**A chapter-by-chapter summary of Data Science for beginners by Andrew Park. Compiled by Ozavize Ayomide Alabi**

**Section 1.**

Decision trees:

Decision trees are built similarly to support vector machines, i.e that are a category of supervised machine learning algorithms that are capable of solving both regression and classification problems. They are used in creating random forests, which is arguably the most powerful learning algorithm

An overview on decision trees.

Decision trees are essentially a tool (algorithm) that supports a decision that will influence all the other decisions that will be made. Decision trees are usually represented as a graph. Branches sprout to individually represent the results and they also have leaves which ae the class labels. Decision trees are one of the most powerful supervised learning methods you can use. A decision tree can perform any common data science task and the results you obtain from the training process are highly accurate.

Advantages of Decision Trees.

1. Decision trees are simple in design and therefore easy to implement.
2. Decision trees is one of the most efficient methods in exploring and determining the most important variables, as well as discovering the connections between them.
3. Decision trees are excellent in clearing up outliers in your data. Decision trees are not affected by outliers.
4. Decision trees can work with both numerical and categorical data.

Disadvantages of working with decision trees:

1. Overfitting. Algorithms tend to sometimes create a very complicated decision trees that will have issues generalizing data due to their complexity which is known as overfitting.
2. Decision trees can have issues with continuous variable. When continuous numerical variables are involved, decision trees lose certain amount of information. A continuous variable is one that is set to a range of numbers. E.g people between the ages of 18 and 26 considered as student age.

Classification and Regression trees

Decision trees are used for both regression tasks as well as classification tasks. However, this does not mean we implement the exact same decision tree in each case, decision trees need to be divided into classification and regression decision trees.

Classification decision trees are implemented when there’s a categorical dependent variable. Regression trees are only implemented when there’s a continuous dependent variable. The result from the training data is the mode of the total relevant observations, so any observations that can not be found will be predicted based on this value, for regression trees the value that results from the training stage is the mean of the total observations so unidentified observations are declared with the mean value.

The overfitting problem

Decision trees can lead to 100% accuracy score for the training set if we do not impose any limits the down side to this is that overfitting creeps in. the decision trees grown many branches and that’s the cause of overfitting. To solve this you need to impose limitation on how much the decision tree can grow and how many branches it can spawn. You can also prune a decision tree to control it.

To limit the size of the decision tree, you need to determine new parameters during the definition of the tree. The parameters include:

1. min\_samples\_split: this is the parameter to specify how many observations a node will require to be able to perform splitting. You can limit a decision tree with higher values.
2. min\_samples\_leaf: this is the parameter to tweak to determine how many observations are required by a node, or in other words a leaf.
3. max\_features: this paremeter is to control the features that are selected randomly. These features are the ones used to perform the best split.
4. max\_depth: this depth parameter consists of the depth value of the decision tree. For limiting overfitting, we are only interested in the maximum depth value.

Note: higher values translate to a high number of splits, therefore a high amount of information. By tweaking this value we control how the training model learns the connections in a sample.

Pruning.

This is a concept in machine learning to nearly eliminate the overfitting issue. This reduces the size of the trees in order to focus the resources on providing highly accurate results. The sections eliminated are those that don’t help the classification process and don’t lead to performance boost. Less complex decision trees lead to a better optimized model. The pruning implementation requires a number of steps and conditions. E.g for a decision tree to be suitable for pruning, it needs to have a high depth value. And the process needs to start at the bottom to avoid negative returns.

Decision Tree Implementation.

Creating a decision tree starts from the root node. The first step is to select one of the data attributes and set up a logical test based on it. Once you have a set of results you can branch out and create another set of tests, which you will use to create the subnode. The level of purity is determined based on the number of cases that sprout from a single class.

K-means Clustering

Unsupervised learning methods are ideal for working with unlabeled data. To be more specific, one of the best techniques if not the best, is to use a type of clustering algorithm. The main idea behind this approach is the cluster analysis which involves reducing the data observations to clusters, or subdivisions of data. Where each cluster contains information that is similar to that of a predefined attribute.

The concept of k-means clustering revolves around building data clusters based on the similarity of the values. The first step is to determine the value of K, which is represented by the total number of clusters we define.

Performance measurements used to accurately interpret the data.

* Homogeneity score: this metric can have a value between zero and one. The idea is that if we have score that is close to the value of one, then the cluster is mostly built from samples that belong to a single class. If the score is close to zero, then we have achieved a low homogeneity.
* Completeness score: this metric compliments the homogeneity measure. Its purpose is to give us information on how the measurements became part of a specific class.
* V-metric or sometimes the V-measure.
* Adjusted Rand index metric: this is score that is used to verify the similarity of the labeling, using a value between zero and one, the Rand index simply determine the relation between the distribution sets.

**Section 2. Python for Data Analysis.**

A step-by-step guide to master the basics of data analysis in python using pandas, numpy and Ipython.

Introduction.

Data analysis places a vital role in everyone’s life today. Companies use data to gain insights on their customers, i.e., to know better about their customers. With lots of competitors trying to steal other companies’ customers, it is important companies use data to understand their customers behavior so they are able to retain customers. Data analysis is not used by companies to retain customers and maximize profits, but governments also use data in governance, in fact they are the largest collectors of data. From collecting data on its citizens to enable them allocate resources properly to collecting data on persons of interest deemed a threat to their nations security buttressing how data can be used to improve the security of a nation.

It is vital that the data analyst has the right skills need for data analysis and also handling the data, because incorrect data can mislead and data can also be manipulated.

There a lot of tools for data analysis, some analyst use Excel which is good but is quite limited, have knowledge on python and its various libraries is need important as python offers a wider range of tools and features for data analysis.

A problem organization may encounter is in protecting the integrity of its data. Measures should be put in place to protect the organization from using contaminated data.

We live in a world where data is being produced in large amounts from automated systems. Learning data analysis through python can help process and extract useful information from data and draw meaningful conclusions from them. One area where data analysis skills will come in handy is in forecasting. Through data analysis predictive models can be built to help your organization meet its objectives.

A good predictive model is only as good as the quality of data introduced into it, the data modeling methods, and more importantly the data set used. Another important aspect of data analysis is data visualization and data visualization is about presentation. A data model is should be good enough for an audience to understand it at first contact, the data analyst should also learn how to plot data on different visualization to help you get a rough idea of the nature of the data you are working with.

When data analysis has been conducted, the data analyst should have a data model complete with visual concepts that will help in predicting outcomes and responses before you can proceed to the testing phase.

**Chap 1 What is Data Analysis?**

in this chapter we are going to see how data analysis can help companies to gain insights and predictions from the data they gather so as to make informed and smart business decisions.

Data analysis is a practice where we take raw data that business in this context has been collecting and then organize and order it to ensure it can be useful. In this process, the information that is the most useful Is extracted and then used from the raw data.

One thing that we need to be careful about when we are working with data analysis, is to be careful about the way they manipulate the data they have because manipulating data in a wrong way can lead to pushing certain conclusions or agendas that are not there.

In the process of going through and organizing data, a trend is likely going to emerge, and sometimes more than one trend. It is important to be able to take some time to highlight these trends, usually in the write-up that is being done on the data.

Spending time with the write-ups of the data, graphs, charts, and other ways to represent and show the data to those who need it the most which will form one of the final steps that come with data analysis.

Data visualization is a key way we can write out data in a manner that is easy to understand and read through. There are many choices of visuals that we can work with. Working with the method that is the best for your needs and the data that we are working with is going to be the best way to determine the visual that is going to be best for you as a data analyst.

When working with a visual of the data to make it easier to work with and understand, it is fine to add in some of the raw data as the appendix. This allows whoever that is going to be working with the data a chance to regularly check the resources and specific number.

It may not be best, or it may not be the best data to work with if it is gathered up from just a small sample or a sample that you worry is not random.

There are many great benefits with working with data analysis, It can help you to learn more about your industry and the customers who are going to purchase your products as a company.

**Chap 2, why python for data analysis?**

Python is a great programming language that is easy to learn, has a lot of power and can handle pretty much all the tasks you need to handle with data analysis and machine learning.

The basics of the python language.

Python is an object-oriented programming language or OOP language with the user in mind, while still providing the functions, extensions and libraries, that will make data analysis and machine learning as easy to work with as possible.

Machine learning is a process where we teach a model or computer how to make decision on its own, and python has a lot of extensions that can make it work with machine learning.

Python is also compatible with other languages. You are able to add on any extension and still write code in python. Combining python with different libraries that are available enable the data analyst to get more complicated task done. Python also relies on classes and objects to help organize information and keep things in line.

How can python help with data analysis?

Python is one of the most famous programming languages, and it works so well with data science. But some of the benefits of using python to help out with your data science model or project include:

* Python is as simple as it gets: The language had two main ideas in mind when it was first stated and these include readability and scalability.
* Python is fast and attractive: python code is leaner and much better looking than others. For example, the python code takes up one-third of the volume that we see with code in java and one-fifth of the code in C++ just to do the same task. The use of common expressions in code writing, rather than going with variable declarations and space in place of ugly brackets can also help python code look much better.
* Data formats are not a worrisome thing in python: python can basically work with any kind of data format. Python comes with libraries such as pandas for helping data scientist to handle all of the parts of not only data analysis, but also the whole process of data science.
* Increased demand of the python library: over the years, there has been an increased demand for programmers who can work with python.
* The vibrant community available: there has been a steady increase in the community of python programmers, these members are willing to help others who may get stuck on their projects and need help moving forward.

In the course of the guidebook, we are going to see that there are many python libraries that can help get more work done in python and data analysis. A few to mention are TensorFlow, Pandas, Numpy, SciPy, Scikit-Learn. Sometimes these libraries can work on their own, sometimes they need to be combined with another library so that they can draw features and functionalities from each other.

**Chap 3, The Steps of Data Analysis.**

This chapter highlights the steps that are important to the process of data analysis. For most businesses, there isn’t going to be a problem with lack of information, many a times the problem it the over-amount of data they have to work with which is going to make it harder to come up with a clear decision based on the data. This means that the business needs to know that the data they have is right for the questions they want to be answered.

To help us get all of this done, a data scientist needs to go through some of the basic steps that are needed to use data to make better decisions overall. Some of the steps that a data scientist can use to help make data analysis more productive for better decision making in the company include:

* Defining your question
* Setting up clear measurement
* Collecting the Data
* Analyzing the Data
* Interpreting the results.

Defining your question

Defining the questions you would like to handle is the first step needed when it comes with working with data analysis. Not doing this would result in confusion and getting lost with the data. You need a clear picture of where you want to go and what you would like to learn from the data and then work from there.

In data analysis, asking the right questions is important but these questions need to be concise, measurable and clear. They need to be designed in such a way that they either qualify or disqualify some of the potential solutions that you are looking for on a specific problem or opportunity.

Setting up Clear Measurements

This involves setting up some clear priorities on your measurements which can be broken into two subsets. The first path is to decide **what we want to measure**. And then we need to **decide how to measure it.**

Using the example of a government contractor, who costs have increased overtime. to answer the question of what we want to measure, we would need to the information on the;

* Number and cost that is present for the current staff and
* The amount of time that these employees are going to spend on necessary business functions.

in trying to answer the above, we would also like to see the number of staff the company can do without say some are been underutilized and also consider the opinions of stakeholders to the company.

The above been said. The second step would be how to measure the above. This would also bring some important questions to consider which will include

* What is the time frame being looked at?
* What Is the unit of measure that is relied on?
* What factors are important to consider in all of this.

Collecting the Data

In the process of collecting the data, there important points to keep in mind:

1. You need to determine what information that is needed to work with. It is advisable to go through some of the existing databases and existing sources you have on hand, because it is simple and easier and can save a lot of money as well. You can now move out to some other sources later, if you need more information.
2. Determine the naming system and file storing system you would like to use. This makes it easier for all team members to collaborate and also prevents wasting of time and money collecting the same kind of information more than once.
3. If you are gathering data through interviews and observations, then you need to develop an interview template ahead of time. This saves time and ensures continuity goes in the process as well.
4. You need to be able to keep the collected data that we have as organized as possible. You can work with a log that has collection dates and add in notes on the sources as needed.

It is important to ensure that the data is the same format, and this can be achieved by storing your data in a standardized database. Also missing value, duplicate values should be taken care off via cleaning and outliers should just be ignored and gotten rid of. **Although if there seems to be a lot of outliers and they appear in the same spot, it is advisable to take a look at this to see what is going on and if this is new information that should be paid attention too.**

Rows or columns with missing values can either be deleted or replaced with the mean of the values in that row or column depending on the number of missing values. A lot of missing values can skew the outcome of the data.

Analyzing the data

After collecting the data, the next step would be to analyze it. And this involves manipulating it, and there are methods by which this can be done. Example, we can plot the data out to see correlations or do a pivot table in excel which would help us filter and sort out the data by different variables and make it easier to calculate the minimum, maximum, mean and even standard deviation.

Also there may be need to go out and collect some more data to work with or revise the original question that is been worked with.

Interpreting the results

After the process of analyzing your data, which will often require some training and testing of your data through the algorithm to make sure it works the way you want. The next process is to go through the data and interprete the result. Note that you are never going to be able to completely prove your hypothesis true at 100% level. But an hypothesis can be rejected.

In the course of interpreting the results, we need to ask some major questions about the data which are going to include:

1. Is the data able to answer the original question that we had in the beginning and how?
2. Does the data help us to defend against the objects that were raised and how?
3. Are there any limitations to the conclusion that we have, or any angles that have not been considered.

By following the steps outlined in the above for data analysis, its going to help your company/business make better informed decisions.

**Chap 4, python libraries**

in this chapter, we are going to look at some of the best python deep learning libraries for data analysis, which would be highlighted below.

Caffe

Caffe is technically not a python library, but it provides some bindings into the python language. And these bindings are used when we are going to deploy the network to the wild, rather than just trying to train a model. It is used pretty much everywhere in all deep learning models that need to be created.

Theano

This is a library that has helped to develop and works with a lot of other deep learning libraries. Theano can be used to define, optimize and evaluate a lot of mathematical expressions that will involve multi-dimensional arrays. Theano is tightly integrated with the numpy library, hence its able to achieve the listed prior. Using the Theano library to build deep learning networks is often seen as the building blocks of these neural networks.

TensorFlow

This is similar to what we can find in the Theano library, TensorFlow is open-sourced and can work with numerical computation with the help of a data flow graph. Originally developed to be used with the research on the Google brain Team within Google’s machine intelligence organization.

One of the major benfits of TensorFlow as compared to Theano, is that it is able to work with distributed computing.

Keras

This is seen as a modular neural network library that is more minimalistic than the previously talked about ones. Can use TensorFlow or Theano as the backend. Its primary goal is to be able to experiment on your models quickly and get from the idea you have over to the result as fast as possible.

Sklearn-Theano

Useful when training your CNN end to end, you just pass your input images through a popular pre-trained architecture that can include options like VGGNet, AlexNet and OverFeat.

Nolearn

This is a good library to work with, good to help out with some GPU experiments, good for performing deep learning on an Amazon EC2 GPU instance. Code with Nolearn is going to be compatible with SciKit-learn.

Digits.

This isn’t considered a true deep learning library. Although it is written out in python and stands for Deep Learning GPU training system. It is more of a web application that can be used for training some of the models of deep learning that can be created with the help of the caffe library.

**Chap 5 Neural Networks**

Convolutional neural networks (CNNs)

This are one of the main categories of deep neural networks that have proven to be very effective in numerous computer science areas like object recognition, object classification and computer vision. ConvNets have been used for many years for distinguishing faces apart, identifying objects, powering vision in self driving cars and robots.

Lately ConvNets have been used effectively in natural language processing problems like sentence classification.

The first CNN was proposed in 1988 by Yann LeCun called LeNet which helped significantly propel the overall field of deep learning. It was primarily used for character recognition problems such as digits and codes.

Standard Neural Networks we use today come with four main operations including **convolution, ReLU non-linearity activation functions, sub-sampling or pooling and classification of their fully-connected layers.**

How convolutional neural networks work.

A CNN structure is normally used for various deep learning problems.

CNNs are used for object recognition, object segmentation, detection and computer vision due to their structure.

The use of CNNs has become populate due to three main factors.

1. The first is the structure of CNNs which eliminates the need for performing manual data extraction as all data features are learned directly by the convolutional Neural Networks.
2. They produce amazing, state-of-art object recognition results.
3. CNNs can easily be retained for many new object recognition tasks to help build other deep neural networks.

A CNN can contain hundreds of layers, which each learns automatically to detect many different features of an image data. Filters are commonly applied to every training image at different resolutions, so the output of every convolved image is used as the input to the following convolution layer

Stride and padding

After specifiying the depth, you also must specify the stride that you slide over the filter. When you have a stride that is one, you move one pixel at a time. By default stride value is one

Parameter Sharing

Matrix Multiplication

**Chap 6 Pandas**

Pandas is built on Numpy and they are meant to be used together. This makes it extremely easy to extract arrays from the data frames. once these arrays are extracted, they can be turned into data frames themselves. E.g

Import pandas as pd

Import numpy as np

Marketing\_filename = ‘regression-datasets-marketing.csv’

Marketing = pd.read\_csv(marketing\_filename, header=None)

In the above code, we are uploading data to a dataframe. We are going to use the ‘values’ method to extract an array.

Input:

marketing\_array = marketing.values

marketing\_array.dtype

out:

dtype(‘float64’)

the dtype () method gives us the type of the array.

Matrix Operations

This includes matrix calculations, such as matrix to matrix multiplication.

coefs = np.array([1., 0.5, 0.5, 0.5, 0.5, 0.5,])

coefs\_matrix = np.column\_stack((coefs, coefs[::1]))

print (coefs\_matrix)

out:

[[1. 0.5]

[0.5 0.5]

…

The np.column\_stack((coefs, coefs[::1]))

Returns the array in a column stack.

Now we can perform the multiplication. Here’s an example of multiplying the array with the vector:

In:

M = 5

np.dot(M,coefs)

out:

array ([5.,2.5, 2.5, 2.5, 2.5])

in both multiplication operations, we used the np.dot function to achieve it.

Slicing and Indexing

Indexing is great for viewing the nd-array by sending an instruction to visualize the slice of columns and rows or the index.

We create a 10 X 10 array. This is initially a two-dimensional array.

>>> import numpy as np

M = np.arrange(100, dtype = int).reshape(10.10)

The above code creates 10 x 10 matrix of numbers from 0 to one integer less than that passed as the first argument. i.e from 0 – 99. Note that we pass the type of the data type as the second argument.

>>> M [2:9:2]

>>>array([[20, 21, 22, 23, 24, 25, 26, 27, 28, 29],

[40, 41, 42, 43, 44, 45, 46, 47, 48, 49],

[60, 61, 62, 63, 64, 65, 66, 67, 68, 69],

[80, 81, 82, 83, 84, 85, 86, 87, 88, 89]])

The above code M[2:9:2] returns an array from rows 2 to 8, but only the ones that are evenly numbered.

We can extract a column from the above array, but only from index 5. With the code below

>>>M [2:9:2, 5]

>>>array([25, 45, 65, 85])

The code below would return an array of from rows 2 to 8, column index 5 to the end.

>>>M[2:9:2,5 :]

array([[25, 26, 27, 28, 29],

[45, 46, 47, 48, 49],

[65, 66, 67, 68, 69],

[85, 86, 87, 88, 89]])

If we try a negative index, doing such would reverse the array.

>>>M[2:9:2,5::-1]

array([[25, 24, 23, 22, 21, 20],

[45, 44, 43, 42, 41, 40],

[65, 64, 63, 62, 61, 60],

[85, 84, 83, 82, 81, 80]])

It is not advisable to make modifications to the original arrays, this might lead to negative side effects. instead use the copy() method and store it in a new variable

N = M[2:9:2,5:].copy

**Chap 7 Jupyter Notebook**

Getting started with Jupyter Notebook (Ipython)

Jupyter Note pad is an open-source web application that permits you to produce and share files containing live code, formulas and visualizations and narrative text.

Jupyter has assistance for 40 various \* languages and python is one of them. The python requirements are python 3.3 or higher or python 2.7 for setting up the jupyter notebook itself.

Setting up jupyter utilizing Anaconda

You can set up python and jupyter utilizing the Anaconda Distribution, which includes python, the jupyter Notebook. We can download the newest python variation.

You can also set up Jupyter Notebook using PIP (Preferred Installer Package)

* Python3 -m pip install jupyter. Use pip install Jupyter on your command prompt
* Command to run the Jupyter notebook. “Jupyter Notebook” which must be done in the directory that contains the executable

Running the command would open the URL to the Jupyter notebook in your web browser. By default <http://localhost:8888>. When the Notepad opens, you will see the Notebook Dashboard which will reveal a list of the notepads, files, and subdirectories in the directory site. It is advisable to begin your Notepad server.

Create a new Notebook

At the top right corner of the dashboard, you will see a new drop-down button. Click it to open the drop-down list and after that click Python 3. This will open a brand-new notebook.

Few Useful Commands

* Command to open a notebook in the currently running notebook server.

--- jupyter notebook notebook\_name.ipynb

by default, the notepad server begins on port 8888. If port 8888 is not available or in usage, the Notepad server browses the next readily available port

--- jupyter note pad – port 9999

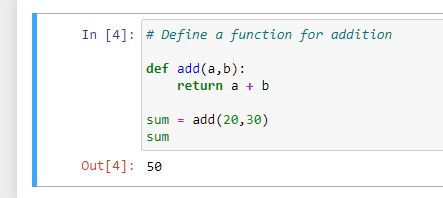
* Command to begin the Notepad server without opening a Web internet browser:

--- jupyter notebook – no-browser

* You can get help messages for other command line arguments using the --help flag

--- jupyter notebook – help

Running your first code.



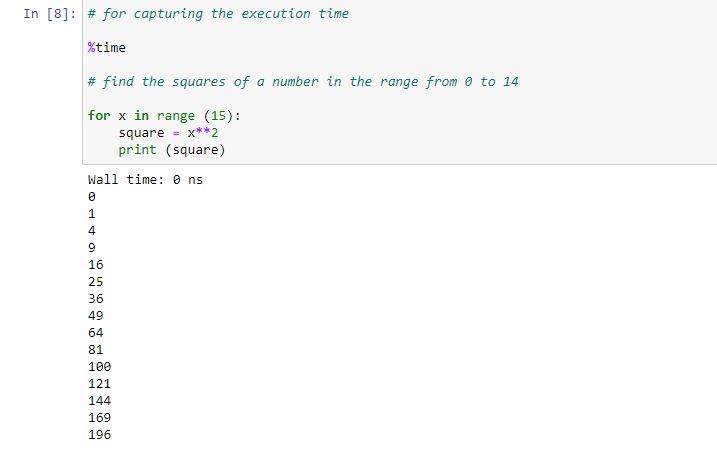
To run a code in the cell, you can press ctrl + Enter in that cell or, click cell on the menu bar to list options available for running your code.

Jupyter Notebook tips and Tricks

Jupyter Notepad previously called IPython Notepad is what makes information analysis in Python more efficient and effective.

Some functions of the Jupyter Notepad

%timeit and %time

We can use the %time command to inspect the running time of a specific cell. i.e see the time it takes to carry out the code in a cell.

Commenting multiple lines of code

To comment multiple lines of code or text in Jupyter, simply highlight them and press ctrl + /

**Chap 8 The PyTorch Library**

this is a Python-based package that works for scientific computing and it relies on the power that it can receive from graphics processing units. This library is known for providing two of the most high level features out of all the other deep learning libraries, which include:

* tensor computation with the support of a strong GPU acceleration,
* being able to build up the deep neural networks on an autograd-system that is tape-based.

The Beginnings of PyTorch.

Pytorch is one of the newest libraries out there that works with Python and can help with deep learning. It makes it easy to build complex neural networks. Its functionality can be extended by adding other Python packages like Cython, SciPy, and Numpy.

Below are some highlights of why professionals of all sorts love the Pytorch library.

* The interface is simple to use. Which makes it simple to operate and run as we do with python
* It is pythonic in nature. Since it is pythonic in nature, it will smoothly integrate to work with the python data science task.
* Computational graphs. The PyTorch library provides with the platform with some dynamic computational graphs.

The Community for PyTorch

The PyTorch community of developers and other professionals is growing daily. And when it comes to artificial intelligence and models of deep learning. It has been used by many tech giants to improve their deep learning models such as Uber, NVIDIA, Twitter and Facebook.

Why use PyTorch with Data Analysis

The TensorFlow library from Google which is very helpful in the field of data science, data analysis, artificial intelligence or deep learning. It is believed that PyTorch is now the biggest competitor to TensorFlow when it comes to data handling. There are many reasons this is happening of which we highlight below:

* Dynamic computational graphs are going to be popular among researchers.
* The second benefit is that this PyTorch comes with a different kind of back-end support. PyTorch is going to use a different backend based on what you are doing. The GPU, CPU and other functional features will all come with a different backend rather than focusing on just one back-end to handle all of these.
* The imperative style is another benefit or working with this kind of library: i.e it is easy to use and very intuitive.
* PyTorch Is easy to extend. In particular it is integrated to work well with the code C++ and it is going to share a bit of the backend with this language when we work on our framework for deep learning.
* PyTorch is going to be seen as a Python approach library.

PyTorch 1.0 How this moves us from research to production.

There are a few downsides that come with the PyTorch. One of these include its lacking when it comes to supporting production.

**Chap 9 Data Visualization and Matplotlib**

in the journey of data science, you would get to a point where its time to actually see the data and visualize it. Data visualized using pictograms is better understood when making comparisms than with writing long paragraphs about it.

Matplotlib, an extension of Numpy and Scipy can help create these visualizations, be it a line graph, pie chart, or histogram.

Data visualization is going to be the presentation of quantitative information in a form that is more graphical, i.e you can use data visualization to take a set of data, whether it is small or large and turn it into visuals that are much easier for the brain to understand and to process.

A combination of more than one of these visuals and even some added information in the mix is known as an infographic. A good data visualization is created when we can add together design, data science, and communication all in one. It is a good idea to look at some of the plots later on and decide which ones can talk about your data the best and which one you should use. The best kinds of data visualisations are those that include a lot of complex ideas that can be communicated in a manner that has efficiency, precision and clarity to name a few. Remember to keep the charts and graphs as simple as possible, because they always make the best types of graphs. Adding a lot of elements that are not needed would simply distract us from the data.

Why is it important to use data visualization

Making decisions without visualization is a poor way to go about, and does not ensure that you know what is going on with the data in front of you, or that you will see the full amount of information and trends that are presented. In addition some of the other reasons that data analysts like to work with these kinds of visualizations include:

* It helps them make better decisions.

You get a more broader and understandable view of the data from visualisations. Hence you are able to make better decisions and ask the right questions.

* It helps data analyst tell a story with a lot of meaning behind it.

This is also employed today in many media outfits

* Data literacy.

Being able to read and then understand the data visualization has become something of a requirement in our modern world.

The history that comes with data visualization.

We highlight Florence Nightingale, a nurse during the Crimean wars and also a data journalist who was known for her rose or coxcomb diagrams. This helped get better conditions in the hospitals which on the long run saved many lives of soldiers who were there

Why we should use data visualization.

Some of the different reasons why you should use data visualization include:

* It can take the data that you have and makes it easier for you to remember and understand rather than just reading the information and hoping it makes sense
* It is going to give you the ability to discover facts that are not known to you, trends and even outliers in the information that could be useful as well
* It makes your life a bit easier because it helps you to visualize relationships and patterns quickly and effectively.
* It ensure that you can ask questions in a better way and make the best decisions for your business.

What is matplotlib?

Most business likely have a ton of information in place an you want to make sure that you are not only choosing the right kind of visualization but that you are also able to put it all together and make the right visual. This is where matplotlib comes into play. The idea behind matplotlib is that a picture is worth a thousand words.

A few challenges that can show up for programmers as regards matplotlib include:

* The library is a pretty large library containing about 70000 lines of code, and it is always expanding.
* Matplotlib is going to be home for more that one type of interface, and is able to interact with a lot of backends as well. The backend deals with the whole process that happens when the chart is rendered, not just structured.
* You would have to know how to use Numpy and Scipy to make things easier. Since matplotlib is going to be part of them.

**Chap 10, Applications of Big Data Analysis**

The applications of big data and big data analysis is benefitting to both big and small companies. Areas where Big Data and Big Data Analysis can be applied include:

eCommerce

there are over 2.6 billion and counting active social media users, which include customers and potential customers for every company out there. The race is on to create more effective marketing and social media strategies powered by machine learning to turn prospective customers into loyal fans. The ability to bridge the gap between execution and big data analysis has been supplemented by artificial intelligence marketing solutions.

Artificial Intelligence (AI) can be defined as a method of you using artificial intelligence consonants like machine learning on available customer data to anticipate customer’s needs and expectations while significantly improving the customers journey. Marketers are able to boost their performance and return on investment in the light of big data insights provided by artificial intelligence marketing solutions. The key elements that make AI as powerful are:

* Big Data: the ability to aggregate and segment a huge dump of data with minimal work.
* Machine Learning: machine learning platforms enable marketers to identify trends or common occurrences and gather effective insights and responses, thereby deciphering the root cause and probability of recuring events.
* Intuitive platform: super fast and easy to operate applications are integral to AI marketing. AI is capable of interpreting emotions and communicating like human.

Predictive Analysis.

AI tech-based solultions are capable of extracting information from data assets to predict future trends. AI has made it possible to model trends. And this predictive analysis models can be reliably used in decision-making to analyse customers purchase behaivour, determine when the customer is more likely to purchase something new or reorder an old purchase. FedEX and Sprint are companies using predictive analysis to identify customers who are at potential risk of deflecting to a competitor.

Smart searches.

Today’s search engines are not only accurate but also much faster unlike before where if you type in women’s flip flops on Nike.com you wouldn’t find what you were looking for. This upgrade has largely been brought on by innovations like 1. Semantic search and 2. Natural language processing that enable search engines to identify links between products and provide relevant search results, recommend similar items. Artificial intelligence technology and big data solutions can rapidly analyse user search patterns and identify key areas that marketing companies should focus on. Examples of tech companies that have Artificial Intel

license-based search algorithms include “RankBrain” which was the first of its kind.

Recommendation Engines

Recommendation Engines has become what is considered a fans favourite, Apple Music for example knows a person’s taste in music better than his/her partner, youtube, Amazon present you with videos, list of products you might be interested in watching, buying. By gathering and analysing consumer data paired with individual profile information and demographics using predictive analysis-based systems, we can continually learn and adapt based on consumer activities such as likes and dislikes on the products in real time.

Product Categorization and pricing

E commerce businesses and marketing solutions have adopted AI for use in categorizing and tagging inventory. These e-commerce businesses have to do with a lot of awful data. Merchants who use this platforms have different ways of naming the same thing. But with the help of artificial intelligence systems such products can be categorized with precision using other associated information on the product. Also business are able to determine what the consumer would be willing to pay for a product, see whether they need to implement a price discount or adjust their prices accordingly with the top marketers of such products.

Customer Targeting and Segmentation.

With AI it is possible to reach customers with high-level personalization, to do this they need to target increasing granular segments. The AI needs cand draw on existing customer data and train machine learning algorithms against the gold standard training sets to identify common properties and significant variables. The granular segments could be as simple as location, gender, and age or as complex as buyer’s persona and past behavior.

Sales and marketing forecast.

one of the most straight forward artificial intelligence applications in marketing is in the development of sales and marketing forecasting models. The high volume of quantifiable data such as clicks, purchases, email responses and time spent on webpages serve as training resources for the machine learning algorithm. Leading business intelligence and production companies in the market are Sisense, RapidMiner and Birst.

Programmatic Advertisement Targeting

Artificial intelligence has made targeting program-based advertisement more efficient. Programmatic advertising can be defined as the automated process of buying and selling ad inventory to an exchange which connects advertisers to publishers. Artificial intelligence is able to serve a particular ad, the probability of an ad turning sales, the receptiveness of the user and the likelihood of engagement with the ad.

Programmatic companies can gather and analyze visiting customers data and behaviors to optimize real-time campaigns and to target the audience more precisely. Google and Facebook serve as the gold standard for efficient and effective advertising and are geared towards providing a user-friendly platform that will all non-technical marketing companies to start, run and measure their initiative and campaigns online.

Visual search and image recognition

There have been significant advancements in artificial intelligence-based image recognition and analysis technology which has resulted in uncanny visual search functionalites. Companies like google with google lens, pinterest with pinterest lens are examples of such that have adopted AI image recognition. Companies like Target and Asos have have also taken advantage of the visual search technology where shoppers can take a picture of products that they would like to purchase while they are out and find similar items on their e-commerce site.

Healthcare Industry.

With the massive availability of healthcare data, big data analysis has caused a big shift in the health care system, the primary concern for the use of bigdata analysis is the relationship between patient outcomes and the type of treatment/preventive measures. AI technology is able to sift through huge amounts.

Some of the medical specialties with increasing big data analysis-based AI research and application include:

* Radiology

AI helps to supplement clinician’s ability to detect changes in an image that can easily be missed by the human eye.

* Electronic health records

AI helps Automate the accurate logging of health-related data in the systems which optimizes information spread and access.

* Imaging

AI research is helping doctors in evaluating the outcome of corrective jaw surgery as well as assessing the cleft palate therapy to predict facial attractiveness.

Entertainment Industry

Companies such as Netflix and Hulu are leveraging on big data analysis and AI to provide their customers more personalized recommendations based on their past history and behavior. Movie producers are also leveraging on machine learning in the development of film trailers and advertisement. Also, in movie productions redundant task are being automated using AI

Marketing and Advertising

A machine learning algorithm developed as a result of big data analysis can be easily trained with text, stills and video as data sources. It can then extract objects and concepts form these sources and recommend efficient marketing and advertising solutions.

Example, a tool called luban’ developed by Alibaba that can create banners at lightning speed was able to create 8000 banners per second during a shopping extravangaza called singles day in china.

Also we see the use AI in collaboration between 20th century fox and IBM’s Watson which created a six minute trailer for a horror movie morgan, after it was trained with video data from a hundred horror movies.

Marketing process can be accelerated exponentially through an AI marketing platform. E.g the AI based marketing platform developed by Albert Intelligence marketing can generate autonomous campaign management strategies, create custom solutions and perform audience targeting. A 183% improvement in customer transaction rate and over 600% higher conversion efficiency was recorded by the company credited to the use of their AI based platform.

Personalization of user experience.

Here we consider Netflix and its AI solution ‘meson’ for recommending personalized videos to its consumers and the dynamic optimizer for helping maintaining video quality of videos that have been compressed, when streaming.

Search and optimization and classification

With the infoglut in the internet, it makes it difficult to search exactly what you are looking for, with AI it is easier to search with much precision what you are looking for on the internet.

**Section 3. Python Machine Learning**

**A Complete Guide for Beginners on Machine Learning and Deep Learning with Python.**

We have barely scratched the surface when it comes to our knowledge on machine learning, as a data scientist, this is where you might want to focus all your energy on as the prospects for it are enormous. Algorithms are the backbone of almost everything that you will do in machine learning.

Machine learning is as much about programming that it is about probability. Machine learning also involves a lot of uncertainty which is the one of the differences between machine learning and programming.

Once you build a model, you train it, test it and eventually deploy it. Since these models would interact with humans, you must consider that some input parameters might be correct, while other might not.

Many mathematical computations are involved in machine learning. So we must learn how to perform specific operations on data based on unique instructions.

As you work with different sets of data, there is the possibility that you will come across massive data sets. The challenge with this is that you must learn how to break down the data into small units that your system can handle and process without any challenges, i.e trying to avoid overworking your learning model.

Eventually when you build a machine model, the aim is to integrate it into some application that people use daily.

**Chap 1, What is machine learning.**

With the advance in technology, AI is now responsible for different tasks like prediction, diagnosis, recognition and so on.

Data are fed into the machines and these machines learn from these data. These data are referred to as training data because they are used to train machines. They analyze this data in order to know the actions they are to perform Machines use various learning mechanisms for analyzing the data according to the actions that they need to perform.

These mechanisms can be broadly classified into two categories:

* Supervised learning
* Unsupervised learning

There are different reasons why machine learning is important:

* It helps us understand a couple of things about human learning.
* It helps increase the accuracy, effectiveness and efficiency of machines.

Machine learning and data mining are intertwined. Data mining is the process of going through tons of data to find bowlers of correlation and relationships that exist within. This is another benefit of machine learning that helps machines find any vital information. Machine learning helps gets machines acclimatized to its environment. To ensure optimum performance.

Applications of machine learning.

Machine learning helps operate a large scale of data that’s available and enables the users to draw helpful predictions based on the given information. Machine learning helps process large amounts of data within a short time frame.

Technical applications of ML and how these apply to problems in the real world include:

1. Virtual personal assistants

Popular virtual personal assistants like Alexa, Siri, Google now help users find necessary information via voice commands. ML then enables the virtual assistants to find the necessary information you need by gathering and refining information. You could ask questions like, what is my schedule for today” , ”are there flights going between London and Germany?” etc.

1. Density estimation

Machine learning can be used to determine the volume of a book (density of words) and then recommend books similar to such a book

1. Latent variables
2. Reduction of Dimensionality

Data obtained tend to have some variables and dimensions and if more than 3 dimensions are involved then the human mind can’t visualize it. ML helps to reduce these data into manageable proportions.

Advantages and Disadvantages of Machine learning.

The disadvantages of machine learning are:

1. Difficulty in identifying a bias in the created machine learning model which will lead to wrong inferences
2. Some social scientist will begin to rely only on machine learning.

Some Advantages include:

1. Human beings cannot process large volumes of data, let alone analyse it. ML automates this task to understand and analyse the data so a conclusion can be reached
2. Machine learning is getting better. With the advent of deep learning systems, the costs of data engineering and pre-processing of data are reducing.

**Chap 2 Machine Learning, concepts and terms**

Machine learning is done by feeding the machine with relevant training data sets. A system with AI can learn, predict and improve the results it provides through training. The models that are based on labeled training data sets are termed as supervised machine learning models. A supervised machine learning model allows the engineer to provide some feedback to the machine. E.g you can feed the machine with the response negative 1 or 0 depending on whether the prediction is wrong or right. The concept of ML, AI and deep learning although used interchangeably by most people, are not the same.

Machine learning, Artificial Intelligence and Deep learning.

Artificial intelligence is a technique used to make machines mimic human behavior. E.g of AI machines include deep blue chess and IBM’s Watson

Machine learning is the use of statistical and mathematical models to help machines learn about human behavior, this is done using past data.

Deep Learning is a subset of machine learning and are the functions, tools the engineer uses to help train the machine. A machine can learn to take the correct option to derive an output. Neural networks and natural language processing are a part of the deep learning ecosystem

Objectives of machine learning system

The system of ML has one of the following objectives:

* Predict a category

This model analyses the input data and then predicts a category under which the output will fall. The prediction in this case is usually a binary answer based on yes or no. e.g is this mail spam or not, the output is attained by referencing a group of data that will indicate whether a certain mail falls under the category of spam or not based on specific keywords. This process is known as classification

* Predict a quantity

This system is used to predict values like rainfall, according go different whether attributes such as temperature, humidity and air pressure etc. this sort of prediction is called regression. The regression algorithm has various subdivisions like linear regression and multiple regression.

* Anomaly detector systems

The purpose of a model in anomaly detection is to detect any outliers in the given set of data. These applications are used in banking and e-commerce systems.

* Clustering systems

This involves clustering a consumer for example by behavioral factors like their age group, region where they live in or programs they like to use, and now suggestion different kind of programs the user might like based on the cluster the said user belongs to during classification.

Categories of Machine learning systems

In the classical world with traditional machines we give the machine a set of instructions and input parameters which the machine will use to process the data and derive an output using specific commands. In ML we never instruct a system. The machine/computer interacts with the historical dataset, develops and algorithm based on it, makes decisions like a human would, analyse the information and then provide an output. This can be done in short periods and provides results with high accuracy.

There are different types of ML algorithms, and they are classified based on the purpose of the algorithm:

* Supervised learning

In this model, the engineer feeds the machine with labeled data. The engineer will determine what the output of the system or specific data sets should be. This type of algorithm is also called a predictive algorithm. We refer to the currency (label) and weight (feature/attribute) where each currency is given an attribute of weight. The SML system is trained with this data set, and when it comes across say a weight of 3 grams, it is able to predict that the coin Is 1INR

* Unsupervised learning.

In this type of model, it is more sophisticated. The system will learn to identify patterns in unlabeled data and produce an output. This algorithm is used to draw any meaningful inference from large data-sets. This model is also called a descriptive model. It is often used in data mining applications that involve large volumes of unstructured input data. Some common algorithms which fall under unsupervised Machine Learning include density estimation, clustering, data reduction and compressing. The concept is that the complexity of data can be reduced if the derived output is useful.

* Reinforced learning.

In this type of model, the system will learn to behave in a specific environment, and take actions based on that environment.

Some common classes of supervised learning algorithms include:

* Logistic Regression, Linear Regression (Numeric prediction), Polynomial Regression
* Regression trees, Gradient Descent, Random Forest
* Decision Trees (Classification), K-Nearest Algorithm (classification), Naïve Bayes
* Support Vector Machines

Some common classes of Unsupervised Learning

* Clustering, compressing density estimation and data reduction
* K-means Clustering, Association Rules (Pattern Detection), Single Value Decomposition
* Fuzzy Means, Partial Least Squares, Hierarchical Clustering, Principal Component Analysis

Steps in building a Machine Learning System

* Define Objective

The first step is to define the purpose you wish to accomplish with your system. The kind of data you use, the algorithm and other factors will primarily depend on the objective or the kind of prediction you want the system to produce

* Collect Data

This is the process of collecting all the relevant data that you will use to train the algorithm

* Prepare Data

This is an important step that is usually overlooked. The cleaner and the more relevant the data you are using is, the more accurate the prediction or the output will be

* Select Algorithm

There are different algorithms that you can choose, like Structured Vector Machine (SVM), k-nearest, Naïve-Bayes. The algorithm you use will primarily depend on the objective you wish to attain with the model.

* Train Model
* Test Model
* Predict

Multiple iteration will be performed and you can also feed the feedback into the system to improve its predictions over time

* Deploy

Once the model is tested and you are satisfied with the way its working, the said model will be sterilized and can be integrated into any application.

**Chap 3 Linear Regression with Python**

Linear regression with one variable

Here we focus on when we have just one variable. We use the set of data car\_price.csv so that we can learn what the price of the car is going to be. We will have the price of the car be our dependent variable and the year of the car going to be the independent variable. To help us make a good prediction on the price of the cars we will need to use the Scikit Learn library from python to help us get the right algorithm for linear regression.

The steps to go about it involves:

* Importing the right Libraries

The codes that we need to get the libraries for this section include:

Import pandas as pd

Import numpy as np

Import matplotlib.pyplot as plt

%matplotlib inline

The final line needs to be there if you are using jupyter notebook.

* Importing the Dataset

This is the next step after importing the appropriate libraries. This involves import the data sets that you want to use for this training algorithm. We are working with the cars\_price.csv dataset. We execute the following script to help you get the data set in the right place.

car\_data = pd.read\_csv(‘path to file on your computer’)

* Analysing the data

Before you use the data that will help with training. It is always best practice to analyse the data for any scaling or any values that are missing. We use the script below to return the first five rows of the data set you want to bring up.

car\_data.head()

the describe function can be used to return all of the statistical details of the data set

car\_data.describe()

we are now going to take the data points and plot them on a graph to see if there is a

relationship between the price and the year of the car and also to see if the linear regression algorithm is suitable for the task at hand.

We run the script below to plot the graph.

plt.scatter(car\_data[‘Year’], car\_data[‘Price’])

plt.title(‘Year vs Price’)

plt.xlabel(‘Year’)

plt.ylabel(‘Price’)

plt.show()

from the code above, we trying to create a scatterplot with the year on x-axis and price on the y-axis. From the figure for the output we see that when there is an increase in the year, the price of the car is going up as well. This shows that there is a linear relationship between year and price

* Going back to data pre-processing

We need to divide the data into features and labels

features = cars\_data.iloc[:,0:1].values

labels = car\_data.iloc[:,1].values

from the scripts above the 0th column is going to contain the feature set, and the first column is going to contain the label.

The iloc indexer for Pandas Dataframe is used for integer-location based indexing / selection by position. The iloc indexer syntax is data.iloc[<row selection>, <column selection>]

We will then be able to divide up the data so that there are 20 percent to the test set and 80 percent to the training. We use the script below to get this done

from sklearn.model\_selection import train\_test\_split

train\_features, test\_features, train\_labels

test\_labels = train\_test\_split (features, labels, test\_size = 0.2,

random\_state = 0)

we then take a look at the data again and see that there is not a huge difference between the values of the years and those of prices. Both of these will end up being in the thousands each, what this means is that you don’t need to do any scaling and can just use the data as you have it here

* How to train the algorithm and get it to make some predictions.

This is where the linear regression class is going to be helpful because it has all the labels and other training features that you need to input and train your models. To do this we work with the script below to help get started,

from sklearn.linear\_model import LinearRegresison

lin\_reg = LinearRegression()

lin\_reg.fit (train\_features, train\_labels)

using the example of the car prices and year from earlier, we are going to see what the coefficient is for only the independent variable. We use the script below to get it,

print(lin\_reg.coef\_)

the output is 204.815, which shows that for each unit change in year the car price is going to increase by 204.815. after training the model the final step to use is to predict the new instance you are going to be working with. The predict method is used with the linear regression class to see this happen. The method takes the test features that you choose and adds them in as input and then predict the output that would correspond with it best. The script below will make this happen

predictions = lin\_reg.predict(test\_features)

the scrip above is going to give us a good prediction of what we are going to see in the future, we can now guess how much a car is going to be worth based on the year it is produced in the future. Now remember that this is not going to be 100 percent accurate, external factors like inflation might come into play. It might be lower or higher but at least it gives us a good way to predict the price of the vehicle in the future.

**Chap 4 List in Python**

We create a list in python by placing items called elements inside square brackets separated by commas and items in a list can be of mixed data types.

e.g of lists

list\_mine = [] #empty list , list1 = [1,2,3,4] #list of integers , list2 = [5, ‘’Happy”, 3.5] # list with mixed data types

Nested list

A nested list is a list as an item in another list e.g nested\_list = [“carrot”, [3,4,5], [‘g’]]

print(nested\_list[0][1])

output: e

from the script above the output is e as e is the value of the index 1 in best.

Python Negative Indexing.

Python allows negative indexing, the last item on the list is index -1, and index -2 is the last and 2nd to last item respectively in a list.

Slicing list in Python

Slicing operator (full colon) is used to access a range of elements in a list

Manipulating Elements in a list using the assignment operator

Items in a list can be changed, meaning lists are mutable

e.g list\_yours = [4,8,5,2,1]

list\_yours [1] = 6

print (list\_yours) # output will be [4,6,5,2,1]

we can also change a range of items in a list.

List\_yours [0:3] = [12,11,10] # will change first item to fourth item in the list

Print(list\_yours) #output will be: [12,11,10,1]

Appending/extending items in a list

The append() method allows adding items to a list, the extend() method can also be used.

e.g list\_yours.append(3)

Removing or Deleting items from a list.

The keyword del Is used to delete elements of the entire python list in python

e.g del list\_mine [1]

multiple items can be deleted with indexing

e.g del list\_mine[0:3]

an entire list can be deleted with: delete list\_mine

the remove() or pop() method can be used to remove the specified item. The pop() method will remove and return the last item if the index is not given. The clear method is used to empty a list.

An empty list can be used to delete and entire or specific elements of a list

List\_mine [1:2] = []

**Chap 5 Modules in Python**

Modules are libraries of functions, object class. Modules allow reusing the same code over and over again. Modules are saved in files, which makes them persistent, reusable and shareable. Importing a module gives you access to all the tools and functions defined in it. E.g we consider two python files, file1.py and file2.py in where file1.py is the main program and file2.py is the module. In file2.py we have a code that defines the following function

def Xfactorial (X)

p =1

for I in range (1, X+1):

p \*=i

return p

to use this function in the main program, we define code statements in file1.py as follows

import file2

A = file2.Xfactorial (3)

How to write and use a module in Python?

Modules are simply files created with the .py extension, when you create a module, all the data objects including functions that are defined becomes the module attributes. And these attributes are accessed and used via the attribute syntax as follows: module.attribute. e.g we define a module say MyModule.py with the following functions:

def Myfct(A)

Print (‘A by 2 is:’, A\*2)

Return A \* 2

The Myfct becomes the attribute of the module. A module can be called any python code you develop and save in a file with a .py extension if you are importing them later to use. After importing a module, the module name is used to access all its attributes. Another way to import a module is by using the ‘from import’ statement. This is used to import a specific attribute from a module directly and we don’t need to use the module name to access that attribute any longer. Using a wild character \* allows copying all names that are assigned to objects in the module.

Python also supports importing a module with an alias, we can now access the attributes using the alias instead of its name e.g

Import myModule as md

Variables initialized in a module are also accessed the same way with using the attribute syntax. The “from import statement” supports importing multiple attributes in one single line. Also we can modify a list attribute in a module locally but it is not so with a variable

>>> from ExModule1 import A, MyList

>>> A = 20

>>> myList [0] = 100

Output:

>>> import ExModule1

>>> print ('A is: ', ExModule1.A)

A is: 9

>>> print ('My list is: ', ExModule.myList)

My list is: [100, 40, 80]

To make a valid modification of a variable in a script and the module it is imported from, we use the import statement like follows:

>>> Import ExModule1

>>> ExModule1.A = 200

Importing a module for the first time in a script implies going through three steps, that are searching for the module, compiling the module and running the module. There are three keywords when it comes to importing modules in python, this includes ‘from’, ‘import’, ‘reload’. To ensure python has gone through all the steps of importing a module for the second time in a script, we use the reload statement, it helps make modifications in a program while it is running without interrupting it. It is function that takes as argument a module that is already loaded in memory. The general form of using reload with import is as follows:

import module\_name

list of statements that use module attributes

reload(module\_name)

list of statements that use module attributes

the reload function overwrites the module source code and re-runs it instead of deleting the file and creating a new one. Note that all imports after the first import use the already loaded module in memory, also the reload function re-runs the module and returns the module object. Below we illustrate how the reload function works:

* We create a module ExModule3.py which contains the following block of code:

My\_message = ‘This is my module first version’

def display ():

print (My\_message)

we then import this in our interactive shell and call the display() attribute, which displays the text on the screen.

>>> import ExModule3

>>> Exmodule3.display()

This is my module first version

While the interactive shell is still running, we go forward to modify the module’s source code:

my\_message = 'This is my module second version edited in the text editor'

def display ():

print (my\_message)

going back to the interactive shell, we try to import and call the function:

>>> import ExModule3

>>> Exmodule3.display()

This is my module first version

From the above, notice that the message did not change. We then use the reload function to re-run the import statement. And calling the function displays the updated message.

>>> reload (ExModule3)

<module 'ExModule3)>

>>> Exmodule3.display()

**Chap 6 Machine learning Training Model.**

In Ml, a model is a mathematical or digital representation of a real-world process. To build a good ML model, developers need to provide the right training data to an algorithm.

An algorithm is a hypothetical set taken before training begins with real-world data.

A linear regression algorithm for example is a set of functions defining similar characteristics or features as defined by linear regression.

The basic purpose of creating any ML model is to expose it to a lot of input as well as the output applicable to it, allowing it to analyse these data and use them to determine the relationship between it and the results.

Simple Machine Learning Training Model in Python.

In Ml having the right data is more important than having the ability to write a fancy algorithm. A good modeling process will protect against over-fitting and maximise performance.

In Ml data are a limited resource, developers should spend their time doing the following:

1. Feeding their algorithm or training their model
2. Testing their model.

The same data used for training should not be used for testing a model as this would lead to over-fitting. The primary aim of training sets is to fit and fine-tune one’s model. Tests sets are new data sets for the evaluation of one’s model. Over-fitting is one of the most important considerations in ML. it describes how well the target functions approximation correlates with the training data provided. An ML model is said to be over-fitted if it fits the training data exceptionally well while generalizing a new data poorly.

We professional talk about tuning models In ML we usually mean working on hyper-parameters. In ML there are two main types of parameters: model parameters and hyper-parameters.

The first defines individual models and is a learned attribute, such as decision tree locations and regression coefficients. The second defines higher-level settings for ML algorithms, such as number of trees in a random forest algorithm or the strength of the penalty used in regression algorithms. Penalty is a way preventing overfitting by limiting the models freedom.

The ML model refers to the model artifact created by the ML training process. The answers or output from a training data is known as target attributes.

Simple Machine learning Python model using linear regression

When building ML model in Python. We need to download and install sci-kit-learn an open-source python library with a wide variety of visualization, cross-validation, pre-processing and machine learning algorithms using a unified user-interface.

The user needs to have SciPy and NumPy installed before installing the sci-kit-learn. To load an external dataset, we can use the panda library, which will allow them to easily load and manipulate datasets.

Having created the training set, we split each row into an input training set and its related output training.

Benefits of splitting datasets include:

1. Gaining the ability to train and test the model on different types of data than the data used for training
2. Testing the model’s accuracy, which is better than testing the accuracy of the out-of-sample training.
3. Ability to evaluate predictions using response values for the test datasets

The linear regression method from python’s sci-kit-learn library will then be used to create and train their model.

Here, the ML model analyses the training data and uses it to calculate the coefficients or weights to assign to the inputs to return the right outputs.

**Chap 7 Conditional or Decision Statements**

In this chapter we will learn various ways to write conditional statements, we will learn basic if – statements, create complex if – statements and write loop to handle items in a list.

e.g., on how to use an if statement and a for loop.

Cars = [‘Toyota’,’Kia’,’Audi’,’Infinity’]

For car1 in cars:

If car1 == ‘kia’

Print (car1.upper())

else:

print(car1.title())

the loop verifies if whether the current value of the car is ‘Kia’. If this I true, then it prints the value in uppercase. If it not true, it prints it in title case.

Output:

Toyota

KIA

Audi

Infinity

Creating Multiple Conditions

When writing code, you may want to verify multiple conditions simultaneously. Atimes you may want only one condition to be satisfied. In such situation, we can use the keyword ‘or’ and ‘and’

Using ‘AND’

This is used to verify that two expressions are both true at the same time. E.g.

>>> score\_1 = 46

>>> score\_2 = 30

>>> score\_1 >=45 and score\_2 >= 45

False

Using ‘OR’

This keyword allows you to check multiple conditions as the AND keyword. The difference is that the or keyword is used when you want to ascertain that one expression is true for multiple conditions.

>>> score\_1 = 46

>>> score\_2 = 30

>>> score\_1 >=45 or score\_2 >= 45

True

We can use the ‘in’ keyword to validate if an item is present in a list. The not keyword can also be used to validate if an item is not part of a list.

IF statements

Here we learn the various types of IF statements and the best conditions to apply them.

1. Simple IF statements:

This follow the basic syntax of

IF condition:

Perform action

If the condition in the statement evaluates to True then the indented code below it is

Executed, but otherwise the indented code is ignored. It is important to indent the

The clause of the IF statement.

e.g person = 21

if person >= 20

print (‘you are qualified for the football match’)

1. If-else statements

We use this when we want to perform certain actions if a particular action isn’t met.

The syntax is as follows:

If conditional test:

Perform action 1

else:

perform action 2

e.g

person =18

if person >= 20:

print(“You are qualified for the football match against Valencia.”)

print(“The match is between Arsenal and Valencia.”)

Print(“The Venue is at the Emirate Stadium in England.”)

else:

print(“Unfortunately, you are not qualified to participate in the match.”)

print(“Sorry, you have to wait until you are qualified.”)

1. The if-elif-else chain

This allows us to test three different conditions based on certain criteria. For instance a cinema hall with different charge rates for different sets of people

Children under 5 years are free

Children between 5 years and 17 years are $30

People older than 18 years is $50

e.g person\_age = 13

if person\_age < 5:

print(“Your ticket cost is $0.”)

elif person\_age < 17:

print(“Your ticket cost is $30.”)

else:

print(“Your ticket cost is $50)

basically the elif statement allows you to test multiple conditions, the elif block is tested if the iF statement returns false. Also if the elif statement returns falls the next line of code is evaluated.

We can have multiple elif blocks also

person\_age = 13

if person\_age < 5:

cost = 0

elif person\_age < 17:

cost =30

elif person\_age < 55

cost = 50

else:

cost = 35

print( “Your ticket cost is %s” %(cost))

in the above code, the cost is introduced in the if-elif-else statement, we then use the placeholder function so we don’t have to repeat the print statement all the time.

Performing multiple Conditions

In situations where we want to check all the conditions available. Here we use multiple if statements without adding the elif or else blocks. This method becomes relevant when more than one of the condition returns true.

e.g

united\_players = [“Rashford,” “Young,” “Pogba,” “Mata,” “De Gea”]

if “Young” in united\_players:

print(“Adding Young to the team list.”)

if “De Gea” in united\_players:

print(“Adding Dea Gea to the team list.”)

if “Messi” in united\_players:

print(“Adding Messi to the team list.”)

print( “\ Team list completed for the match against Manchester City!”)

from the above code, the subsequent if statements are executed regardless of the out come of the preceding ones

**chap 8 Essential libraries for machine learning in python**

nowadays developers prefer the usage of python, the features of python that have helped scientists to use it in machine learning include:

* Consistent syntax
* Being flexible
* Shorter time in development.
* Can develop sophisticated models and has engines that could help in predictions

Python boasts of having a series of very extensive libraries, libraries refer to a series of routines and sort of function with different languages.

Examples of essential libraries include:

* Scikit -learn : one of the best and trendy library In ML, has the ability to supporting learning algorithms, especially unsupervised and supervised ones

e.g of scikit -learn include:

* K-means
* Decision trees
* Linear and logistic regression
* Clustering

Scikit learn helps in task related to data mining, i.e. Classification, clustering, and even regression analysis.

* TensorFlow: this is a form of algorithm which involves deep learning. They are applied in the identification of objects, and helps in different apps that deal with the recognition of voice. Search engines such as Google depend on this type of library.
* Theano: its vital task are to help with anything related to numerical computation, it is a faster python library that can perform calculation of intensive data up to 100 times. It plays other roles such as:
  + Definition of mathematical expressions
  + Assist in the optimization of mathematical calculation
  + Promotes the evaluation of expression related to numerical analysis
* Pandas: helps in the provision of data structures that are high level and quality, contains in built methods that make it capable of performing tasks such as grouping and timing analysis. Pandas can collect data from sources such excel, csv and sql databases. Its main function is to help in data manipulation. It consists of two structures that enable it perform its functions correctly. That’s the series which has only one dimensional and data frame that boast of two dimensional. It has the power to export a wide range of data. Pandas is effective in the following areas:
  + Splitting of data
  + Merging of two or more types of data
  + Data aggregation
  + Selecting or subletting data
  + Data reshaping

Applications of pandas in a real-life situation will enable you to perform the

following:

* + You can quickly delete some columns or even add some texts found within the Data frame
  + It will help you in data conversion
  + Pandas can reassure you of getting the misplaced or missing data
  + It has a powerful ability, especially in the grouping of other programs according to their functionality.
* Matplotlib: this is a sophisticated data analysis technique that helps in data visualization. It is used when you want to share findings in the form of chars/graphs with different stakeholders. It is a library that has good looks when it comes to to graphics and images. The library can also export graphics and can change these graphics into PDF, GIF and so on. The following task can be undertaken with much ease:
  + Formation of line plots
  + Formation of scatter plots
  + Creation of beautiful bar charts and building up of histograms