Mini-project report

Main task:

Apply three classification algorithms (logistic regression, neural network and soft-margin SVM) on a machine learning problem of predicting spam emails. The goal of this task is to make accuracy of the prediction as high as possible, the loss of each algorithm as low as possible.

Dataset background:

1. **Url:** http://archive.ics.uci.edu/ml/datasets/spambase
2. **Name:** UCI Spambase Data Set
3. **Abstract:** Classifying Email as Spam or non-Spam
4. **Brief introduction:** This is a pre-processed dataset and the raw data are already featurized. There are 4600 samples,1813 Spam (39.4%) and 2788 non-spam (60.6%), each has 57 features and 1 nominal class label. The detail of the 57 features is in the document of this data set (spambase/spambase.Documentation).

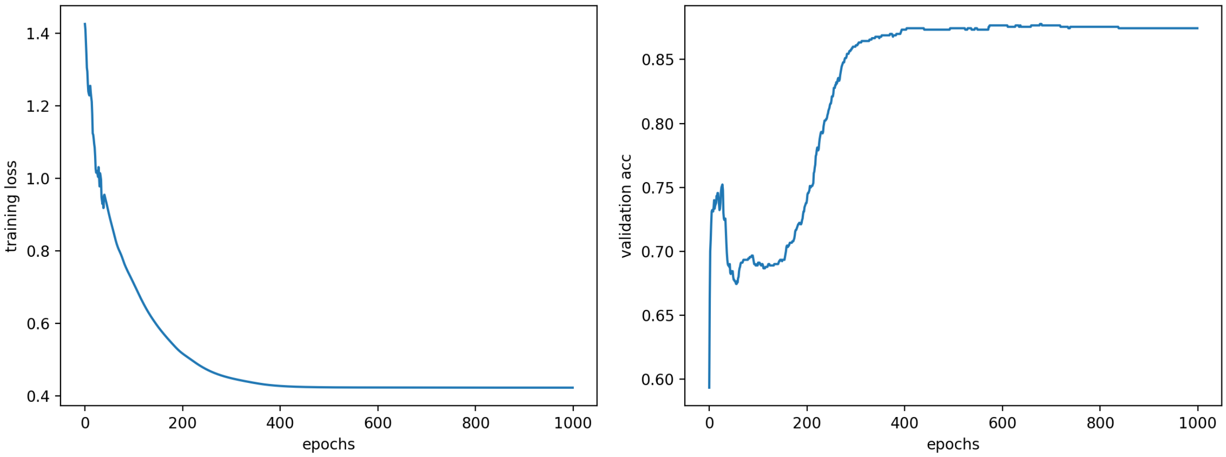
Data preparation:

Load the data set from spambase/spambase.data line by line and split each line to 57 features adding a trivial feature x0 = 1 as input X and 1 label as y. The raw label is 1(spam) or 0(non-spam). Shuffle X and y and split them into 2800 train samples, 900 validation samples and 900 test samples(nearly 3:1:1).

Logistic regression:

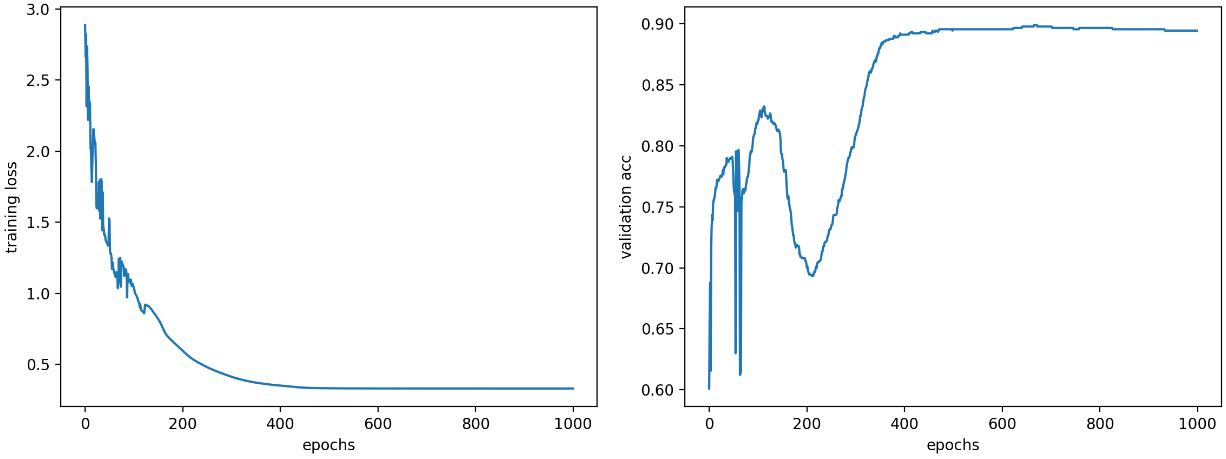
INITIAL HYPERPARAMETERS:
Alpha: 1e-4 with exponential decay with epoch (0.99^epoch * alpha)
batch_size: 10
MaxIter: 1000
L2 regularization decay: 1e-2

At epoch 677. We have best validation accuracy: 87.78% and test accuracy: 86.33%. And the curve for training loss and validation accuracy as shown below.



TUNE HYPERPARAMETERS ALPHA:

Search for best alpha from 1e-4 to 1e-3: Best alpha is 3e-4. And the curve for training loss and validation accuracy after tuning alpha is shown below.



From the plot above, we found that after slightly increase the value of alpha, in the beginning stage of the training(epoch 0-200), the performance fluctuates dramatically, but with decay of alpha, the performance increase and tend to converge to its best possible value. At epoch 665, it has best validation accuracy: 89.89%, and test accuracy: 87.89%.

Neural Network:

STRUCTURE:

Fully connected. 4 layers, one input layer, one hidden layers and one output layer.

HYPERPARAMETERS:

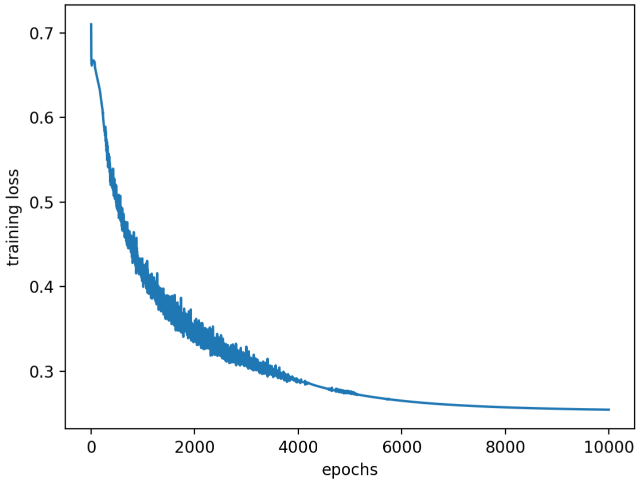
Alpha: 1e-3 with exponential decay with epoch ((0.996^epoch/10) * alpha)

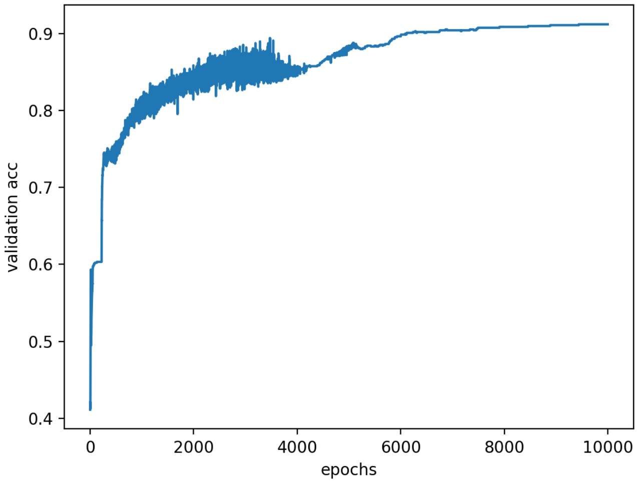
batch_size: 100

MaxIter: 10000

L2 regularization decay: 1e-2

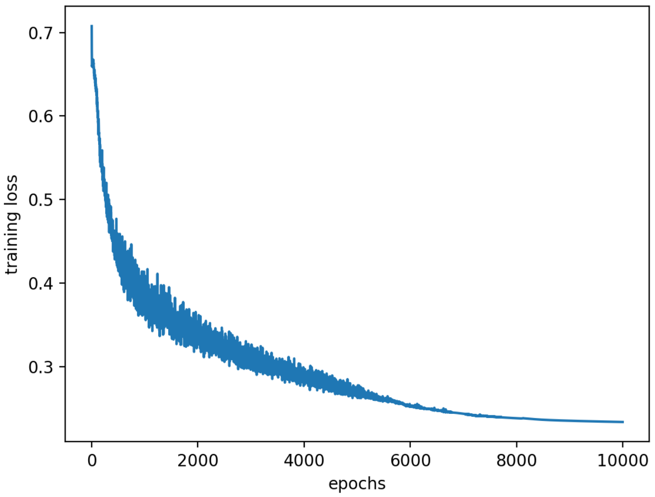
At epoch 9440. We have best validation accuracy: 91.22% and test accuracy: 90.00%. The performance is clearly better than using logistic regression. And the curve for training loss and validation accuracy as shown below.

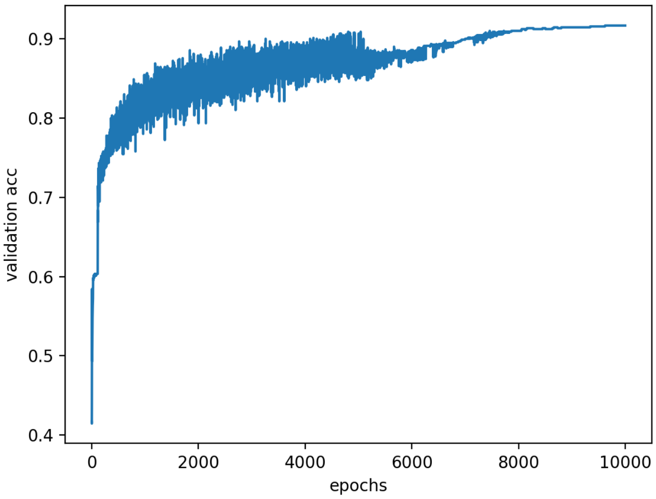




TUNE HYPERPARAMETERS ALPHA:

Search for best alpha from 1e-3 to 1e-2: Best alpha is 2e-3. And the curve for training loss and validation accuracy after tuning alpha is shown below.





Same as tuning alpha in logistic regression, as the value of alpha is slightly increased, there are more fluctuation in the early stage of the training (roughly from epoch 0 to 6000) compared to the initial training with alpha=1e-3. Then the performance of the training and validation accuracy tend to converge as the alpha decay with epoch. Since this is the 3 layers neural net structure, so it should be slightly hard for two weights go to its optimization, so the training epoch is larger. At epoch 9621, the validation accuracy: 91.67%, the test accuracy: 90.89%.

Soft-margin SVM

CONVERTING THE RAW LABELS Y:

0 to -1; 1 to 1

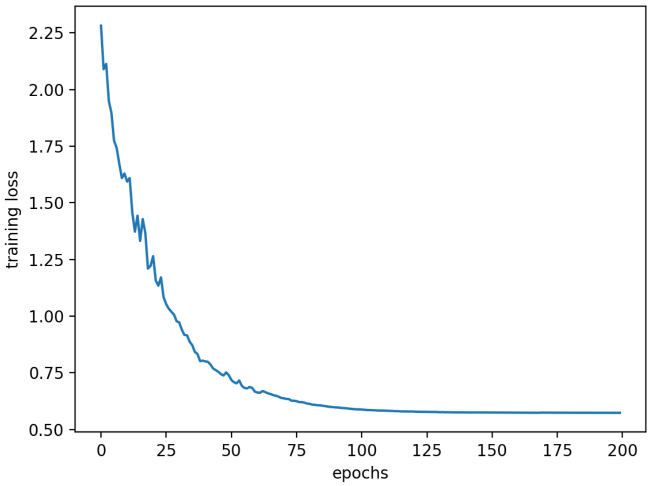
HYPERPARAMETERS:

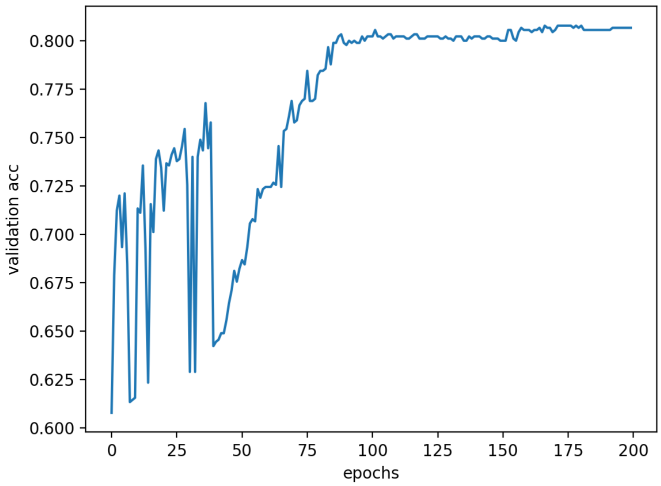
Alpha: 1e-4 with exponential decay with epoch (0.96^epoch * alpha)

batch_size: 10

MaxIter: 200

L2 regularization decay: 1e-2

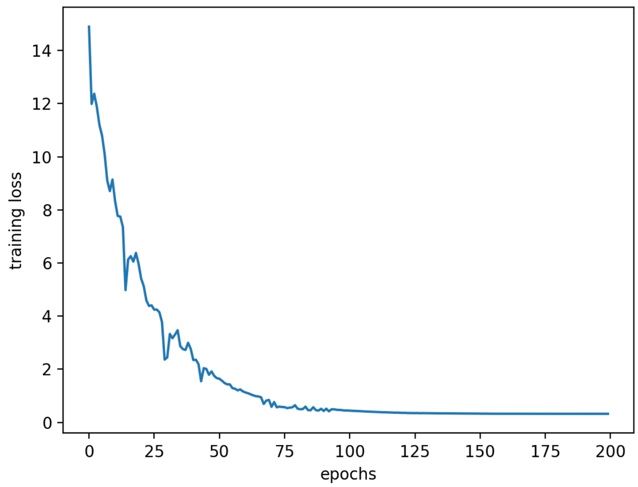


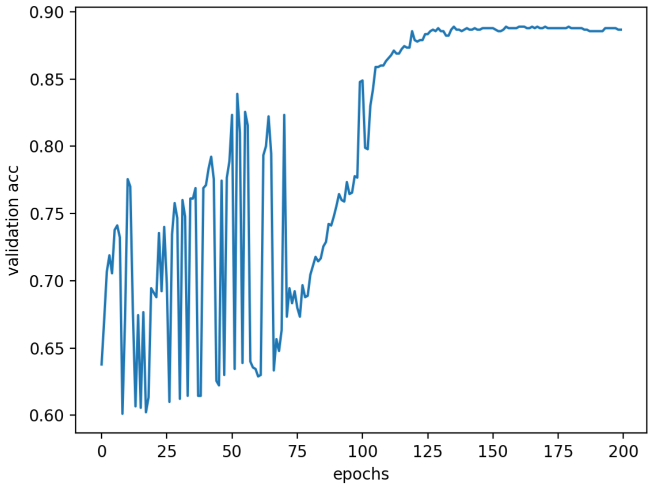


At epoch 166. We have best validation accuracy: 80.78% and test accuracy: 81.44%. The performance is clearly worse than using neural network even worse than using logistic regression. And the curve for training loss and validation accuracy as shown below.

TUNE HYPERPARAMETERS ALPHA:

Search for best alpha from 1e-4 to 1e-3: Best alpha is 8e-4. And the curve for training loss and validation accuracy after tuning alpha is shown below.





Same as tuning alpha in logistic regression and neural network, as the value of alpha is slightly increased, there are more fluctuation in the early stage of the training (roughly from epoch 0 to 75) compared to the initial training with alpha=1e-4. However the final performance increases dramatically. At epoch 135, best validation accuracy: 88.89% and test accuracy: 87.22%. Which means it can have the similar performance as logistic regression.

Conclusion:

For the machine learning of two classes prediction, the model with one linear layer such as logistic regression and SVM can have approximately the same performance, and a more complicated model which has more layers such as linear neural network can have better performance.