Libraries for Machine Learning

Libraries for Machine Learning

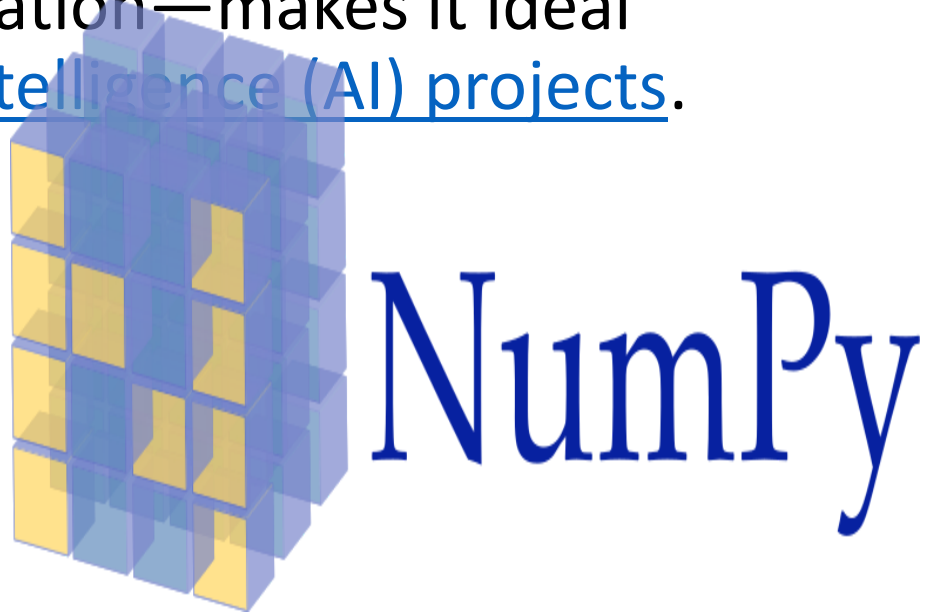
- The programming language Python was most widely used languages today .
- It’s efficient and easy to learn, and one of its greatest features is its open-source libraries available for users.
- The libraries allow users to choose from frameworks that they can build off of to produce new machine learning (ML) models.

What is a Python library?

- Python libraries are collections of modules that contain useful codes and functions, eliminating the need to write them from scratch.
- Python libraries that help machine learning developers, as well as professionals working in data science, data visualization, and more.
- Python is the preferred language for machine learning because its syntax and commands are closely related to English, making it efficient and easy to learn.
- Compared with C++, R, Ruby, and Java, Python remains one of the simplest languages, enabling accessibility, versatility, and portability. It can operate on nearly any operating system or platform.

NumPy

- NumPy is an open-source numerical and popular Python library. It can be used to perform a variety of mathematical operations on arrays and matrices. It's one of the most used scientific computing libraries, and it's often used by scientists for data analysis. Additionally, its ability to process multidimensional arrays—handling linear algebra and Fourier transformation—makes it ideal for machine learning and artificial intelligence (AI) projects.



SciPy

- SciPy is a free and open-source library that’s based on NumPy.
- It can be used to perform scientific and technical computing on large sets of data.
- Similar to NumPy, SciPy comes with embedded modules for array optimization and linear algebra
- SciPy is ideal for image manipulation and provides basic processing features of non-scientific high-level mathematical functions.
- It’s easy to use and fast to execute. It also includes high-level commands that play a role in data visualization and manipulation.

Scikit-Learn

- Scikit-learn is a very popular machine learning library that is built on NumPy and SciPy.
- It supports most of the classic supervised and unsupervised learning algorithms, and it can also be used for data mining, modeling, and analysis.
- Scikit-learn’s simple design offers a user-friendly library for those new to machine learning.

Pandas

- The Library is written in Python Web framwork and is used for data manipulation for numerical data and time series.
- It uses data frames and series to define three-dimensional and two-dimensional data respectively. It also provides options for indexing large data for quick search in large datasets.
- It is well known for the capabilities of data reshaping, pivoting on user-defined axis, handling missing data, merging and joining datasets, and the options for data filtrations.
- Pandas is very useful and very fast with large datasets. Its speed exceeds that of Numpy when the records are more than 50k. It is the best library when it comes to data cleaning

Matplotlib

- Matplotlib is a data visualization library that’s used for making plots and graphs.
- It’s an extension of SciPy and is able to handle NumPy data structures as well as complex data models made by Pandas.
- Although its expertise is limited to 2D plotting, Matplotlib can produce high-quality and publish-ready diagrams, graphs, plots, histograms, error charts, scatter plots and bar charts.

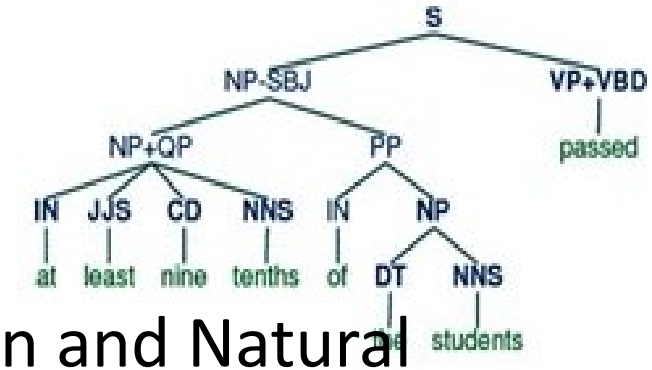
Seaborn

- Seaborn is an open-source data visualization and plotting Python library.
- It’s based on the plotting library Matplotlib and includes the extensive data structures of Pandas.
- On its own, Seaborn provides a high-level and feature-heavy interface to draw accurate and informative statistical graphs.
- It’s used in ML and DL projects

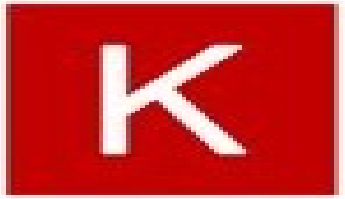
TensorFlow

- TensorFlow is a library developed by the Google Brain team for the primary purpose of Deep Learning and Neural Networks.
- It allows easy distribution of work onto multiple CPU cores or GPU cores, and can even distribute the work to multiple GPUs.
- TensorFlow uses Tensors for this purpose. Tensors can be defined as a container that can store N-dimensional data along with its linear operations.
- Although it is production-ready and does support reinforcement learning along with Neural networks

Natural Language Toolkit (NLTK)



- NLTK is the widely used library for Text Classification and Natural Language Processing. It performs word Stemming, Lemmatizing, Tokenization, and searching a keyword in documents.
- The library can be further used for sentiment analysis, understanding movie reviews, food reviews, text-classifier, checking and censoring the vulgarised words from comments, text mining, and many other human language-related operations.
- The wider scope of its uses includes AI-powered chatbots which need text processing to train their models to identify and also create sentences important for machine and human interaction in the upcoming future.



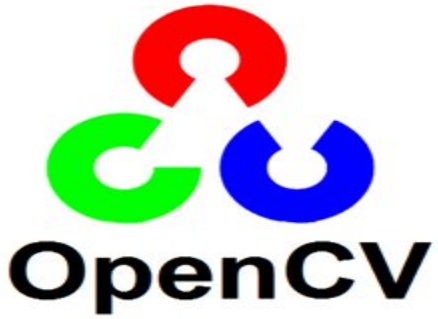
Keras

- Keras provides a Python interface of Tensorflow Library especially focused on AI neural networks.
- The earlier versions also included many other backends like Theano, Microsoft cognitive platform, and PlaidMl.
- Keras contains standard blocks of commonly used neural networks, and also the tools to make image and text processing faster and smoother

PyTorch

- Pytorch is a Facebook-developed ML library that is based on the Torch Library (an open-source ML library written in Lua Programming language).
- The project is written in Python Web Development, C++, and CUDA languages. Along with Python, PyTorch has extensions in both C and C++ languages.
- It is a competitor to Tensorflow as both of these libraries use tensors but it is easier to learn and has better integrability with Python.

OpenCV



- OpenCV is an open-source platform dedicated to computer vision and image processing. This library has more than 2500 algorithms dedicated to computer vision and ML.
- It can track human movements, detect moving objects, extract 3d models, stitch images together to create a high-resolution image, exploring the AR possibilities.
- It is used in various CCTV monitoring activities by many governments, Also the major camera companies in the world use OpenCv for making their technology smart and user-friendly.

Core Libraries for Machine Learning

- **scikit-learn**: For traditional ML algorithms (e.g., regression, classification, clustering)
- **TensorFlow**: Deep learning and neural networks
- **Keras**: High-level API for TensorFlow for building neural networks easily
- **PyTorch**: Popular for deep learning and research projects
- **XGBoost**: Gradient boosting for structured/tabular data
- **LightGBM**: Efficient gradient boosting for large datasets

Data Manipulation and Analysis

- **NumPy**: Numerical computing with arrays and matrices
- **Pandas**: Data manipulation and analysis
- **SciPy**: Scientific computing and mathematical operations

Data Visualization

- **Matplotlib**: Basic data visualization and plotting
- **Seaborn**: Statistical data visualization built on Matplotlib
- **Plotly**: Interactive and dynamic visualizations
- **Bokeh**: Interactive visualizations for large datasets

Data Preprocessing and Feature Engineering

- **scikit-learn**: Data preprocessing, scaling, encoding
- **Pandas**: Data cleaning and transformation
- **Feature-engine**: Advanced feature engineering
- **Category Encoders**: Encoding categorical variables

Model Evaluation and Experimentation

- **scikit-learn**: For metrics and model evaluation
- **MLflow**: For experiment tracking and model management
- **Optuna**: Hyperparameter optimization
- **Natural Language Processing (NLP)**
- **NLTK**: Natural language processing tasks
- **spaCy**: Industrial-strength NLP
- **Transformers (Hugging Face)**: Pretrained models for NLP tasks
- **TextBlob**: Simple NLP tasks like sentiment analysis

Computer Vision

- **OpenCV**: Image processing and computer vision
- **Pillow**: Image manipulation
- **scikit-image**: Image analysis and transformations
- **Additional Libraries**
- **Shap** and **LIME**: For explainable AI (model interpretability)
- **Statsmodels**: For statistical modeling
- **PyCaret**: Automated machine learning (AutoML)
- **Dask**: Parallel computing for large datasets