UNIVERSITY AT BUFFALO

Programming Assignment 3

CSE 574 (Machine Learning)

Classification and Regression Project

Submitted By:

Nishant Kapoor

Manasa Challa

Mahek Sangwan

Problem 1: Logistic Regression Implementation

For the first part, we had to implement the logistic regression model which is basically used for classification. In this part, we were given two functions to implement. One of the functions was used to create object and other was used for prediction. A good accuracy was obtained using the given model and dataset. 2 columns error and error_grad were returned in this case. The following results were obtained after implementing the first part:-

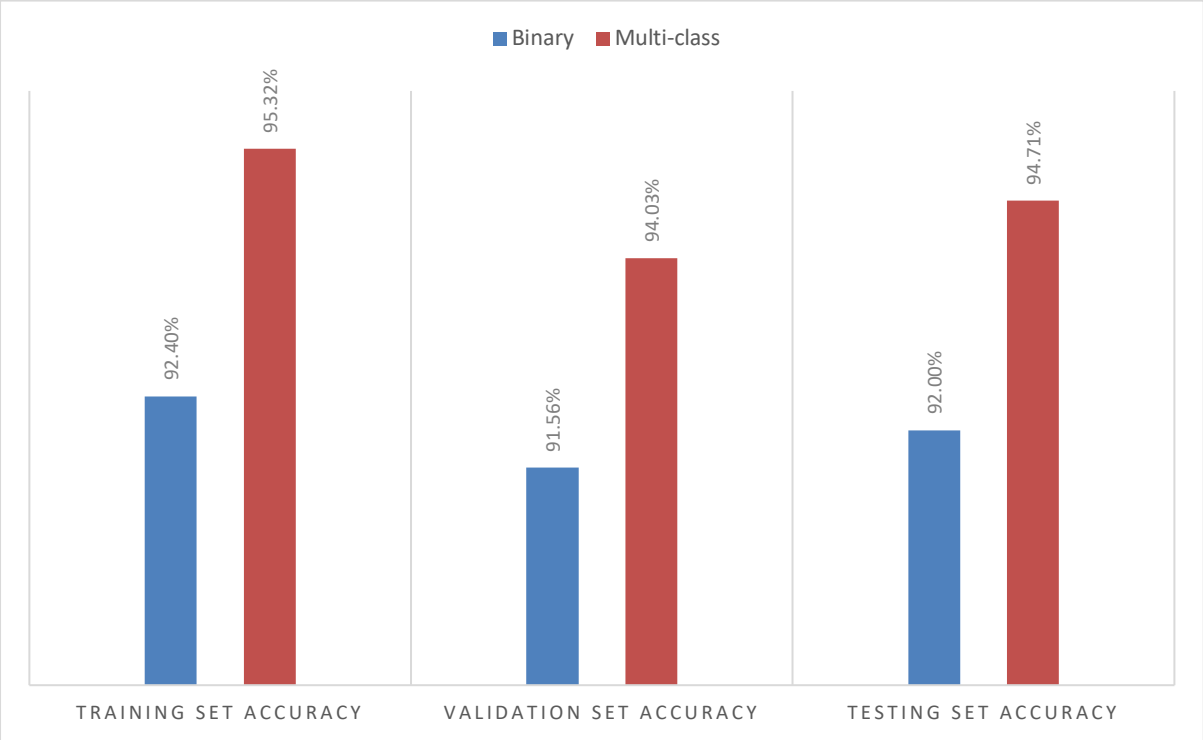
- Training set Accuracy : 92.404%
- Validation set Accuracy : 91.56%
- Testing set Accuracy : 92.0%

Problem 2: Extra-Credit Evaluation

For calculating multi-class logistic regression, we employ a single classifier to implement the model. For modelling more than one class, a single classifier can be used here. The accuracy measured for the multi logistic model is as follows:-

- Training set Accuracy: 95.32%
- Validation set Accuracy: 94.03%
- Testing set Accuracy: 94.71%

By looking at the trends, we reach to the conclusion that multi-class logistic regression gives us more accurate values and estimation than observed in a binary classifier logistic function.



Problem 3: SVM

SVM is referred to as a supervised or labelled model used to find hyper-plane in N dimensional model space. We calculated different values for different type of arrangements in the support vector machine model.

1. Implementation using Linear Kernel
   - Training set Accuracy: 97.286%
   - Validation set Accuracy: 93.64%
   - Testing set Accuracy: 93.78%

We observe that the Linear Kernel SVM delivers a very good accuracy of 93.78%.

2. Implementation using Radial Basis Function, Gamma = 1
   - Training set Accuracy: 100.0%
   - Validation set Accuracy: 15.48%
   - Testing set Accuracy: 17.14%

After a deep and thorough observation, we came to some excellent accuracy ie, 100% in case of training data but less precision was observed with testing and validation data. This results into overfitting of curve. Overfitting happens when a model is trained too much and is having a lot of complexities.

3. Implementation Using Radial Basis Function , keeping Gamma default
   - Training set Accuracy: 98.982%
   - Validation set Accuracy: 97.89%
   - Testing set Accuracy: 97.87%

Implementation Using Radial Basis Function – Default Gamma with changing C values -We measured the set for different C values. We found that accuracy for testing data is directly proportional to increasing C values At peak values of C ie, close to 100 The accuracy escalates by a small amount.

TABLE :-----

C	Training	Validation	Testing
1	98.982%	97.89%	97.87%
10	99.988%	98.34%	98.45%

20	97.045%	98.31%	98.441%
30	98.23%	96.24%	97.83%
40	99.547%	97.02%	97.31%
50	99.67%	97.344%	98.4%
60	98.13%	96.23%	97.566%
70	98.972%	96.934%	97.21%
80	98.065%	97.042%	97.07%
90	99.22%	97.1%	98.3%
100	99.85%	97.88%	97.9%