CSE 574:Introduction to Machine Learning
Programming Assignment 3:Classification and Regression

Group Members (Group 26):
1. Nischala Manjunath
2. Reshma Raghavan
3. Sai Srinath Sundar

Logistic Regression

We use Logistic Regression to classify hand-written digit images into corresponding labels by building 10 binary-classifiers.
Entropy error which is a scalar value is calculated using the equation:

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}$$

The gradient of error function, error grad, a column vector of size (D + 1) × 1 is obtained by using the equation:

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\mathbf{x}_n$$

Results:

Dataset	Accuracy
Train	92.27%
Validation	91.52%
Test	91.87%

Support Vector Machines

In this we learn the SVM model and computed the accuracy of prediction with respect to training data, validation data and test data.

SVM Classifier using linear kernel:
All other parameters are kept default. We learn the SVM model which classifies the data points as belonging to each of the classes.

Dataset	Accuracy
Train	97.286%
Validation	93.64%
Test	93.78%

SVM Classifier using rbf kernel and gamma=1.0:
We use the radial basis function with value of gamma set to 1, all other parameters are default.

Dataset	Accuracy
Train	100%
Validation	15.48%
Test	17.14%

We observe that high values of gamma cause over-fitting thereby producing low accuracy for validation and test datasets due to low variance and the flexibility of the decision boundary. For each class we are learning multiple decision boundaries with respect to training data points.

SVM Classifier using rbf kernel and gamma set to default:
All other parameters are set to default.

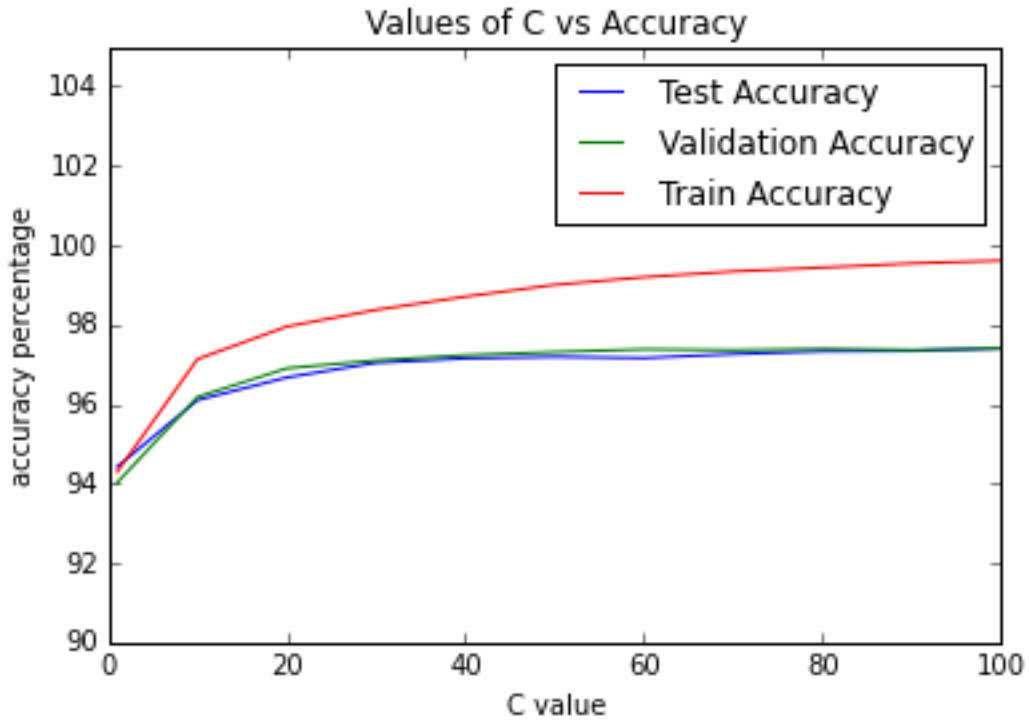
Dataset	Accuracy
Train	94.294%
Validation	94.02%
Test	94.42%

The prediction accuracies improve for the validation and test datasets by setting the value of gamma to default due to higher variance which causes smoother, less complex and non-linear decision boundaries than by setting gamma=1.

SVM Classifier using RBF kernel and gamma set to default and varying value of C:

Cost	Accuracy		
	Train	Validation	Test
1	94.294%	94.02%	94.42%
10	97.132%	96.18%	96.1%
20	97.952%	96.9%	96.67%
30	98.372%	97.1%	97.04%
40	98.706%	97.23%	97.19%
50	99.002%	97.31%	97.19%
60	99.196%	97.38%	97.16%
70	99.34%	97.36%	97.26%
80	99.438%	97.39%	97.33%
90	99.542%	97.36%	97.34%
100	99.612%	97.41%	97.4%

A plot of the accuracy with respect to varying C values with a value of gamma as default is given as:



We observe that for higher values of C, the accuracy for each of the datasets increases. This mitigates the adverse effects of miscalculations, which are the results of very small values of gamma. The C parameter takes into account, the number of outliers in calculating the support vectors. Though accuracy increases with C, we risk losing the generalization properties of the classifier by increasing C to a very high value due to it trying to fit all the training data points. Optimum value is one that gives the highest accuracy for validation data.

Conclusion:

Thus on implementing both Logistic Regression and Support Vector Machines we can infer that Support Vector Machines performs better than Logistic Regression provided it has the right tuning parameters.