Week 3

Libraries, APIs, Datasets and Models

Libraries for Data Science

Libraries:

collection of functions and methods that allows to perform many actions without writing code from scratch

Python Libraries:

Scientific Computing Libraries in Python

Visuaization Libraries in python

High-level ML and Deep learning

Deep Leaning Libraries in python

Libraries used in othetr languages.

Scientific Computing Libraries in Python

Pandas provides data structures and tools to storing , manipulating and cleaning data.Provides tools to work with different types of data. They also help us to perform analysis. Primarily it uses 2d structure called DataFrame. provides easy indexing.

Numpy:

Arrays and matrices.

mathematical formulas can be applied easily.

pandas is built on top of numpy

Visualization Libraries:

Enables to create graphs, charts, maps, etc,

matplotlib- (plots, graphs, most popular)

seaborn built using matplotlib. (plots: heat maps, time series, violin plots),

ML and Deep learning Lib

scikit-learn(High Level): (ML:Regression , classification, clusering) - built on top of numpy, pandas, matplotlib.

keras (High Level): Deep learning Neural Networks

Can use GPUs.

define model and specify the parameters. it will take care or a lot of stuff for us.

TensorFlow(low level): (Deep Leaning:Production and deployment)

Pytorch(low level): (Deep leaning: regression, classification for experimentation)

Apache Spark = General purpose cluster- computing framework.

Data is processed in simultaneously in more than 1 computer.

Has similar functions like numpy , pandas and scikit learn.

Other language Libs:

Apache Spark Data processing jobs:

python, R, scala and sql.

Scala lib: Used in Data Engineering and Data Science

Vegas for statistical data viz. Uses files as well as spark dataframes

BigDL - Used of deep learning.

R Libs: Has builtin functionality for data science. But also has libs

ggplot2 data viz

Keras and TensorFlow

API

Application Programmable Interface: Allows communication between two pieces of software.

pandas is not completely built in python. It uses other language codes.

Our program sends the data to pandas but pandas sends the data to the other language code and that code processes the data. This helps to abstract a lot of complexity.

Example:

Tensorflow which is written in C++ can accept data from a program written in python, js, c++, java and go , julia, matlab, scala, R.

REST API: Represntational state transfer API.

Allows communication over internet to get data from web servers.

client gets data from server.

terms:

client - our computers

data - data

resource - machine that contains data.

endpoint - path at which we can request specific data

http methods - the type of request we are making.

HTTP = hyper text transfer protocol.

Example for Rest api:

Whatson Text to speech api.

Whatson language-translator api

Datasets - Powering Data Science

What is a data set?

Collection of structured data is called a dataset.

Contains text, numbers, media

Examples: Tabular data like csv.

Data from weather station contains a row of data containing different measurements at a particular time.

Hierarchical data(tree),

network data(graphs) social network data

raw data file - images, audio, video

previously data was proprietary (private)

confidentail

private or personal info

commercially sensitive

now its open source.

publically available

companies

scientific institutions

government

organizations

open data is used in Data Science, ml,AI,

Data Sources

datacatalogs.org

Government portals

data.un.org

data.gov

europeandataportal.eu/en/

Kaggle dataset

datasetsearch.research.google.com

open data distribution is restrictive depending on the liecensing terms. previously people used open sourse software license to share data but it did not specify a few rules

cdla was created by Linux Foundation project(open data license)

cdla.io

CDLA-Sharing- Premission to use and modify data; publication only under the same terms as original

No need to share the models you created using the data.

CDLA-Permissive- Premission to use and modify data; no obligations.

## Reading: Additional Sources of Datasets

<https://author-ide.skills.network/render?token=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJtZF9pbnN0cnVjdGlvbnNfdXJsIjoiaHR0cHM6Ly9jZi1jb3Vyc2VzLWRhdGEuczMudXMuY2xvdWQtb2JqZWN0LXN0b3JhZ2UuYXBwZG9tYWluLmNsb3VkL0lCTURldmVsb3BlclNraWxsc05ldHdvcmstRFMwMTA1RU4tU2tpbGxzTmV0d29yay9sYWJzL0xhYnNfVjQvQWRkaXRpb25hbF9Tb3VyY2VzX09mX0RhdGFTZXRhLm1kIiwidG9vbF90eXBlIjoiaW5zdHJ1Y3Rpb25hbC1sYWIiLCJhZG1pbiI6ZmFsc2UsImlhdCI6MTY4MTE5MDI3NH0.pCOdngRvzD3dwNmtCiKSI583HU6EWXEWWOBISRcpHIM>

Sharing Enterprise Data - Data Asset exchange:

IBM created dax for finding quality data with a licenses

Provides

IBM Research data sets

Trusted 3rd party data sets

Ready for use in enterprise applications.

all types of data

uses CDLA data license.

makes easy to find data

also provides tutorial notebooks that has

Data cleaning

pre-processing

Exploratory analysis.

some notebooks are advanced

creating charts

training ML models

Integrating deep learning(Model Asset Exchange)

Statistical analysis

Time-series analysis

<https://developer.ibm.com/>

(Needs IBM whatson account-- Not free)

What is a ML model?

Data contains a wealth of info that can be used to solve problems.

ML use algo to identify patterns in data

Model training is the process by which the model learns the patterns

After a model is trained, it can be used for predictions.

Types of Ml are:

Supervised Learning

Unsupervised Learning

Reinforcement Learning

Supervised Learning - most commonly used.

Human provides the input data and outputs.

The model identifies the relationships and dependencies between them.

Regression - To predict real numerical values

Ex: home sales prices, stock market prices

Calassification- To classify data into certain category.

Ex: email spam filtering, fraud detection, image classification

Unsupervised Learning:

Data is labeled

Model tries to identify patterns without external help.

Clustering is an Example - grouping

Anomaly detection identifies outliers in datasets.

Grouping items together based on the ecommerce basket content.

Reinforcement Learning:

Conceptually similar to human learning process.

gets a reward when a task is completed.

Example:Maze , robot learning to walk, chess, go, etc.,

uses trial and error technique and improves overtime.

Deep Learning:

Specialized type of learning

tries to loosely emulate show how the human brain works.

Applications:

Natural Language processing

Image , audio and video analysis

Time series forecasting.

Requires large datasets of labeled data and is compute intensive.

Requires special purpose hardware.

You can build models from scratch or download from the public model repos

Built using frameworks such as:

Tensorflow

Keras

pyTorch

provide python API and support C++ and JS

Popular Model repos:

Most frameworks provides a model zoo

TensorFlow, PyTorch, Keras, and ONNX model zoo

Using models to solve a problem:

Using an Teddy bear image to figure out what image it is.

Steps

1. collect and prep data - can be time consuming and labour intensive process. label the raw data by drawing a bounding box aound the object and label them.
2. Build model: start from scratch or get an existing model and modify it for your needs.
3. Train the mode

These are iterative steps:

Requres data, expertise, time, and resources.

1. Deploy Model

Model Asset Exchange

Free open sourse resource.

Pretained model are ready to use models to reduce the time consumption.

Steps:

Data + Model+ Compute resources + Domain Expertise

Research, evaluate, test , train and validate.

We get validated model

MAX is free open source repo for ready to use customizable deep learning microservices.

Pre = or custom-trained deep learning model

Fully tested, can deployed in mins

Approved for personal and commercial use.

On Max, find models for a variety of domains.

Oject Detections

Image, Video, audio, text classification.

Named entity recognition

Image to text translation

Human pose detection

And more

Components of model-serving micro service

Pre-trainned deeplearning model, code that pre- process the input code to post-process output, public API to connect apps.

They are created by running input through validated model and delpoyed with REST API so third party apps can access the model though the API.

After Implement, package, document and test steps are complete we will have model-serving microservice that can be sent to a local maching or private , hybrid or public cloud.

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

Docker container program that makes it easy to build and deploy apps easily,

Docker Images are published on github,Users can download and customize for personal or commercial environments.

Use the Kubernetes open source system to automate the deployment, scaling, and management of these Dicker images. Red Hat Openshift is a popular enterprise grade kubernetes platform.

Available in Ibm whatson, google cloud platform, aws, azure,

<https://ml-exchange.org/models>