DAY-1:-23-02-2024

- **TASK-1:** Review of core python concepts data types,operators ,control flows functions, modules, packages
  - ❖ IDE-integrated development environment (google colab)
  - ❖ Github repository creation (folder- Data analysis)

NUMPY~

- ❖ Numpy is used for numerical or scientific computation in arrays, vectors matrices
- ❖ To import the numpy package we just have to use the import statement which is as follows: import numpy as np(our wish).

FUNCTIONS OF NUMPY [1st DAY]

- **ARRAY CREATION:**

➢ **np.array()-** creates an array from a list

Eg- var=np.array([elements])

➢ **np.zeros()-** creates an array filled with zeros of the specified size

➢ **np.ones()** - creates an array filled with ones of the specified shape.

➢ **np.arange()**-Generates numbers up to the given value and we can also define the starting value,ending value and step value as the parameter to this function.

Eg- var=np.arange(0,100,5)

❖ For all these functions the default data type is “float” and we can change This is using **“dtype”.** Used to define the datatype for the variables.

- **ARRAY MANIPULATION:**

➢ **reshape()**-Reshapes the array into desired shaped array by defining no:of rows and columns

➢ **slicing**-Slicing is done with “:” operator.It prints the elements from the given range.

➢ **Transpose**-Transpose is done as follows:”**array_variable.T**” As the name itself says it returns the transpose of the given array or matrix.

➢ **np.split(a,n)**-Split function splits the given array(a) into n arrays

➢ **np.dot(a,b)**-Gives the dot product of matrix a and b (multiplication of matrix) with the help of a function.

➢ **np.linalg.eig(a)**-Gives the eigenvalues and vectors of given matrix.

These are required to perform the further operations and manipulations For data processing.

❖ **np.loadtxt(path,dtype)**-Loads any file from the path that we've given and we can give the datatype as we want like"dtype=int" in order to print the data of the file.

Eg-
```
data= np.loadtxt("/content/drive/MyDrive/dataset/1st
           day.txt",dtype=int)
```

❖ **np.savetxt(path,file)**-Loads the file that we've given into the path

Eg -
```
d=np.savetxt("/content/date.txt",data)
```

❖ **np.random.rand()**-Randomly produces numbers between 0 and 1 on Executing every time.

❖ **np.random.randint(a,b)**-Randomly produces between a and b

❖ **type(a)**-Gives the type of the matrix a

❖ **ndim** and **shape** are used to find the dimension of array and shape the array respectively

➢ Invoked as **"Array_variable.ndim"** and **"Array_var.shape"**

❖ Element multiplication is done using **"*"**operator and Matrix multiplication i done using **"@".**

➢ **EX:**a*b and a@b

❖ **sum()**-Returns the sum of all the elements in the matrix

➢ **EX:**a.sum() where a is the matrix

❖ **max()-**Returns the maximum element in the matrix

❖ **cumsum()**-Returns the cumulative sum of the elements

❖ The above three functions have special feature called "axis" if axis=1 respective operation is done along the column and if axis=0 then the operations are done along the rows

- **np.vstack**(a,b) and **np.hstack**(a.b)-Used to stack the given two matrices a and b

vertically and horizontally respectively.

- **np.dstack**(a)-It used for some changes in matrix a.The changes are:

➢ Number of rows become number of groups

➢ Number of column become number of rows

➢ Number of groups become number of columns

DAY-2

24-02-2024

a=[1,2,3,4,5]
mean=3
median=3
median=((1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2)/5
=10/5
=2
Variance=2
StandardDeviation=sqrt(variance)

Statistical Operations:

1) Mean.

→"result=np.mean(array_name)"

2) Median.

→"result=np.median(array_name)"

3) Variance.

→"result=np.var(array_name)"

4) Standard deviation.

→"result=np.std(array_name)"

PANDAS:-

- Pandas is a data manipulation package in python for tabular data.That is,data in the form of rows and columns,also known as DataFrames.

- TASK-2:Twitter developer portal

>>To import pandas : “import pandas as pd”

Operations:

1) Series : A Series is a one-dimensional array of data. It can hold data of any type: string, integer, float, dictionaries, lists, booleans, and more.

→”var_name=pd.Series(array_1,array_2)”

2) Read the data in the csv file into the program.

→”var_name=pd.read_csv(“path_of_csv_dataset”)”

3) Read the data in the text file into the program.

→”var_name=pd.read_csv(“path_of_txt_dataset”)”

4) Read the data in the excel file into the program.xc

→”var_name=pd.read_excel(“path_of_dataset”,sheet_name=0)”

5) Display the data at the specified location in a dataset.

→”var_name.loc[number_of_location]”

6) Calculate mean from the data in a dataset.

→”result=var_name[‘label_of_data’].mean()
var_name=var_name.fillna(result)”

7) Remove duplicate values from the dataset.

→”var_name=var_name.duplicates()”

8) Display specified number of rows from first in a dataset.

→”var_name.head(number)”

9) Display the specified number of rows from last in a dataset.

→”var_name.head(number)”

10) Display the number of rows and columns in a dataset as “(rows,columns)”

→”var_name.shape”

11) Create a file from the data.

→”var_name.to_csv(“new_file_name.format_type”)”

12) Count the repeated data in a dataset.

→”var_name.groupby([‘label1_name’])[‘label2_name’].count()”

DAY-3

26-02-2024

- Matplotlib and seaborn for data visualisation.
- KAGGLE
- Machine Learning for Kids
- Teaching Machine

MATPLOTLIB :

Matplotlib is a comprehensive library for creating static, animated, and interactive visualisations in Python.

->”**matplotlib.pyplot**” is a collection of command style functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.

->To import matplotlib : “import matplotlib.pyplot as plt”

Dataset : The data in the form of a file is called dataset.

Operations with data set:

1) loading the dataset into the program.

→”data_name=np.loadtxt("path of .txt file",dtype=data_type)”

2) Saving the dataset from the program.

→”data_name=np.savetxt("/content/new_file_name.txt",data_file_name)”

3) Plotting using the dataset.

→”plt.plot(array_name)”

4) Scatter plot the data.

→”plt.scatter(x,y,color=’color_name’)

→To give title : “plt.title(‘title_name’)”

→To display graph: “plt.show()”

→To display label : “plt.legend(loc=’best’)”

5) Plot the data.

→”plt.plot(x,y,color=’color_name’)

→To give title : “plt.title(‘title_name’)”
→To display graph : “plt.show()”
→To display label : “plt.legend(loc=’best’)”

6) Display data in the form of a pie chart.
→”plt.pie(array_1,array_2,colors=array_of_colors,startangle=angle)”

Seaborn :

Seaborn is a library for making statistical graphics in Python.
It builds on top of matplotlib and integrates closely with pandas data structures.
>>It contains the builtin datasets.
>>To import seaborn : “import seaborn as sns”

Operations:

1) Load the dataset into the program.
→”var_name=sns.load_dataset(“dataset_name”)”

2) Scatterplot the data in a dataset.
→”sns.scatterplot(x=”label1_name”.y=”label2_name”,data=var_name)”

3) Violinplot the data in a dataset.
→”sns.violinplot(x=”label1_name”.y=”label2_name”,data=var_name)”

4) Heatmap from the data in a dataset.

->**Heatmap** : a technique of data visualisation that makes use of colour in order to exhibit how a value of interest varies on the basis of the values of the two other variables. To sum it up, using different colours to represent data gives you a general view of the numerical data.

->**Correlation matrix** : The correlation matrix is a matrix that shows the correlation between variables.
→correlation_matrix=”var_name.corr()”
→sns.heatmap(correlation_matrix,annot=True,cmap=”coolwarm”)

Machine Learning :

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Steps in Machine Learning :

1)Uploading The File
2) Training the data
3) Testing the data.

Types of Machine Learning :

1) Supervised Machine Learning(Labelled data-Features)
2) Unsupervised Machine Learning(UnLabeled data-Without Features)
3) Semi Supervised Machine Learning(Combination of both Labelled and Unlabeled data)
4)Re-Enforcement(Extraction of Data)

Neural Network :

->A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain.

->It is a type of machine learning process, called **deep learning**, that uses interconnected nodes or neurons in a layered structure that resembles the human brain.

Layers of Neural Network:

1. Inner layer
2. Hidden layer
3. Outer layer

DAY-4

27-02-2024:

CNN(Convolutional Neural Network) : a Convolutional Neural Network or CNN or CovNet is a type of artificial neural network, which is widely used for image/object recognition and classification.

Deep Learning thus recognizes objects in an image by using a CNN.

>>CNNs have an input layer, an output layer, numerous hidden layers, and millions of parameters, allowing them to learn complicated objects and patterns.

>>Filters are weights and biases that are randomly generated vectors in the network. Instead of having individual weights and biases for each neuron , CNN uses the same weights and biases for all neurons.

Layers in CNN :

>Input Layer
>Hidden Layer
>Convolutional Layer
>Activation Layer
>Pooling Layer
>Output Layer

1) Input Layer : It contains the input images or sequences of images.

2) Convolutional Layer : This layer often contains input vectors, such as an image, filters, such as a feature detector, and output vectors, such as a feature map. The image is abstracted to a feature map, also known as an activation map, after passing through a convolutional layer.

3) Activation Layer : It is also known as “padding and stride”

>>the activation function performs element wise operation on input.

>>It is a term used in convolutional neural networks to describe how many pixels are added to an image when it is processed by the CNN kernel.

If the padding in a CNN is set to zero, for example, every pixel value-added will have the value zero. If the zero padding is set to one, a one-pixel border with a pixel value of zero will be added to the image.

4) Pooling Layer : Its purpose is to gradually shrink the representation’s spatial size to reduce the number of parameters and computations in the network. The pooling layer treats each feature map separately.

>>**Max-pooling :** It chooses the most significant element from the feature map. The feature map’s significant features are stored in the resulting max-pooled layer. It is the most popular method since it produces the best outcomes.

>>**Average-pooling :** It entails calculating the average for each

region of the feature map.

5) Output layer : The probability score into the output class.

Activation Function :

→Used for to decide which neuron has to be activated or not based on calculation of weighted sum and bias

→When the errors occurred in the output layer it modifies input layer by using weights and biases this method is called back propagation

→**Activation function** is present in all the algorithms

→There are 4 types of activation functions. They are

1. Sigmoid
2. Tanh
3. Softmax
4. Relu

Sigmoid:

→The shape of sigmoid is ‘S’

→It is non-linear

→Usually the sigmoid function is used in the output layer.

→The range of sigmoid is 0-1.

→Mainly used for binary classifications.

Tanh:

→Range -1 to 1.

→t is non linear.

→It is mainly used in hidden layers.

DAY-5

28-02-2024

Linear Regression :

>>It is a type of machine learning algorithm.

>>It is a Supervised Learning Algorithm(labelled data).

>>Learns from the labelled data sets and maps the data points to the

most optimised linear functions.
>>These points can be used for the prediction of new datasets.

Types of variables:

1) Independent Variable.
2) Dependent Variable.

- ❖ Find the mean of both dependent and independent variables(x’ and y’)
- ❖ Find differences between each x point and x-mean(x-x’).
- ❖ Find differences between each y point and x-mean(y-y’).
- ❖ Find sum of squares of (x-x’)
- ❖ Find the product of (x-x’) and (y-y’)

Equation of regression line
y=a+bx
a=y’-bx’
b=sum of products of (x-x’) and (y-y’) / sum of squares of (x-x’)

Logistic regression :

->For binary classification.
->It is a Supervised Learning Algorithm.

DAY-6

29/2/2024

Decision trees:

-->it provides an effective method for making decisions because they lay out problem and all the possible outcomes
-->have nodes and leaves
–>node:condition have true and false branches
-->leaf:result-showing the dataset that is true and false branch condition
–>it recurrently(continuously)split the data until it get the pure nodes

Random forest:

–>collection of many decision trees
–>keywords:bootstrapping,aggregation
–>to get easily output
–>not a recurrently
–>accuracy more

1.Bootstrapping: Random forest create new datasets from original data set

This is called **bootstrapping**
Conditions:

- select rows that are different and cover all the rows
- select features that are different and cover all features
- selecting same data(row)and same features have no use

Day-7

1/3/2024

Big data concept:

Big data is vast and complex datasets that cannot be effectively managed or processed using traditional data processing tools
>volume,velocity,variety,veracity

1.volume:

It is immense amount of data generated and collected from various source such as social media,sensors,transactions and more
>with the proliferation of digital technologies ,data is being produced at an unprecedented rate,leading to massive volumes of information that traditional databases struggle to handle

2.velocity:

It denotes the speed at which data is generated collected and processed
>data is produced rapidly in real time or near real time from sources like social media updates,sensor data from iot devices,online transactions,and more
>efficient processing of this streaming data is crucial for real time analytics and decision making

3.variety:

It is diversity of data types and sources
>data comes in various formats including structured data (like databases),semi-structured data(like XML-extensible markup language,JSON-javascript object notation),and unstructure data(images,videos)
>big data platforms must be able to handle this diverse range of data types efficiently

4.veracity:

Veracity emphasises the quality and reliability of the data
>due to sheer volume and variety of data sources,ensuring data quality and reliability can be challenging
>big data analytics often involves dealing with incomplete,inconsistent,or inaccurate data,requiring robust mechanisms for data cleaning,validation and quality assurance
>those four V’s collectively define the challenges and opportunities associated with big data
>effectively managing,analysing big data

Statistical analysis with scipy and stat model:

- **Hypothesis**
- **Regression analysis**
- **ANOVA testing(analysis of variance)**

- ➢ scipy is an **open-source scientific computing** library for python that build on numpy.it provides many additional functionalities compared to numpy,including optimization,integration,interpolation,eigenvalues problems,signal and image processing,statistical distributions,and much more
- ➢ **stat model:** it is a python library that provides classes and functions for estimating and testing statistical models.it is built on top of numpy,scipy and matplotlib,and it is integrated with pandas for data handling. Stat models includes a wide range of statistical models and tests,marking it a powerful tool for statistical analysis and hypothesis testing
- ➢ machine learning framework
- ➢ open source for building data and deploying machine learning models

1.Hypothesis testing:

- **T-statistic:**The T-Static is a measure of how many standard deviations a sample mean is away from the hypothesized population mean,relative to their variable T in their sample.It is calculated as the difference between the sample mean and population mean / standard error of the sample mean.
  >>The larger absolute value of T -static of the stronger the evidence makes their null hypothesis.

P-value:The P-value is the probability of observing the test statistics (of one or more extreme).If the Null is true.If quantifies the strength of the evidence against the null hypothesis.A p-value(typically less than 0.05) indicates that the observed data is unligthy under the assumption of null hypothesis is true

- A p-value(typically less than 0.05) indicates that the observed data is unligthy under the assumption of null hypothesis is true leading to their rejection of their null hypothesis in favor of their alternative hypothesis
- Conversely, a large p-value suggests that the observed data is under the null hypothesis leading to the failure to reject their null hypothesis.

➢ Hypothesis testing : IF p-value is less than the predetermined significant errors such as 0.05.It is typically interpreted as sufficient evidence to reject their null hypothesis.

➢ If the p-value is greater than the significance level their not enough evidence to reject their null hypothesis

➢ Together t-static and p-value determining whether the observed sample data provide enough evidence to support a client hypothesis about their population