CSE 574 PROJECT 3

Vishnu Varshath Harishankar
Person Number 50291399
ubit vharisha

1 Introduction

This project asks to classify mnist and usps data numerical images into digits 0,1,2,3,..,9 by training different machine learning models on mnist data set. The results of the individual classifiers are then combined to make the final decision and then check if the combined result is better than the individual classifiers. For the task of combining the result simple majority voting is used.

2 Data Partition and preprocessing

MNIST data has 28 X 28 gray scale images. each pixel is considered as a feature. Therefore we get a total of 784 features for each image. We split the data into training , validation and testing. USPS images are not 28 X 28 therefore we reshape them so that they can be used with the model trained with 784 features. The entire USPS dataset will be tested on the model trained with MNIST data.

3 Logistic Regression

For our multi class classification problem , logistic regression with softmax function is used with cross entropy as loss function and mini batch stochastic gradient descent as optimizer. Softmax gives the probabilities that the data will belong to ten of the classes. The one with the maximum probability is chosen as the obtained target. A batch size of 1000 is chosen and the weights are updated after every batch with the gradient calculated. The learning rate which is used to update the weights is taken as 0.01. A total of 100 epochs was run to obtain the desired accuracy. The accuracy for MNIST and USPS data is shown below
The final accuracy observed for MNIST was 0.9 and for USPS was 0.3498165237

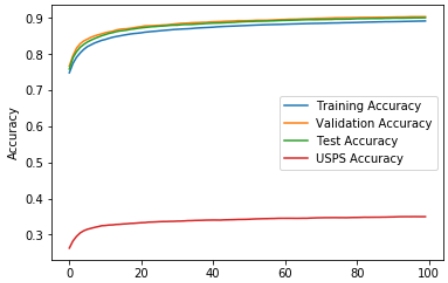

Figure 1: Accuracy including USPS

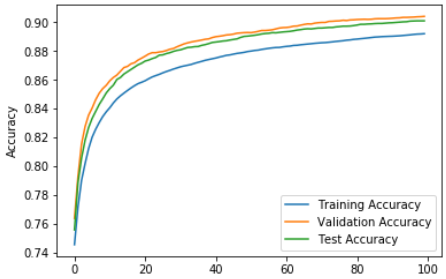

Figure 2: Accuracy of mnist alone

It can be seen from figure 1 that the accuracy of MNIST data is way higher than USPS data set. It is because the model is trained to look at data from MNIST dataset. Therefore when tested on USPS data the accuracy is low supporting the no free lunch theorem.

3.1 Confusion matrix

The confusion matrix showing how each data of MNIST and USPS was classified is shown below. It can be seen that MNIST data has high true positive(indicated by the diagonal elements) where as for USPS data has low true positives thereby giving a less accuracy.

	0	1	2	3	4	5	6	7	8	9
0	956	0	3	3	0	2	8	1	7	0
1	0	1102	2	4	1	2	4	0	20	0
2	11	7	888	18	15	0	17	21	45	10
3	6	1	18	896	1	33	6	15	22	12
4	2	6	5	0	901	0	10	2	8	48
5	15	6	5	42	15	728	17	10	43	11
6	16	3	6	2	13	16	897	1	4	0
7	3	20	29	4	11	0	0	918	4	39
8	9	10	10	30	8	26	13	13	838	17
9	13	8	6	11	41	16	0	25	6	883

Figure 3: MNIST

	0	1	2	3	4	5	6	7	8	9
0	602	4	375	56	255	112	103	42	145	306
1	234	298	130	350	286	52	41	299	292	18
2	219	25	1176	138	67	75	93	93	90	23
3	108	3	121	1261	21	236	31	58	101	60
4	65	86	35	62	1030	121	40	127	290	144
5	182	20	214	184	45	1032	126	71	89	37
6	382	13	346	106	105	218	696	25	75	34
7	198	214	317	451	74	78	35	300	286	47
8	226	31	146	208	129	573	118	43	441	85
9	53	188	165	469	155	85	15	365	336	169

Figure 4: USPS

4 Neural network

The following subsections discuss DNN and CNN and how the models perform with MNIST and USPS data

4.1 DNN

Dense neural network is modeled with the parameters given the table below. The model was trained with MNIST training data and validated on MNIST. The trained model was then tested on MNIST testing data and USPS data. The accuracy is show below. It was seen that the model performed better while testing on MNIST dataset but performed poorly on USPS data set. DNN performed better than logistic regression in both the data set

Parameter	Value
No of layers	Input layer, hidden layer, output layer
No of Hidden layer	1
Input neurons	784 because we have 784 features
output neurons	10 because this is a ten class classifier problem
Hidden layer neurons	100
Activation Function at hidden layer	sigmoid
Activation Function at output layer	Softmax
Optimiser	Adam

Table 1: Parameters for DNN

MNIST Testing :
Testing Accuracy: 0.949

USPS testing :
Testing Accuracy: 0.573

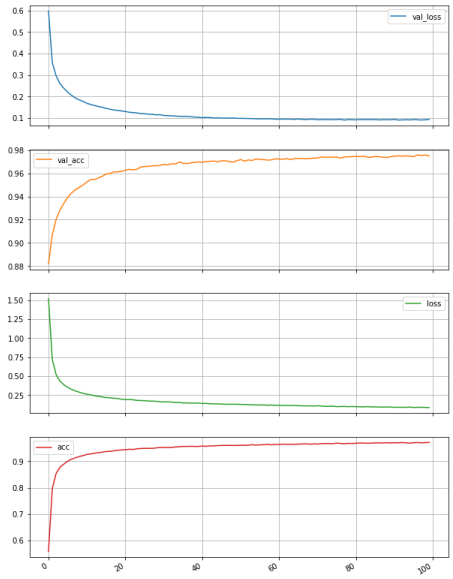

Figure 5: DNN

4.2 Confusion matrix

The confusion matrix for MNIST and USPS is given below. It can be seen that MNIST has majority of the values in the major diagonal therefore the accuracy is more but that is not the case for USPS and the data is scattered with a lot of true negatives and false positives telling that prediction is not accurate and hence giving less accuracy

	0	1	2	3	4	5	6	7	8	9
0	969	0	1	0	0	3	5	3	5	4
1	0	1123	0	0	0	1	3	5	2	4
2	2	4	1013	8	5	2	1	14	1	0
3	1	1	2	986	0	12	1	6	9	14
4	0	0	1	0	956	1	3	1	5	13
5	2	0	0	4	0	856	4	0	4	2
6	4	2	2	0	6	6	937	0	2	0
7	1	1	7	8	2	2	0	993	5	5
8	1	4	6	2	2	7	4	1	939	0
9	0	0	0	2	11	2	0	5	2	967

Figure 6: MNIST

	0	1	2	3	4	5	6	7	8	9
0	318	26	21	10	6	6	67	12	73	1
1	1	229	12	5	2	0	6	21	1	5
2	280	514	1503	323	153	384	554	143	271	133
3	318	119	82	1119	101	194	66	699	600	579
4	54	233	13	1	845	1	38	11	41	88
5	407	215	208	482	182	1319	293	94	447	53
6	94	17	80	16	34	38	757	18	78	9
7	274	501	47	22	522	37	99	879	165	807
8	21	106	29	12	97	17	15	108	295	160
9	233	40	4	10	58	4	105	15	29	165

Figure 7: USPS

4.3 Convolutional Neural Network

With the parameters given below in the table the CNN model was trained and tested on MNIST data and tested on USPS data.

Parameter	Value
No of layers	Input layer, hidden layer, output layer
No of Hidden layer	1
Input neurons	64
Hidden layer neurons	32
Activation Function at hidden layer	relu
Activation Function at output layer	Softmax
Optimiser	adam
Loss	categorical cross entropy

Table 2: Parameters for CNN

The parameters were tuned to obtain highest accuracy by

1.Tuning the kernel size
2.Altering the number of hidden layers

In either of the case the MNIST data performed well better than USPS data same as observed in logistic regression and dense neural network

1. Tuning the kernel size: The kernel size was changed from 2 to 5 and it can be seen from figure 8 that as the kernel size increased the accuracy increased in both MNIST and USPS dataset
2. Altering the number of hidden layers: The number of hidden layers was changed from 2 to 7 and it can be seen from figure 9 that increasing the number of hidden layers increased the performance initially but then decreased. This can be related to overfitting problem

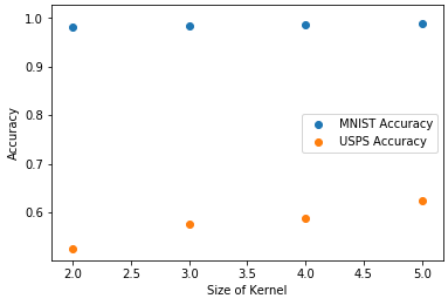

Figure 8: Accuracy vs kernel size

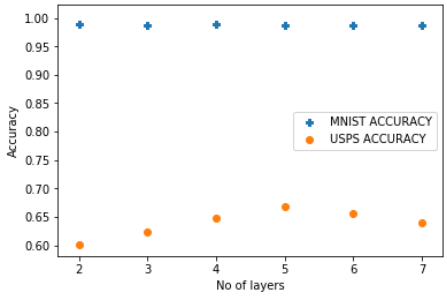

Figure 9: Accuracy VS layers

4.4 Confusion matrix

The confusion matrix is shown below and it can be seen as in other models that MNIST has higher true positives as majority of the predictions is along the major diagonal unlike USPS data

	0	1	2	3	4	5	6	7	8	9
0	974	0	2	0	2	1	6	1	2	3
1	2	1131	4	0	0	0	3	3	0	3
2	0	2	1009	4	0	0	0	6	1	0
3	1	0	0	996	0	8	0	2	1	0
4	0	0	2	0	973	0	2	1	1	8
5	2	0	0	3	0	873	2	1	0	6
6	0	1	2	0	0	1	939	0	1	0
7	1	0	5	1	0	0	0	1004	2	2
8	0	1	8	6	1	7	6	1	965	8
9	0	0	0	0	6	2	0	9	1	979

Figure 10: MNIST

	0	1	2	3	4	5	6	7	8	9
0	972	92	44	4	10	4	309	17	30	7
1	6	764	3	2	59	1	35	94	3	83
2	127	252	1682	90	19	50	149	463	62	205
3	17	41	54	1555	1	64	11	391	229	155
4	270	603	24	8	1507	5	55	60	53	121
5	64	15	89	313	62	1754	80	40	480	49
6	17	17	4	0	3	3	1177	9	8	2
7	6	188	19	4	104	8	2	819	39	508
8	85	19	80	24	224	109	176	106	1072	542
9	436	9	0	0	11	2	6	1	24	328

Figure 11: USPS

5 Support vector machine

The svm model was trained for mnist data with the following parameters tuned

5.1 Gamma

Gamma is the parameter that specifies how well the model needs to fit the training data. Gamma was changed to 0.01,0.05,0.09,1. Gamma was also set to the default value. It can be seen that as the gamma was increased the accuracy decreased this is because the training data was over fit. Thus when gamma is 1 data was over fit highly resulting in very low accuracy

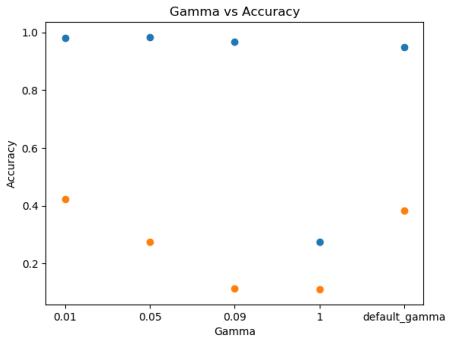

Figure 12: svc gamma

5.2 Linear vs Radial Basis Function kernel

The linear model trained faster than RBF but the accuracy was lesser than that of RBF for both the MNIST data ans USPS data. This can be seen in the graph. Therefore choice of linear vs RBF needs to be based on the dataset and for this set RBF seems to work better with the cost of more time

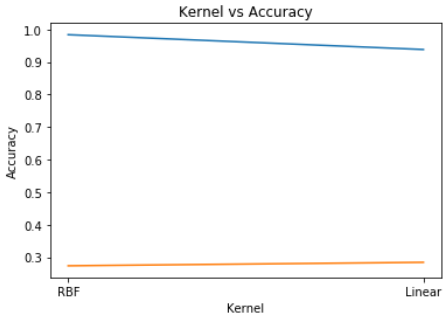

Figure 13: svc linear

5.3 Confusion Matrix

The confusion matrix for MNIST and USPS dataset is shown below The number of samples which are predicted correctly was higher for MNIST Dataset than USPS data set.

	0	1	2	3	4	5	6	7	8	9
0	974	0	5	0	0	2	3	0	3	2
1	0	1128	0	0	0	0	2	4	0	2
2	1	3	1014	2	2	0	0	6	2	0
3	0	1	0	997	0	5	0	1	3	6
4	0	0	1	1	967	1	2	0	1	9
5	2	1	0	2	0	877	3	0	2	1
6	0	0	1	0	4	3	947	0	2	1
7	1	1	7	3	0	0	0	1008	2	5
8	2	1	4	3	1	2	1	1	954	1
9	0	0	0	2	8	2	0	8	5	982

Figure 14: MNIST

	0	1	2	3	4	5	6	7	8	9
0	650	82	51	19	13	53	155	48	59	11
1	1	441	1	1	11	0	0	82	0	15
2	531	352	1734	185	181	217	647	733	359	424
3	23	107	29	1308	22	23	12	236	210	311
4	139	147	10	0	1062	2	34	21	26	138
5	298	134	126	465	313	1683	301	278	1044	151
6	22	31	14	0	5	6	832	2	22	1
7	22	670	27	16	283	8	2	583	15	588
8	3	23	5	2	67	6	4	10	259	168
9	311	13	2	4	43	2	13	7	6	193

Figure 15: USPS

6 Random forest

Random forest model was trained with different number of trees and the results are shown in the graph and it can be seen that as the number of trees increased the accuracy was better for both the mnist and usps dataset. However after some limit the increase in accuracy was very small

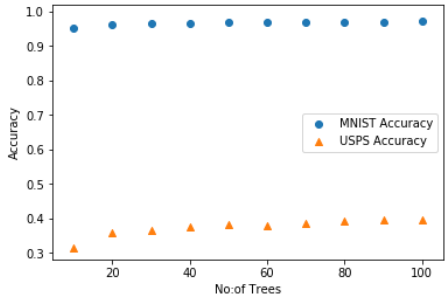

Figure 16: Random forest

6.1 Confusion Matrix

The confusion matrix show the same trend as seen in other models. The MNIST data has higher true positives with majority of the values in the main diagonal in contrast to USPS data

	0	1	2	3	4	5	6	7	8	9
0	969	0	6	0	2	5	8	1	5	5
1	0	1122	0	0	0	0	3	3	0	5
2	2	2	996	10	2	1	1	22	3	2
3	0	3	7	970	0	13	0	3	9	9
4	0	0	2	0	943	5	4	3	5	12
5	1	1	0	9	0	857	7	0	4	2
6	3	3	5	0	7	4	932	0	2	1
7	1	1	8	12	2	1	0	987	5	3
8	3	2	7	8	4	4	3	2	931	6
9	1	1	1	1	22	2	0	7	10	964

Figure 17: MNIST

	0	1	2	3	4	5	6	7	8	9
0	634	35	110	40	7	131	294	35	60	22
1	8	566	45	12	183	33	49	333	59	249
2	296	125	1184	146	77	182	248	406	211	291
3	50	117	109	1189	44	157	51	268	236	332
4	475	79	64	61	1046	39	127	45	126	214
5	159	106	189	349	173	1268	360	203	942	116
6	62	30	31	4	23	39	739	42	87	19
7	116	925	256	171	378	125	110	645	96	547
8	3	12	3	6	44	13	10	7	157	90
9	197	5	8	22	25	13	12	16	26	120

Figure 18: USPS

7 Majority voting

On doing simple majority voting following accuracy was obtained for MNIST and USPS data
Accuracy of MNIST : 0.9806
Accuracy of USPS :0.4577

8 Inference

1. No free lunch theorem : This theorem says that the model trained for one dataset will function properly for that data set only and will fail for others. From all the models that's been trained it can be concluded that they are in support of free lance theorem as the model trained with MNIST gave high accuracy for MNIST and low accuracy for USPS
2. Overall Performance : The accuracy for each of the model is given below

Method	mnist	usps
LR	0.90	0.349
DNN	0.94	0.573
CNN	0.9839	0.552
SVM	0.9835	0.431
RF	0.9576	0.376

   From the table and observing all the confusion matrix it can be seen that CNN has the overall highest accuracy
3. Majority voting : Majority Voting accuracy is better than logistic model , DNN and RF

9 Conclusion

Thus different machine learning techniques were applied to classify two datasets and ensemble of their results is calculated to get combined result. It is observed that the combined result performed better than few machine learning models