Chapter 1.

MACHINE LEARNING.

Date 18/01/2023

Machine Learning:

Machine Learning is a Branch of AI and Computer Science which focuses on the use of data and Algorithm to imitate the way that the humans learn, & improve from experience without being explicitly programmed which gradually improves its accuracy.

The primary goal or aim of ML is to allow the computers to learn automatically without human intervention or assistence & adjust actions accordingly.

ML focuses on the development of Computer-program that can access data and use it learn for themselves.

Traditional Programming: Input + Program = Output

Traditional Programming is a manual process - meaning a person or programmer creates the program. But without anyone programming the logic, one has to manually formulate or code rules.

We have the I/p data & someone coded a program that uses that data & runs on a computer to produce the desired O/p.

Definition:

According to Tom Mitchell "Machine Learning is the study of algorithms that improve their Performance P at Some Task T with Experience E".

A well defined Learning task for a system is given by <P, T, E>

Imp

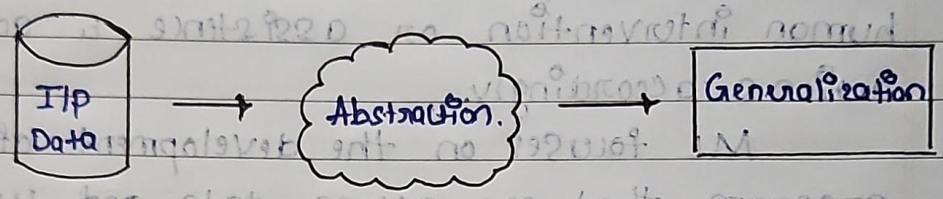
Define ML? Explain the Concept of ML with neat Diagram.

According to Herbert Simon "Learning is any process by which a system improves performance from experience"

The Basic ML process can be divided into 3 parts:

1 Data Input
2 Abstraction
3 Generalization.

Process of ML:



1 Data Input:

Past Data or Information is utilized as basis for future decision making.

2. Abstraction: (Training the Model)

During the ML process, knowledge is fed in the form of I/p data. However the data cannot be used in the original shape & form. The model may be in any one of the follow form:

* Computational Blocks like if/else Rules
* Mathematical equations.
* Specific DS like Trees or Graphs.

The choice of model used to solve a specific learning problem is a human task. The decision related to the choice of model is taken based on multiple aspects.

3. Generalization :

The 1st part of ML process is Abstraction, i.e, abstract the knowledge which comes as I/p data in the form of a model.

However this abstraction process or more popularly training the model, is just one part of ML. The other key part is to tune up the abstracted knowledge to a form which can be used to take future decisions. This is achieved as a part of generalization.

This part is quite difficult to achieve. This is because the model is trained based on the finite set of data, which may posses a limited set of characteristics.

classmate
Date 22/03/2023
Page

4. Explain the various applications of ML.

Imp. Machine Learning has wide range of applications in various fields including the follow.

1 Image Recognition
2 Speech Recognition
3 Natural Language Processing
4 Predicitive Analysis
5 Autonomous Vechicle
6 Medical Diagnosis
7 Robotics.

1 Image Recognition:
It is one of the most common ML applications, which classifies the objet as a digital images. ML algorithms are used in many image recognition applications such as Face detection or Recognition and Character recognition.

2 Speech Recognition:
Speech Recognition (SR) is a translation of spoken words into text. It is also known as "automatic Speech Recognition" (ASR)

3 Natural Language Processing:
ML algorithms are used in Natural Language Processing (NLP) applications such as language translations, Chatbots & Speech to text (detection) conversion.

4 Predicitive Analytics:
Such as Weather forecasting, stock market prediction & customer churn prediction.

5. Autonomous Vechicle:

ML Algorithms are used in Autonomous vechicles to help them navigate & make decisions.

6 Medical Diagnosis:

Cancer Detection whether it is malegent or not & Diagnosis of other diseases.

7 Robotics:

To help robots learn how to interact with the environment & perform tasks.

5. Explain various Tools that support ML :

Imp.

There are many tools available to support ML ranging from programming languages to specialized software packages.

1.
2
3
4
5
6

1 Scikit-learn :

It is probably the most popular & easy to implement ML library. It is written in Python & provides a wide array of tools like classification, clustering, regression analysis etc.

It offers a simple tools for Data-mining & analysis of data.

2 NLTK :

It is an open-source ML library that is for the purpose of Natural Language Processing. NLTK stands for Natural Language Tool Kit.

3. PyTorch :

Is an open source Deep Learning Framework that was developed by Facebook AI. It offers 2 main important features like tensors & Deep Neural Network.

4. Apache Spark :

Is a open Source big data platform & it is an improvement over the older big data platform like Hadoop. & provides support for Fault Tolerance.

5 SAS :

Statistical Analysis System (SAS) is a stable trusted & an efficient statistical analysis tool offered by the SAS institute.

6 Shogun :

Is a popular open-source ML library that is written in C++. Since it is written in C++, it offers rapid prototyping & provides support in Python, Ruby & C#.

7. Keras :

Is a powerful API that is used for building powerful neural networks.

classmate
Date 09/04/2023
Page

along with example.
Discuss the Broad Classification of Data used in ML

1 Qualitative Data
2 Quantitative Data

1. Qualitative Data:

In ML, quantitative data refers to the data that is non-numeric and describes the qualities or attributes of a particular object or entity,

Qualitative data is subjective in nature & cannot be measured using traditional statistical method, So it is non-measurable one.

For example:

If we consider the performance of students in terms of 'Good', 'Average' & 'Poor' then it falls under the category of qualitative data. also, color of a car, flavour of food & so on.

Qualitative data is also called as Categorical data,

Qualitative data can be sub-divided into:

1 Nominal:

Nominal data is one which has no numeric value, but a named value. It is used for assigning named values to attributes. & it cannot be quantified.

Example: 1) Blood Group: A, B, AB+, O ...
2) Nationality: Indian, American, British etc
3) Gender: Male, Female, other

A special case of nominal data is when only 2 labels are possible, Ex: Pass or Fail as a result of an examination. The sub-type of nominal data is called 'Dichotomous'.

2 Ordinal:

In addition to processing the properties of nominal data, can also be naturally ordered. this means ordinal data also assigns named values to attributes but, unlike nominal data they can be arranged in a sequence of increasing or decreasing value. So that we can say whether a value is better than or greater than another value.

Ex: 1) Customer Satisfaction : 'Very Happy', 'Happy', 'Unhappy'
2) Grades : A, B, C etc.
3) Hardness of material : 'Very Hard', 'Hard', 'Soft'

2 Quantitative Data:

In ML, quantitative data refers to numerical data that can be measured or counted.

Quantitative data is also termed as numerical data, & it can be measured using scale of measurement. Ex: Marks, age, height, weight etc.

There are 2 Types:

1 Interval data:

In Interval Data, not only the order is known, but the exact diff b/w values is also known. In otherwords, the diff b/w any 2 consecutive values is the same. Ex: Temprature measured in Cesius or time measured in min or sec. So the diff b/w 12°C and 18°C degree is measured & is 6°C.

2 Ratio Data:

Ratio Data represents numeric data for which exact value can be measured. Absolute zero is available for ratio data. Also, these variables can be added, sub, mul, & divided.

Ex: Height, Weight, Salary, age etc.

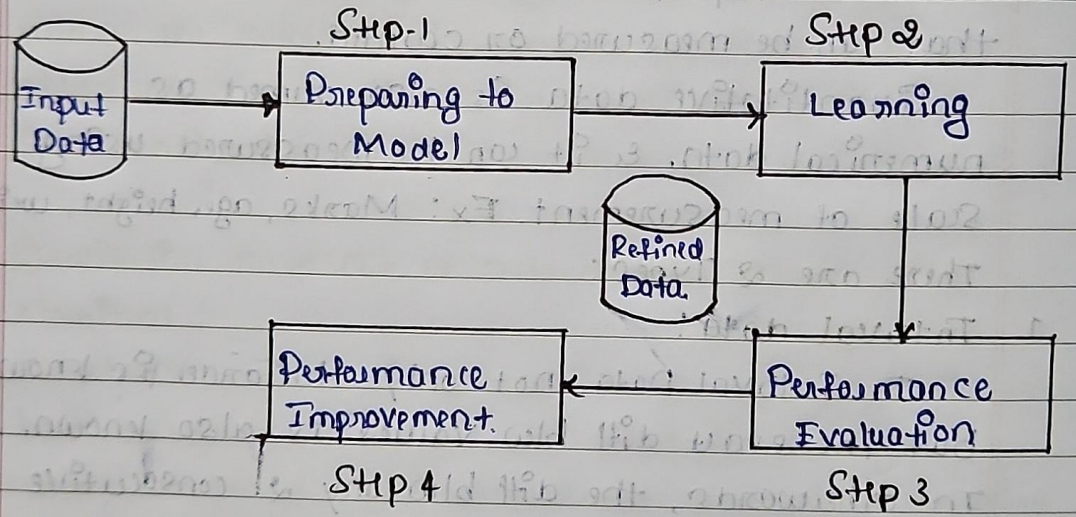
4. Explain the Various Activities of ML with Diagram.

Activities of ML:

The 1st step in machine learning activity starts with data. In case of Supervise Learning, it is the labelled training data set followed by the test data which is not labelled. In case of unsupervised Learning there is no question of labelled data but the task is to find the patterns in input data.

A through review & explaination of the data is needed to understand the type of the data, the quality of the data & the relationship b/w the diff data elements.

Once the data is prepared for modeling, then the learning tasks start off.



1 The input Data 1st Divided into 2 parts- Training & Test Data. This step is applicable for only supervised learning.

2. Consider diff models or Learning Algorithms for selection- "Train the model based on the Training data for supervised learning problem & apply to unknown data. Directly apply the choosen Unsupervised model on the input data for supervised Learning problem.

3. After the model is selected, trained & applied on input data, then the performance of the model is evaluated. Based on option available, specific actions can be taken to improve the performance of the model if possible.

Handling missing values in Data Preprocessing:

Data Preprocessing is a process of cleaning & transforming raw data to make it suitable for machine learning models. In most cases, the raw data collected from various sources may be incomplete, noisy, or inconsistent & may contain missing or irrilevant data. these issues can lead to inacurate results & make it difficult to train ML models.

So in a dataset, there are one or more data elements may have missing values in multiple records. Thue are multiple strategies to handle missing value of data elements.

Some of those strategies are:

1 Deletion
2 Imputation
3 Regression
4 Interpolation
5 ML based methods.

1. Delete or eliminate records having missing values:

This involves the removing the rows or columns that contain missing values from the dataset. However this strategy should be used with caution, as it can results in a loss of information.

2 Imputting missing values:

This involves filling the missing values with a substitute or replacement value. the substitute value can be choosen based on the characteristics of data, such as the mean, median, or mode of the available data. Imputation method includes mean imputation, median imputation & mode imputation.

3 Regression:

This involves using a regression model to predict the missing values based on the available data.

Regression models can be used to estimate the missing values based on the relationships b/w the variables in the data set.

4 Interpolation:

This involves estimating the missing values based on the values of neighboring data points. this method includes Linear interpolation & KNN imputation

5. ML based methods:

These methods involve training a machine learning model to predict the missing values based on the available data.

this method can be usefull for handling missing data in large data sets with complex relationship b/w the variables.

5. Explain the Concept of Dimentionality Reduction in Preprocessing?

Dimentionality Reduction:

In general, High-dimentional data sets need a high amount of computational space & time. At the same time, not all features are useful they reduce the performance of Machine Learning Algorithms.

Most of the Machine Learning algorithms perform better if the dimentionality of data set, i.e the number of features in the data set is reduced.

'Dimentionality Reduction' is a technique used in data preprocessing to reduce the number of features in a dataset while retaining the most important information.

It is used to overcome the Curse of Dimentionality, where the performance of ML models effected as the num of features increases. It involves reducing num of features by removing redundant, irrelavent or noisy features from the data set.

It is classified into 2 Categories:

1. Feature Selection:

This involves selecting a subset of original features that are most relevant to the task at hand.

2. Feature Extraction:

PCA

This involves creating a new feature by combining or transforming the original feature. It includes Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), Non-Negative Matrix Factorization (NMF) & t-Distributed Stochastic Neighbor Embedding (t-SNE).

PCA is a statistical technique to convert a set of correlated variables into a set of transformed, uncorrelated variables called Principal Components.