Introduction to Machine Learning in R

Last Updated : 31 Mar, 2023

The word **Machine Learning** was first coined by Arthur Samuel in 1959. The definition of machine learning can be defined as that machine learning **gives computers the ability to learn without being explicitly programmed**. Also in 1997, Tom Mitchell defined machine learning 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 considered to be the most interesting field of computer science.

How Machine Learning Works?

1. Clean the data obtained from the dataset
2. Select a proper algorithm for building a prediction model
3. Train your model to understand the pattern of project
4. Predict your results with higher accuracy

Classification Of Machine Learning

Machine learning implementations are classified into 3 major categories, depending on the nature of learning.

1. **Supervised Learning** Supervised learning as the name itself suggests that under the presence of supervision. In short in supervised learning we try to teach the machine with the data using labels and which already have the correct answer in it. After this, the machine will create an example set of data so that the supervised algorithm analyses the training data and produce the correct output of the labeled data. For example, if we create a set of data of fruits then we will be labeling as the fruit having a round shape with a dip upside and red in color then it is termed as an apple. Now when we ask the machine to identify the apple from the basket of fruits then it will use the previous labeling and identify an apple. Supervised Learning is classified into two categories as below:
    - **Classification:** A classification problem is when the output variable is a category, such as "Red" or "Orange" or "countable" or "not countable".
    - **Regression:** A regression is used when the output variable is real value, such as "rupees" or "height".
2. **Unsupervised Learning** Unsupervised learning is the training of machines using information that is not labeled and it works without any guidance. Here the main task of the machine is to separate the data using the similarities, differences, and patterns without any prior supervision. Hence, the machine is restricted to find the hidden structure in unlabeled data by own-self. For example, if we provide a group of cats and dogs which are never seen before. Then the machine will differentiate the group of cats and dogs according to their behavior and nature. Now when we provide the pictures of dogs and cats according to the classification made by the machine it will provide the result. Unsupervised Learning is classified into two categories as below:
    - **Clustering:** A clustering problem is where the machine identify the inherent groupings in the data, such as grouping customers according to visits in the shop.
    - **Association:** An association problem is where we can find the relation between two events or items, such as people buying item A also tends to buy B.

3. **Reinforcement Learning** The reinforcement learning method is all about taking suitable action to maximize reward in a particular situation. It is supervised by various machines to take the best possible path to solve the problem in a specific situation. The difference between reinforcement learning and supervised learning is that in supervised learning the data has a key of the correct answer which it uses to find the answer but in reinforcement, the agent decides what to do perform the given task. For example, while traveling from one place to another we always consider the shortest and best part possible to reach the destination. Some main points in reinforcement learning:
    - **Input:** The input should be from the initial stage where the model actually starts.
    - **Output:** There are multiple outputs to any problem.
    - **Training:** As the training is dependent on input, the model will return the state and the user will decide to reward or discard the model based on its output.

R language is basically developed by statisticians to help other statisticians and developers faster and efficiently with the data. As by now, we know that machine learning is basically working with a large amount of data and statistics as a part of data science the use of R language is always recommended. Therefore the R language is mostly becoming handy for those working with machine learning making tasks easier, faster, and innovative. Here are some top advantages of R language to implement a machine learning algorithm in R programming.

Advantages to Implement Machine Learning Using R Language

- It provides good explanatory code. For example, if you are at the early stage of working with a machine learning project and you need to explain the work you do, it becomes easy to work with R language comparison to python language as it provides the proper statistical method to work with data with fewer lines of code.
- R language is perfect for data visualization. R language provides the best prototype to work with machine learning models.
- R language has the best tools and library packages to work with machine learning projects. Developers can use these packages to create the best pre-model, model, and post-model of the machine learning projects. Also, the packages for R are more advanced and extensive than python language which makes it the first choice to work with machine learning projects.

Popular R Language Packages Used to Implement Machine Learning

- **lattice:** The lattice package supports the creation of the graphs displaying the variable or relation between multiple variables with conditions.
- **DataExplorer:** This R package focus to automate the data visualization and data handling so that the user can pay attention to data insights of the project.
- **Dalex(Descriptive Machine Learning Explanations):** This package helps to provide various explanations for the relation between the input variable and its output. It helps to understand the complex models of machine learning
- **dplyr:** This R package is used to summarize the tabular data of machine learning with rows and columns. It applies the “split-apply-combine” approach.
- **Esquisse:** This R package is used to explore the data quickly to get the information it holds. It also allows to plot bar graph, histograms, curves, and scatter plots.
- **caret:** This R package attempts to streamline the process for creating predictive models.

- **janitor:** This R package has functions for examining and cleaning dirty data. It is basically built for the purpose of user-friendliness for beginners and intermediate users.
- **rpart:** This R package helps to create the classification and regression models using two-stage procedures. The resulting models are represented as binary trees.