CSE 574: Introduction to Machine Learning
PA-3: Classification and Regression

By,
Satya Chandu Dheeraj Balakavi
Pravin Umamaheswaran
Mithun Nagesh

Logistic Regression:
Logistic Regression is an approach to learning functions of the form f: X → Y, or P(Y|X) in the case where Y is discrete-valued, and X = <X1 ... Xn> is a vector containing discrete or continuous variables. In this project, we consider the case of a simple logistic regression, where Y is a boolean variable. A simple Logistic regression is similar to linear regression but is intended for use with binary outcomes instead of continuous outcomes.

Equations used in the computation of cross-entropy error and gradient of the error (wrt w),

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

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

Where,

$\mathbf{x} = [1, x_1, x_2, \cdots, x_D]$ is the input vector

$\{t_1, t_2, \cdots, t_N\}$ is the corresponding label vector

$\mathbf{w} = [w_0, w_1, w_2, \cdots, w_D]$ is the weight vector

and,

$y_n = \sigma(\mathbf{w}^T\mathbf{x}_n)$ for $n = 1, 2, \cdots, N.$ is the Logistic Regression hypothesis

Results:

Training set Accuracy:92.306%
Validation set Accuracy:91.57%
Testing set Accuracy:91.83%

Analysis:

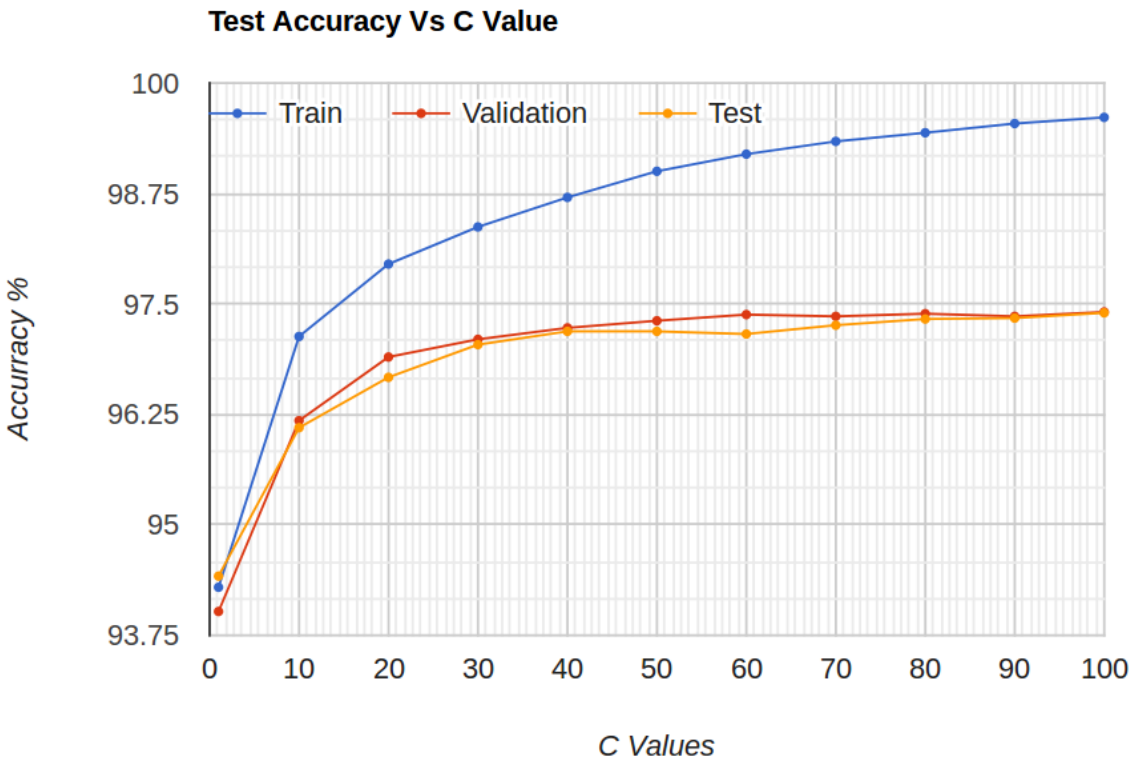
In this project we have shown that the Logistic Regression can be a good technique to use when the outcome variable is dichotomous. The effectiveness of the model was proved by running it against the Training, Validation and Test data, which yielded an accuracy rates of around 92% in all the three cases.

Support Vector Machines:

Support Vector machines are supervised learning models . In this part of the project we use SVM toolkit to perform the classification on our dataset . We makes use of different kernel methods for performing this classification and then compute the accuracy of the prediction based on the different parameters / kernel functions used.

Results:

			ACCURACY %		
Model	**Gamma**	**C**	**Training**	**Validation**	**Testing**
Linear	NA	NA	97.286	93.640	93.780
Radial Basis fn. with Gamma	1	NA	100.000	15.480	17.140
Radial Basis fn. No Gamma	0	NA	94.294	94.020	94.420
	0	1	94.294	94.020	94.420
	0	10	97.132	96.180	96.100
	0	20	97.952	96.900	96.670
	0	30	98.372	97.100	97.040
	0	40	98.706	97.230	97.190
	0	50	99.002	97.310	97.190
	0	60	99.196	97.380	97.160
	0	70	99.340	97.360	97.260
	0	80	99.438	97.390	97.330
	0	90	99.542	97.360	97.340
	0	100	99.612	97.410	97.400



Analysis:

In this part of the experiment we examined the accuracy of prediction of the SVM models with different parameter values and two different kernel functions namely linear and radial basis functions. From the test readings that we obtained we could infer that the ***radial basis functio***n fares better than the ***Linear model*** yielding a better accuracy except in the case with **gamma -1** where we see an overfitting problem. The radial basis function fares better than the linear model since we are trying to classify images of digits and individual pixels are not very informative . In this case the data is not very well linearly separable and so we get a better performance when we use non linear kernel function - RBF.

We can also infer that the behaviour of the SVM - RBF model is very sensitive to the gamma value as for the gamma value of 1 the SVM model and performs overfitting on the data.

In the last part we examined the performance of the SVM with RBF as the kernel function by varying the C value of the SVM and we found that the the accuracy of the classifier increases with increase in C and after a particular value of C there is no much change in the accuracy. As we increase the value of the C we are increasing the number of support vectors which increases the complexity of the decision surfaces and hence the accuracy on the training data increases. As C is increased we also found that the running time increased.

SVM vs Logistic Regression: From the test results it can be inferred that the SVM performs a better classification of the digits when compared with the Logistic regression. The given data for classification is non linear and also the dimensions of the data is large where SVM outperforms logistic regression.