Page:
Date: / /

Assignment A1

Aim: Data Wrangling I.

Perform the following operations using Python on any open source dataset (e.g., data.csv)

1] Import all the required Python Libraries.

2] Locate an open source data from the web (e.g. https://www.kaggle.com). Provide a clear description of the data and its source (i.e. URL of the website)

3] Load the dataset into pandas data frame

4] Data PreProcessing: check for missing values in the data using pandas insull(), describe() function to get some initial statistics. Provide variable descriptions. Types of variables etc. check the dimensions of the data frame.

5] Data formating and Data Normalization: Summarize the types of variables by checking the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the data set. If variables are not in the correct data type, apply proper type conversions.

6] Turn categorical variables into quantitative variables in Python.

In addition to the codes and outputs, explain every operation that you do in the above steps and explain everything that you do to import/read/scrape the data set.

Theory: Data Wrangling:

Data wrangling is the process of

gathering, collecting, and transforming Raw data into another format for better understanding, decision-making, accessing, and analysis in less time. Data wrangling is also known as Data munging.

Data wrangling is a very important step. The below example will explain its importance as:

Books selling website wants to show top-selling books of different domains, according to user preference. For example, a new user search for motivational books, then they want to show those motivational books which sell the most or having a high rating, etc.

But on their website, there are plenty of raw data from different users. Here the concept of Data Munging or Data wrangling is used. As we know Data is not wrangled by system. This process is done by Data scientists. So, the data scientist will wrangle data in such a way that they will sort that motivational books that are sold more, or have high ratings or user buys this book with these package of books, etc. on the basis of that, the new user will make choice. This will explain the importance of Data wrangling.

Data wrangling in Python

Data wrangling is a crucial topic for data science and data analysis. Pandas framework of Python is used for Data wrangling. Pandas is an open-source library specifically developed for data analysis and data science. The process like data sorting or filtration, Data grouping, etc.

Data wrangling in Python deals with the below functionalities:

1. Data exploration: In this process, the data is studied, analyzed and understood by visualizing representations of data.

2. Dealing with missing values: most of the datasets having a vast amount of data contain missing values of NaN, they are needed to be taken care of by replacing them with mean, mode, the most frequent value of the column or simply by dropping the row having a NaN value.

3. Reshaping data: In this process, data is manipulated according to the requirements, where new data can be added or pre-existing data can be modified.

4. Filtering data: Some times datasets are comprised of unwanted rows or columns which are required

to be removed or filtered.

5. other: After dealing with the raw dataset with the above functionalities we get an efficient dataset as per our requirements and then it can be used for a required purpose like data analyzing, machine learning, data visualization, model training, etc.

Conclusion:

Data wrangling in ml is huge necessity in the recent times because of the huge amount of data that gets processed every day making user services more efficient. Data wrangling provides its importance in the world of data science.

Assignment No. 2

Page :
Date : / /

Name: Sahil Nivrutti Gadge
Roll No: 19C0021
Class: TE (1st shift)

Aim: Data wrangling – II

Perform the following operations using Python on "Academic Performance" dataset of students.

1) Scan all variables for missing values and/or inconsistencies, use any of the suitable techniques to deal with them.
2) Scan all numeric variables for outliers. If there are outliers, use any of the suitable techniques to deal with them.
3) Apply data transformations on at least one of the variables. The purpose of this transformation should be one of the following reasons: To change the scale for better understanding of the variable, to convert a non-linear relation into a linear one, or to decrease the skewness and convert the distribution into a normal distribution.

Reason and document your approach properly.

Date :

Theory:

* Dealing with missing values:
Here, missing values or NAN present in college which are going to be taken care by replacing them with column mean value

Dataset:

Data =

Roll	name	gender	M_1	M_2
1	A	M	43.0	45.0
2	B	F	45.0	NaN
3	C	F	47.0	37.0
4	D	M	NaN	46.0
5	E	F	35.0	NaN

```
# Calculate mean for M1

Import pandas as pd
df = pd.DataFrame(data)
mean_val = df['M1'].mean()

# Replacing NaN by mean
df['M1'].fillna(value = mean_val, inplace = True)

# Display Data
df
```

O/P →

Roll	name	gender	M_1	M_2
1	A	M	43.0	45.0
2	B	F	45.0	NaN
3	C	F	47.0	37.0
4	D	M	34.0	46.0
5	E	F	35.0	NaN

* Skewness of Data:

The skewness is measure of symmetry or asymmetry of data distribution and kurtosis measure whether data is heavy-tailed or light-tailed in a normal distribution. Data can be positive-skewed or negative-skewed.

* Positive skewed → Data pulled towards right side
* negative skewed → Data pushed towards left side
* dataframe.skew():

Pandas dataframe.skew() function return unbiased skew over requested axis Normalised by N-1. skewness in a measure of the asymmetry of the probability distribution of a real valued random variable about its mean

default skewness:

```
df.skew()
```

O/P → Roll = 0.000000
M_1 = -0.252435
M_2 = -0.095741
dtype = float64

Linear Regression:

Linear Regression is a statistical method for modeling relationship between a dependent variable with a given set of independent variables.

Simple linear regression is a approach for Predicting a response using a single feature

Let consider dataset where, we have a value of response y for every feature and x:

∴

x	0	1	2	3	4	5	6	7	8	9
y	1	3	2	5	7	8	8	9	10	12

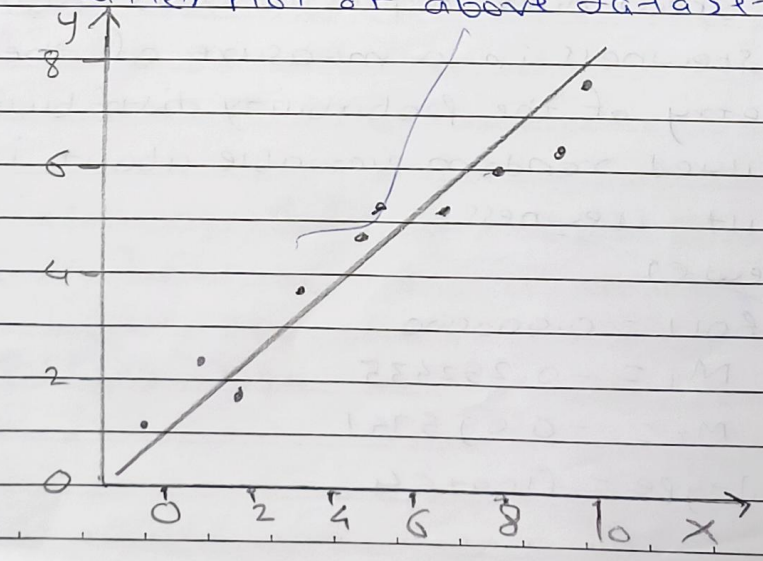
for generally, we define

x as feature vector i.e. $x = [x_1, x_2, \cdots, x_n]$

y as Response vector i.e. $y = [y_1, y_2, \cdots, y_n]$

∴ for n observations i.e. $n = 10$.

A scatter plot of above dataset looks like:



Conclusion:

Thus the data wrangling and data transformation Practical was Successfully Performed.

Page:
Date: / /

Data science & Big data analysis Laboratory
Assignment - 3

Title- Basic statistics - measures of central tendencies & variance

Aim- Perform the following operations on any open source dataset (eg. data.csv)

① Provide summary statistics (mean, median, minimum, maximum, standard deviation) for a dataset (age, income, etc) with numeric values group by one of the qualitative variable. For example, if your categorical variable is age groups & qualitative variable is income, then provide summary statistics of income grouped by the age groups. Create a list that contain a numeric value for each response to categorical variable.

② write a python program to display some basic statistical details like Percentile, mean, standard deviation, etc. of the species of 'Iris-setosa', 'Iris-versicolor' & 'Iris-verginica' of iris.csv dataset.
Provide the codes with outputs & explain everything that you do in this step.

Requirements- Data sets, Python editor

Theory- Descriptive Statistics - measures of central tendency & variability -

Descriptive statistics are broken down into measures of central tendency & measures of variability. measures of central tendency include mean, median & mode, while measure of variability include standard deviation, variance, minimum & maximum variables, Kurtosis & skewness.

mean (Arithmetic)

The mean (average) is most popular & well known measure of central tendency. It can use with both discrete & continous data, although its use is most often with continous data. The mean is equal to the sum of all values in data set divided by number of values in data set. so, if we have values in data set & they have values ..... the sample mean denoted by 'x bar' ($\bar{x}$)

Formula is $\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$

eg.

	Name	marks
0	Raj	90
1	Rinci	95
2	Ram	98

$\bar{x} = \frac{90+95+98}{3} = 94.33$

An important property of mean is that includes every value in your data set as part of calculation. In addition, the mean is only measure of central tendency where sum of deviations of each value from the mean is always zero.

Median-

The median is middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers & skewed data. In order to calculate median, suppose we have data below:

65, 55, 89, 56, 35, 14, 56, 55, 87, 45, 92.

we first need to rearrange that data into order of magnitude -

14, 35, 45, 55, 55, 56, 67, 87, 89, 92

Our media mark is middle mark - in this case, 56 (highlighted). It is middle mark because there are 5 scores before & after it. This works fine when you have an odd number of scores, but what happen when you have even number of scores? Well, simply to take middle two scores & average the result. for example -

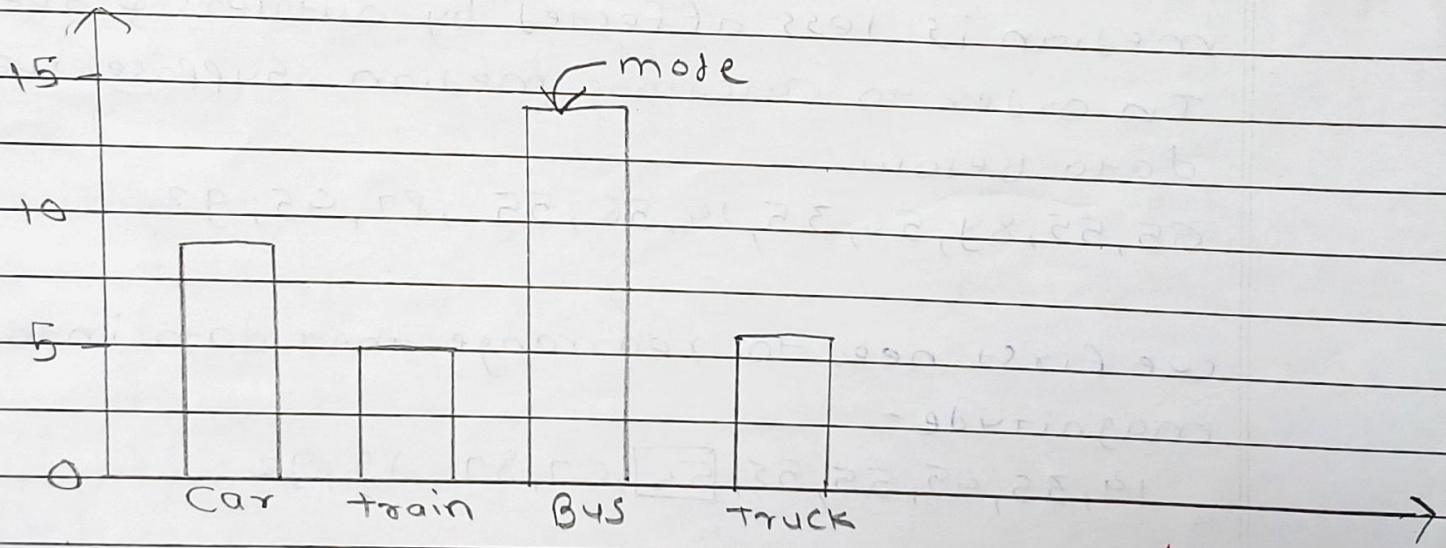
14, 35, 45, 55, 55, 56, 56, 67, 87, 89

Take a average of 55 & 56 i.e 55.5 is a median

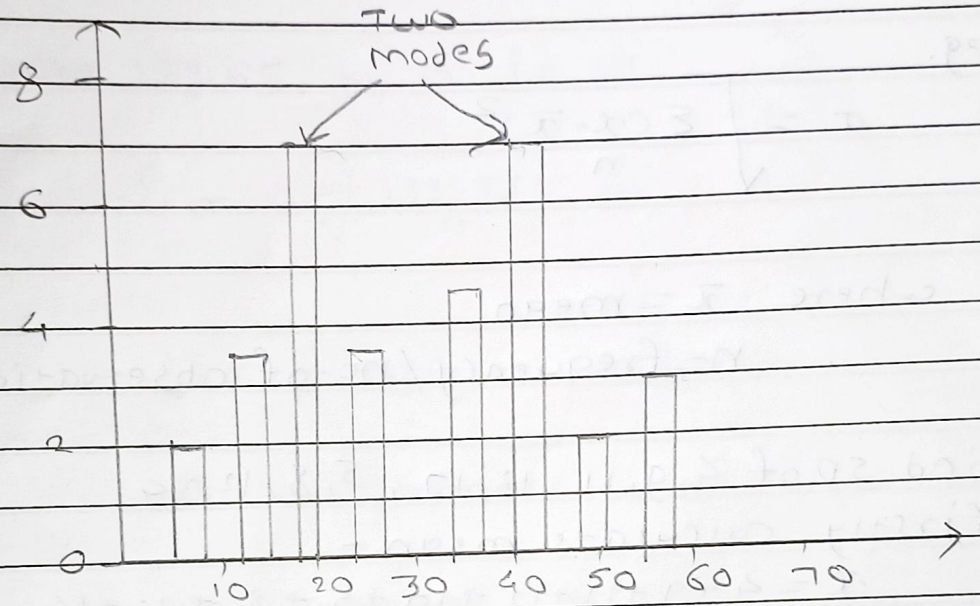
mode -

The mode is the most frequent score in our data set on a histogram it represents the highest bar chart or histogram. You can, therefore, sometimes consider the mode as being the most popular option.

Normally the mode is used for categorical data where we wish to know which is most common category, as illustrated below -



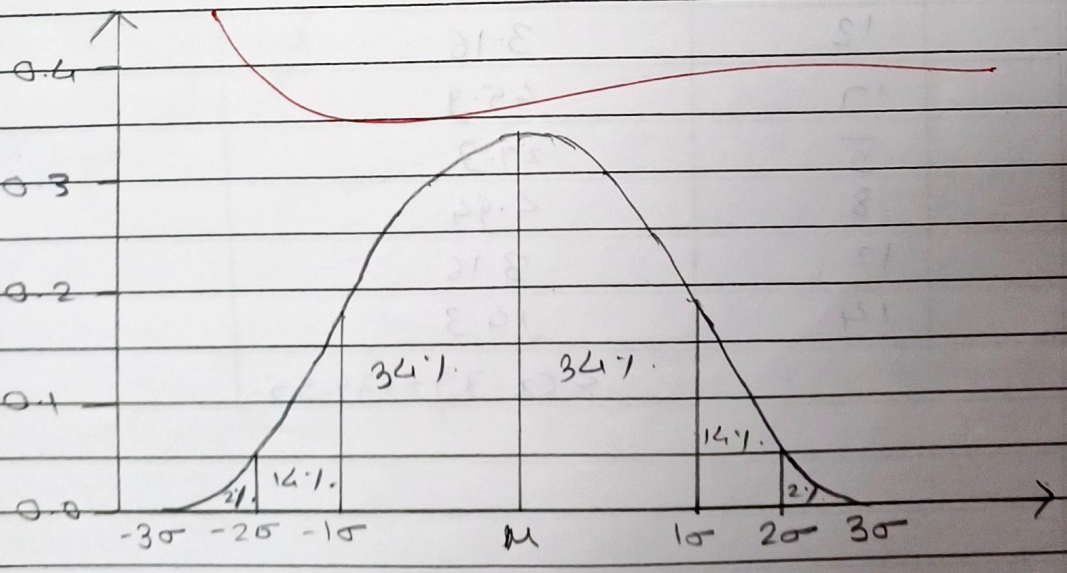
We can see above that most common form of transport, in this particular data set, is the bus. However, one of the problem with the mode is that it is not unique, so it leaves it with problems. when we have two or more values that share highest frequency, such as below



Standard Deviation -

A standard deviation (σ) is a measure of how dispersed data is in relation to the mean. Low standard deviation means data is clustered around the mean & high standard deviation indicates data is more spread out.

Graph denotation -



eg.

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$$

where, $\bar{x}$ = mean

n = frequency / no. of observations

Find SD of 4, 9, 11, 12, 17, 5, 8, 12, 14

firstly calculate mean -

$$\bar{x} = \frac{4+9+11+12+17+5+8+12+14}{9}$$

$$\bar{x} = 10.22$$

now subtract $\bar{x}$ from x individually & square it.

x	$(x-\bar{x})^2$
4	38.7
9	1.49
11	0.6
12	3.16
17	45.9
5	27.3
8	4.94
12	3.16
14	14.3
	$\Sigma(a-\bar{x}) = 139.55$

Page :
Date : / /

Divide 139.55 by n i.e. 9

$$\therefore \sigma = \sqrt{\frac{139.55}{9}}$$

$$\therefore \sigma = 3.94$$

$$\sigma = \sqrt{\text{variance}}$$

Or variance $= \sigma^2$

$\therefore$ variance $= (3.94)^2 = 15.52$

Conclusion - Hence, I perform basic statistics (mean, mode, median, & standard deviation) on given dataset & also I plot histograms using this dataset. I can able to generate correct output

Page :
Date : / /

Assignment No - 04

Title: Data analytics - 1

Aim: Create a linear regression model using Python / R to predict home prices using Boston Housing Dataset. The Boston Housing Dataset contains information about various houses in Boston through different parameters. There are 506 samples and 14 feature variable in this dataset.

Theory: Regression -

Regression analysis is a one of the most important fields in statistics and machine learning. There are many regression methods available.

Regression searches for relationships among vehicles i.e. you can observe several employees of salaries depend on the features, such as experience, level of education, role, city, they work in and so on.

This is a regression problem where data related to each employee represent one observation. The prescription is that the experience, education, role and city are the independent features, while the salary depends on them.

Generally, in regression analysis, you usually consider some phenomenon of interest and have a number of observation. Each observation has two or more features. Following the assumption that are of the feature depends on the others, you

try to establish a relation among them.
In other words, you need to find a function that maps some features/variable to others sufficiently well.

Linear Regression:

Linear regression is probably one of the most important and widely used regression techniques. It is among the simplest regression methods. one of its main advantage is ease of interpreting Results.

Linear Regression is a statistical method for modelling relationships between a dependent variable with a given set of independent variable.

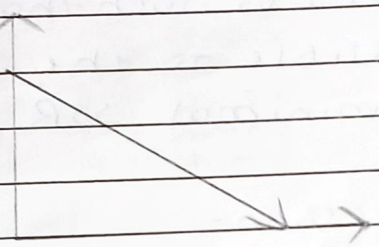
Types of Linear Regression:

1) Positive Linear Relation:-

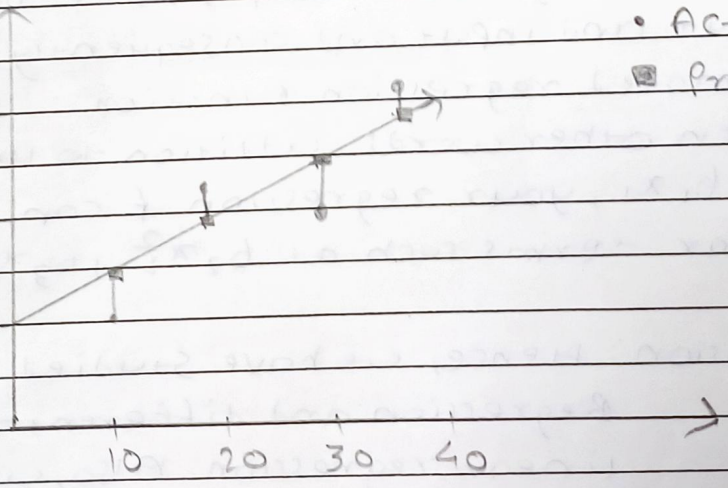
A linear regression will be called Positive if both independent and dependent variable increases. It can be understood with the help of following diagram.

ii) Negative Linear Regression:

A linear regression will be negative if independent variable decreases with increase in dependent variable It can be understood with the help of following diagram:



For example:



multiple Linear Regression:

It is the extension of simple linear regression that predicts a response using two or more features. multiple or multivalved linear regression is a case of linear regression with two or more independent variables

If there are just two independent variables, the estimated regression function is $f(x_1, x_2) = b_0 + b_1x_1 + b_2x_2$. It represents a regression plane in a three dimensional space. The goal of regression is to determine the values of the weights b_0, b_1, and b_2 such that this plane is as close as possible to the actual responses and yield the minimal SSR.

Polynomial Regression -

You can record polynomial regression as a generalized case of linear regression. You assume that the polynomial dependence between the output and input and consequently, the polynomial estimated regression function.

In other words, addition to linear terms like b_1x_1, your regression f can include non-linear terms such as $b_2x_1^2$, $b_3x_3^2$ and so on.

Conclusion: Hence, we have studied linear Regression and different types of linear regression. Also, we have successfully created linear regression model using Python/R to predict home prices using Boston Housing Dataset.

Page :
Date : / /

Assignment No - 5

Title: Data analytics - II

Aim: Implement logistic regression using Python/R to perform classification on Social_Network_Ads.csv dataset

Compute confusion matrix to find TP, FD, TN, FN, Accuracy, Error rate, Precision, Recall on the given dataset.

Theory: Logistic Regression

Logistic regression can be used for various classification problems such as spam detection, Diabetes prediction, if a given customer will purchase a particular product or will they churn another competitor, whether the user will click on a given advertisement link or not & many more examples are in the bucket. Logistic Regression is a Statistical method for producing binary classes. The outcome or target variable is ~~dichotomi~~ dichotomous in nature. Dichotomous mean there are only two possible classes. For example, it can be used for Cancer detection problems. It computes the probability of an event occurence.

It is a special case of linear regression where the target variable is categorical in nature. It uses a log of odds as the dependent variable. Logistic regression Predicts the Probability of occurence of a

binary event utilizing of log it function

- Linear Regression Equation

$$y = B_0 + B_1X_1 + B_2X_2 + \cdots + B_nX_n$$

where y is dependent variable and x_1, x_2 & x_n are explanatory variables.

- Sigmoid function:

$$P = 1/1 + e^{-(B_0 + B_1X_1 + B_2X_2 \cdots B_nX_n)}$$

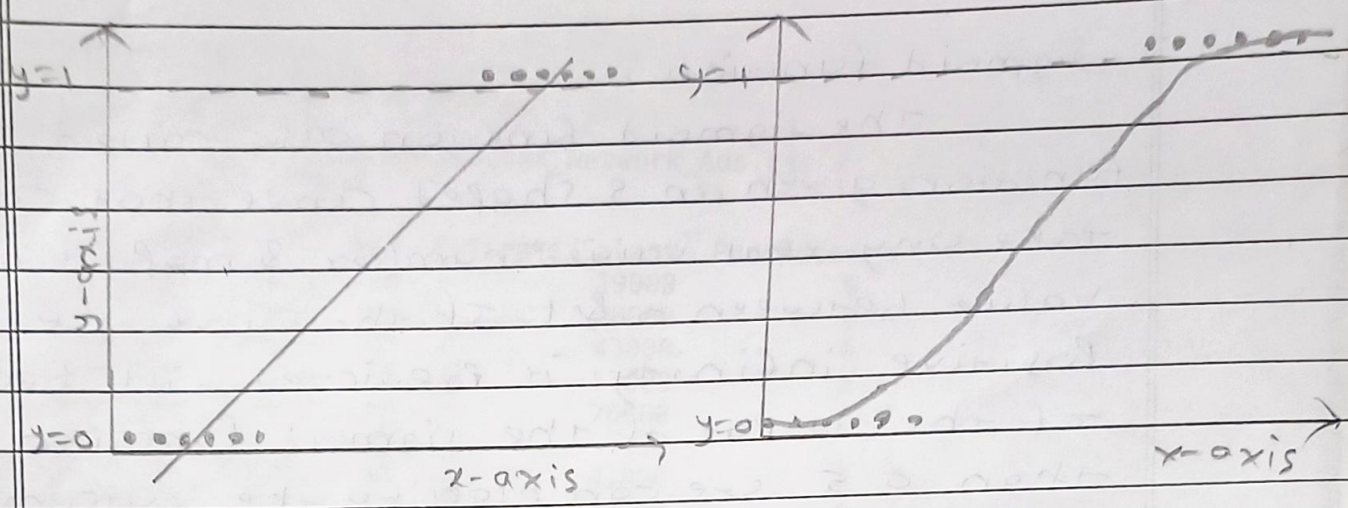
$$P = 1/1 + e^{-y}$$

Properties of Logistic Regression

1. The dependent variable in logistic regression follows Bernoulli Distribution
2. Estimation is done through maximum likelihood.
3. No R square model fitness is calculated through concordance, KS-statistics

Linear Regression vs Logistic Regression:

Linear regression gives you a continuous output but logistic regression provides a constant output. An example of the continuous output is house price & stock price. Linear Regression is estimated using maximum likelihood Estimation (MLE) approach.



Linear Regression logistic Regression

maximum likelihood estimation VS least square method

The MLE is a likelihood maximization method, while LLS is a distance minimizing approximation method. maximizing the likelihood function determines the parameters that are most likely to produce the observed data from a statistical point of view, MLE sets the mean & variance as parameters in determining the specific parametric values for a given model. This set of given parameters can be used for predicting the data needed in a normal distribution. ordinary least squares estimates are computed by fitting a regression line on given data points that has a minimum sum of the squared deviations. Both are used to estimate the parameters of a linear regression model.

Sigmoid function:

The sigmoid function also called logistic function given an s shaped curve that can take any real valued number & map it into a value between 0 & 1. If the curve goes to positive infinity it predicted will become 0. If the output of the sigmoid function is more than 0.5. we can classify the outcome as 1 or YES and if it less than 0.5, we can classify it as 0 or NO. If the output is 0.75 we can say in terms of Probability as there is a 75 percent chance that patient will suffer from cancer

$$f(x) = \frac{1}{1 + e^{-(x)}}$$

Types of logistic Regression

1. Binary logistic Regression -

The target variable has only two possible outcomes such as spam or NOT spam, cancer or NO cancer.

2. ordinal logistic Regression -

The target variable has three or more ordinal categories

3. Multinomial logistic Regression

The target variable has 3 or more categories

Conclusion: we succesfully implemented logistical Regression using Python on given dataset.

Assignment No-6

Title: Data analytics III

Aim: Implement simple Naive Bayes classification algorithm using Python/R on iris.csv dataset. Compute confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, Precision, Recall on the given dataset.

Theory:

Naive Bayes classification Algorithm:

It is a classification algorithm based on Baye's Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to out-perform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating Posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:

$$\text{Posterior probability} = \frac{\text{likelihood} \times \text{class prior probability}}{\text{predictor prior probability}}$$

$$\text{i.e. } P(c|x) = \frac{P(x|c)\,P(c)}{P(x)}$$

$$P(c|x) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

Working of Naive Bayes Algorithm

Step 1: We need to convert any dataset we have into a frequency table

Step 2: Create likelihood table by finding the probabilities. for example, for weather data set.

Frequency table

Weather	No	Yes
overcast		4
Rainy	3	2
Sunny	2	3
Total	5	9

Likelihood table

Weather	No	Yes		
overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Step 3: Now, use Naive Bayesian equation to calculate the Posterior probability for each class. The class with the highest Posterior probability is the outcome of prediction.

Steps in Naive Baye's are:

i) seperate by class ii) summarize the dataset

iii) summarize data by class iv) Gaussian Probability density function

v) class Probabilities

(i) Seperate by class: we seperate our training data by class to calculate the probability of data by the class they belong to. we can create a dictionary object where each key is the class value and then add all the records as the value

(ii) Summarize the dataset - we need the mean and standard deviation of a dataset.

mean = $\frac{\sum_{i=1}^{n} x_i}{N}$ Standard deviation = $\sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}}$

$\bar{x}$ = mean N = Total number of observations

x_i = observations

(iii) Summarize data by class - we require statistics from our training dataset organized by class. The dataset is first split by class, then statistics are calculated on each subset.

(iv) Gaussian Probability Density function - A Gaussian distribution can be summarized using mean and standard deviation. Gaussian probability function

$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \cdot \frac{e^{-x^2}}{2 \cdot \sigma^2}$ σ = standard deviation

(v) Class Probabilities - Statistics calculated from training data is used to calculate probabilities for new data. They are calculated separately for each class.

P(class / data) = P(X | class) * P(class)

The result is no longer strictly a probability of the data belonging to a class. The value is maximised, meaning that the calculation for the class that results in the largest value is taken as the prediction. This is a common implementation simplification as we are often more interested in the class prediction rather than the probability.

The input variables are treated seperately, giving the name 'naive'.

First the total number of training records is calculated from the counts stored in summary. This is used in the calculation of the probability.

Next, probabilities are calculated for each input value in the row using Gaussian probability density function and the statistics for that column and that class probabilities are multiplied together as they are accumulate

Conclusion: Thus Naive Bayes classification has been implemented.