CSE 474/574: Introduction to Machine Learning

University at Buffalo

Keming Kuang (kemingku)

personal #: 50161776

November 21, 2018

Contents

1 Description

1.1 Objective

In this project, we will learn to implement an ensemble of four classifiers for a given task. The classification task contains 28*28 grayscale handwritten digit image with 10 labels: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9.

1.2 Task

With given MNIST data and USPS data, we will partition the data into dataset and target value. By applying Logistic Regression, Neural Network, SVM, and Random Forest, we will train the model with given data and compare the accuracy. During the project, we will answer the following questions.

Q1: We test the MNIST trained models on two different test sets: the test set from MNIST and a test set from the USPS data set. Do your results support the ?No Free Lunch? theorem?

Q2: Observe the confusion matrix of each classifier and describe the relative strengths/weaknesses of each classifier. Which classifier has the overall best performance?

Q3: Combine the results of the individual classifiers using a classifier combination method such as majority voting. Is the overall combined performance better than that of any individual classifier?

1.3 Plan of Work

Step 1: Data Process Processing data from MNIST and USPS into Raw Data and Target Data. Data Processed from MNIST will be partitioned into training data, validation data, testing data.

Step 2: Softmax Regression By applying Muti-class Logistic Regression, we will use softmax to generalize multi-class classification. Since it's a logistic regression, we use gradient decent to learn the softmax model. I will also study the hyper-parameter and their influence to performance.

Step 3: Neural Network Neural NetWork are a a classifier of models that inspired by biological neural network. It has great capable of modeling nonlinear relationships between dataset in parallel. In this project, I will use multilayer deep neural network and study the performance with relation to hidden layers, hidden nodes, learning rate, and activation function.

Step 4: SVM Support Vector Machine is one of the high-performance in machine learning algorithms. With its capable of Maximal Margin Classifier, we can calculate the distance from the margin to the closest data and find the best optimal line that separates classes. In this project, I will study the performance with relation to linear kernel, radial basis function and gamma.

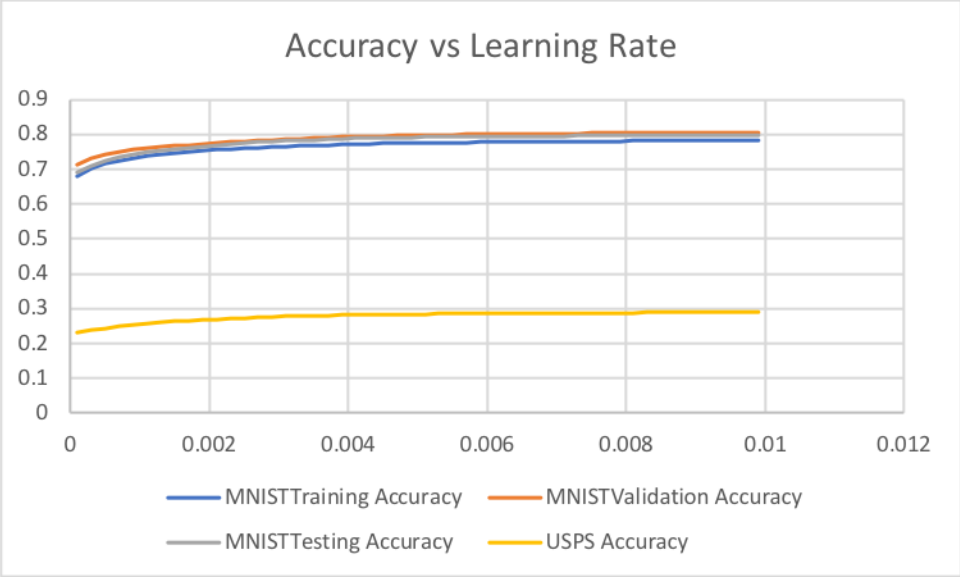
Step 5: Random Forest As a supervised machine learning process, Random Forest builds an ensemble of Decision Trees. Since it builds multiple decision trees, we will mainly focus on the effect of the number of trees and the accuracy.

Step 6: Compare Data Set After training both MNIST test set and USPS data, I will discuss the difference and similarity between MNIST and USPS data set.

2 Logistic Regression

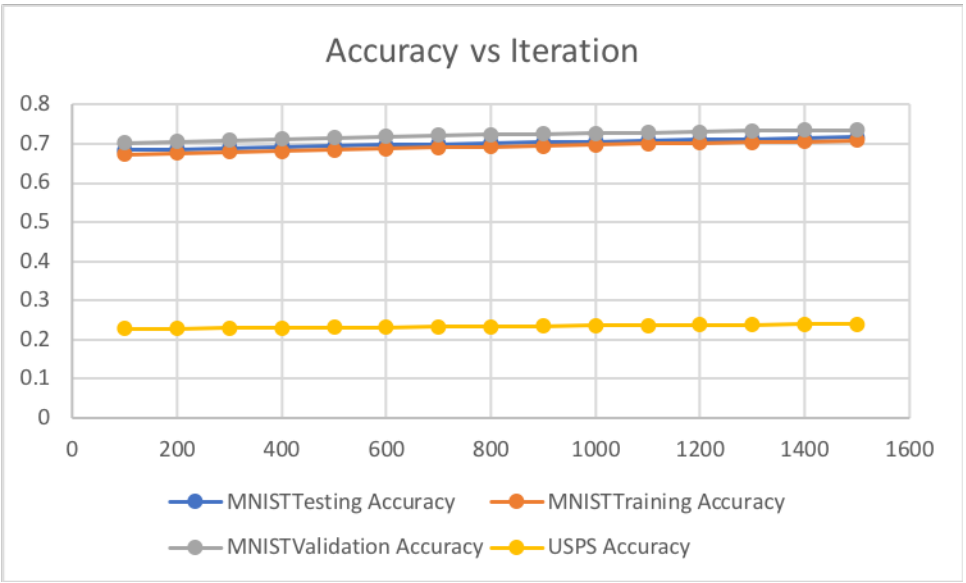
Different from the binary classes logistic regression in our last lab, we will use softmax to learn the machine learning model to generalize multi-classes classification. I will also study the relation between hyper-parameter and performance.
By default, I will use
Learning Rate = 0.0001
Iteration = 400
Lambda = 1

2.1 Learning Rate



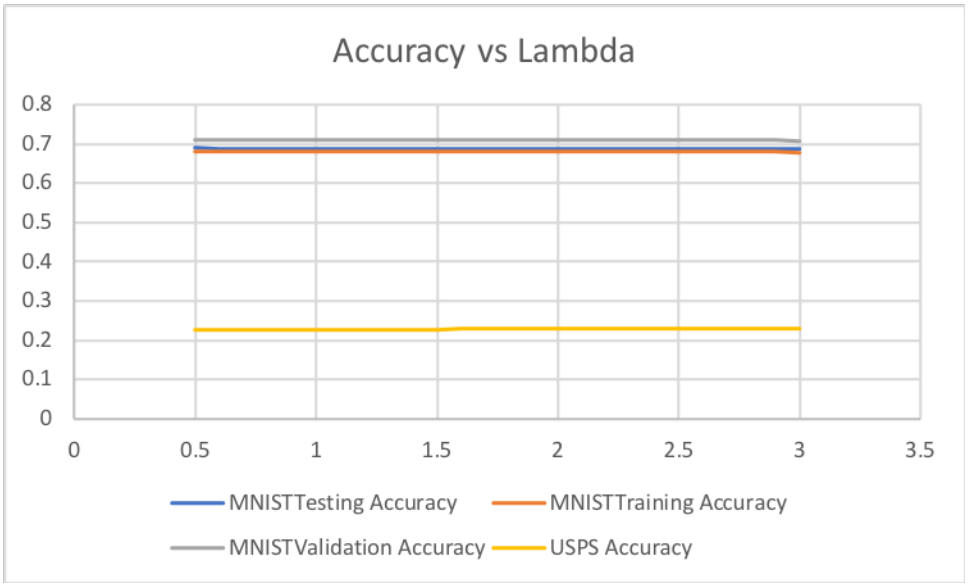
All sets of accuracy gradually increase from learning rate 0.001 to 0.004. However, MNIST training data, validation data, testing data tend to have similar accuracy, and USPS has the smallest accuracy. From above, changing in learning rate doesn't significant influence the accuracy in the learning rate area from 0.001 to 0.01.

2.2 Iteration



From the data we could see all sets of accuracy for MNIST training data, validation data, testing data, and USPS testing data stay almost the same. The four sets of accuracy all have slight altered during the increasing of Iteration with change less than 5 significant digits. Overall, MNIST training data, validation data, testing data have similar accuracy larger than USPS accuracy.

2.3 Lambda



From the data we could see all sets of accuracy for MNIST training data, validation data, testing data, and USPS testing data stay almost the same. The four sets of accuracy all have slight altered during the increasing of Lambda with change less than 5 significant digits. Overall, MNIST training data, validation data, testing data have similar accuracy larger than USPS accuracy.

3 Neural Network

In this section, I will train the gathered data set in a Neural Network classifier. The objective is finding the similarity between the digit samples training the model. We will also study the model accuracy by testing model parameters.
By default, I will use
Hidden Layers = 1
Hidden Nodes = 32
Learning Rate = 0.01
Activation Function = sigmoid

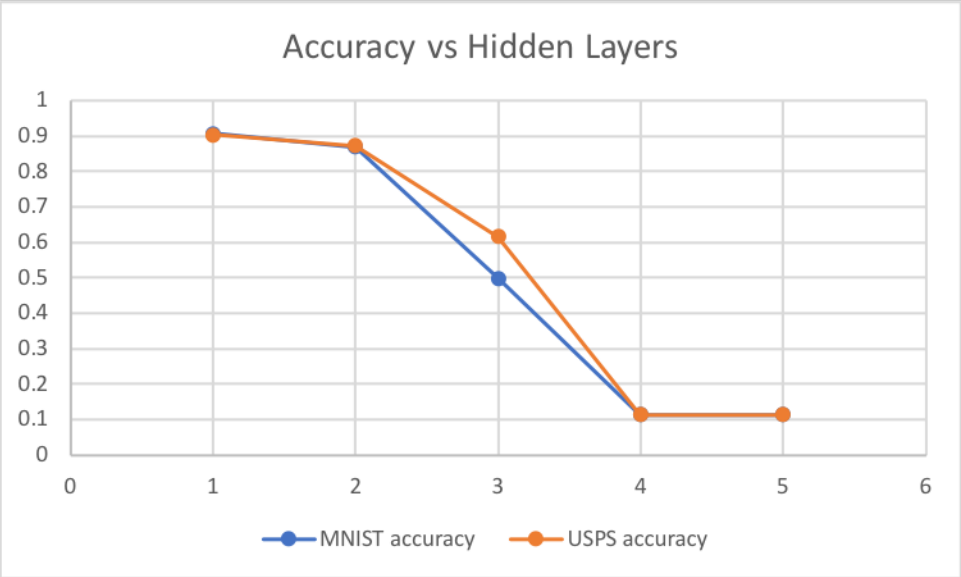
Hidden Layers Since Neural Network is inspired by biological neural network, hidden layer works similar to the neural layer that simulate all types of activity sending into human brain. A hidden layer work as a middle station that takes a set of weighted input and produce an output with process by activation function. It applies any function from the previous layer and transforms the input into usable elements for output layers.

Hidden Nodes While hidden layer doing its job to transform data, hidden nodes are constructed inside the hidden layers. To create non-linear decision boundaries for overall function, hidden nodes contain non-linear activation function, and increase the complexity of decision boundaries.

Learning Rate By Applying the learning rate in the Neural Network, we can control how much the co-effiecients changes or learns each time it is updated.

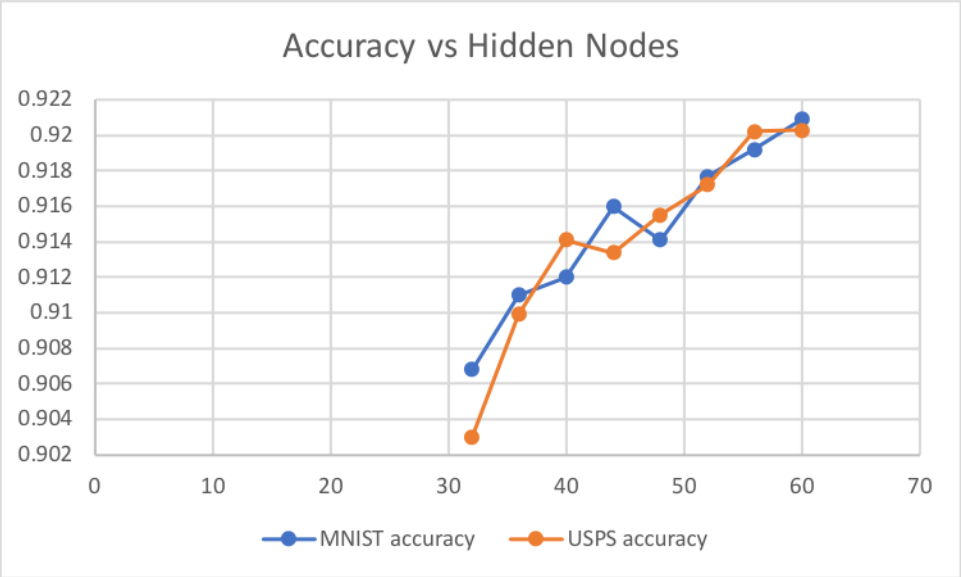
Activation Function Since we need to solve non-linear decision, we need to find the proper activation function for our model to fit different problem environment and setting the right hyper-parameter.

3.1 Hidden Layers



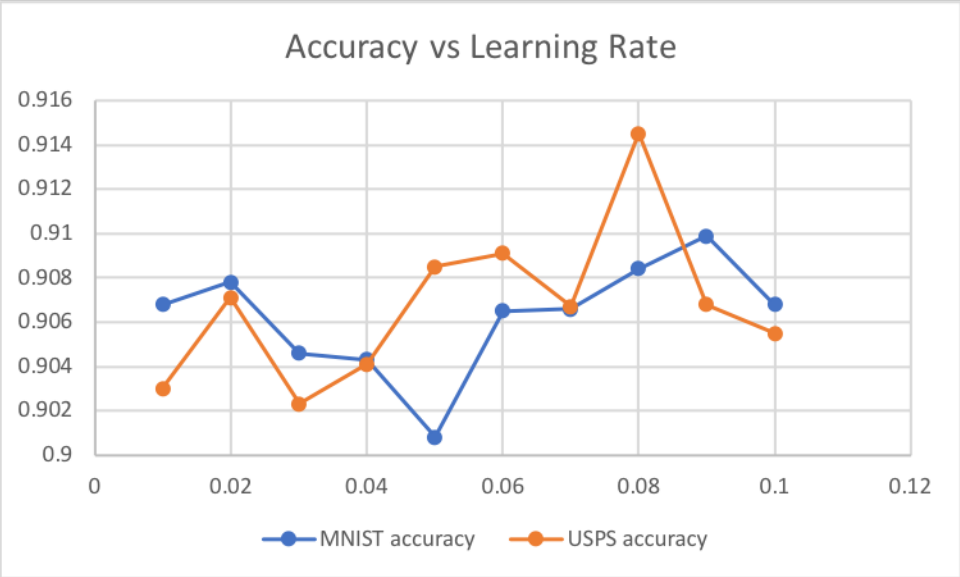
Both MNIST and USPS accuracy largely decrease in the similar trend from hidden layers 2 to 4. After hidden layer 4, both MNIST and USPS accuracy stay the same. This could be a result from largely increasing the complexity of decision boundaries from hidden layers.

3.2 Hidden Nodes



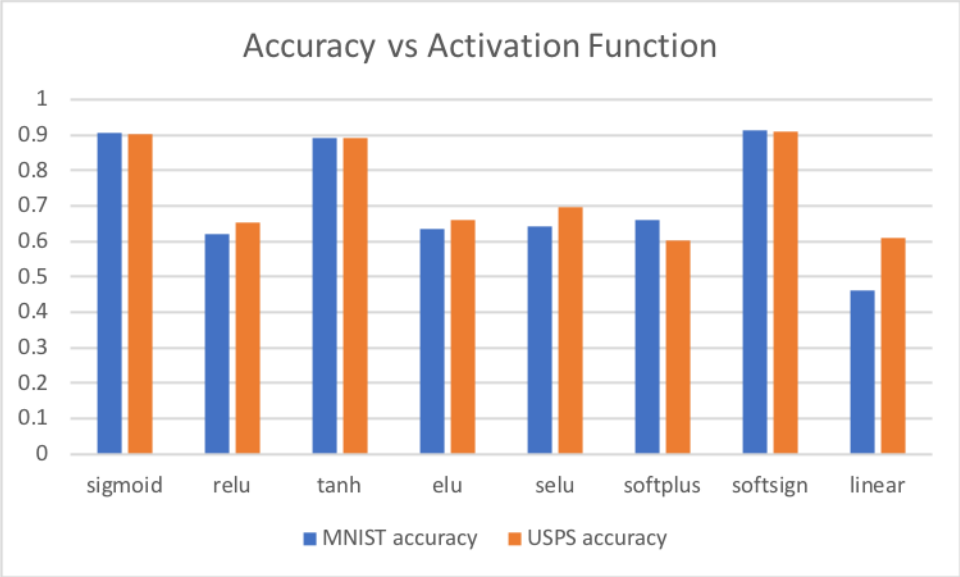
MNISt and USPS accuracy gradually increases with the growing of hidden nodes. Noticing that the increasing of accuracy doesn't stay constant as it slightly decrease in periods.

3.3 Learning rate



Both MNIST and USPS accuracy have slight change in a area of 0.902 to 0.914. USPS tends to have higher accuracy from learning rate 0.04 to 0.08. Both sets of accuracy show great unstable during the growing of learning rate.

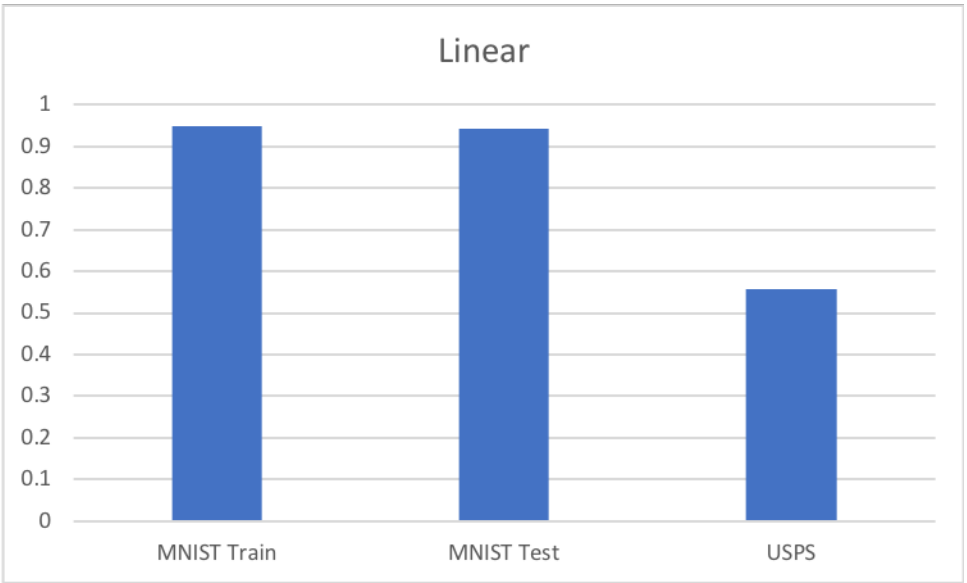
3.4 Activation



Overall, Sigmoid, Tanh, and Softsign have the highest accuracy for both MNIST and USPS. Both data set accuracy have similar performance on the activation function except linear has large difference.

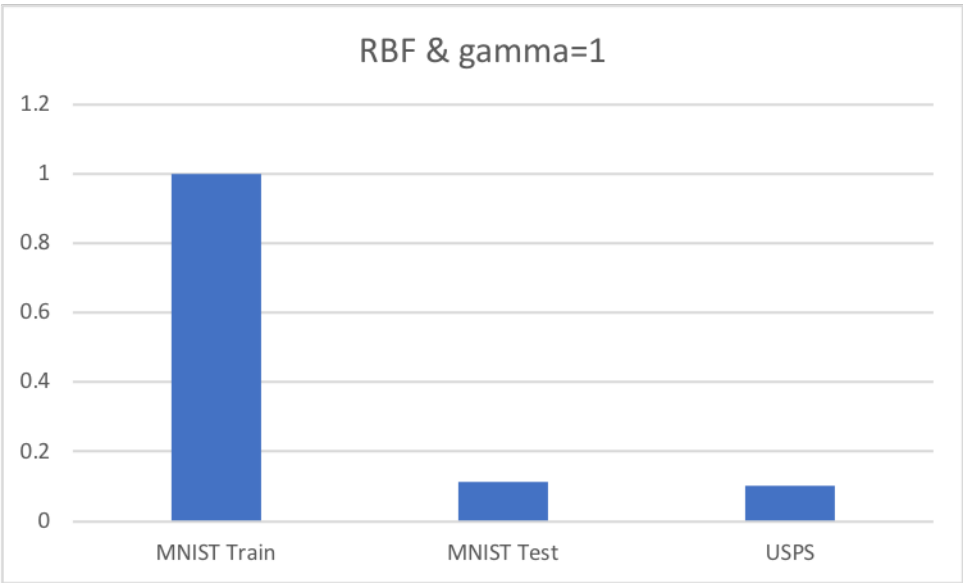
4 SVM

4.1 Linear



By applying linear kernel, MNIST testing data, USPS testing data tends to have the highest accuracy comparing to other data set tested. MNIST training data and testing data have larger accuracy than USPS.

4.2 RBF with gamma set to 1



By applying Radius Basic Function with value of gamma set to 1, MNIST training accuracy is 100 percent, and MNIST test accuracy and USPS accuracy is 0.1135113 and 0.100005. Both MNIST testing data and USPS have smallest value in accuracy comparing to other data set tested.

4.3 RBF

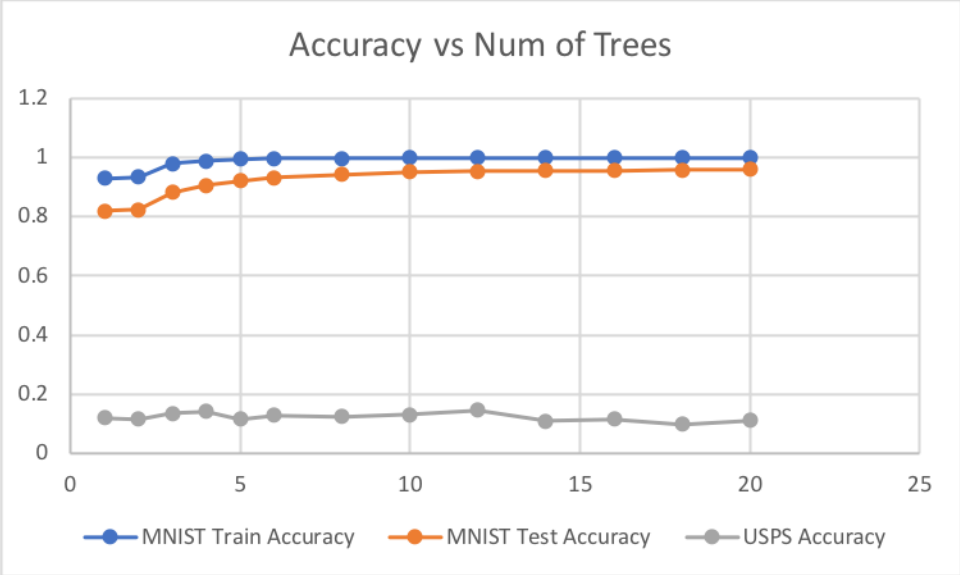
By applying Radius Basic Function with other value set to be default, MNIST training data and testing data have larger accuracy than USPS.

5 Random Forest

By applying Random Forest, it builds an ensemble of Decision Trees. The number of trees will be tested.

Number of Trees Random forest builds the decision trees by picking one sample of observation and picking one sample of features. If the number of tree is small, and the observation is too large, then some observations prediction may miss. if the number of tree is large, and the observation is to small, the predictive power decreases.

5.1 Number of Trees



MNIST training data and testing data have overall larger accuracy than USPS accuracy. MNIST accuracy sets also gradually increase from number of trees 1 to 5. While MNIST accuracy sets tend to have similar value, USPS accuracy have slightly change over the growing of number of trees, but its accuracy overall stays the roughly unchanged. The overall negative change on these sets of accuracy may result from low number of trees since some observation prediction may be missing.

6 Summary

As a result, we learn how to solve the task of classification with self implement Softmax Regression, Neural Network, SVM, Random Forest. By training the linear regression model with the right data set and setting proper hyper-parameters, we can reduce the loss and get higher accuracy. We also learn how these different classifier works on the same question but have different performance. By using different algorithm, we could process the problem in different way and thus get different result.

6.1 Question Answer

Q1: Yes, my test result supports "No Free Lunch" Theorem.
As a on-going studying topic, "No Free Lunch" Theorem illustrates that if one theorem has significant performance on one type of problem solving, then it doesn't support its universality. In our project 3, Random Forest and SVM have relatively much higher accuracy on MNIST data set comparing to Neural Network. Logistic Regression has significant accuracy on binary classes classification, but it has relatively lower accuracy on multi-classes classification. This experimental results claim that practical implication is true. The best choice is to choice the best algorithm based on the problems.

Q2: Overall, SVM has the best performance.
Even it has relatively low USPS accuracy, it has the highest MNIST accuracy and USPS accuracy among all classifiers.

7 Reference Website

https://blog.csdn.net/wn314/article/details/79972988
https://blog.csdn.net/wusecaiyun/article/details/49737499
https://blog.csdn.net/yzf0011/article/details/71521001
https://en.wikipedia.org/wiki/Radial_basis_function_kernel
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
http://rasbt.github.io/mlxtend/user_guide/data/mnist_data/
https://scikit-learn.org/stable/modules/model_evaluation.html
http://sklearn.apachecn.org/cn/0.19.0/datasets/mldata.html
https://keras.io/zh/activations/
https://keras.io/zh/layers/core/
https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
https://keras.io/optimizers/?fbclid=IwAR1w-XlK0wSbiWqoNpXkpts4sbvANEDCweGekTDH6dIEfx6z75UeuaiSEH0
https://www.kdnuggets.com/2016/07/softmax-regression-related-logistic-regression.html?fbclid=
IwAR07sE8fP1aO2EFVDDeqQ9qTCVAssdk1T6oHer-a5rDxICU5GP0OGTuMOgc
https://stats.stackexchange.com/questions/233658/softmax-vs-sigmoid-function-in-logistic-classifier
https://www.techopedia.com/definition/33264/hidden-layer-neural-networks