CSE474/574 Introduction to Machine Learning
Programming Assignment 3
Handwritten Digits Classification
Using Logistic Regression and Support Vector Machine (SVM)

TEAM MEMBERS:

ADITHYA RAMAKRISHNAN (5009 8106)

SHIYAMSUNDAR SOUNDARA RAJAN (5009 7590)

VIVEKANANDH VEL RATHINAM (5009 8075)

The project is about implementing Logistic Regression and use Support Vector Machine toolbox to classify hand-written digit images and compare the performance of these methods using various parameters involved.

LOGISTIC REGRESSION:

Logistic regression is a type of probabilistic statistical classification model. It is also used to predict a binary response from a binary predictor, used for predicting the outcome of a categorical dependent variable (i.e., a class label) based on one or more predictor variables (features). That is, it is used in estimating empirical values of the parameters in a qualitative response model.

Logistic regression can be binomial or multinomial. Binomial or binary logistic regression deals with situations in which the observed outcome for a dependent variable can have only two possible types (for example, "dead" vs. "alive"). Multinomial logistic regression deals with situations where the outcome can have three or more possible types (e.g., "disease A" vs. "disease B" vs. "disease C")

We have implemented logistic regression varying the following optimization methods:

1. Gradient Descent

2. Newton Raphson Method

We have also implemented logistic regression for classifying multinomial varibles using the same two optimization methods.

We use gradient descent and Newton Raphson method to find the minimum loss function. The accuracy rates for the various datasets are as follows:

1. Gradient Descent:

Training Set Accuracy	93.000000
Validation Set Accuracy	91.420000
Test Set Accuracy	91.77000

2. Newton Raphson Method:

Training Set Accuracy	92.480000
Validation Set Accuracy	90.550000
Test Set Accuracy	91.13000

From the above two tables we can infer that logistic regression using gradient descent has better accuracy compared Newton Raphson method.

MULTICLASS LOGISTIC REGRESSION:

Multinomial logistic regression is a classification method that generalizes logistic regression to multiclass problems, i.e. with more than two possible discrete outcomes. That is, it is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables (which may be real-valued, binary-valued, categorical-valued, etc.).

We have implemented logistic regression varying the following optimization methods:

1. Gradient Descent

2. Newton Raphson Method

We have also implemented logistic regression for classifying multinomial variables using the same two optimization methods.

We use gradient descent and Newton Raphson method to find the minimum loss function. The accuracy rates for the various datasets are as follows:

1. Gradient Descent:

Training Set Accuracy	93.588000
Validation Set Accuracy	92.420000
Test Set Accuracy	92.760000

SUPPORT VECTOR MACHINE

More formally, a support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

SVM model and compute accuracy of prediction with respect to training data, validation data and testing using the following parameters:
1. Using linear kernel (all other parameters are kept default).
2. Using radial basis function with value of gamma setting to 1 (all other parameters are kept default).
3. Using radial basis function with value of gamma setting to default (all other parameters are kept default).
4. Using radial basis function with value of gamma setting to default and varying value of C.

1. USING LINEAR KERNEL:

Training Set Accuracy	98.84
Validation Set Accuracy	91.15
Test Set Accuracy	90.85

Linear Kernel is useful if the original data is already high dimensional, and if the original features are individually informative. Of course, not all high dimensional problems are linearly separable. As in our project as we are handling the digit image data set, images are high dimensional, but individual pixels are not very informative, so image classification typically requires non-linear kernels. Therefor the accuracy is reasonable.

2. USING RADIAL BASIS FUNCTION WITH VARYING GAMMA VALUES

Intuitively, the gamma parameter defines how far the influence of a single training example reaches, with **low values meaning 'far' and high values meaning 'close'**.

GAMMA VALUE	**1**	**DEFAULT**
Training Set Accuracy	100	93.66
Validation Set Accuracy	10	90.35
Test Set Accuracy	15.75	90.2

As we can see the gamma with lower value gives very poor accuracy compared to default gamma value.

3. **USING RADIAL BASIS FUNCTION WITH VALUE OF GAMMA SETTING TO DEFAULT AND VARYING VALUE 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**.

VALUES OF C	TESTING SET ACCURACY
1	90.2
10	93.25
20	93.55
30	93.9
40	**94.15**
50	93.9
60	93.95
70	94.1

80	94
90	93.95
100	93.65

From the table we can infer that the testing set accuracy seems to increase till C = 40 and oscillates for further values above it till 100. Thus SVM seems to give the best testing set accuracy for a default gamma value and C = 40. As explained above higher values of C tries to classify all the samples correctly.

Thus **SVM seems to give a higher accuracy of 94.15%** for the testing test compared to all other methods.