SCTR's Pune Institute of Computer Technology
(PICT) Pune

B.E. Machine Learning and Data Science (Honors)

SUBMITTED BY

NAME : AISHWARYA KHANDERAO KAMBLE

ROLL NO : 42428

PRN NO : 72016920J

ABC ID : 594-165-436-179

BRANCH : E&TC

Under the guidance of
Prof. H. B. Mali

DEPARTMENT OF ELECTRONICS AND TELECOMMUNICATION ENGINEERING

ACADEMIC YEAR 2022-23

DEPARTMENT OF ELECTRONICS AND TELECOMMUNICATION ENGINEERING

SCTR's Pune Institute of Computer Technology (PICT), Pune
Maharashtra 411043

CERTIFICATE

This is certified that ML&DS laboratory experiments submitted by Mr. Aishwarya Kamble has satisfactorily completed the curriculum-based B.E. Machine learning and Data Science honors experiments under the guidance of Prof. H. B. Mali towards the partial fulfillment of final year Electronics and Telecommunication Engineering Semester VII (Honors Degree), Academic Year 2022-23 of Savitribai Phule Pune University.

Prof. H. B. Mali

Principal

LAB MANUAL

Academic Year: 2022-23

Class: B.E. **Date: 09/11/2022**

Subject: Machine Learning and Data Science (Honours)

Semester: I

Lab Expt. No	Problem Statement.
1.	Creating & Visualizing Neural Network for the given data. (Use python) Note: download dataset using Kaggal. Keras, ANN visualizer, graph viz libraries are equired.
2.	Recognize optical character using ANN
3.	Implement basic logic gates using Hebbnet neural networks
4.	Exploratory analysis on Twitter text data Perform text pre-processing, Apply Zips and heaps law, Identify topics
5.	Text classification for Sentimental analysis using KNN Note: Use twitter data
6.	Write a program to recognize a document is positive or negative based on polarity words using suitable classification method.

Machine Learning Data Science Laboratory (410501)

BE Sem I Honors in ML&DS

Academic Year: 2022-23

Lab Assignment No.1

Title: Creating and visualizing neural networks for the given data.

Objectives:

1. To handle given data for creating and visualizing neural network.
2. To analyze data using a python programming language.

Software Requirement:

Windows /Linux

Theory:

Neural network was inspired by the design and functioning of human brain and components

Definition:

Information processing model that is inspired by the way biological nervous system (i.e the brain) process information, is called Neural Network. Neural Network has the ability to learn by examples. It is not designed to perform fix /specific task, rather task which need thinking (e.g. Predictions). ANN is composed of large number of highly interconnected processing elements(neurons)working in unison to solve problems. It mimic human brain. With advanced in deep learning, you can now visualize the entire deep learning process or just the Convolutional Neural Network you've built.We are going to build simple neural network using Keras and then use ANN visualizer to visualize our neural network

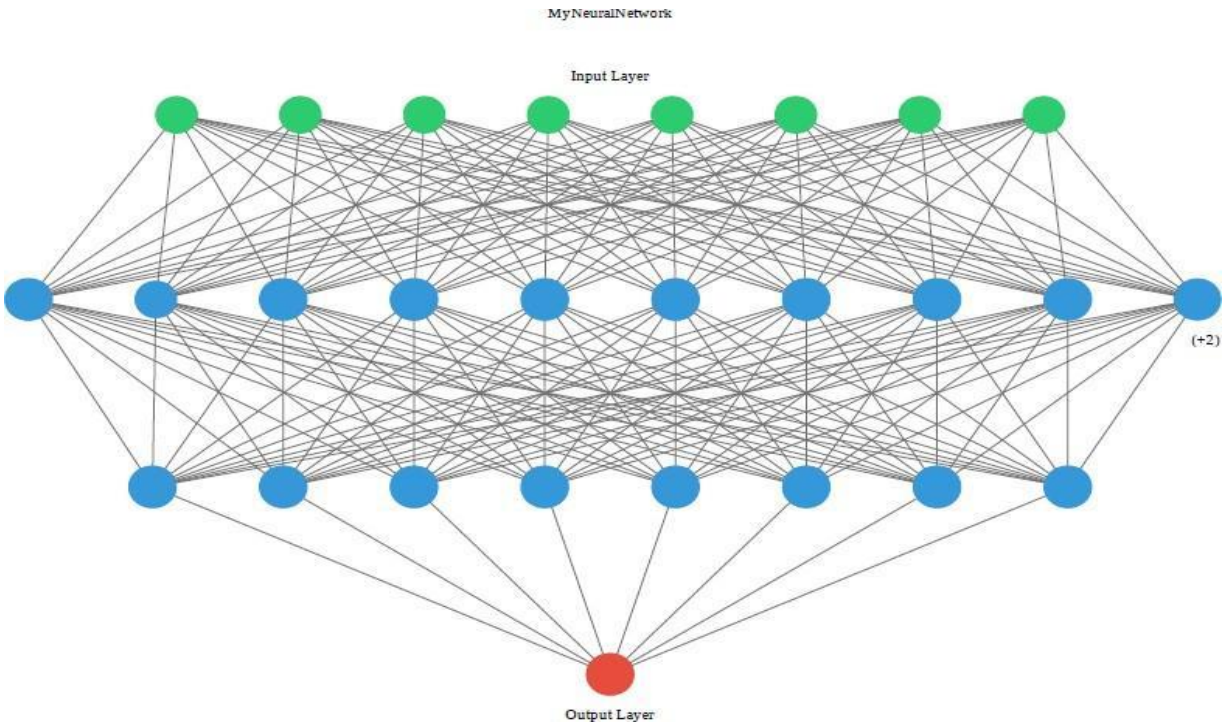
ANN Visualizer:

It is a python library that enables us to visualize an Artificial Neural Network using just a single line of code. It is used to work with Keras and makes use of python's graphviz library to create a neat and presentable graph of the neural network you're building

Visualize a Neural Network in Python using Graphviz

- Import module.
- Create a new object of Diagraph.
- Add **node ()** and **edge ()** into graph object.
- Save the source code with render () object.

Output :



Conclusion:

Here, we studied creating and visualizing neural network for the given data using
python

Machine Learning Data Science Laboratory (410501)

BE Sem I Honors in ML&DS

Academic Year: 2022-23

Lab Assignment No.2

Title: Recognize Optical Character using ANN

Objective:

To recognize the optical characters using ANN

Theory:

What is Optical Character Recognition?

Optical character recognition or **optical character reader** (**OCR**) is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (for example: from a television broadcast).[1]

Widely used as a form of data entry from printed paper data records – whether passport documents, invoices, bank statements, computerized receipts, business cards, mail, printouts of static-data, or any suitable documentation – it is a common method of digitizing printed texts so that they can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as cognitive computing, machine translation, (extracted) text-to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision.

Early versions needed to be trained with images of each character, and worked on one font at a time. Advanced systems capable of producing a high degree of recognition accuracy for most fonts are now common, and with support for a variety of digital image file format inputs.[2] Some systems are capable of

reproducing formatted output that closely approximates the original page including images, columns, and other non-textual components.

What is keras_ocr?

keras-ocr provides out-of-the-box OCR models and an end-to-end training pipeline to build new OCR models. Using this we get pre trained data and weights so our time and effort is saved.

Conclusion:

Here, we studied how to recognize optical characters using ANN.

Machine Learning Data Science Laboratory (410501)

BE Sem I Honors in ML&DS

Academic Year: 2022-23

Lab Assignment No.3

Aim: Implement basic logic gates using Mc-Culoch-Pitts or Hebbnet neural networks.

Objectives:

1.The student will be able to obtain the fundamentals and different architecture of neural networks.

2. The student will have a broad knowledge in developing the different algorithms for neural networks.

Software Requirements:

Ubuntu 18.04 /windows

Hardware Requirements:

Pentium IV system with latest configuration

Outcomes: The students will be able to,

- Describe the relation between real brains and simple artificial neural network models.
- Understand the role of neural networks in engineering.
- Apply the knowledge of computing and engineering concept to this discipline.

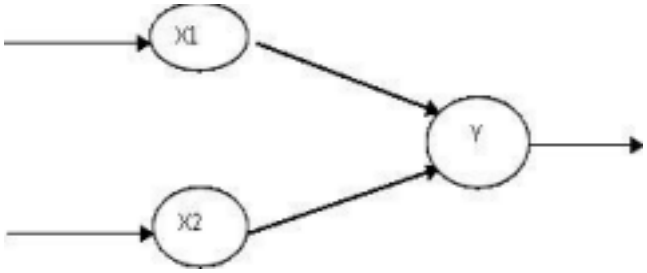
Theory:

Neural network was inspired by the design and functioning of human brain and components. Definition: Information processing model that is inspired by the way biological nervous system (i.e the brain) process information, is called Neural Network.

Neural Network has the ability to learn by examples. It is not designed to perform fix /specific task, rather task which need thinking (e.g. Predictions).

ANN is composed of large number of highly interconnected processing elements(neurons) working in unison to solve problems. It mimic human brain. It is configured for special application such as pattern recognition and data classification through a learning process. ANN is 85-90% accurate.

Basic Operation of a Neural Network:



X1 and X2 – input neurons.

Y- output neuron

Weighted interconnection links- W1 and W2.

Net input calculation is :

Yin= x1w1+x2w2

Output is :

y=f(Yin)

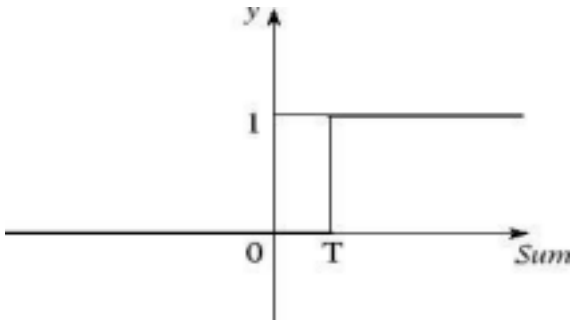
Output= function

The McCulloch-Pitts Model of Neuron:

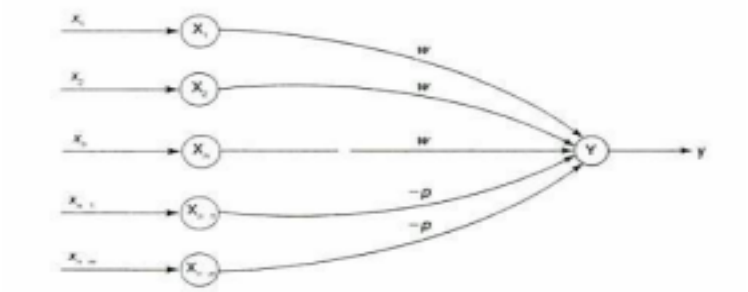
The early model of an artificial neuron is introduced by Warren McCulloch and Walter Pitts in 1943. The McCulloch-Pitts neural model is also known as linear threshold gate. It is a neuron of a set of inputs I1,I2,I3…Im and one output y . The linear threshold gate simply classifies the set of

inputs into two different classes. Thus the output y is binary. Such a function can be described mathematically using these equations:

$$y = f(Sum). \quad Sum = \sum_{i=1}^{N} I_i W_i,$$



W1,W2…Wm are weight values normalized in the range of either (0,1) or (-1,1) and associated with each input line, Sum is the weighted sum, and T is a threshold constant. The function f is a linear step function at threshold T as shown in figure



A simple M-P neuron is shown in the figure.

It is excitatory with weight (w>0) / inhibitory with weight –p (p<0).

In the Fig., inputs from x1 to xn possess excitatory weighted connection and Xn+1 to xn+m has inhibitory weighted interconnections.

Since the firing of neuron is based on threshold, activation function is defined as

$$f(y_{in}) = \begin{cases} 1 & if\ y_{in} \geq \theta \\ 0 & if\ y_{in} < \theta \end{cases}$$

For inhibition to be absolute, the threshold with the activation function should satisfy the following condition:θ >nw –p

Output will fire if it receives ―k‖ or more excitatory inputs but no inhibitory inputs where kw≥θ>(k-1) w

- The M-P neuron has no particular training algorithm.

- An analysis is performed to determine the weights and the threshold. - It is used as a building block where any function or phenomenon is modelled based on a logic function.

Activation function Yin is as follows: $Yin=x_1w_1+x_2w_2$

Input:

Input1	Input2	Output
0	0	0
0	1	0
1	0	0
1	1	1

Conclusion: Mc-Culloch pits Model is implemented for XOR function by using the thresholding activation function. Activation of M-P neurons is binary (i.e) at any time step the neuron may fire or may not fire. Threshold plays major role here.

Machine Learning Data Science Laboratory (410501)

BE Sem I Honors in ML&DS

Academic Year: 2022-23

Lab Assignment No.4

Aim:

Exploratory analysis on Twitter Text Data. Perform text preprocessing. Apply Zipf's and Heaps law , Identify topics.

Objectives:

1. The student will be able to perform text preprocessing
2. The student will learn the concept of Zipf's law.

Software Requirements: Windows with python and Jupyter notebook.

Outcomes:

The students will be able to perform text preprocessing.And will learn the Zipf's law concept.

Theory:

Data Preprocessing :

Data preprocessing is a technique which is used to transform the raw data in a useful and efficient format.

Steps Involved in Data Preprocessing:

1. Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, data

cleaning is done. It involves handling of missing data, noisy data etc.

- **(a). Missing Data:**
  This situation arises when some data is missing in the data. It can be handled in various ways.
  Some of them are:
    1. **Ignore the tuples:**
       This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

    2. **Fill the Missing values:**
       There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

- **(b). Noisy Data:**
  Noisy data is meaningless data that can't be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :
    1. **Binning Method:**
       This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segment is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

    2. **Regression:**
       Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

    3. **Clustering:**

This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2. Data Transformation:

This step is taken in order to transform the data in appropriate forms suitable for the mining process. This involves following ways:

1. **Normalization:**
   It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

2. **Attribute Selection:**
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

3. **Discretization:**
   This is done to replace the raw values of numeric attributes by interval levels or conceptual levels.

4. **Concept Hierarchy Generation:**
   Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

3. Data Reduction:

Since data mining is a technique that is used to handle huge amounts of data. While working with a huge volume of data, analysis became harder in such cases. In order to get rid of this, we use data reduction techniques. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

1. **Data Cube Aggregation:**
   Aggregation operation is applied to data for the construction of the data cube.

2. **Attribute Subset Selection:**
   The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use the level of significance and p- value of the attribute.The attribute having p-value greater than significance level can be discarded.

3. **Numerosity Reduction:**
   This enables us to store the model of data instead of whole data, for example: Regression Models.

4. **Dimensionality Reduction:**
   This reduces the size of data by encoding mechanisms.It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction is called lossless reduction, else it is called lossy reduction. The two effective methods of dimensionality reduction are:Wavelet transforms and PCA (Principal Component Analysis).

Zipf's Law:-

Zipf's law is a law about the frequency distribution of words in a language (or in a collection that is large enough so that it is representative of the language). To illustrate Zipf's law let us suppose we have a collection and let there be V unique words in the collection (the vocabulary). For each word in the collection we need to compute the freq(word) = how many times a word occurs in the collection. Then we

rank the words descending by their frequency (most frequent words have rank 1, next frequent word has rank 2, ...).**Zipf's law**, in probability, assertion that the frequencies *f* of certain events are inversely proportional to their rank *r*. The law was originally proposed by American linguist George Kingsley Zipf (1902–50) for the frequency of usage of different words in the English language; this frequency is given approximately by $f(r) \cong 0.1/r$. Thus, the most common word (rank 1) in English, which is *the*, occurs about one-tenth of the time in a typical text; the next most common word (rank 2), which is *of*, occurs about one-twentieth of the time; and so forth. Another way of looking at this is that a rank *r* word occurs $1/r$ times as often as the most frequent word, so the rank 2 word occurs half as often as the rank 1 word, the rank 3 word one-third as often, the rank 4 word one-fourth as often, and so forth. Beyond about rank 1,000, the law completely breaks down.

Zipf's law purportedly has been observed for many other statistics that follow an exponential distribution. For example, in 1949 Zipf claimed that the largest city in a country is about twice the size of the next largest, three times the size of the third largest, and so forth. While the fit is not perfect for languages, populations, or any other data, the basic idea of Zipf's law is useful in schemes for data compression and in allocation of resources by urban planners.

Heaps' law:-

Heaps' law (also called **Herdan's law**) is an empirical law which describes the number of distinct words in a document (or set of documents) as a function of the document length

Conclusion: Learnt text preprocessing technique and Zipf's law.

Machine Learning Data Science Laboratory (410501)

BE Sem I Honors in ML&DS

Academic Year: 2022-23

Lab Assignment No.5

Title: Text classification for Sentiment analysis using KNN

Objectives:

1. To handle Twitter Data for performing computing.
2. To analyze data using R programming tools.

Theory:

Sentiment analysis refers to the use of natural language processing, text analysis, and computational linguistics to systematically identify, extract, quantify, and study effective states and subjective information. Sentiment analysis is widely applied to customer materials such as reviews and survey responses. The most common type of sentiment analysis is 'polarity detection' and involves classifying customer materials/reviews as positive, negative or neutral.

Text Processing

With the increasing importance of computational text analysis in research, many researchers face the challenge of learning how to use advanced software that enables this text analysis. Text processing has a direct application to Natural Language Processing, also known as NLP. NLP is aimed at processing the languages spoken or written by humans when they communicate with one another. This is different from the communication

between a computer and a human where the communication is either a computer program written by a human or some gesture by a human like clicking the mouse at some position. NLP tries to understand the natural language spoken by humans and classify it, analyze it as well if required to respond to it. Python has a rich set of libraries which cater to the needs of NLP. The Natural Language ToolKit (NLTK) is a suite of such libraries which provides the functionalities required for NLP..

Twitter Data

Twitter is an online microblogging tool that disseminates more than 400 million messages per day, including vast amounts of information about almost all industries from entertainment to sports, health to business etc. One of the best things about Twitter—indeed, perhaps its greatest appeal-is in its accessibility. It's easy to use both for sharing information and for collecting it.Twitter provides unprecedented access to our lawmakers and to our celebrities, as well as to news as it's happening. Twitter represents an important data source for the business models of huge companies as well. All the above characteristics make twitter a best place to collect real time and latest data to analyse and do any sought of research for real life situations.

DATASET DESCRIPTION

We are given a Twitter US Airline Sentiment dataset that contains around 14,601 tweets about each major U.S. airline. The tweets are labelled as positive, negative, or neutral based on the nature of the respective Twitter user's feedback regarding the airline. The dataset is further segregated into training and test sets in a stratified fashion. Train set contains 11,680 tweets whereas the test set contains 2,921 tweets.Our task is to develop and train a k-nearest neighbors classifier on the training set and use it to predict sentiment classes of the tweets present in the test set.

Pre-Processing

Raw tweets scraped from twitter generally result in a noisy dataset. This is due to the casual nature of people's usage of social media. Tweets have certain special characteristics such as retweets, emoticons, user mentions, etc. which

have to be suitably extracted. Therefore, raw twitter data has to be normalized to create a dataset which can be easily learned by various classifiers. We have applied an extensive number of pre-processing steps to standardize the dataset and reduce its size. We first do some general pre-processing on tweets which is as follows.

- Convert the tweet to lower case.
- Replace 2 or more dots (.) with space.
- Strip spaces and quotes (" and ') from the ends of tweet.
- Replace 2 or more spaces with a single space.

Special twitter features as follows.

URL:

Users often share hyperlinks to other webpages in their tweets. Any particular URL is not important for text classification as it would lead to very sparse features. Therefore, we replace all the URLs in tweets with the word URL. The regular expression used to match URLs is ((www\.[\S]+)|(https?://[\S]+)).

User Mention

Every twitter user has a handle associated with them. Users often mention other users in their tweets by @handle. It replaces all user mentions with the word USER_MENTION. The regular expression used to match user mention is @[\S]+.

K-Nearest Neighbours

K-Nearest Neighbours is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection.It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data).

KNN algorithm is used to classify by finding the K nearest matches in training data and then using the label of closest matches to predict. Traditionally,

distance such as euclidean is used to find the closest match.KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Feature Extraction

In the feature extraction step, we will need to represent each tweet as a bag-of-words (BoW), i.e. an unordered set of words with their positions ignored and all of the emphasis placed on the respective frequencies of each word.

For example, consider these two tweets:

T1 = Welcome to machine learning, machine!

T2 = kNN is a powerful machine learning algorithm.

The bag-of-words representation (ignoring case and punctuation) for the above two tweets are:

Vocabulary	welcome	to	machine	learning	knn	is	a	powerful	algorithm
T1	1	1	2	1	0	0	0	0	0
T2	0	0	1	1	1	1	1	1	1

In order to create this bag-of-words representation, we would first need to extract out the unique words from all of our tweets in the training dataset.

Conclusion:

Hence, we studi ed On Twitter Data performs computing using Business Intelligence analytical tools electively.

Machine Learning Data Science Laboratory (410501)

BE Sem I Honors in ML&DS

Academic Year: 2022-23

Lab Assignment No.6

Title: Yelp reviews polarity

Objectives:

1. To handle given data for Text Classification
2. To analyze Text using python programming language.

Software Requirement:

Windows /Linux

Theory:

The Yelp reviews polarity dataset is constructed by considering stars 1 and 2 negative, and 3 and 4 positive. For each polarity 280,000 training samples and 19,000 testing samples are take randomly. In total there are 560,000 trainig samples and 38,000 testing samples. Negative polarity is class 1, and positive class 2. Definition:

Information processing model that is inspired by the way biological nervous system (i.e. the brain) process information, is called Neural Network.

Neural Network has the ability to learn by examples. It is not designed to perform fix /, rather task which need thinking (e.g. Predictions).

ANN is composed of large number of highly interconnected processing elements(neurons)working in unison to solve problems. It mimic human brain.

With advanced in deep learning, you can now visualize the entire deep learning process or just the Convolutional Neural Network you've built.

We are going to build simple neural network using Keras and then use ANN visualizer to visualize our neural network.

Conclusion

Handled given data for Text Classification and analyzed Text using python programming language.