# IPR CSE 555 Problem Set 2: Linear Discriminant Functions And Support Vector Machines

**Aniruddha D. Karlekar**
Department of Computer Science
University at Buffalo
Buffalo, NY 14214
akarleka@buffalo.edu

## Abstract

In this problem set, you will train a support vector classifier using the the MNIST training data set and report performance using MNIST testing data set. Then you will derive the primal-dual relationship of the 1-norm soft-margin classification problem, in which process you will demonstrate your understanding of several key concepts, such as maximal margin and support vector. The emphasis is on getting hands dirty with SVM and understanding the theory. We will train our model using the training data sets ("train-images-idx3-ubyte.gz" and "train-labels-idx1-ubyte.gz") and test the performance using the test data set ("t10k-images-idx3-ubyte.gz" and "t10k-labels-idx1-ubyte.gz").

## 1 Task 1

Write code to train a multi-class support vector classifier with dot-product kernel and 1-norm soft margin using the MNIST training data set. Then report the performance using MNIST test data set. There is a hyper-parameter that sets the trade-off between the margin and the training error --- tune this hyper-parameter through cross-validation.

### Method:

1. We first Read the MNIST training dataset that has 60000 images, and thetest data with 10000 images are read.

2. The SVM classifier object is created using the svm.SVC feature of sklearns library.

3. The SVM classifier object and parameters are passed into the Gridsearch() function. This runs the SVM classifier algorithm over all the parameters and returns the optimal ones.

4. The SVM classifier object is run on the test dataset by using the optimized parameters.

5. Values of the images of digits of test dataset are predicted.

6. Accuracy of the model is computed by checking predicted values against label values.

Note: After getting the optimal parameters I ran the model on 20000 samples since running

on the entire dataset takes a huge amount of time.

## 2 Task 2

The MNIST database (Modified National Institute of Standards and Technology database) is a large database of handwritten digits that is commonly used for training various image processing systems. The database is also widely used for training and testing in the field of machine learning.

# Part 2

1) Since there are inequalities in the constraints, we will be using Lagrangian inequalities.

Primal Lagrangian of the given problem will be

$$\alpha(w, b, \epsilon, \alpha, u) = \frac{1}{2} w^T \cdot w + c \sum_{i=1}^{N} \epsilon_i - \sum_{i=1}^{N} \alpha_i [y_i (w^T x_i + b) + \epsilon_i] - \sum_{i=1}^{N} u_i \epsilon_i$$

$\frac{1}{2} w^T \cdot w$ is the regularization term

$c$ is the penalization of the slack variable.

The optimal solution or the saddle point can be determined by:

→ maximum wrt the dual variable

→ minimum wrt the primal variable

We now differentiate the Lagrangian wrt the primal variables i.e. slack variables, bias and normal vectors. and setting to zero.

$$\frac{\partial L}{\partial \epsilon} = c - \alpha_i - u_i = 0$$

$$\therefore \alpha_i - u_i = c \quad \rightarrow (i)$$

$$\frac{\partial L}{\partial b} = - \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\therefore \sum_{i=1}^{N} \alpha_i y_i = 0 \quad \rightarrow (ii)$$

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0$$

$$\therefore w = \sum_{i=1}^{N} \alpha_i y_i x_i \longrightarrow (i.i)$$

Margin $w = \sum_{i=1}^{N} \alpha_i y_i x_i$

Substituting in primal,

$$L(w, b, \epsilon, \alpha, u) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{\lambda=1}^{N} y_i y_i \alpha_i \alpha_j \langle x_i x_j \rangle$$

where $\sum_{i=1}^{N} \alpha_i y_i = 0$ and $\alpha_i \geq 0$. $u_i \geq 0$

## 2) Margin in Primal/Dual Formulation

→ Margin in Primal Formulation.

For two features $(x_1, y_1)$, $(x_2, y_2)$ in two classes $(+1, -1)$ we have

$$wx_1 + b = 1$$

$$wx_2 + b = -1$$

$$\therefore w(x_1 - x_2) = 0$$

Margin $M = \|x_1 - x_2\|$

Let $x_2$ be a point in Region of -1 and $X_1$ be the closest point to $X_2$ but in the other region $(+1)$. So we get $X_1 = X_2 + rw$ where $r$ is a constant.

$$wx_1 + b = 1$$

$$w.(X_2 + rw) + b = 1 \quad (\text{subs})$$

$$r\|w\|^2 + wx_2 + b = 1$$

$$r\|w\|^2 - 1 = 1$$

$$r = \frac{2}{\|w\|^2}$$

$$\therefore M = \|x_1 - x_2\| = \|r w\| = \frac{2}{\|w\|^2}\|w\|$$

$$M = \frac{2}{\|w\|}$$

→ Margin for dual formulation.

$$\max \alpha_0(\alpha_i) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle$$

where $\sum_{i=1}^{N} \alpha_i y_i = 0$ and $\alpha_i \geq 0$

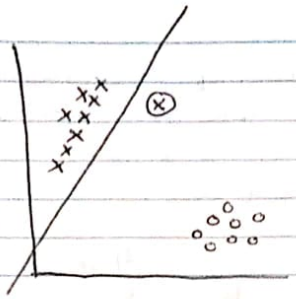
Differentiating wrt $\alpha$ and equating to zero.

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \quad \text{where } 0 < \alpha_i < c$$

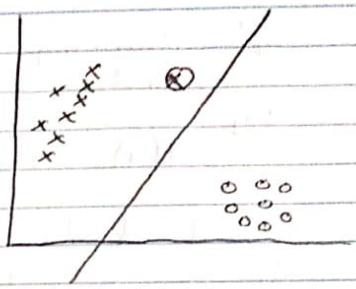
$$w = \sum_{i=1}^{N} \alpha_i y_i x_i$$

## 3) Benefits of Maximizing the Margin.

Maximizing the margin makes the model more robust. SVM maximizes the margin. Maximizing the margin is beneficial because it reduces uncertainty that comes with predictions that fall close to the margin. It provides a safety margin and reduces the possibility of misclassificati



Non-maximized margin



Maximized margin

consider the point * (circled). It is misclassified if the margin isn't maximized

## 4) Characterize the Support Vectors

Differentiating the dual problem and equating to zero. The non-zero $\alpha_i$s are the support vectors.

$$\max \alpha_0(\alpha_i) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i \cdot x_j \rangle \quad ; \alpha_i \geq 0$$

Deriving wrt $\alpha$

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \quad ; 0 \leq \alpha_i \leq c$$

Solving for $\alpha_i$,

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i$$

After training and obtaining $w$, assuming a random point '$u$' we can classify it by

$$f(x) = w.u = \sum_{i=1}^{N} \alpha_i y_i x_i . u$$

Most of the weights $w_i$ i.e. the $\alpha_i$s will be 0. only the support vectors have non zero $\alpha_i$s.
Support vectors reduce dimensionality of the solution.

5) Benefits of solving the dual Problem instead of Primal Problem.

→ Regularizing the sparse support vector in dual problem is sometimes more intuitive than regularizing the vector of regression coefficients. Dual problem always adapts to the amount of available data.
→ Finding an initial feasible solution is easier for the dual problem.
→ It is easier to optimize for the dual problem than it is for the primal problem when the no. of data points is smaller than the no. of dimensions.
→ Being a convex problem the dual problem always converges. The same isn't true for primal problem.

**Result:**
After successfully implementing the SVM, we were able to classify the MNIST dataset with an accuracy of over 89%.

## References

[1] https://en.wikipedia.org/

[2]Professor Wen Dong's class notes and ppt

[3] http://cs229.stanford.edu/notes/cs229-notes3.pdf