

TY B.Tech. (CSE) – II [2020-21]
4CS462: PE-2 - Data Mining Lab.
Assignment No. [06]
Date: 08/10/2021

GroupID: DM21G07

Group Members:	2018BTECS00063	Aryan Mali
	2018BTECS00094	Shreya Singh
	2018BTECS00099	Ganesh Kasar

Title: Classifiers

Aim: Design and implement the following classifiers:

- a) Regression classifier.
- b) Naïve Bayesian Classifier.
- c) k-NN classifier (Take $k = 1, 3, 5, 7$)
- d) Three-layer Artificial Neural Network (ANN) classifier (use backpropagation). Plot error graph (iteration vs error).

Introduction:

Classification is the process of predicting the class of given data points. Classes are sometimes called targets/ labels or categories. Classification predictive modelling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

For example, spam detection in email service providers can be identified as a classification problem. This is a binary classification since there are only 2 classes as spam and not spam. A classifier utilizes some training data to understand how given input variables relate to the class. In this case, known spam and non-spam emails have to be used as the training data. When the classifier is trained accurately, it can be used to detect an unknown email.

Classification belongs to the category of supervised learning where the targets are also provided with the input data. There are many applications in classification in many domains such as in credit approval, medical diagnosis, target marketing etc.

Theory:

Regression Classifier:

Machine learning, more specifically the field of predictive modelling is primarily concerned with minimizing the error of a model or making the most accurate predictions possible, at the expense of explainability. In applied machine learning, we will borrow, reuse and steal algorithms from many different fields, including statistics and use them towards these ends.

As such, linear regression was developed in the field of statistics and is studied as a model for understanding the relationship between input and output numerical variables, but has been borrowed by machine learning. It is both a statistical algorithm and a machine learning algorithm.

Naive Bayes Classifier

The fundamental Naive Bayes assumption is that each feature makes an:

- independent
- equal

contribution to the outcome.

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

where A and B are events and $P(B) \neq 0$.

Basically, we are trying to find the probability of event A, given the event B is true. Event B is also termed as evidence.

$P(A)$ is the probability of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance (here, it is event B).

$P(A|B)$ is a posteriori probability of B, i.e. probability of event after evidence is seen.

k-nearest neighbours algorithm

In statistics, the k-nearest neighbours' algorithm (k-NN) is a non-parametric classification method first developed by Evelyn Fix and Joseph Hodges in 1951 and later expanded by Thomas Cover. It is used for classification and regression. In both cases, the input consists of the k closest training examples in a data set. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours (k is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of that single nearest neighbour.

In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbours.

k-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically.

Artificial neural network

Artificial neural networks (ANNs), usually simply called neural networks (NNs), are computing systems inspired by the biological neural networks that constitute animal brains.

An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal to other neurons. An artificial neuron receives a signal then processes it and can signal neurons connected to it. The "signal" at a connection is a real number,

and the output of each neuron is computed by some non-linear function of the sum of its inputs. The connections are called edges. Neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Neurons may have a threshold such that a signal is sent only if the aggregate signal crosses that threshold. Typically, neurons are aggregated into layers. Different layers may perform different transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times.

Procedure:

Use/extend the data analysis tool (menu-driven GUI) developed in Assignment No. 2

to perform the following classification task :

1. Design and implement the following classifiers:

- a) Regression classifier.
- b) Naïve Bayesian Classifier.
- c) k-NN classifier (Take $k = 1, 3, 5, 7$)
- d) Three-layer Artificial Neural Network (ANN) classifier (use back propagation). Plot error graph (iteration vs error).

2. Tabulate the results in the confusion matrix and evaluate the performance of above

classifier using the following metrics :

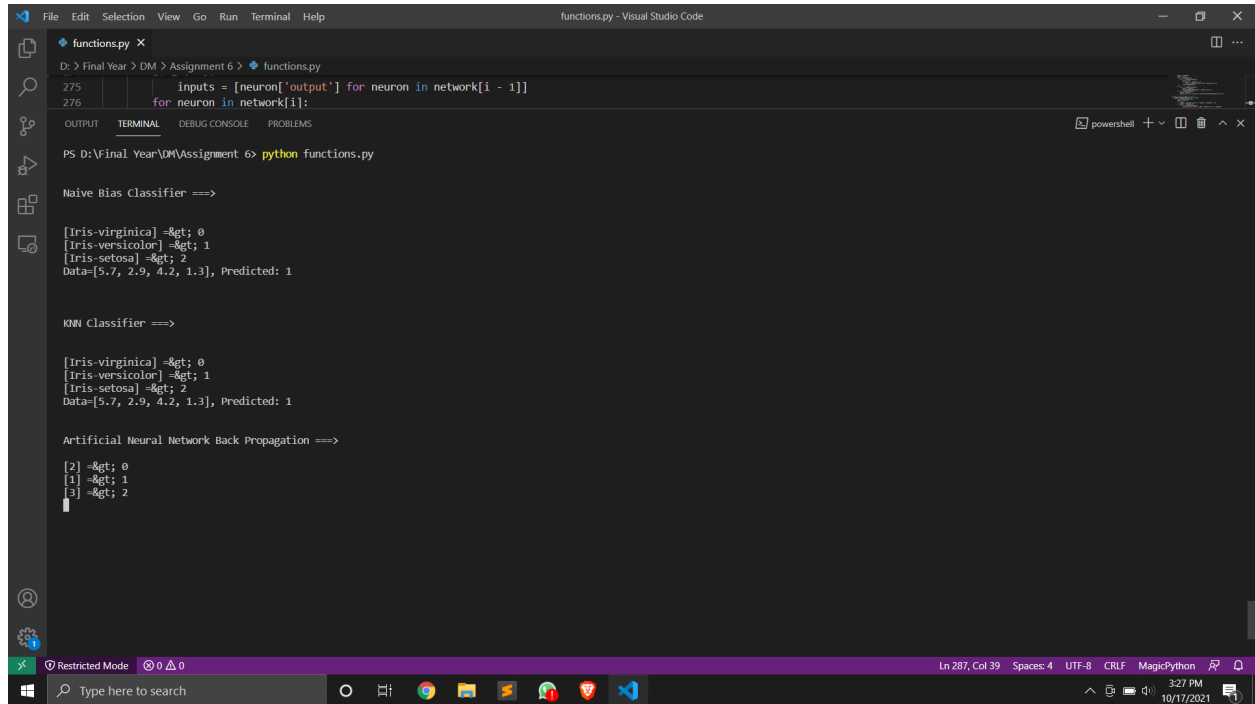
- a) Recognition rate
- b) Misclassification rate
- c) Sensitivity
- d) Specificity
- e) Precision & Recall

3. Use the following data sets from the UCI machine learning repository :

- a) IRIS

b) Breast Cancer

Results:



```
functions.py
D:\> Final Year > DM > Assignment 6 > functions.py
275     inputs = [neuron['output'] for neuron in network[i - 1]]
276     for neuron in network[i]:
OUTPUT TERMINAL DEBUG CONSOLE PROBLEMS
PS D:\Final Year\DM\Assignment 6> python functions.py

Naive Bias Classifier ==>

[Iris-virginica] => 0
[Iris-versicolor] => 1
[Iris-setosa] => 2
Data=[5.7, 2.9, 4.2, 1.3], Predicted: 1

KNN Classifier ==>

[Iris-virginica] => 0
[Iris-versicolor] => 1
[Iris-setosa] => 2
Data=[5.7, 2.9, 4.2, 1.3], Predicted: 1

Artificial Neural Network Back Propagation ==>

[2] => 0
[1] => 1
[3] => 2
```

Conclusion:

Implemented

- Regression classifier.
- Naïve Bayesian Classifier.
- k-NN classifier (Take $k = 1, 3, 5, 7$)
- Three-layer Artificial Neural Network (ANN) classifier (use backpropagation). Plot error graph (iteration vs error)