CS 6375
ASSIGNMENT 2

Group Members:

SaiVikas Meda: SXM190011
Sneha Elizabeth Sam: SES190004

Number of free late days used: 0

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

www.stackoverflow.com
https://scikit-learn.org/stable/modules/preprocessing.html
https://www.analyticsvidhya.com/blog/2016/07/practical-guide-data-preprocessing-python-scikit-learn/
https://www.kaggle.com/statinstilettos/neural-network-approach/data
https://analyticsindiamag.com/5-ways-handle-missing-values-machine-learning-datasets/
https://archive.ics.uci.edu/ml/datasets.php

OUTPUT FOR DATASET IN TABULAR FORM:

Dataset Used: HandWritten Digit recognition dataset

Recordings:

		TRAINING DATA		**TEST DATA**	
Act. Fn	**L.R.**	**T.E.**	**Avg Err.**	**T.E.**	**Avg Err.**
Sigmoid	0.02	1727.325448	0.05140695	435.057409	0.051798715
Sigmoid	0.05	1727.33967	0.05140738	435.06236	0.051799305
Sigmoid	0.25	1727.33487	0.0514072	435.060185	0.05179900
Tanh	0.02	38807.26543	1.154943	1510.96296	0.17989795
Tanh	0.05	33699.48765	1.002931	1510.9768	0.1798996
Tanh	0.25	39297.46917	1.169532	1458.97685	0.1737084
ReLu	0.02	5852.32098	0.173171036	435.060185	0.0517990457
ReLu	0.05	5852.32098	0.173171036	435.060185	0.0517990457
ReLu	0.25	5852.32098	0.173171036	435.060185	0.0517990457

KEY:	
Act. Fn.	Activation Function
L.R.	Learning Rate
T.E.	Total Error
Avg Err	Avg Error. (T.E / data rows)

Dataset Used: Loanprediction dataset

Recordings:

		TRAINING DATA		**TEST DATA**	
Act. Fn	**L.R.**	**T.E.**	**Avg Err.**	**T.E.**	**Avg Err.**
Sigmoid	0.02	23.97683323	0.062439669	9.56991604	0.0996866
Sigmoid	0.05	21.1233095	0.05500861	9.91794243	0.1033119
Sigmoid	0.25	25.6680022	0.066843	11.00076916	0.114663
Tanh	0.02	33.73233	0.0878446	12.000032	0.125
Tanh	0.05	45.7156833	0.11905	11.471	0.119542
Tanh	0.25	60.4727813	0.1574812	12.0	0.125
ReLu	0.02	135.5	0.35286	14.8499714	0.15468
ReLu	0.05	135.5	0.35286	14.8499714	0.15468
ReLu	0.25	135.5	0.35286	14.8499714	0.15468

REPORT BASED ON FINDINGS:

I performed and deduced to conclusion from the results of two different dataset that I used.

1. Handwritten digit recognition dataset
2. Loanprediction dataset

As we notice from the tables above sigmoid Fns gives the lowest Total error when compared to tanh and ReLu activation functions for both the datasets.

Sigmoid is the most commonly used classifier in Neural Networks. As, we notice from the graph of sigmoid, X values between -2 to 2, Y values are very steep. This trends to bring the activations to either side of the curve. Unlike linear function, the output of the activation function is always in the range (0,1). Which is essential in predicting the target to have bound range.

For a sigmoid fn the Y values varies very less for the changes in X at the ends of the graph. Due to that we don't observe much change in the total of error train and test data.

There is a high change in train error and test error during tanh fn because the gradient of tanh is stronger than gradient of sigmoid or Relu fns. This results in high value update in weight vectors of neurons.

For ReLu fn the gradient will be zero, because of this there will be no updates in the weights vectors during back propagation.

a sigmoid works well for a classifier (see the graph of sigmoid, doesn't it show the properties of an ideal classifier?) because approximating a classifier function as combinations of sigmoid is easier than maybe ReLu,

Assumptions and Logic:

The project files contains NeuralNet.py along with 2 datasets

1. Handwritten digit recognition dataset. - Numbers.csv
    a. This files contains more than 40k lines of data in it.
    b. 80% of data is used for training and 20% for testing.

2. Loanprediction Dataset:
    a. Loan_X_train.csv and Loan_Y_train.csv data are used for training purposes where X_train contains the columns data and Y_train contains the TARGET.
    b. Loan_X_test.csv and Loan_Y.test.csv this files contain the data used to test the trained Neural network.

If any column has NULL value in handwritten dataset they are filled with mean value of the column in preprocessing method.

All column values are normalized to [0-1] range. {(Value - Min)/(Max-Min) }

The Neural networks is trained after the preprocessing method.

In Loanprediction dataset columns(Gender, Married, Education, self-employed,) are converted into 0,1 values and (ApplicantIncome CoapplicantIncome, LaonAmount, Loan_Amount_Term) columns are normalized.

The logic are LoanPrediction dataset is committed in the NeuralNet.py file with a comment added on top at each section. To run the logic for LoanPredict uncomment the logic in init , preprocess and predict methods and comment the logic are Handwritten dataset.

Note: HandWritten recognition dataset takes more time to process as it contains huge data to read and train.