CSE 574 Introduction to Machine Learning
Programming Assignment 3

Team 66:
Weiyi Jiang 50207995
Shih-Chia Chen 50207079

1. Results of Logistic Regression Classification

Accuracy of Logistic Regression Classification

Training Set	Validation Set	Testing Set
84.87%	83.74%	84.20%

The Logistic Regression Classification considers all data to find a decision boundary which generates the smallest error, so if the size of dataset is large, then its performance will be not very good. We can find out that compared with SVM, the accuracy of Logistic Regression Classification is lower since SVM learns the decision boundary with maximum margin.

2. Results of Direct Multi-class Logistic Regression

Accuracy of Direct Multi-class Logistic Regression

Training Set	Validation Set	Testing Set
93.10%	92.39%	92.54%

Compared with the accuracy of one-vs-all Logistic Regression Classification, we can find that the accuracy of Direct Multi-class is higher. That’s because multi-class classify treats ten classes as a whole in one error and error gradient function, instead of building ten independent classifiers. That will reduce the error rate of prediction.

3. Results of Support Vector Machines

Accuracy of SVM under different conditions

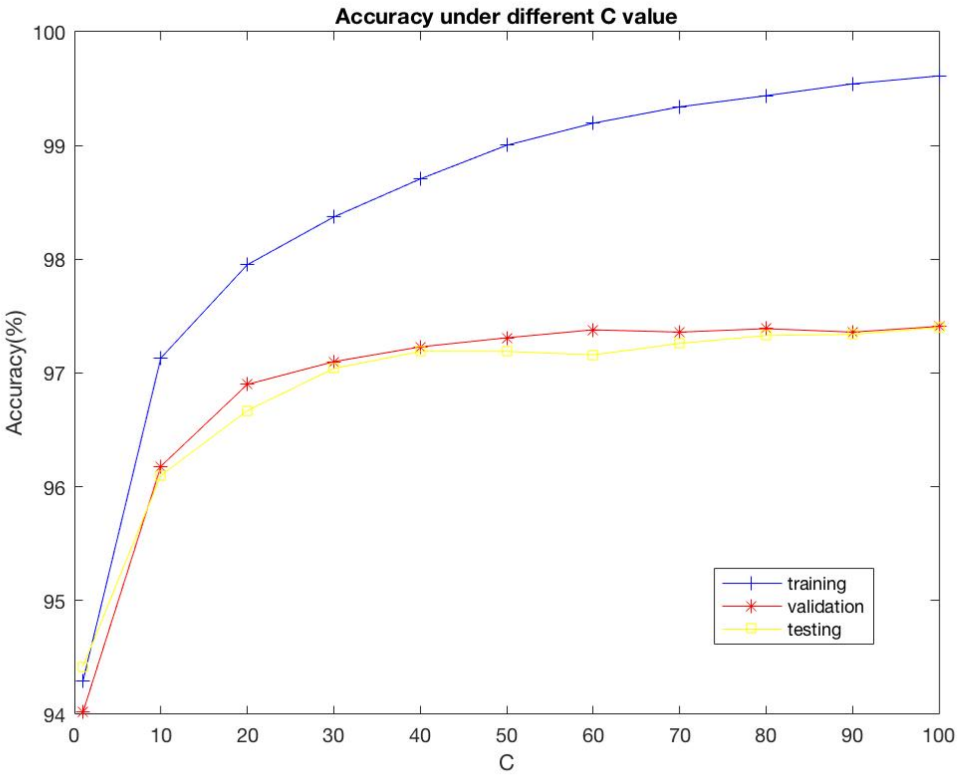
	Training Set	Validation Set	Testing Set
Linear Kernel	97.29%	93.64%	93.78%
Rbf, gamma = 1.0	100.0%	15.48%	17.14%
Rbf, gamma = 0.0	94.29%	94.02%	94.42%

From the table above we can observe that for SVM, using a radial basis function will produce higher prediction accuracy than linear kernel since the Gaussian kernel maps inputs to an infinite dimensional space and models the input examples better than the linear kernel. However, the training time of the SVM with Gaussian kernel is much longer than the SVM with linear kernel for this dataset.

In addition, we use two different Gamma values in rbf SVM training. Gamma describes the kernel coefficient for rbf. If Gamma is too high (gamma = 1.0), then a typical overfitting case will appear. We get 100% accuracy on the training set but very low accuracy on validation and testing set. When Gamma is 0(close to the default value 1 / n_features), the performance of the SVM is good.

Accuracy of SVM under different C

Rbf, gamma=default	Training Set	Validation Set	Testing Set
C = 1.0	94.29%	94.02%	94.42%
C = 10.0	97.13%	96.18%	96.10%
C = 20.0	97.95%	96.90%	96.67%
C = 30.0	98.37%	97.10%	97.04%
C = 40.0	98.71%	97.23%	97.19%
C = 50.0	99.00%	97.31%	97.19%
C = 60.0	99.20%	97.38%	97.16%
C = 70.0	99.34%	97.36%	97.26%
C = 80.0	99.44%	97.39%	97.33%
C = 90.0	99.54%	97.36%	97.34%
C = 100.0	99.61%	97.41%	97.40%



C is the penalty parameter of the error term in training example. If the C value is small, more errors will be accepted in the training phase and as the C value becomes larger, less errors will be accepted and less examples will be misclassified. As we can observe from the plot above, with C increasing, we can also obtain higher prediction accuracy. However, when C is high (> 70), we can find that the accuracy of the training data is close to 100%, so there may be a risk of overfitting. In practice, we should select appropriate C to increase the prediction accuracy but also avoid overfitting.