MACHINE LEARNING CS60050

Instructor: Dr. Aritra Hazra
Department of Computer Science Engineering
Indian Institute of Technology, Kharagpur
Scribed by: Arpan Dam ; 20CS91R06

Notes for class on 12th February 2021

1 SVM optimisation problem

For SVM the optimisation problem is :

$$Min\frac{1}{2}W^TW \tag{1}$$

subject to constraints

$$y_i(W^Tx_i + b) >= 1 \tag{2}$$

$$\forall(x_i, y_i) \tag{3}$$

Where W is the coefficient of the hyper plane of SVM, b is the bias and (x_i, y_i) is the coordinate of the training data.
The primal form of the above mentioned inequality constraint optimization problem(according to Lagrange multiplier method) is given by

$$L_p = \frac{1}{2}W^TW - \Sigma_{i=1}^{n}\alpha_i(y_i(W.x_i + b) - 1) \tag{4}$$

where α_i 's are called Lagrange multipliers.L_p is called the primal form of the Lagrangian optimization problem. It can be seen from Figure (1) that the orange data points which are closest to the hyper plane only contribute to equation (4) as for only these 4 orange data points $\alpha_i > 0$ and for all the other data points $\alpha_i = 0$.

Dual Formulation of Optimization Problem

To minimize the Lagrangian, we must take the derivative of L_p with respect to W, b and set them to zero.

$$\frac{\delta L_p}{\delta W} = 0 \Rightarrow W = \Sigma_{i=1}^{n}\alpha_i.y_i.x_i \tag{5}$$

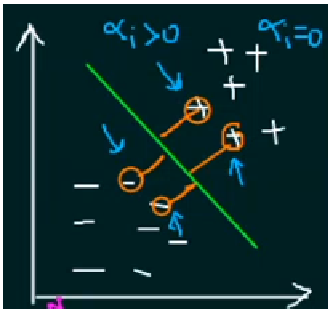

Figure 1: Lagrange multipliers values for training data

$$\frac{\delta L_p}{\delta b} = 0 \Rightarrow \Sigma_{i=1}^{n} \alpha_i . y_i = 0 \tag{6}$$

From above two equation we get dual form of Lagrangian L_D as

$$L_D = \Sigma_{i=1}^{n} \alpha_i - \frac{1}{2} \Sigma_{i,j} \alpha_i . \alpha_j . y_i . y_j . x_i . x_j \tag{7}$$

There are key differences between primal (L_p) and dual (L_D) forms of Lagrangian optimization problem as follows.

- L_p involves a large number of parameters namely W, b and α_i 's. On the other hand, L_D involves only α_i 's, that is, Lagrange multipliers.
- L_p is the minimization problem as the quadratic term is positive. However, the quadratic term in L_D is negative sign, Hence it is turned out to be a maximization problem.
- L_p involves the calculation of W.x, whereas L_D involves the calculation of $x_i.x_j$. This, in fact, advantageous, and we will realize it when we learn Kernel-based calculation.

2 Classification in SVM :

Let the equation of the hyper plane for separating the two class is

$$W^T x + b = 0 \tag{8}$$

By training the value of W and b will be found. Let X^{new} be a text data point which is required to be classified. So if $W^T x^{new} + b > 0$ then x^{new} will be classified as positive class and if $W^T x^{new} + b < 0$ then x^{new} will be classified as negative class.

3 Issues of SVM

- Classes are not linearly separable.
- Multi class classification.

4 SVM classification for non separable data

A linearly not separable data can be classified using Linear SVM with soft margin. Figure (2) shows a set of data which are linearly separable and figure (3) shows a set of data which are linearly non-separable.

4.1 Linear SVM for Linearly Not Separable Data

- A linear SVM can be refitted to learn a hyperplane that is tolerable to a small number of non-separable training data.
- The approach of refitting is called soft margin approach (hence, the SVM is called Soft Margin SVM), where it introduces slack variables to the inseparable cases
- More specifically, the soft margin SVM considers a linear SVM hyperplane (i.e., linear decision boundaries) even in situations where the classes are not linearly separable.
- For soft margin we rewrite the optimization problem as follows.

$$minimise \frac{1}{2} W^T W \tag{9}$$

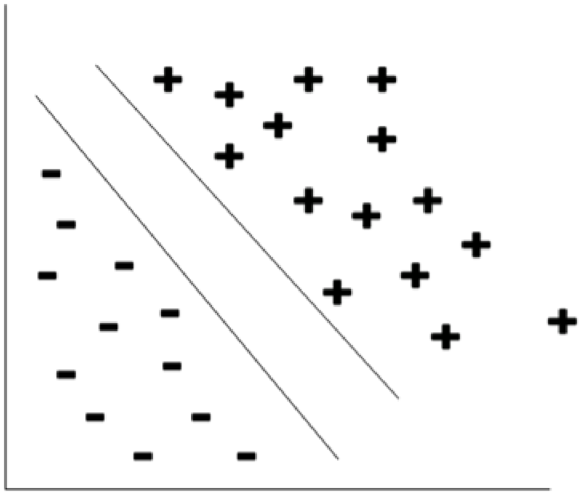

Figure 2: Linearly separable

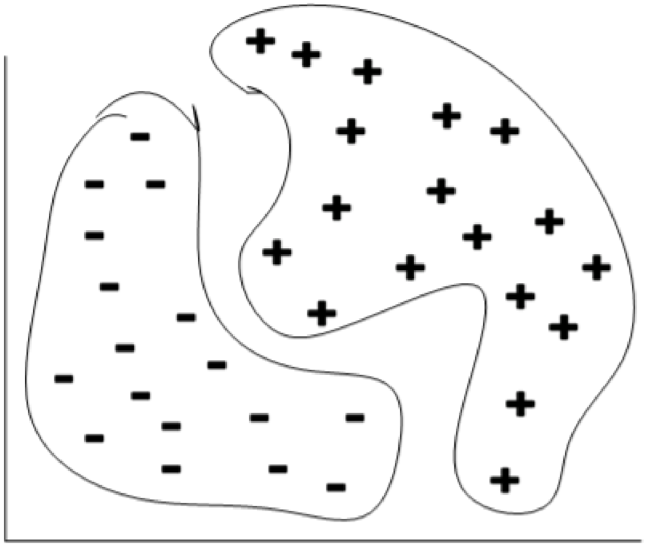

Figure 3: Linearly non-separable

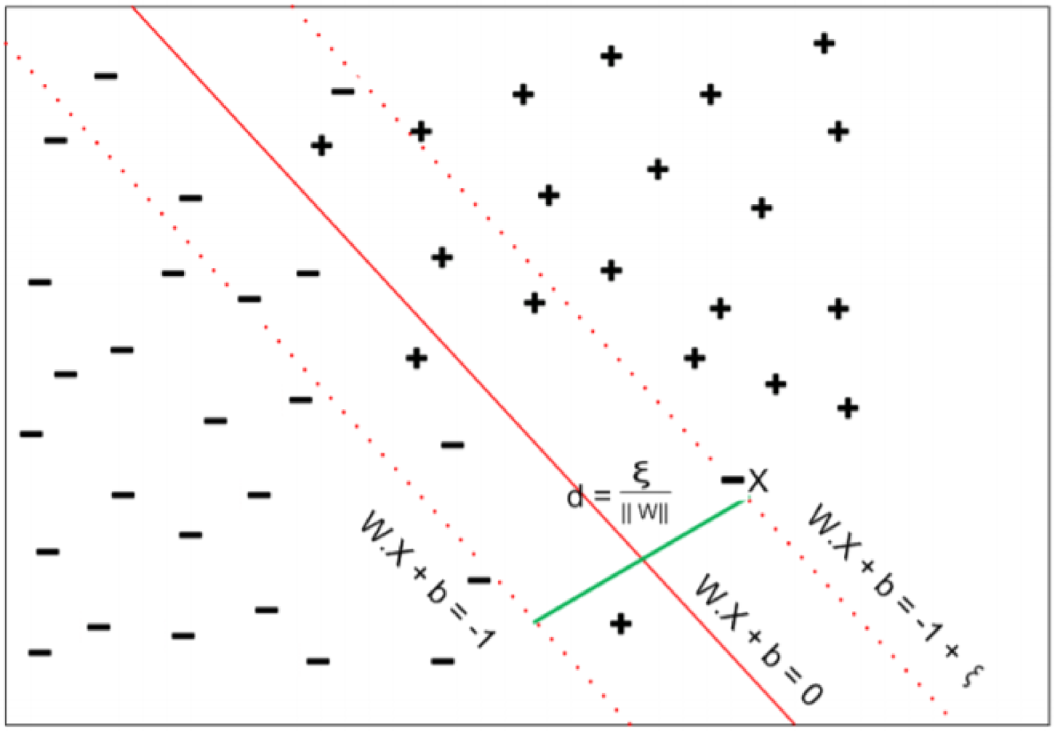

Figure 4: : Interpretation of slack variable ξ

subject to $W.x_i + b >= 1 - \xi_i$ if $y_i = +1$

and $W.x_i + b <= -1 + \xi_i$ if y_i=-1

where $\xi_i >= 0$

- Thus, in soft margin SVM, we are to calculate W, b and ξ as a solution to learn SVM.
- Figure (4) shows an interpretation of ξ, the slack variable in soft margin SVM
- ξ provides an estimate of the error of decision boundary on the training example X.
- The soft margin SVM should impose a constraint on the number of such non linearly separable data it takes into account.This is so because a SVM may be trained with decision boundaries with very high margin thus, chances of misclassifying many of the training data.If the increase margin is increased, more points will be misclassified.Thus, there is a trade-off between the length of margin and training error.To avoid this problem, it is therefore necessary to modify the objective function, so that penalizing for margins with a large gap, that is, large values of slack variables.
- The modified objective function can be written as

$$f(W) = \frac{1}{2}W^TW + c\Sigma_{i=1}^{n}\xi \tag{10}$$

  where c is user specified parameters representing the penalty of misclassifying the training data.
- The Lagrange multiplier method to solve the inequality constraint optimization problem is as follows:

$$L = \frac{1}{2}W^TW + c.\Sigma_{i=1}^{n}\xi - \Sigma_{i=1}^{n}\alpha_i(y_i(W.x_i + b) - 1 + \xi_i) - \Sigma_{i=1}^{n}\lambda_i.\xi_i \tag{11}$$

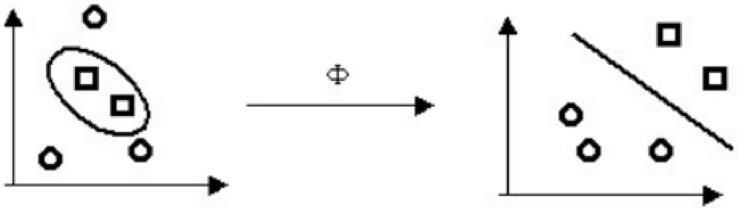

Figure 5: Transformation of a non-linearly separable problem into a linearly separable problem.

Here, α_i 's and λ_i 's are Lagrange multipliers.
The inequality constraints are:

$$\xi_i >= 0 \tag{12}$$

$$\alpha_i >= 0 \tag{13}$$

$$\lambda_i >= 0 \tag{14}$$

$$\Sigma_{i=1}^{n} \alpha_i (y_i (W.x_i + b) - 1 + \xi_i) = 0 \tag{15}$$

$$\lambda_i . \xi_i = 0 \tag{16}$$

4.2 Non-Linear SVM

In order to work with non-linear decision boundaries the key idea is to transform x_i to a higher dimension space Figure (5) using a transformation function , so that in this new space the samples can be linearly divided. In a nonlinear SVM, the trick is to transform non-linear data into higher dimensional linear data.

5 Multiclass Classification:

When there are multiple classes then for multiclass classification, the same principle is utilized after breaking down the multi classification problem into multiple binary classification problems.
A single SVM does binary classification and can differentiate between two classes. So that, according to the two breakdown approaches, to classify data points from m classes data set:

- In the One-to-Rest approach, the classifier can use m SVMs . Each SVM would predict membership in one of the m classes. In figure (6) there are 4 classes so here 4 SVMs are being used.

- In the One-to-One approach, two pairs of classes are selected at a time and a binary classifier trained for them. This is done for every possible pair of classes thus there are $\frac{m(m-1)}{2}$ of them where m is the total number of classes. Figure(7) shows an example of one-to-one approach

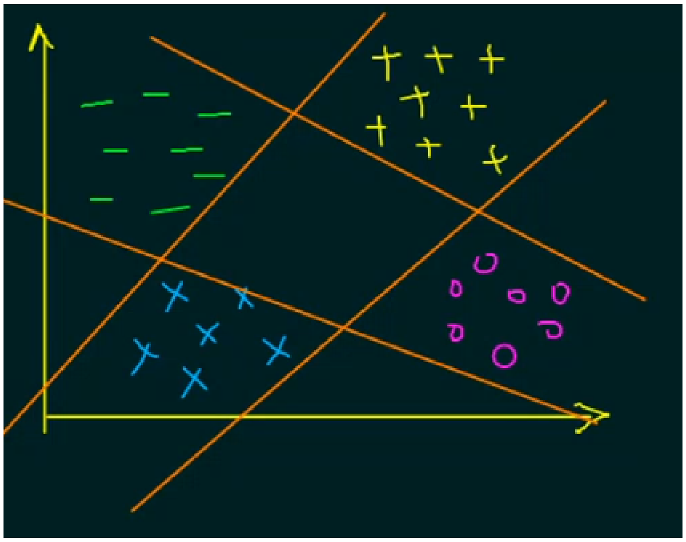

Figure 6: One-to-Rest SVM Multiclass classification

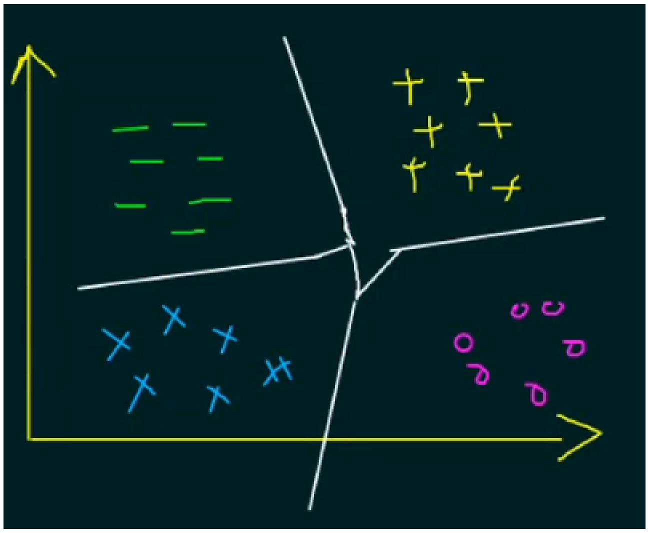

Figure 7: One-to-One SVM Multiclass classification