ML Assignment 3

Neeharika Immaneni - NI2452

neeharika@utexas.edu

1 Introduction

The goal of the assignment was to use Sequential Minimal Optimization (SMO) algorithm for training Support Vector Machines (SVM) on the MNIST data set. We are provided with 60000 training and 10000 test images each of size 28 x 28. SVM is a supervised binary classification algorithm, which involves constructs a hyper plane that can linearly separate the data into two with maximum separation. For data that is not linearly separable, we project them to a higher dimensional space and then try to separate them. To avoid data over fitting, we introduce slack variables (soft margin) which allows errors while fitting the model. Many experiments were done which have been elaborated in detail below.

2 Method

SVM computes a linear classifier that can be applied to binary classification problems, where we find the hyper plane the separates the data points into two with maximum margin. We ultimately predict, y = 1 if f(x) $\geq$ 0 and y = -1 if f(x) $<$ 0. I have solved the problem in dual space, where f(x) is computed as follows:

$$f(x) = \sum_{i=1}^{m} \alpha_i y^{(i)} < x^{(i)}, x > +b$$

The SMO algorithm gives an efficient way of solving the dual problem of the support vector machine optimization problem. We solve the following using by computing the Lagrange multipliers.

$$max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - 1/2 \sum_{i=1}^{m} \sum_{j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j < x^{(i)}, x^{(j)} >$$

subject to $0 <= \alpha_i <= C, i = 1, ..., m$

$$where \sum_{i=1}^{m} = \alpha_i y^{(i)} = 0$$

The positively labeled data stay on one side of the hyper plane, and the negatively labelled instances on the other. Since most data is not perfectly separable, so we add some slack to the problem leading the to complete, soft-margin, SVM optimization problem.

The SMO algorithm selects two α parameters, α_i and α_j and optimizes the objective value jointly for both these $\alpha^{'}$s. Finally it adjusts the bias parameter based on the new $\alpha^{'}$s. This process is repeated until the $\alpha^{'}$s converge.

The SMO algorithm uses KKT conditions to check for convergence to the optimal point. For this problem the KKT conditions are:

$$\alpha_i = 0 => y^{(i)}(w^T x^{(i)} + b) >= 1$$
$$\alpha_i = C => y^{(i)}(w^T x^{(i)} + b) <= 1$$
$$0 < \alpha_i < C => y^{(i)}(w^T x^{(i)} + b) = 1$$

Any $\alpha_i\ 's$ that satisfy these properties for all i will be the optimal solution to the optimization problem. The SMO algorithm iterates until all these conditions are satisfied (to within a certain tolerance) thereby ensuring convergence.

The model was trained using different kernel functions (polynomial SVM and radial basis SVM) and also different polynomial degrees were tried to study variation in accuracy.

3 Experiments

3.1 Kernel functions

The model was trained using a polynomial kernel of degree 1,2,3 and a radial basis kernel of by varying the gamma parameter on 10000 training images. The variation in accuracy is plotted and shown in Figure 1. Polynomial - 1,2,3 represent the polynomial functions with degree 1, 2 and 3 respectively. RBF(1) and RBF(2) represent the radial basis function with sigma value of 0.5 and 0.1 respectively. It can be seen that RBF outperforms the polynomial classification technique since it is a squared exponential kernel and it defines a function space that is a lot larger than that of the linear kernel or the polynomial kernel. The training time is also more for the RBF kernel. As we increase the degree of the polynomial (increasing dimension of space), we see that the accuracy also increases proportionally.

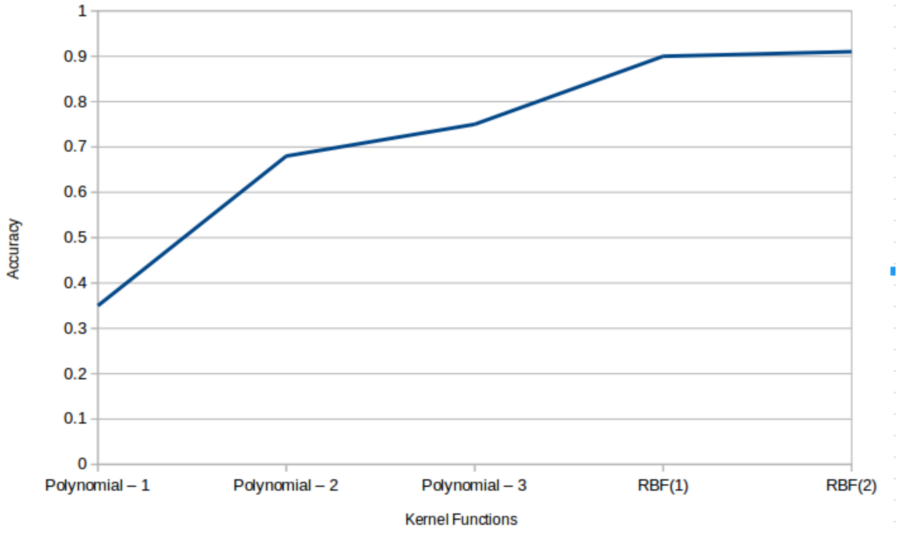

Figure 1: Kernel Functions vs Accuracy

3.2 Data split

Different strategies were tries such as one vs one, one vs rest. The strategies I tried are listed below.

a. 0 vs rest(1—-9)

b. 7 vs rest

c. 4 vs 9

d. 0 vs 8

e. (0,8,3) vs (1,7,9)

For all these strategies, the model was trained on 10000 images using the parameters C(margin) = 5.0 and tolerance = 0.00001 using RBF kernel. The accuracy for 0 vs rest and 7 vs rest seems to be the highest of all. Since 4,9 and 0,8 look similar, the accuracy is less for this split of the data. All the accuracy values are plotted which can be seen in Figure 2.

3.3 Size of training data

With increase in size of the training data, the accuracy of the model also increases linearly. Using the first 5000 test images, gave better accuracy rather than testing it on the entire test set of 10K images.The training size versus accuracy values are plotted which can be seen in Figure 3

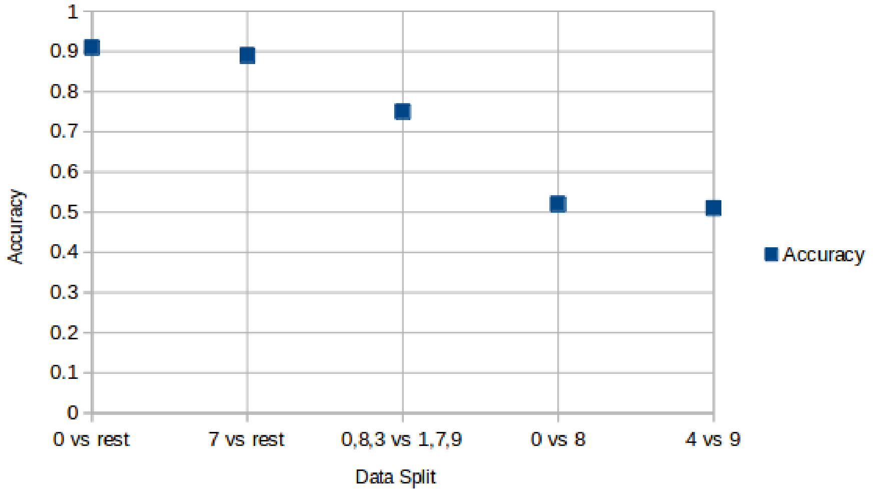

Figure 2: Training Samples vs Accuracy

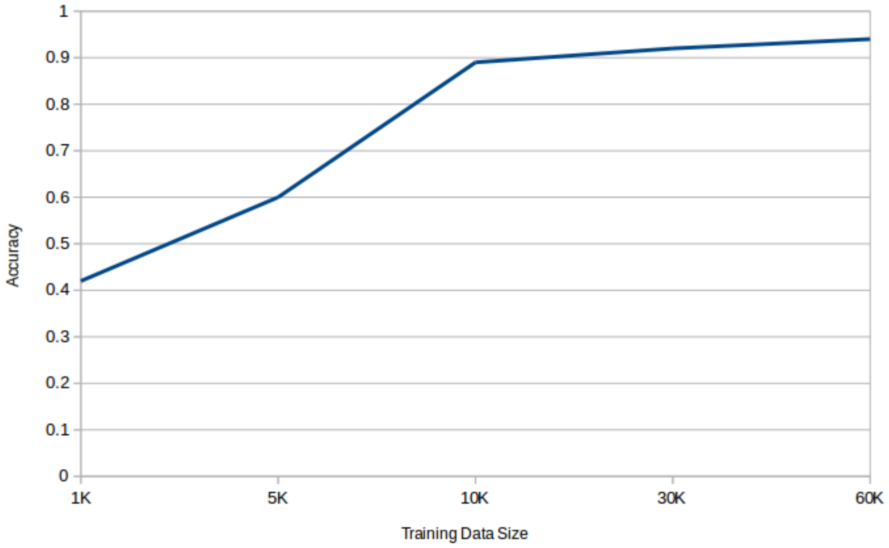

Figure 3: Training Data Size vs Accuracy

Training data	Test data	c=1	c=5	c=10	c=infinity
30000	5000	0.93	0.91	0.88	0.51

Table 1: Accuracy of SVM model by varying margin

3.4 Hard and Soft Margin

The margin specifies the percentage of training data that can be mis-classified in the SVM model. It allows a trade off between learning simple functions and fitting the data exactly. All the training was performed with a soft margin of 5 and for training it on the hard margin, it was increased to infinity. Hard margin SVMs work only when the data is perfectly linearly separable. The variation in accuracy is shown in the table 1.

3.5 Accuracy

Comparing this with Assignment 1, we can see that PCA performs better than SVM on the MNIST data set for digit classification. With SVM, I could achieve a maximum of 95% accuracy using RBF kernel, training on 60000 images.

4 Discussions and Conclusion

A maximum accuracy of **0.95** was achieved using the radial basis function by projecting the data to a higher dimensional space and using a soft margin of 1. Compared to PCA, the training time is high and the accuracy is lesser, since we take into consideration all the 784 dimensions and the process us repeated in an iterative fashion until it converges. Increasing the tolerance value, would result in faster convergence but the accuracy dropped. Hence a tolerance value of 0.00001 was maintained for all the various experiments performed.