**IBM Data Science Coursera Certificate – Tools for Data Science (Course 2)**

**Week 3**

**Notes:**

**Categories of Data Science Tools**

A picture containing text, screenshot, font, number

Description automatically generated

**Libraries of Data Science**

* Libraries are a collection of functions and methods that allow you to perform many actions without writing the code.
* **Python Libraries:**
* Scientific Computing Libraries
* Visualization Libraries in Python
* High-Level Machine Learning and Deep Learning Libraries
* Deep Learning Libraries in Python
* Libraries used in other languages
* **Scientifics Computing** Libraries contain built-in modules providing different functionalities.

Modules are also called frameworks.

Example**:**

**Pandas** offers data structures tools and tools for effective data cleaning, manipulation, and analysis. It provides tools to work with different types of data. The primary instrument of Pandas is a two-dimensional table consisting of columns and rows, called a **Data Frame.** Pandas can also provide easy indexing so you can work with your data

**NumPy** libraries are based on arrays and matrices, allowing you to apply mathematical functions to the arrays. Pandas is built on top of NumPy.

* **Visualization Libraries:**

These libraries enable you to create graphs, charts, and maps.

**Matplotlib** package is the most well-known library for data visualization. They are popular for making graphs and plots, and the graphs are easily customizable.

**Seaborn:** Another high-level visualization library is Seaborn. It is based on matplotlib. This library generates heat maps, time series, and violin plots

* **High-Level Machine Learning and Deep Learning Libraries**

**Scikit-learn** library contains tools for statistical modeling including regression, classification, clustering, and so on. It is built on NumPy, SciPy, and matplotlib. It is simple to get started. In this high-level approach, you define the model and specify the parameter types you want to use.

**Keras** allows you to build the deep learning models and

* **Deep Learning Libraries**

**TensorFlow** is a low-level framework used in large-scale production of deep learning models. It is designed for production and deployment but can be unwieldy for experimentation.

**Pytorch** is used for experimentation, making it simple for researchers to test ideas.

**Apache Spark:** is a general-purpose cluster-computing framework allowing you to process data using compute clusters.

The data is processed in parallel in more than one computer simultaneously. The Spark library has similar functionality to the following: Pandas, Numpy, and Scikit-learn.

Apache Spark data processing jobs can be done in: Python, R, Scala, and SQL.

* **Scala Libraries:**

Scala is predominately used in data engineering and data science.

**Vegas** is a Scala Library for statistical data visualizations. With Vegas, you can work with data files as well as Spark Data Frames. For deep learning, you can use **big DL.**

**Summary:**

* Libraries usually contain built-in frameworks (called modules) providing different functionalities that can be used directly. You can use data visualization methods to communicate with others and display meaningful results of an analysis.
* For machine learning, the Scikit-learn library contains tools for statistical modeling, including regression, classification, clustering and so on.
* TensorFlow is a low-level framework used in large-scale production of deep learning models. And,
* Apache Spark is a general-purpose cluster-computing framework allowing you to process data using compute clusters

**APIs (Application Programming Interference)**

APIs allow communication between two pieces of software. It basically works as intermediator between two software to exchange information.

**REST API:** **Re Stands for Representation, S stands for State, and T stands for transfer**.

RE – Representation

S – State

T – Transfer

They allow you to communicate through the internet and take advantage of resources like storage, data, artificially intelligent algorithms, and much more. **In Rest API, your program is the client**. The API communicates with a web service you can call through the internet. Though there are rules regarding Communication, Input or Request, and Output or Response.

In REST API, your program is a client, and the API communicates with a web service you can call through the internet.

* You or your code is a client.
* While web service is the resource.
* And the REST API is the API and communicates through the internet.

The Rest APIs get all the information from the request sent by the client. Then, the request is sent using an HTTP message that contains a **JSON file**. The file contains instructions for what operation is to be performed by the web service. This operation is transmitted to the web service via the internet. And the service performs the operation. Similarly, the web service returns a response through an HTTP message, where the information is returned using a **JSON file**. And this information is transmitted back to the client.

**Summary:**

An **application programming interface (API)** allows communication between two pieces of software, An API is the part of the library you see while the library contains all the components of the program. And REST APIs allow you to communicate through the internet and take advantage of resources like storage, data, artificially intelligent algorithms, and much more.

**Datasets Powering Data Science**

Hierarchical or network data structures are typically used to represent relationships between data.

Hierarchical data is organized in a tree-like format, whereas network data is stored as a graph.

Resources : You can find a comprehensive list of available data portals worldwide on the Open Knowledge Foundation’s datacatalogs.org website. The United Nations, the European Union, and many other governmental and intergovernmental organizations maintain data repositories providing access to a wide range of information

**Where to find open data :**

Open data portal list from around the world

* <http://datacatalogs.org/>

Governmental, intragovernmental, and organizational websites

* <http://data.un.org/>
* <https://www.data.gov/>
* <https://www.europeandataportal.eu/en/>

Kaggle

* [www.kaggle.com](http://www.kaggle.com)

and many more.

The Linux Foundation created the **Community Data License Agreement**, or CDLA.

Two licenses were initially created for sharing data: CDLA-Sharing and CDLA-Permissive.

**The CDLA-Sharing license** grants you permission to use and modify the data. The license stipulates that if you publish your modified version of the data, you must do so under the same license terms as the original data.

**The CDLA-Permissive license** also grants you permission to use and modify the data. However, you are not required to share changes to the data

**Open datasets and sources**

**Government Data:**

* <https://www.data.gov/>
* <https://www.census.gov/data.html>
* <https://data.gov.uk/>
* <https://www.opendatanetwork.com/>
* <https://data.un.org/>

**Financial Data Sources:**

* <https://data.worldbank.org/>
* <https://www.globalfinancialdata.com/>
* <https://comtrade.un.org/>
* <https://www.nber.org/>
* <https://fred.stlouisfed.org/>

**Crime Data:**

* <https://www.fbi.gov/services/cjis/ucr>
* <https://www.icpsr.umich.edu/icpsrweb/content/NACJD/index.html>
* <https://www.drugabuse.gov/related-topics/trends-statistics>
* <https://www.unodc.org/unodc/en/data-and-analysis/>

**Health Data:**

* <https://www.who.int/gho/database/en/>
* <https://www.fda.gov/Food/default.htm>
* <https://seer.cancer.gov/faststats/selections.php?series=cancer>
* <https://www.opensciencedatacloud.org/>
* <https://pds.nasa.gov/>
* <https://earthdata.nasa.gov/>
* <https://www.sgim.org/communities/research/dataset-compendium/public-datasets-topic-grid>

**Academic and Business Data:**

* <https://scholar.google.com/>
* <https://nces.ed.gov/>
* <https://www.glassdoor.com/research/>
* <https://www.yelp.com/dataset>

**Other General Data:**

* <https://www.kaggle.com/datasets>
* <https://www.reddit.com/r/datasets/>

### **Propriety datasets and sources**

Proprietary datasets contain data primarily owned and controlled by specific individuals or organizations. This data is limited in distribution because it is sold with a licensing agreement.  
Some data from private sources cannot be easily disclosed, **like public data.**

National security data, geological, geophysical, and biological data are examples of propriety data. Copyright laws or patents usually bind this type of data. Proprietary datasets that mainly contain sensitive information are less widely available than open datasets.

Some standard propriety dataset sources are listed below.

**Health Care:**

<https://www.sgim.org/communities/research/dataset-compendium/proprietary-datasets>

**Financial Market data:**

<https://datarade.ai/data-categories/proprietary-market-data>

**Google Cloud based datasets:**

<https://cloud.google.com/datasets>

**Data Asset eXchange:**

It is also called DAX which is a curated collection of open datasets.

It also provides tutorials notebooks that walk through the basics of data cleaning, pre-processing, and exploratory analysis.

**Model Asset eXchange:**

Certain data sets include advanced notebooks that explain how to perform more complex tasks, like creating charts, training machine-learning models, integrating deep learning, and running statistical analysis and time-series analysis can be done using MAX.

**Advance Notebooks:**

* Creating charts
* Training ML models
* Integrating deep learning
* Statistical Analysis
* Time-series analysis

The Data Asset eXchange and the Model Asset eXchange are both available on the IBM Developer website. With these resources, developers can create end-to-end analytic and machine learning workflows and to consume open data and models with confidence under clearly defined license terms.

Source: <https://developer.ibm.com/exchanges/data/>

**Machine Learning Models – Learning from Models to Make Predictions**

Machine learning (ML) uses algorithms – also known as “models” - to identify patterns in the data. The process by which the model learns these patterns from data is called “model training”.

Once a model is trained, it can then be used to make predictions. When the model is presented with new data, it tries to make predictions or decisions based on the patterns it has learned from past data.

Machine Learning models can be divided into three basic classes:

* Supervised Learning – most commonly used machine learning
* Unsupervised Learning,
* and Reinforcement Learning.

**Supervised Learning:**

The most commonly used type of machine learning is Supervised Learning and the data is labelled by human.

In Supervised Learning, a human provides input data and correct outputs. The model tries to identify relationships and dependencies between the input data and the correct output.

Supervised Learning model identifies relationships and dependencies between the input data and the correct output

Supervised Learning learning comprises two types of models:

* Regression
* and classification.

**Regression** models are used to predict a numeric (or “real”) value.

For example, if information is given about past home sales, such as geographic location, size, number of bedrooms, and sales price, you can train a model to predict the estimated sales price for other homes with similar characteristics.

**Classification** models are used to predict whether some information or data belongs to a category (or “class”).

For example, for a set of emails along with a designation you can classify whether they are to be considered as spam or not. And so, you can train an algorithm to identify unsolicited emails.

A diagram of machine learning

Description automatically generated with medium confidence

**Unsupervised Learning:**

In Unsupervised Learning, **the data is not labeled by a human**. The models must analyze the data and try to identify patterns and structure within the data based on its characteristics.

**Clustering** is an example of this learning style. Clustering models are used to divide each record of a dataset into one of a similar group.

An example of a clustering model could be providing purchase recommendations for an e-commerce store, based on past shopping behavior and the content of a shopping basket.

Another example is anomaly detection that identifies outliers in a dataset, such as fraudulent credit card transactions or suspicious online log-in attempts.

**Reinforcement Learning:**This is loosely based on the way human beings and other organisms learn

* Conceptually similar to human learning processes
* Examples: mouse and maze, robot learning to walk, chess, go and other board games of skill

**Deep Learning:**

Deep learning is a specialized type of machine learning. It refers to a general set of models and techniques that loosely emulate the way the human brain solves a wide range of problems.

It is commonly used to analyze natural language (both spoken and text), images, audio, video, to forecast time series data and much more.

It requires large datasets of **labeled data** to train a model, is compute intensive, and usually requires special purpose hardware to achieve acceptable training times.

Now you can build a custom Deep Learning model from scratch or use pre-trained models from public model repositories.

Deep Learning models are implemented using popular frameworks such as **TensorFlow**, **PyTorch** and **Keras**.

The learning frameworks provide a Python API and many supports other programming languages, such as C++ and JavaScript.

You can download **pre-trained state-of-the-art models** from repositories that are commonly referred to as **model zoos**. Popular model zoos include those provided by TensorFlow, PyTorch, Keras, and ONNX.

**Summary:**

Machine learning (ML) uses algorithms – also known as “models” ‒ to identify patterns in the data.

The process by which the model learns data patterns is called “model training”.

Types of ML are Supervised, Unsupervised, and Reinforcement. Supervised learning comprises two types of models, regression and classification.

And deep learning refers to a general set of models and techniques that loosely emulate the way the human brain solves a wide range of problems.

**Model Asset eXchange (MAX):**

The Model Asset eXchange is a free open source repository for ready-to-use and customizable deep learning microservices.

***MAX is a free open source resource for DEEP LEARNING MODELS.***

These microservices are configured to use pre-trained or custom-trainable deep learning models to solve common business problems.

These models have been fully tested and can be quickly deployed in local and cloud environments.

All models in MAX are available under permissive open source licenses, making it easier to use them for personal and commercial purposes, which reduces the risk of legal liabilities.

**Model-Serving Microservice**

Each microservice includes pre-trained deep learning model, code that pre-processes the input before it is analyzed by the model, code that post-processes the model output, and a standardized public API that makes the services functionality available to applications.

Pre-Trained Deep Learning Model -> Code to pre-process the input -> code post-process the output -> public API to connect apps

Model-serving microservices are created by running inputs through a validated model and then applying the output to a REST API.

After implementing the package, document, and test steps are complete, you will have a model-serving microservice that can then be sent to a Local machine, or a Private, Hybrid, or Public cloud.

MAX model-serving microservices are built and distributed as open source Docker images.

**Docker:** is a container platform that makes it easy to build and deploy applications.

Docker image source is published on GitHub and can be downloaded and customized for use in personal and commercial purposes.

**Kubernetes** is an open source system to automate the deployment, scaling, and management of these docker images.   
  
**RedHat OpenShift (a Kubernetes)** is a popular enterprise-grade Kubernetes platform.

It is also available on IBM Cloud, Google Cloud Platform, Amazon Web Services, and Microsoft Azure.

**Summary:**

The Model Asset eXchange is a free open source repository for ready-to-use and customizable deep learning microservices.

To reduce time to value, consider taking advantage of pre-trained models for certain types of problems.

MAX model-serving microservices are built and distributed on GitHub as open source Docker images.

Red Hat OpenShift is a Kubernetes platform used to automate deployment, scaling, and management of microservices.

Ml-exchange.org has multiple predefined models.

The CodePen tool lets users edit front-end languages.