CSE474 Assignment3 Report

Yunhan Wang, Zhaoyang Pan, Yicheng Luo

Binary Logistic Regression(BLR):

Training set Accuracy: 92.772%
Training set Error: 7.228%

Validation set Accuracy: 91.36%
Validation set Error: 8.64

Testing set Accuracy: 91.97999999999999%
Testing set Error: 8.02%

The difference between training error and test error:
The main reason for the difference between training and testing errors is overfitting. The model may memorize the training set, resulting in high accuracy during training but poor generalization to unseen data. The model shows good performance on both training and test sets. Regularization techniques or adjusting model complexity can be considered to improve generalization ability.

Multi-class Logistic Regression(MLR):

Training set Accuracy: 93.448%
Training set Error: 6.552%

Validation set Accuracy: 92.47999999999999%
Validation set Error: 7.52%

Testing set Accuracy: 92.55%
Testing set Error: 7.45%

The difference between training error and test error:
The difference between training and testing errors can be attributed to overfitting. However, in this case, the small difference between training and test errors indicates relatively low overfitting. The small difference between training and test errors indicates that the model is performing well.

Multi-class strategy: This approach directly trains the model to classify instances into multiple classes. The reported accuracy value may represent the model's ability to handle multiple categories simultaneously.

One-vs-all strategy: In this strategy, a binary classifier is trained for each category, treating one category as positive and the remaining categories as negative. The final prediction is then made by combining the outputs of these binary classifiers.

Support Vector Machine(SVM):

1. **Using a linear kernel**

   Training data Accuracy: 97.29%

   Validation data Accuracy: 93.64%

   Testing data Accuracy: 93.78%

   The linear kernel gives almost the same results as our linear model.

2. **Using radial basis function with value of gamma setting to 1**

   Training data Accuracy: 100.00%

   Validation data Accuracy: 15.48%

   Testing data Accuracy: 17.14%

When gamma is 1, the accuracy of the test data and the accuracy of the validation data are very low.
But the training data is almost 100%, so higher gamma is helpful for the final result.

3. **Using radial basis function with value of gamma setting to default**

   Training data Accuracy: 98.98%

   Validation data Accuracy: 97.89%

   Testing data Accuracy: 97.87%

   So the higher the gamma, the higher the accuracy of the result.

4. **Using radial basis function with value of gamma setting to default and varying value of C (1, 10, 20, 30, · · · , 100) and plot the graph of accuracy with respect to values of C in the report**

   radial basis function with value of gamma setting to default and varying value of C

   when c is 1.0

   Training data Accuracy: 98.982%

   Validation data Accuracy: 97.89%

   Testing data Accuracy: 97.87%

   when c is 10.0

   Training data Accuracy: 99.988%

   Validation data Accuracy: 98.45%

Testing data Accuracy: 98.34%

when c is 20.0

Training data Accuracy: 100.0%

Validation data Accuracy: 98.44000000000001%

Testing data Accuracy: 98.31%

when c is 30.0

Training data Accuracy: 100.0%

Validation data Accuracy: 98.44000000000001%

Testing data Accuracy: 98.31%

when c is 40.0

Training data Accuracy: 100.0%

Validation data Accuracy: 98.44000000000001%

Testing data Accuracy: 98.31%

when c is 50.0

Training data Accuracy: 100.0%

Validation data Accuracy: 98.44000000000001%

Testing data Accuracy: 98.31%

when c is 60.0

Training data Accuracy: 100.0%

Validation data Accuracy: 98.44000000000001%

Testing data Accuracy: 98.31%

when c is 70.0

Training data Accuracy: 100.0%

Validation data Accuracy: 98.44000000000001%

Testing data Accuracy: 98.31%

when c is 80.0

Training data Accuracy: 100.0%

Validation data Accuracy: 98.44000000000001%

Testing data Accuracy: 98.31%

when c is 90.0

Training data Accuracy: 100.0%

Validation data Accuracy: 98.44000000000001%

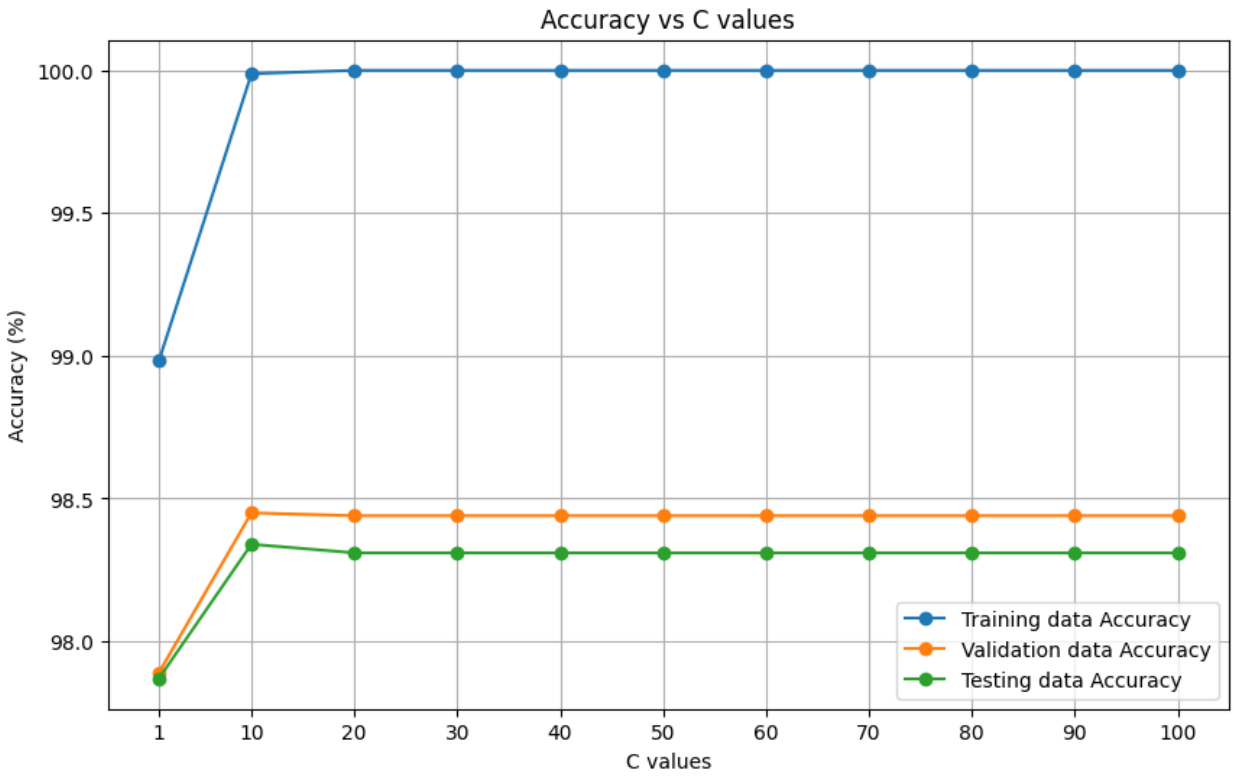
Testing data Accuracy: 98.31%

when c is 100.0

Training data Accuracy: 100.0%

Validation data Accuracy: 98.44000000000001%

Testing data Accuracy: 98.31%



We set the C value from 1 to 100, increased by 10, and used all the C values to train the SVM. As we can see the above accuracies do not change after C reaches 20.

To be more specific, the above figure shows that when C is 10, the Training data Accuracy, Validation data Accuracy, and Testing data Accuracy is the highest, which is Training data Accuracy: 99.988%, Validation data Accuracy: 98.45%, Testing data Accuracy: 98.34%.Therefore, we can say that when C is 10 and kernel is ‘rbf’, the performance of SVM is the best.