**ML Toolbox**

Compartment 1: **Data**

Data usually constitutes of input variables needed to form predictions. Data can be either structured (data is defined and labelled) or non-structured data (data is not defined and not labelled).

In a tabular dataset, each row generally represents single observation of a given feature and each column generally represents a feature/variable/dimension or any attribute.

Each column is also known as Vector. They store your X and y values (where X is usually various features and Y is usually correct value output of data or label) and multiple vectors(columns) are represented as matrices.

In Supervised learning, last column is generally our target value which is usually represented by Y.

Compartment 2: **Infrastructure**

It consists of platforms and tools to process data using a web application (such as Jupyter Notebook) and a programming language (such as Python, C++, C, R, MATLAB etc.). I will be using Jupyter notebook and Python for my further post when we will start with the coding part. And then there are libraries. Now what are libraries? Libraries are a collection of pre-compiled programming routines used in Machine Learning that enable you to manipulate data and execute algorithms with minimal use of code. Some libraries are: Numpy (efficiently load and work with large datasets including managing metrics), Pandas (helps in cleaning the data, perform calculations, extract data from CSV file and enables your data to be represented on a virtual spreadsheet which you can control through code), Scikit-learn (which gives access to various algorithms like linear regression, Bayes Classifier, Support Vector Machine etc.) etc. Now there are some specialized libraries such as Seaborn and Matplotlib or a standalone software program like Tableau, which supports a range of visualization techniques including charts, graphs, maps, and other visual options.

Compartment 3: **Algorithms**

Now that the development environment is set up and you’ve chosen your programming language and libraries, you can next import your data directly from a CSV file. You can find hundreds of interesting datasets in CSV format from kaggle.com. After registering as a Kaggle member, you can download a dataset of your choice. Best of all, Kaggle datasets are free, and there’s no cost to register as a user. The dataset will download directly on your device as a CSV file, which means you can use Microsoft Excel to open and even perform basic algorithms such as linear regression on your dataset.

This final compartment stores the machine learning algorithms. Beginners typically start out using simple supervised learning algorithms such as linear regression, logistic regression, decision trees, and k-nearest neighbors. Beginners are also likely to apply unsupervised learning in the form of k-means clustering.

|  |
| --- |
| **ML Toolbox** |
| Data |
| Infrastructure |
| Algorithms |

**Advanced ML Toolbox**

Compartment 1: **Big Data**

Big data is usually in Petabytes. Big data is used to describe a dataset that, due to its values, variety, volume, and velocity, defies conventional methods of processing and would be impossible for a human to process without the assistance of advanced technology. Big data is also less likely to fit into standard rows and columns and may contain numerous data types, such as structured data and a range of unstructured data, i.e. images, videos, email messages, and audio files.

With great and big data comes great noise and complicated data structures and hence we require a process called scrubbing (A process of refining your dataset before building the model).

Compartment 2: **Infrastructure**

Given that advanced learners are dealing with up to petabytes of data, robust infrastructure is required. Instead of relying on the CPU of a personal computer, the experts typically turn to distributed computing and a cloud provider such as Amazon Web Services (AWS) or Google Cloud Platform to run their data processing on a virtual graphics processing unit (GPU). As a specialized parallel computing chip, GPU instances are able to perform many more floating-point operations per second than a CPU, allowing for much faster solutions with linear algebra and statistics than with a CPU.

GPU chips were originally added to PC motherboards and video consoles such as the PlayStation 2 and the Xbox for gaming purposes. They were developed to accelerate the rendering of images with millions of pixels

whose frames needed to be continuously recalculated to display output in less than a second.

Compartment 3: **Advanced Algorithms**

The third compartment of the advanced toolbox containing machine learning algorithms. TensorFlow is the machine learning library of choice for deep learning/neural networks. It supports numerous advanced techniques including automatic calculus for back-propagation/gradient descent. Popular alternative libraries for neural networks include Torch, Caffe, and the fast-growing Keras. Written in Python, Keras is an open-source deep learning library that runs on top of TensorFlow, Theano, and other frameworks, which allows users to perform fast experimentation in fewer lines of code.

|  |
| --- |
| **ML Toolbox** |
| Big Data |
| Infrastructure |
| Advanced Algorithms |