CSE 574
Project 2
Report

Name- Ankit Kumar Sinha
UB Person#-50286874

Task:

1. Performing Linear Regression on each of the 4 datasets:
(a) Human Observed Dataset with feature concatenation.
(b) Human Observed Dataset with feature subtraction.
(c) GSC Dataset with feature concatenation.
(d) GSC Dataset with feature subtraction.
2. Performing Logistic Regression on each of the 4 datasets:
(a) Human Observed Dataset with feature concatenation.
(b) Human Observed Dataset with feature subtraction.
(c) GSC Dataset with feature concatenation.
(d) GSC Dataset with feature subtraction
3. Perform Neural Network on each of the 4 datasets.

Data Preprocessing:

1. Dataset was read from 3 different files.
    1.1 DifferentPair.csv had pairs which were not having similar features and were of the different writers.
    1.2. SamePair.csv had pairs which were having more similar features and were of the same writers.
    1.3. Feature.csv had features of all the writer of the dataset.
2. The above step was done for both Human Observed Dataset and GSC Dataset.
3. Then for both the Dataset, features were concatenated for each writer of the pair selected.
4. Similarly for both the Dataset, features were subtracted for each writer of the pair selected.
5.Divide the dataset into 3 parts:
    5.1. Training Data- 80% of the dataset is used for training the model and finding the optimal weights.
    5.2. Validation Data- next 10% of the dataset is used for validation of the model for tuning the hyperparameters on an unseen dataset to achieve the optimal hyperparameters.
    5.3 Testing Data- last 10% of the dataset is used for testing the defined model on unseen data for finding the accuracy and error in it.

Dataset:

1. Human Observed Dataset:
    - It has 9 features for each writer and is prepared by human document examiner.
    - The dataset contains 791 same pairs and 293,032 different pairs of writers. So to balance it an equal number of writers from both files were selected for better convergence.
    - Features were concatenated to form 18 feature and also subtracted to form 9 features then fed into the model for training and testing.
2. Human Observed Dataset:

- It has 512 features for each writer and is prepared by human document examiner.
- The dataset contains 71,531 same pairs and 762,557 different pairs of writers. So to balance it an equal number of writers from both files were selected for better convergence.
- Features were concatenated to form 1024 feature and also subtracted to form 512 features then fed into the model for training and testing.

Methodology and Result:

1. Linear Regression for Human Observed Dataset:
   - Concatenation-
     - Fig shows the result of concatenation.

```
Training Erms:0.49975104940790493
Validation Erms:0.4996069665342921
Testing Erms:0.49987481321277005
Training Accuracy:52.843601895734594
Validation Accuracy:48.734177215189874
Testing Accuracy:48.10126582278481
```

   - Subtraction-
     - Fig shows the result of subtraction.

```
Training Erms:0.49986802713525325
Validation Erms:0.5207512623131108
Testing Erms:0.5250077879939247
Training Accuracy:52.44865718799368
Validation Accuracy:47.46835443037975
Testing Accuracy:45.56962025316456
```

2. Logistic Regression for Human Observed Dataset:
   - Concatenation-
     - Fig shows the result of concatenation.

```
Training Erms:0.7048563943718974
Validation Erms:0.6843642527044694
Testing Erms:0.6980986836772425
Training Accuracy:50.31595576619273
Validation Accuracy:53.164556962025316
Testing Accuracy:51.265822784810126
```

   - Subtraction-
     - Fig shows the result of subtraction.

```
Training Erms:0.6429040186840879
Validation Erms:0.7115680669648201
Testing Erms:0.7115680669648201
Training Accuracy:49.447077409162716
Validation Accuracy:49.36708860759494
Testing Accuracy:49.36708860759494
```

2. Neural Network for Human Observed Dataset:
   - Concatenation-
     - Fig shows the result of concatenation.

```
Errors: 150  Correct :165
Testing Accuracy: 52.38095238095238
Testing Erms: 0.6900655593423543
```

- Subtraction-
    - Fig shows the result of subtraction.

```
Errors: 152  Correct :163
Testing Accuracy: 51.74603174603175
Testing Erms: 0.6946507630023036
```

4. Linear Regression for GSC Dataset:

- Concatenation-
    - Fig shows the result of concatenation.

```
Training Erms:0.5521196089384836
Validation Erms:0.5564190938698723
Testing Erms:0.5555477238186821
Training Accuracy:52.330357142857146
Validation Accuracy:52.39456754824875
Testing Accuracy:52.894924946390276
```

- Subtraction-
    - Fig shows the result of subtraction.

```
Training Erms:0.49243655165801614
Validation Erms:0.491853004122092
Testing Erms:0.49112113165276383
Training Accuracy:61.99107142857143
Validation Accuracy:62.544674767691205
Testing Accuracy:64.2601858470336
```

5. Logistic Regression for GSC Dataset:

- Concatenation-
    - Fig shows the result of concatenation.

```
Training Erms:0.7054000891085043
Validation Erms:0.7213679325193149
Testing Erms:0.6976935349875347
Training Accuracy:50.24107142857143
Validation Accuracy:47.96283059328091
Testing Accuracy:51.32237312365976
```

- Subtraction-
    - Fig shows the result of subtraction.

```
Training Erms:0.7062223445912746
Validation Erms:0.706854017360353
Testing Erms:0.7093776027432548
Training Accuracy:50.125
Validation Accuracy:50.03573981415297
Testing Accuracy:49.678341672623304
```

6. Neural Network for GSC Dataset:

- Concatenation-
    - Fig shows the result of concatenation.

```
Errors: 1362  Correct :1437
Testing Accuracy: 51.339764201500536
Testing Erms: 0.6975688912107496
```

- Subtraction-
    - Fig shows the result of subtraction.

```
Errors: 1450  Correct :1349
Testing Accuracy: 48.19578420864595
Testing Erms: 0.7197514556522554
```

Conclusion:

Here the data is classified into binary classification, where 0 means different writer and 1 means the same writer.

- Human Observed Dataset Vs GSC Dataset-
  - The GSC dataset is a more rich dataset, has more features and also has more instances than the substracted dataset.
  - The Human Observed dataset has fewer features and less the instances.
  - So the model Performs well on GSC Dataset in comparison to Human Observed Dataset because it gets more distinguition features between the two writers and can differentiate well.
- Concatenation Vs Subtraction of Dataset-
  - In concatenation, there are more number of feature to train as features of both writer are used together, so the model can learn on more information and perform better.
  - In subtraction, there are less number of feature to train as the features of both writer are subtracted, so the model can learn on more information and perform better.

- Linear Regression Vs Logistics Regression:
  - In Linear regression, the model is trained without any activation function applied to the predicted value.
  - In Logistic Regression, the model is trained with sigmoid activation function applied to the predicted value.
  - The Equation for Predicted Output -
    - For Linear Regression- $PredictedTarget = W^T \phi(x)$
    - For Logistic Regression- $PredictedTarget = \sigma(W^T \phi(x))$
  - The Equation to update weights are-
    - For Linear Regression- $Wupdated = W - \eta\Delta W$,
      - where $\Delta W = -\phi(x) * (ActualTarget - W^T \phi(x))$
    - For Logistic Regression- $Wupdated = W - \eta\Delta W$,
      - Where $\Delta W = Feature(\sigma(W^T \phi(x) - ActualTarget)$
  - The Equation for Predicted Output -
    - For Linear Regression- $Loss = 0.5(ActualTarget - PredictedTarget)^2$
    - For Logistic Regression- $Loss = -(ActualTarget * log(PredictedTarget) + (1 - ActualTarget)log(1 - PredictedTaget))$

From the implementation and resource point of view, for large Dataset, it can be concluded that the Neural Network performs better than Logistic Regression and Logistic Regression Regression Performs better than Linear regression. Whereas for the smaller dataset, it can be concluded that Neural network is not able to train appropriately in comparison to logistic and linear regression.