CSE 474 ASSIGNMENT 3

Group 14:
Andrew Frauens
Aakanksha Raika
Dominique Hickson

Contents

Logistic Regression:

In your report, record and discuss classification results and accuracy.

Our implementation of Logistic Regression was 84.92% with the training set, 83.74% accurate with the validation set, and 84.24% accurate with the testing set when it created the submitted pickle file.

Support Vector Machine:

In your report, provide justification for the above selection of hyperparameters, as well as plots and discussion of your results.

	Training Accuracy	**Validation Accuracy**	**Testing Accuracy**
gamma default	94.294	94.02	94.42
gamma = 1	100	15.48	17.14

Figure 1: Percent accuracy of hyperparameter gamma

As shown in figure 1, setting gamma equal to 1 leads to a huge amount of over fitting. The accuracy on the training data was 100, but the accuracy for both the validation and testing data was terrible. Since gamma being 1 makes the accuracy on the testing data so bad, it's evident that the default gamma is significantly better. This is not even accounting for how much longer it takes for gamma = 1 to finish its runtime.

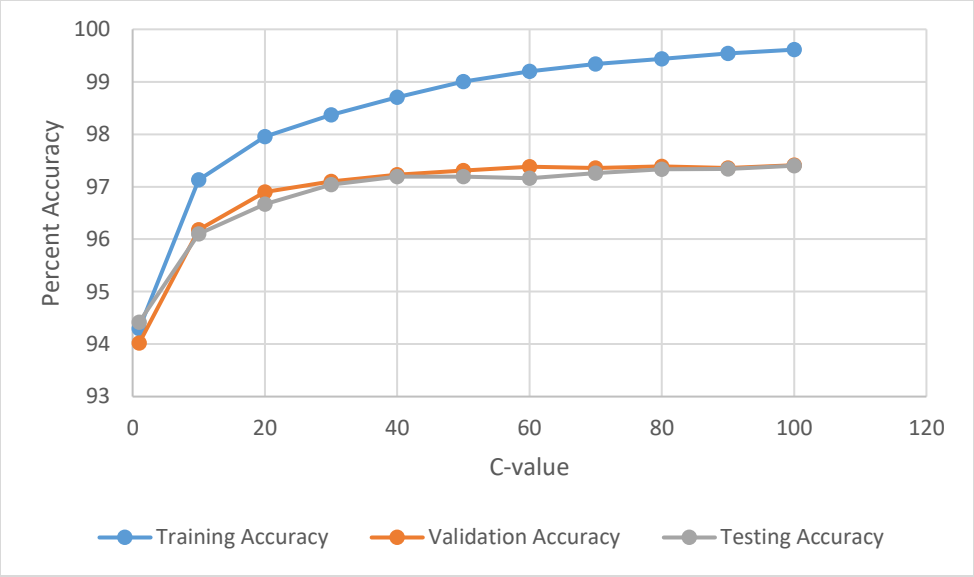

Figure 2: Accuracy of each C value

In Figure 2, the trend that the accuracy increases as the value of C increases for all of the subsets of data tested. The separation between the blue training data line and the other 2 lines shows that there is overfitting occurring which should be avoided, however since the overall accuracy remains high even for the validation and testing data, the evidence still suggests that the highest value should be chosen.

By looking at the data collected, especially the testing data's accuracies, the optimal hyperparameters are C = 100, and gamma = default. Looking at the accuracy compared to the testing data is important because it should theoretically be the most representative of the real world data since it was not used during training at all. The accuracy on the training data is primarily useful for noticing how much overfitting is occurring.

Discuss results, comparing the selections of linear kernel and radial basis function kernel.

	Training Accuracy	**Validation Accuracy**	**Testing Accuracy**
linear	97.286	93.64	93.78
Radial basis function	94.294	94.02	94.42

Figure 3 Linear basis function compared to radial

Since the default basis function is the radial basis function, the default and C =1.0 also have identical results. From the data in Figure 3, it's clear that the radial basis function is more accurate for both the validation and testing data accuracy. The fact that the linear method had a higher training accuracy serves to indicate that the method is more prone to overfitting.

Optimal Hyperparameter Selection

Basis function: radial C = 100 Gamma = default