# Unit 6
# Applications of Machine Learning

By Suchita S. Patil

# Content:

- Introduction to machine learning libraries,
- applications in structured data,
- applications in unstructured data – Image, Text, Speech.

## Machine learning (ML)

- Machine learning (ML) is a branch of artificial intelligence (AI) that enables computers to “self-learn” from training data and improve over time, without being explicitly programmed.
- “Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed.”
- Machine learning algorithms are able to detect patterns in data and learn from them, in order to make their own predictions.

# Introduction to machine learning libraries

- Machine learning libraries are an important tool for building and deploying machine learning models.
- They provide a range of functions and algorithms that can be used to train and test models as well as make predictions and decisions based on data.
- Machine Learning libraries (Pandas, Numpy, Matplotlib, OpenCV, Flask, Seaborn, etc.) are defined as an interface of a set of rules or optimized functions that are written in a given language to perform repetitive work like arithmetic computation, visualizing datasets, reading of images, etc.
- This saves a lot of time for the developer and makes the developer's life easier as the developers can directly use the libraries' functions without knowing the algorithms' implementation.

# Libraries of Machine Learning

Following are some of the most popular Machine Learning Libraries:

- Pandas
- Numpy
- Matplotlib
- Scikit learn
- Seaborn
- Tensorflow
- Theano
- Keras
- PyTorch
- OpenCV
- Flask

# 1. Pandas

- Pandas is an open-source Python library that provides flexible, high-performance, and easy-to-use data structures like series and data frames.
- Python is a helpful language for data preparation but lags in data analysis and modeling.
- To overcome this lag, Pandas helps complete the entire data analysis workflow in Python without switching to other domain-specific languages like R.
- Pandas enables users to read/write datasets in various formats TEXT, CSV, XLS, JSON, SQL, HTML, and many more.
- It performs highly for data mining, reshaping, sub-setting, data alignment, slicing, indexing, and merging/joining data sets.
- But pandas are inefficient when it comes to memory utilization.
- It creates too many objects to make data manipulation easy, which utilizes high memory.

# 2. NumPy

- **NumPy** is the most fundamental data handling library, popularly used for scientific computing with Python.
- It allows the user to handle a large N-dimensional array with the ability to perform mathematical operations.
- NumPy is famous for its runtime execution speed, parallelization, and vectorization capabilities.
- It is helpful for matrix data manipulation like reshaping, transposing, and fast mathematical/logical operations.
- Other operations include sorting, selecting, basic linear algebra, discrete Fourier transform, and more.
- NumPy consumes lesser memory and provides better runtime behavior.
- But it depends on Cython, making NumPy difficult to integrate with other C/C++ libraries.

# 3. Matplotlib

- Matplotlib is a data visualization library with numpy, pandas, and other interactive environments across platforms.
- It produces high-quality visualization of data. Matplotlib can be customized to plot charts, axis, figures, or publications, and it is easy to use in **jupyter notebooks**.
- Once the user becomes familiar with it, implementing the code for matplotlib is pretty easy, although it may appear daunting to some.
- But it takes a lot of practice to use matplotlib efficiently.

# 4. Sci-kit learn

- Sci-kit learning is the heart of classical machine learning, which is completely focused on modeling the data instead of loading, manipulating, or summarizing the data.
- You name any task, and sci-kit learns you can perform it efficiently.
- One of the most simple and efficient libraries for **data mining and** analysis, sci-kit learn is an open-source library built on NumPy, SciPy & Matplotlib.
- It was developed as a part of the google summer code project, which now has become a widely accepted library for machine learning tasks.
- Sci-kit learns can prepare classification, regression, clustering, dimensionality reduction, model selection, feature extraction, normalization, etc.
- One drawback of sci-kit learning is it is not convenient to utilize categorical data.

# 5. Seaborn

- The seaborn library is built on top of the matplotlib. Seaborn makes it easy to plot data visualizations.
- It draws pretty information-generating graphs with fewer lines of code.
- Seaborn has special support for categorical and multivariate data to show aggregate statistics.

# 6. Tensorflow

- Developed by the Google brain team for its internal use, TensorFlow is an open-source platform to build and train **machine learning models**.
- ML researchers, developers, and production environments widely accept and utilize Sci-kit Learn as a prominent platform.
- Tensorflow performs various tasks, including model optimization, graphical representation, probabilistic reasoning, and statistical analysis.
- Tensors are the basic concept of this library, which provides a generalization of vectors and matrices for high-dimensional data.
- People use TensorFlow to build deep neural networks and perform numerous machine learning tasks.

## 7. Theano

- Developed by Montreal Institute for Learning algorithm (MILA), theano is a Python library that enables users to evaluate mathematical expressions with N-Dimensional arrays.
- Yes, this is similar to the Numpy Library. The only difference is Numpy is helpful in machine learning, while theano works well for deep learning.
- In addition, Theano provides faster computational speed than a CPU and detects and resolves many errors.

# 8. Keras

- '**Deep neural networks made easy**' should be this library's tagline.
- Keras is user-friendly and designed for humans, which follows the best process to reduce cognitive load.
- Keras provides easy and fast prototyping. It is a high-level neural networks API written in Python and runs on top of CNTK, TensorFlow, and MXNET. Keras provides a large number of already pre-trained models. It supports recurrent and convolutional networks and the combination of both networks too. Users can easily add new modules, making Keras suitable for high-level research. The performance of Keras completely depends on under-the-hood backends (CNTK, TensorFlow, and MXNET)

# 9. PyTorch

- PyTorch was initially developed by Facebook's artificial intelligence team, which later combined with caffe2.
- Till TensorFlow came, PyTorch was the only **deep learning** framework in the market.
- It is so integrated with Python that it can be used with other trending libraries like numpy, Python, etc.
- Furthermore, PyTorch allows users to export models in the standard ONNX (Open Neural Network Exchange) to directly access ONNX platforms, runtimes, and more.

# 10. OpenCV

- OpenCV is a computer vision library built to provide central infrastructure for computer vision applications and improve machine perception.
- This library is free for commercial use.
- OpenCV provides applicable algorithms for various tasks such as face detection, object identification, tracking moving objects, and camera movement analysis.
- In addition, OpenCV is useful for combining two images, which can produce high-resolution images, follow eye movements, extract 3D models of objects, and much more.
- It can perform on different platforms; its C++, Java, and Python interfaces can support Windows, macOS, iOS, Linux, and Android.

## 11. Flask

- A group of international Python enthusiasts developed a flask in 2004.
- The Flask can be the best Python web application framework if you want to develop web applications.
- It relies on the Jinja template engine and the Werkzeug WSGI toolkit.
- It is compatible with the Google app engine and contains the development server and debugger. Some other libraries:- Scrapy, Plotly, Bokeh, Spacy, Dask, Gensim, and Data.
- Table, Caffe, NLTK, FastAI, Gluon, and the list can continue.

# Structured and Unstructured Data

## What Is Structured Data?

Structured data is typically stored in tabular form and managed in a relational database (RDBMS). Fields contain data of a predefined format. Some fields might have a strict format, such as phone numbers or addresses, while other fields can have variable-length text strings, such as names or descriptions.

## What Is Unstructured Data?

Unstructured data includes various content such as documents, videos, audio files, posts on social media, and emails. These data types can be difficult to standardize and categorize.

Unstructured data often consists of data collections rather than a clear data element—for example, a document with thousands of words addressing multiple topics.

# Structured Data Pros and Cons

## Pros of structured data:

- **Easy to use for business users**—structured data can be used by business users who understand the subject matter related to the data. It is useful for entry level users with access to basic tools like Excel, and can be even more useful for power users familiar with SQL or business intelligence (BI) tools.
- **Extensive tools support**—structured data is several decades old and most data management and analytics tools support it. There is a huge variety of RDBMS, data analytics, and big data management tools for structured datasets.
- **Instantly usable**—structured data can be used, with no further processing, by a variety of business processes. For example, customer data in structured form can be visualized and manipulated by a CRM system.

**Cons of structured data:**

- **Data preparation**—data often needs to undergo complex transformations before it can enter a flexible data store.
- **Not flexible**—structured data requires users to create schema data definitions in advance. It is difficult to change the structure over time, and because there is a fixed, predefined structure, data can only be used for its intended purpose. This limits the use cases that can be served by structured data.
- **High overhead**—structured data is often stored in data warehouses, which can store structured data at large scale and enable fast access for user queries. A data warehouse is a complex system requiring significant resources to run, operate and maintain.
- **Complex data structures**—as organizations grow, the number of databases, tables, and fields grows exponentially. It becomes difficult to manage structured data, and it is common to have overlaps between datasets, redundant data, and stale or low quality data.

# Unstructured Data Pros and Cons

## Pros of unstructured data:

- **Native format**—unstructured data can be stored in its native format until needed, with no pre-processing.
- **Flexible**—unstructured data can be used for many different purposes and can contain a much wider variety of data, including textual data, images, videos, and source code.
- **Low overhead**—unstructured data can be stored and processed at much lower cost using elastically scalable data lakes.

## Cons of unstructured data:

- **Lack of visibility**—it is difficult to tell what is stored in a data lake and whether the data is useful. Data lakes can turn into "data swamps" with large amounts of data, which is not useful for the organization, yet incurs costs to store and manage it.
- **Requires advanced analytics**—there is typically a need for data science skills and advanced algorithms to analyze and extract insights from unstructured data. This also means it is not useful for most business users, who do not have the skills to perform advanced analytics.
- **Requires dedicated tools**—retrieving and processing unstructured data requires specialized tooling and expertise.

# Applications in structured data

Common examples of applications that rely on structured data include:

- Customer Relationship Management (CRM)
- Invoicing systems
- Product databases
- Contact lists etc.

# Customer Relationship Management (CRM)

## What is CRM (customer relationship management)?

CRM (customer relationship management) is the combination of practices, strategies and technologies that companies use to manage and analyze customer interactions and data throughout the customer lifecycle. The goal is to improve customer service relationships and assist with customer retention and drive sales growth.

## Why CRM benefits businesses

The benefits of CRM systems apply to all types of organizations, ranging from small businesses to large corporations. They include the following:

- **Enhanced customer service.** Having customer information, such as past purchases and interaction history, easily accessible helps customer support representatives provide better and faster customer service.
- **Trend spotting.** Collection of and access to customer data let businesses identify trends and insights about their customers through reporting and visualization features.
- **Automation.** CRM systems can automate menial, but necessary, sales pipeline and customer support tasks.

# Invoicing systems

Invoice data capture involves entering invoice details like invoice number, supplier name and address, project details, PO number, and other critical details for tracking goods and services provided by vendors and suppliers.

Typically, businesses collect this data manually using spreadsheets or paper ledgers. Depending on the size of the business, a procurement team, manager, or c-level executive manages this process. They pull data from structured, semi-structured, and unstructured sources into a more unified format.

Automated invoice data capture involves using turnkey or custom applications to automatically scrape invoices for the relevant information. These tools use optical character recognition (OCR) to scan invoices into digital copies or have vendors input their invoice details into a self-service portal.

As a result, the data moves seamlessly through their procurement systems, saving them time and money. Automated invoice data capture reduces errors and increases data transparency by removing people from the manual data capture part of the process.

Reducing errors in your invoice processing process goes beyond cost savings. Without mistakes or delays, your vendors, customers, clients, and even your team will have a more positive experience with your organization, keeping them around longer.

## Unstructured data includes various content such as

- Documents,
- Ideos,
- Audio files,
- Posts on social media
- Emails.

# Applications in Unstructured data

There are various applications come under unstructured data.