.

Dataset

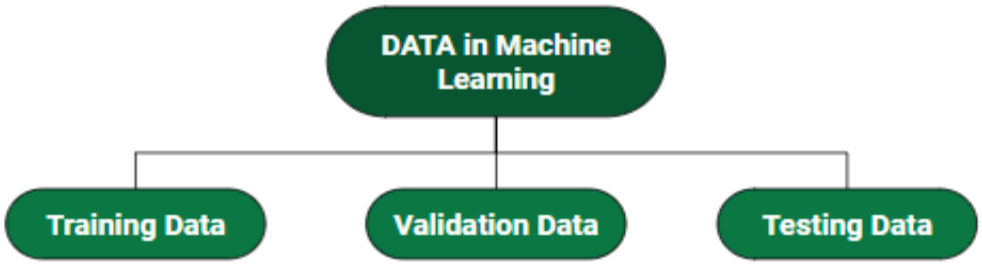
A collection of instances is a dataset and when working with machine learning methods we need datasets for different purposes.

Splitting of Data in Machine Learning

Training Data: The part of data we use to train our model. This is the data which your model actually see (both input and output) and learn from.

Validation Data: The part of data which is used to do a frequent evaluation of model, fit on training dataset along with improving involved hyperparameters (initially set parameters before the model begins learning). This data plays it's part when the model is actually training.

Testing Data: Once our model is completely trained, testing data provides the unbiased evaluation. When we feed in the inputs of Testing data, our model will predict some values (without seeing actual output). After prediction, we evaluate our model by comparing it with actual output present in the testing data. This is how we evaluate and see how much our model has learned from the experiences feed in as training data, set at the time of training.



Features

Important pieces of data that help us understand a problem. These are fed in to a Machine Learning algorithm to help it learn. A **feature** is an input variable—the x variable in simple linear regression. A simple machine learning project might use a single feature, while a more sophisticated machine learning project could use millions of features, specified as:

$$x1,x2,...xN$$

In the spam detector example the features could include the following:

- words in the email text
- sender's address
- time of day the email was sent
- email contains the phrase "one weird trick."

Model

The representation (internal model) of a phenomenon that a Machine Learning algorithm has learnt. It learns this from the data it is shown during training. The model is the output you get after training an algorithm.

A model defines the relationship between features and label. For example, a spam detection model might associate certain features strongly with "spam". Let's highlight two phases of a model's life:

- **Training** means creating or **learning** the model. That is, you show the model labeled examples and enable the model to gradually learn the relationships between features and label.
- **Inference** means applying the trained model to unlabeled examples. That is, you use the trained model to make useful predictions (y'). For example, during inference, you can predict medianHouseValue for new unlabeled examples.

Regression vs. classification

A **regression** model predicts continuous values. For example, regression models make predictions that answer questions like the following:

- What is the value of a house in California?
- What is the probability that a user will click on this ad?

A **classification** model predicts discrete values. For example, classification models make predictions that answer questions like the following:

- Is a given email message spam or not spam?
- Is this an image of a dog, a cat, or a hamster?