LANGUAGES/TOOLS

There are several programming languages and tools commonly used for machine learning (ML). Here are some of the most popular ones:

Programming Languages:

1. **Python:**

The increasing adoption of machine learning worldwide is a major factor contributing to its growing popularity. There are 69% of machine learning engineers and Python has become the favourite choice for data analytics, data science, machine learning, and AI.

Python is the preferred programming language of choice for machine learning for some of the giants in the IT world including Google, Instagram, Facebook, Dropbox, Netflix, Walt Disney, YouTube, Uber, Amazon, and Reddit. Python is an indisputable leader and by far the best language for machine learning today and here's why:

- **Extensive Collection of Libraries and Packages**

  Python's in-built libraries and packages provide base-level code so machine learning engineers don't have to start writing from scratch. Machine learning requires continuous data processing and Python has in-built libraries and packages for almost every task.

  This helps machine learning engineers reduce development time and improve productivity when working with complex machine learning applications. The best part of these libraries and packages is that there is zero learning curve, once you know the basics of Python programming, you can start using these libraries.

  1. Working with textual data – use NLTK, SciKit, and NumPy
  2. Working with images – use Sci-Kit image and OpenCV
  3. Working with audio – use Librosa
  4. Implementing deep learning – use TensorFlow, Keras, PyTorch
  5. Implementing basic machine learning algorithms – use Sci-Kit- learn .
  6. Want to do scientific computing – use Sci-Py
  7. Want to visualise the data clearly – use Matplotlib, Sci-Kit, and Seaborn

- **Code Readability**
- **Flexibility**

2. **R Programming Language**

R is an open-source programming language making it a highly cost-effective choice for machine learning projects of any size.R is an incredible programming language for machine learning written by a statistician for statisticians. R language can also be used by non-programmer including data miners, data analysts, and statisticians.

A critical part of a machine learning engineer's day-to-day job roles is understanding statistical principles so they can apply these principles to big data. R programming language is a fantastic choice when it comes to crunching large numbers and is the preferred choice for machine learning applications that use a lot of statistical data. With user-friendly IDE's like RStudio and various tools to draw graphs and manage libraries – R is a must-have programming language in a machine learning engineer's toolkit.

R has an exhaustive list of packages for machine learning –

1. MICE for dealing with missing values.
2. CARET for working with classification and regression problems.
3. PARTY and rpart for creating data partitions.
4. randomFOREST for creating decision trees.
5. dplyr and tidyr for data manipulation.
6. ggplot2 for creating beautiful visualisations.
7. Rmarkdown and Shiny for communicating insights through reports.

3. Java and JavaScript

Though Python and R continue to be the favourites of machine learning enthusiasts, Java is gaining popularity among machine learning engineers who hail from a Java development background as they don't need to learn a new programming language like Python or R to implement machine learning. Many organizations already have huge Java codebases, and most of the open-source tools for big data processing like Hadoop, Spark are written in Java.

- Java has plenty of third party libraries for machine learning. JavaML is an inbuilt machine learning library that provides a collection of machine learning algorithms implemented in Java.
- Also, you can use Arbiter Java library for hyper parameter tuning which is an integral part of making ML algorithms run effectively or you can use Deeplearning4J library

which supports popular machine learning algorithms like K-Nearest Neighbor and Neuroph and lets you create neural networks or can also use Neuroph for neural networks

- Java Virtual Machine is one of the best platforms for machine learning as engineers can write the same code on multiple platforms. JVM also helps machine learning engineers create custom tools at a rapid pace and has various IDE's that help improve overall productivity. Java works best for speed critical machine learning projects as it is fast executing.

4. Julia

Julia is a high-performance, general-purpose dynamic programming language emerging as a potential competitor for Python and R with many predominant features exclusively for machine learning.

Why use Julia for machine learning?

- Julia is particularly designed for implementing basic mathematics and scientific queries that underlies most machine learning algorithms.
- Julia code is compiled at Just-in-Time or at run time using the LLVM framework. This gives machine learning engineers great speed without any handcrafted profiling techniques or optimization techniques solving all the performance problems.
- Julia's code is universally executable. So, once written a machine learning application it can be compiled in Julia natively from other languages like Python or R in a wrapper like PyCall or RCall.
- Scalability, as discussed, is crucial for machine learning engineers and Julia makes it easier to be deployed quickly at large clusters. With powerful tools like TensorFlow, MLBase.jl, Flux.jl, SciKitlearn.jl, and many others that utilise the scalability provided by Julia, it is an apt choice for machine learning applications.
- Offer support for editors like Emacs and VIM and also IDE's like Visual studio and Juno.

5. LISP

Founded in 1958 by John McCarthy, LISP (List Processing) is the second oldest programming language still in use and is mainly developed for AI-centric applications. LISP is a dynamically typed programming language that has influenced the creation of many machine learning programming languages like Python, Julia, and Java. LISP works on Read-

Eval-Print-Loop (REPL) and has the capability to code, compile, and run code in 30+ programming languages.

The first AI chatbot ELIZA was developed using LISP and even today machine learning practitioners can use it to create chatbots for eCommerce.

Machine Learning Tools

What is Machine Learning Tool?

Machine learning tools are artificial intelligence-algorithmic applications that provide systems with the ability to understand and improve without considerable human input. It enables software, without being explicitly programmed, to predict results more accurately.

Machine learning tools (Caffee 2, Scikit-learn, Keras, Tensorflow, etc.) are defined as the artificial intelligence algorithmic applications that give the system the ability to understand and improve without being explicitly programmed as these tools are capable of performing complex processing tasks such as the awareness of images, speech-to-text, generating natural languages, etc. These tools are used for applications in which training wheels (where the individual schedules input and the desired output) are used the termed as supervised algorithm while the tools without training wheels are unsupervised algorithms and the selection of these machine learning tools entirely depends upon the type of algorithm that needs to be used for the application.

Machine Learning Tools consists of:

- Preparation and data collection
- Building models
- Application deployment and training

Local Tools for Telecommunication and Remote Learning

We can compare machine learning tools with local and remote. You can download and install a local tool and use it locally, but a remote tool runs on an external server.

1. Local Tools

You can download, install and run a local tool in your local environment.

Characteristics of Local Tools are as follows:

- Adapted for data and algorithms in memory.
- Configuration and parameterisation execution control.
- Integrate your systems to satisfy your requirements.

Examples of Local Tools are Shogun, Golearn for Go, etc.

2. Remote Tools

This tool is hosted from the server and called to your local environment. These instruments are often called Machine Learning as a Service (MLaaS).

- Customized for larger datasets to run on a scale.
- Execute multiple devices, multiple nuclei, and shared storage.
- Simpler interfaces provide less configuration control and parameterizing of the algorithm.

Examples of these Tools Are Machine Learning in AWS, Prediction in Google, Apache Mahout, etc.

TOOLS FOR MACHINE LEARNING

Given below are the different tools for machine learning:

1. TensorFlow

This is a machine learning library from Google Brain of Google's AI organization released in 2015. Tensor Flow allows you to create your own libraries. We can also use C++ and python language because of flexibility. An important characteristic of this library is that data flow diagrams are used to represent numerical computations with the help of nodes and edges. Mathematical operations are represented by nodes, whereas edges denote multidimensional data arrays on which operations are performed. TensorFlow is used by many famous companies like eBay, Twitter, Dropbox, etc. It also provides great development tools, especially in Android.

2. Keras

Keras is a deep-learning Python library that can run on top of Theano, TensorFlow. Francois Chollet, a member of the Google Brain team, developed it to give data scientists the ability to run machine learning programs fast. Because of using the high-level, understandable interface of the library and dividing networks into sequences of separate modules, rapid prototyping is possible. It is more popular because of the user interface, ease of extensibility, and modularity. It runs on CPU as well as GPU.

3. Scikit-learn

Scikit-learn, which was first released in 2007, is an open-source library for machine learning. Python is a scripting language of this framework and includes several models of machine learning such as classification, regression, clustering, and reduction of dimensionality.

Scikit- learn is designed on three open-source projects — Matplotlib, NumPy, and SciPy. Scikit-learn provides users with n number of machine learning algorithms. The framework library focuses on data modeling but not on loading, summarizing, manipulating data.

4. Caffe2

Caffe2 is an updated version of Caffe. It is a lightweight, open-source machine learning tool developed by Facebook. It has an extensive machine learning library to run complex models. Also, it supports mobile deployment. This library has C++ and Python API, which allows developers to prototype first, and optimization can be done later.

5. Apache Spark MLlib

Apache Spark MLlib is a distributed framework for machine learning. The Spark core is developed at the top. Apache sparks MLlib is nine-time faster than disk-based implementation. It is used widely as an open-source project which makes focuses on machine learning to make it easy. Apache Spark MLlib has a library for scalable vocational training. MLlib includes algorithms for regression, collaborative filters, clustering, decisions trees, pipeline APIs of higher levels.

6. OpenNN

OpenNN is developed by the artificial intelligence company Artelnics. OpenNN is an advanced analytics firmware library written in C++. The most successful method of machine learning is the implementation of neural networks. It is high in performance. The execution speed and memory allocation of this library stand out.

7. Amazon SageMaker

Amazon SageMaker is a fully managed service that allows data researchers and developers to build, train and implement machine learning models on any scale quickly and easily. Amazon SageMaker supports open-source web application Jupyter notebooks that help developers share live code. These notebooks include drivers, packages, and libraries for common deep learning platforms and frameworks for SageMaker users. Amazon SageMaker optionally encrypts models both during and during transit through AWS Key Management Service, and API requests are performed over a secure connection to the socket layer. SageMaker also stores code in volumes that are protected and encrypted by security groups.

Issues

Although machine learning is being used in every industry and helps organizations make more informed and data-driven choices that are more effective than classical methodologies, it still has so many problems that cannot be ignored. Here are some common issues in Machine Learning that professionals face to inculcate ML skills and create an application from scratch.

1. Inadequate Training Data

The major issue that comes while using machine learning algorithms is the lack of quality as well as quantity of data. Although data plays a vital role in the processing of machine learning algorithms, many data scientists claim that inadequate data, noisy data, and unclean data are extremely exhausting the machine learning algorithms. For example, a simple task requires thousands of sample data, and an advanced task such as speech or image recognition needs millions of sample data examples. Further, data quality is also important for the algorithms to work ideally, but the absence of data quality is also found in Machine Learning applications. Data quality can be affected by some factors as follows:

- **Noisy Data-** It is responsible for an inaccurate prediction that affects the decision as well as accuracy in classification tasks.
- **Incorrect data-** It is also responsible for faulty programming and results obtained in machine learning models. Hence, incorrect data may affect the accuracy of the results also.
- **Generalizing of output data-** Sometimes, it is also found that generalizing output data becomes complex, which results in comparatively poor future actions.

2. Poor quality of data

As we have discussed above, data plays a significant role in machine learning, and it must be of good quality as well. Noisy data, incomplete data, inaccurate data, and unclean data lead to less accuracy in classification and low-quality results. Hence, data quality can also be considered as a major common problem while processing machine learning algorithms.

3. Non-representative training data

To make sure our training model is generalized well or not, we have to ensure that sample training data must be representative of new cases that we need to generalize. The training data must cover all cases that are already occurred as well as occurring.

Further, if we are using non-representative training data in the model, it results in less accurate predictions. A machine learning model is said to be ideal if it predicts well for generalized cases and provides accurate decisions. If there is less training data, then there will be a sampling noise in the model, called the non-representative training set. It won't be accurate in predictions. To overcome this, it will be biased against one class or a group. Hence, we should use representative data in training to protect against being biased and make accurate predictions without any drift.

4. Overfitting and Underfitting

Overfitting:

Overfitting is one of the most common issues faced by Machine Learning engineers and data scientists. Whenever a machine learning model is trained with a huge amount of data, it starts capturing noise and inaccurate data into the training data set. It negatively affects the performance of the model. Let's understand with a simple example where we have a few training data sets such as 1000 mangoes, 1000 apples, 1000 bananas, and 5000 papayas. Then there is a considerable probability of identification of an apple as papaya because we have a massive amount of biased data in the training data set; hence prediction got negatively affected. The main reason behind overfitting is using non-linear methods used in machine learning algorithms as they build non-realistic data models. We can overcome overfitting by using linear and parametric algorithms in the machine learning models.

Methods to reduce overfitting:

- Increase training data in a dataset.
- Reduce model complexity by simplifying the model by selecting one with fewer parameters
- Ridge Regularization and Lasso Regularization
- Early stopping during the training phase
- Reduce the noise
- Reduce the number of attributes in training data.
- Constraining the model.

Underfitting:

Underfitting is just the opposite of overfitting. Whenever a machine learning model is trained with fewer amounts of data, and as a result, it provides incomplete and inaccurate data and destroys the accuracy of the machine learning model.

Underfitting occurs when our model is too simple to understand the base structure of the data, just like an undersized pant. This generally happens when we have limited data into the data set, and we try to build a linear model with non-linear data. In such scenarios, the complexity of the model destroys, and rules of the machine learning model become too easy to be applied on this data set, and the model starts doing wrong predictions as well.

Methods to reduce Underfitting:

- Increase model complexity
- Remove noise from the data
- Trained on increased and better features
- Reduce the constraints
- Increase the number of epochs to get better results.

5. Monitoring and maintenance

As we know that generalized output data is mandatory for any machine learning model; hence, regular monitoring and maintenance become compulsory for the same. Different results for different actions require data change; hence editing of codes as well as resources for monitoring them also become necessary.

6. Getting bad recommendations

A machine learning model operates under a specific context which results in bad recommendations and concept drift in the model. Let's understand with an example where at a specific time customer is looking for some gadgets, but now customer requirement changed over time but still machine learning model showing same recommendations to the customer while customer expectation has been changed. This incident is called a Data Drift. It generally occurs when new data is introduced or interpretation of data changes. However, we can overcome this by regularly updating and monitoring data according to the expectations.

7. Lack of skilled resources

Although Machine Learning and Artificial Intelligence are continuously growing in the market, still these industries are fresher in comparison to others. The absence of skilled resources in the form of manpower is also an issue. Hence, we need manpower having in-depth knowledge of

mathematics, science, and technologies for developing and managing scientific substances for machine learning.

8. Customer Segmentation

Customer segmentation is also an important issue while developing a machine learning algorithm. To identify the customers who paid for the recommendations shown by the model and who don't even check them. Hence, an algorithm is necessary to recognize the customer behavior and trigger a relevant recommendation for the user based on past experience.

9. Process Complexity of Machine Learning

The machine learning process is very complex, which is also another major issue faced by machine learning engineers and data scientists. However, Machine Learning and Artificial Intelligence are very new technologies but are still in an experimental phase and continuously being changing over time. There is the majority of hits and trial experiments; hence the probability of error is higher than expected. Further, it also includes analyzing the data, removing data bias, training data, applying complex mathematical calculations, etc., making the procedure more complicated and quite tedious.

10. Data Bias

Data Biasing is also found a big challenge in Machine Learning. These errors exist when certain elements of the dataset are heavily weighted or need more importance than others. Biased data leads to inaccurate results, skewed outcomes, and other analytical errors. However, we can resolve this error by determining where data is actually biased in the dataset. Further, take necessary steps to reduce it.

Methods to remove Data Bias:

- Research more for customer segmentation.
- Be aware of your general use cases and potential outliers.
- Combine inputs from multiple sources to ensure data diversity.
- Include bias testing in the development process.
- Analyze data regularly and keep tracking errors to resolve them easily.
- Review the collected and annotated data.
- Use multi-pass annotation such as sentiment analysis, content moderation, and intent recognition.

11. Lack of Explainability

This basically means the outputs cannot be easily comprehended as it is programmed in specific ways to deliver for certain conditions. Hence, a lack of explain ability is also found in machine learning algorithms which reduce the credibility of the algorithms.

12. Slow implementations and results

This issue is also very commonly seen in machine learning models. However, machine learning models are highly efficient in producing accurate results but are time-consuming. Slow programming, excessive requirements' and overloaded data take more time to provide accurate results than expected. This needs continuous maintenance and monitoring of the model for delivering accurate results.

13. Irrelevant features

Although machine learning models are intended to give the best possible outcome, if we feed garbage data as input, then the result will also be garbage. Hence, we should use relevant features in our training sample. A machine learning model is said to be good if training data has a good set of features or less to no irrelevant features.