CSE 574 - Introduction to Machine Learning

*Programming Assignment 3  
Classification and Regression*

Submitted by :

1. Madhur Gupta (UBIT Name : madhurgu)
2. Mohit Arora (UBIT Name : marora3)
3. Abhijit Pattanaik (UBIT Name : abhijitp)

**Logistic Regression**

For this part of the assignment we used multiclass logistic regression to classify the each input based on the probabilities computed.  
Logistic regression is implemented by using gradient descent optimization on error function.   
For the given data, following accuracies were obtained:

|  |  |
| --- | --- |
| Type of Data | Accuracy |
| Training data | 92.27% |
| Test data | 91.52% |
| Validation data | 91.87% |

**Support Vector Machine**

Following results were obtained for various settings of the SVC parameters

**a) Using Linear Kernel with default values**

|  |  |
| --- | --- |
| Type of Data | Accuracy |
| Training data | 97.286% |
| Test data | 93.78% |
| Validation data | 93.64% |

**b) Using RBF Kernel with Gamma set to default and 1**

The gamma parameter defines how far the influence of a single training example reaches, with low values meaning ‘far’ and high values meaning ‘close’. The gamma parameters can be seen as the inverse of the radius of influence of samples selected by the model as support vectors.

From the below table we can see that for default value of gamma we get higher accuracies on test data and validation data.

|  |  |  |
| --- | --- | --- |
| Type of Data | Accuracy for Gamma = 1 | Accuracy for Gamma default |
| Training data | 100.0% | 94.294% |
| Test data | 17.14% | 94.42% |
| Validation data | 15.48% | 94.02% |

**c) Using RBF kernel for various values of C**

The C parameter trades off misclassification of training examples against simplicity of the decision surface. A low C makes the decision surface smooth, while a high C aims at classifying all training examples correctly by give the model freedom to select more samples as support vectors.

|  |  |  |  |
| --- | --- | --- | --- |
| C | Training Data | Test Data | Validation Data |
| 1 | 94.29% | 94.42% | 94.02% |
| 10 | 97.13% | 96.10% | 96.18% |
| 20 | 97.95% | 96.67% | 96.90% |
| 30 | 98.37% | 97.04% | 97.10% |
| 40 | 98.71% | 97.19% | 97.23% |
| 50 | 99.00% | 97.19% | 97.31% |
| 60 | 99.20% | 97.16% | 97.38% |
| 70 | 99.34% | 97.26% | 97.36% |
| 80 | 99.44% | 97.33% | 97.39% |
| 90 | 99.54% | 97.34% | 97.36% |
| 100 | 99.61% | 97.40% | 97.41% |

From the above graph we can see that as the value of C increase accuracies on all types of data increases.

**References:**

1. http://scikit-learn.org/stable/auto\_examples/svm/plot\_rbf\_parameters.html