Handwritting Comparison

1 2 3 4		Anunay Rao anunayra@buffalo.edu	
5	1	Introduction	
6 7 8 9 10 11 12	The project is to apply the different machine learning methods namely, Linear Regression, Logistic Regression and Neural Network on the dataset to develop a predictive model and compare the performance of each model. The task is to compare the handwriting and so the target value 0 denotes that the handwriting samples that we are comparing are from different writers whereas the target value 1 denotes that the handwriting samples are from the same writer.		
13 14	2	Preprocessing of the Dataset	
15 16 17		lly we have been provided with two datasets namely, Human Observed et and GSC Dataset.	
18	2.1	Human Observed Dataset	
19	This c	lataset consists 3 CSV files:	
20 21 22 23 24 25 26 27 28	•	 same_pair.csv (img_id_A, img_id_B, target): where img_id_A and img_id_B are two id's of images from the same writer and so the target value corresponding to all the samples in this file is 1. This file contains 791 samples. diffn_pairs.csv (img_id_A, img_id_B, target): where img_id_A and img_id_B are two id's of images from the different writer and so the target value corresponding to all the samples in this file is 0. This file contains 293032 samples HumanObserved-Features-Data.csv (img_id,f1,f2,f9): where img_id is id of a particular image and fl to f9 are human observed features corresponding to that image. 	
29 30 31 32 33 34 35	has 22 order featur gettin (absol	since the same_pairs.csv has only 791 samples whereas diffn_pairs.csv 93032 samples we need to make the probability of each class equal in to apply any machine learning method. Further each image has 9 res so we will create 2 datasets one with concatenating the features thus g 18 features for one image pair and other with subtracting the features lute value) thus getting 9 features for one image pair, so we need to w the steps given below in order to get a proper dataset.	
36 37 38 39	1.	Copy the features corresponding to all the the image pairs in same_pairs.csv from HumanObserved-Features-Data.csv and and concatenate them to get 18 features and 1 as the target value in file named final_cat.csv. Shuffle the rows in diffn_pairs_csv	

39
40
40
3. Copy the features corresponding to the first 791 image pairs in the diffn_pairs.csv
41
42
41
42
43
44
44
44
45
46
47
47
48
49
49
40
40
41
42
41
42
42
43
44
44
45
45
46
47
47
48
49
49
40
40
41
42
41
42
43
44
44
45
46
47
47
48
49
49
40
41
41
42
42
44
45
46
47
47
47
48
49
49
49
40
41
41
42
42
44
45
46
47
47
47
48
48
49
49
49
49
49
40
41
41
42
41
42
42
42
44
45
46
47
47
47
47
48
48
49
49
49
49
40
41
41
42
42
44
44
45
46
47
47
48
48
48
49
49
49
49
49
49
49
49
49
49
49
49
40
41
41
42
42
44
44
44
45
46
47
47
48
48
49
49
49
49
49
49
49
49
49
49
49
40
40
40
40
41
<

- 43 4. Shuffle the rows in the file final_cat.csv to get the final data.
- 44 Now to get the Dataset with $\overline{9}$ features i.e. feature subtraction subtract the col 5. 45 index[9] with col index[0], [1] with [10],......[8] with [17]. This file will be 46 final sub.csv where col index[10] will be the target value.
- 47

56 57

58

48 2.1 **GSC** Dataset

49 This dataset also contains 3 CSV files:

- 50 same pair.csv (img id A, img id B, target): where img id A and img id B are 51 two id's of images from the same writer and so the target value corresponding to all 52 the samples in this file is 1. This file contains 71531 samples. 53 54
 - diffn_pairs.csv (img_id_A, img_id_B, target): where img_id_A and img_id_B are • two id's of images from the different writer and so the target value corresponding to all the samples in this file is 0. This file contains 762557 samples
 - GSC-Features.csv (img id,f1,f2.....,f9): where img id is id of a particular • image and f1 to f512 are 512 features corresponding to that image which are extracted from the image with GSC algorithm.

59 Now since the same pairs.csv has only 71531 samples whereas diffn pairs.csv has 762557 samples we need to make the probability of each 60 class equal in order to apply any machine learning method. Further each 61 62 image has 512 features so we will create 2 datasets one with concatenating the 63 features thus getting 1024 features for one image pair and other with subtracting the features (absolute value) thus getting 512 features for one 64 image pair. These two datasets can be obtained in the similar fashion as 65 described for Human Observed dataset. 66

67 For linear regression task, we need to compute the inverse of the matrix and and it is found that for GSC dataset it yields singular matrix which implies 68 69 that some columns have all the entries as zero which needs to be processed so 70 that we can compute the inverse. So, one way is to add some noise to the 71 diagonal elements of the matrix or delete the columns with zero value. I have 72 opted for the second method and found that there are 6 columns with zero 73 value in feature concatenation namely column index [450], [452], [456], 74 [457], [962], [964] and [968] in case of feature subtraction there are 3 75 columns namely column index [450], [452] and [456].

76 77

3 **Performance Metric**

- 78 Linear Regression: We will evaluate the solution obtained by using Root
- Mean Square (RMS) error defined as $E_{RMS} = \sqrt{2E(w^*)/N_V}$ where w^* is the 79 solution and N_V is the size of dataset. Accuracy is not a good performance 80 81 metric for this linear regression tasks.

82 Logistic Regression and Neural Network: We will evaluate the performance 83 of these two models by accuracy which is defined as:

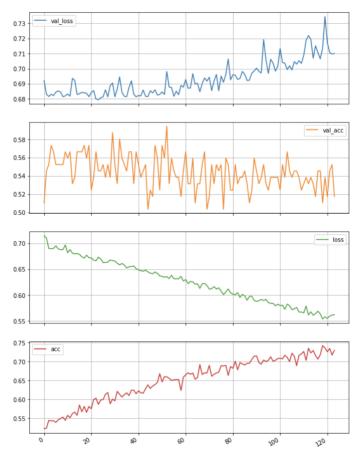
84
$$Accuracy = \frac{correct}{correct + wrong} \times 100$$

85 As here accuracy makes more sense than Erms for calculating the performance. 86

```
87
            Hyper-parameters values and Results:
      4
 88
 89
      4.1 Linear Regression
 90
      \lambda for closed form: 0.03 (Regularization Term)
 91
      Training Percent = 80
 92
      Validation Percet = 10
 93
      Testing Percent = 10
 94
      \lambda for gradient descent solution: 1.8 (Regularization Term)
 95
      learning rate : 0.01
 96
 97
      4.1.1 Linear Regression on Human Observed Dataset with
 98
 99
      Feature Concatenation:
      # of Gaussian Basis Function: 18
100
101
102
      Results:
103
      Closed Form:
104
      E rms Training = 0.4967303666643748
105
      E rms Validation = 0.4936395796664404
      E rms Testing = 0.49767401628116004
106
107
108
      Gradient Descent:
      E_rms Training = 0.49957
109
      E rms Validation = 0.49914
110
111
      E rms Testing = 0.49795
112
113
      4.1.2 Linear Regression on Human Observed Dataset with
114
      Feature Subtraction:
115
      # of Gaussian Basis Function: 9
116
117
      Results:
118
119
      Closed Form:
      E rms Training = 0.4991599889825706
120
121
      E rms Validation = 0.497113631312705
      E_rms Testing = 0.4973578173205024
122
123
124
      Gradient Descent:
125
      E rms Training = 0.50009
126
      E_rms Validation = 0.49994
127
      E_rms Testing = 0.49513
128
129
      4.1.3 Linear Regression on GSC Dataset with Feature
130
```

```
131
      Concatenation:
132
      # of Gaussian Basis Function: 10
133
      Results:
134
      Closed Form:
      E_rms Training = 0.5070384151105187
135
136
      E_rms Validation = 0.5052882346175389
      E rms Testing = 0.506047865254507
137
138
139
      Gradient Descent:
140
      E rms Training = 0.68516
      E rms Validation = 0.67824
141
142
      E rms Testing = 0.68111
143
      4.1.4 Linear Regression on GSC Dataset with Feature
144
      Subtraction:
145
      # of Gaussian Basis Function: 10
146
      Results:
147
      Closed Form:
148
      E rms Training = 0.7082241711685279
      E rms Validation = 0.7010008622415624
149
150
      E_{\rm rms} Testing = 0.704159509748694
151
152
      Gradient Descent:
153
      E_rms Training = 0.70822
      E rms Validation = 0.701
154
155
      E_rms Testing = 0.70416
156
157
158
      4.2 Logistic Regression:
159
      Training Percent = 80
      Validation Percet = 10
160
161
      Testing Percent = 10
162
      learning rate : 0.01
163
164
      4.2.1 Logistic Regression on Human Observed Dataset with
      Feature Concatenation:
165
166
      Results:
167
      Training Accuracy = 56.962025316455694
168
      Validation Accuracy = 55.69620253164557
169
      Testing Accuracy = 45.859872611464965
170
171
172
      4.2.2 Logistic Regression on Human Observed Dataset with
```

```
173
      Feature Subtraction:
174
      Results:
175
      Training Accuracy = 49.32806324110672
      Validation Accuracy = 51.265822784810126
176
177
      Testing Accuracy = 54.140127388535035
178
179
      4.2.3 Logistic Regression on GSC Dataset with Feature
180
      Concatenation:
181
      Results:
182
      Training Accuracy = 55.83049366535605
      Validation Accuracy = 54.65538934712708
183
184
      Testing Accuracy = 54.771059070255156
185
186
      4.2.4 Logistic Regression on GSC Dataset with Feature
187
      Subtraction:
188
      Result:
189
      Training Accuracy = 77.25993883792049
      Validation Accuracy = 76.69509296798546
190
191
      Testing Accuracy = 76.77735057672143
192
193
      4.3 Neural Network Implementaion:
194
      Training Percent = 81
195
      Validation Percet = 9
      Testing Percent = 10
196
197
      drop out = 0.2
198
      first_dense_layer_nodes = 256
199
      second dense layer nodes = 1
200
      Activation function first layer = ReLu
201
      Activation function second layer = sigmoid
202
      Optimizer = rmsprop
203
      Loss = binary_crossentropy
204
      model batch size = 128
205
206
207
      4.3.1 Neural Network on Human Observed Dataset with
208
      Feature Concatenation:
209
      num epochs = 10000
210
      early_patience = 100
211
      input_sizze = 18
212
213
      Results:
214
      Errors: 66
215
      Correct :91
      Testing Accuracy: 57.961783439490446
216
```



217 218 219 *Figure 1: Showing Validation loss, Validation Accuracy, Training Loss and Training Accuracy (top to bottom) against number of epochs*

221 4.3.2 Neural Network on Human Observed Dataset with

222 Feature Subtraction:

- 223 num_epochs = 10000
- 224 early_patience = 100
- 225 input_sizze = 9
- 226

227 **Results:**

- 228 Errors: 19
- 229 Correct :138
- 230 Testing Accuracy: 87.89808917197452
- 231

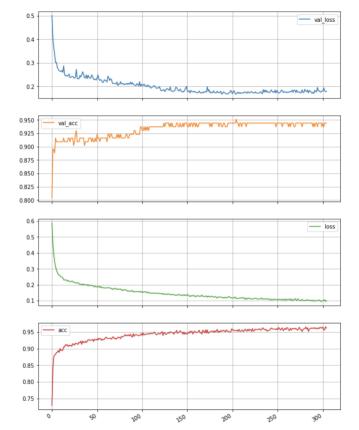


Figure 2:Showing Validation loss, Validation Accuracy, Training Loss and Training Accuracy (top to bottom) against number of epochs

- 232 233 234
- 235
- 236

237 4.3.3 Neural Network on GSC Dataset with Feature

- 238 Concatenation:
- 239 num_epochs = 50
- 240 early_patience = 10
- 241 input_size = 1024
- 242

243 **Results:**

- 244 Errors: 950
- 245 Correct :13355
- 246 Testing Accuracy: 93.35896539671444
- 247

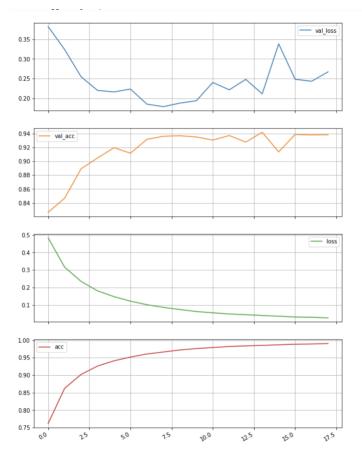


Figure 3:Showing Validation loss, Validation Accuracy, Training Loss and Training Accuracy (top to bottom) against number of epochs

252 4.3.4 Neural Network on GSC Dataset with Feature

- 253 Subtraction:
- 254 num_epochs = 50
- 255 early_patience = 10
- 256 input_size = 512
- 257
- 258 Results:
- 259 Errors: 2503
- 260 Correct :11802
- 261 Testing Accuracy: 82.50262146102762
- 262

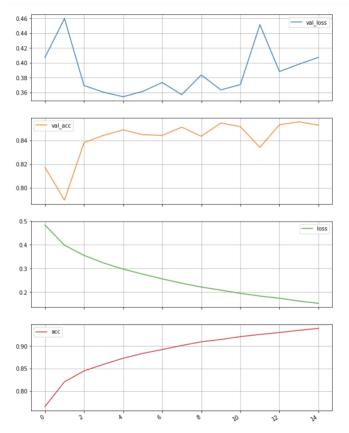


Figure 4:Showing Validation loss, Validation Accuracy, Training Loss and Training Accuracy (top to bottom) against number of epochs

266 Conclusion:

267

	Human Observed	GSC Dataset
	Dataset	
Logisitic Regression	45.859	54.771
(Concatenation)		
Logisitc Regression	54.140	76.777
(Subtraction)		
Neural Network	57.961	93.358
(Concatenation)		
Neural Network	87.898	82.502
(Subtraction)		

268

Table 1: Testing Accuracy values for different models with different configuration

As seen from the above data,

- Neural Network performed best on GSC Dataset with feature concatenation.
- Neural Network performed best on Human Observed data with feature subtraction.

274		
275	• For Human Observed Dataset, Feature Subtraction setting is better than	
276	concatenation setting.	
277		
278	• For GSC Dataset, Logistic Regression performed better in subtraction	
279	setting as compared to concatenation setting. Whereas neural network	
280	performed better for Concatenation setting as compared to subtraction	
281	setting.	
282		
283	• When compared to Logistic regression, Neural network performed better	
284	on both the datasets in all the settings whether it is concatenation or	
285	subtraction.	
286		
287	References	
200	[1] Tawarda Data Sajanaa (2018) Maghina Lanming Tawarda Data Sajanaa [anlina] Availahla a	

 [1] Towards Data Science. (2018). Machine Learning – Towards Data Science. [online] Available at: https://towardsdatascience.com/machine-learning/home

 [2] Brownlee, J. (2018). Machine Learning Mastery. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/