Introduction to ML - CS 419

# Lecture 9: Non Separable - Support Vector Machine

February 10, 2023

*Lecturer: Abir De* *Scribe: Course Team*

## 1 Revision

### 1.1 Classification Problem

Our dataset is of the form $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$.
The $x_i$ are data points, usually in $\mathbb{R}^k$ for some $k \in \mathbb{N}$ and $y_i$s are discrete.

Goal of Support Vector Machine is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a **hyperplane**.

SVM can be used for linearly separable as well as non-linearly separable data. Linearly separable data is the hard margin whereas non-linearly separable data poses a soft margin.

### 1.2 Separable Case of SVM

This is applicable when there is no overlap between data points. In such a scenario, we will be able to classify the data points perfectly such that the resultant classification error is zero.

$$\mathbf{w}^T x + \mathbf{b} > 1 \;\Rightarrow\; y = +1$$
$$\mathbf{w}^T x + \mathbf{b} < -1 \;\Rightarrow\; y = -1$$

For these two classes we need to train 2 parameters w & b to correctly classifying a training dataset.

Both the classes can be divided by a hyperplane. The convex hull of the positive and the negative points do not intersect. We want a function $\psi$ which can solve:

$$\{\mathbf{w}^*, \mathbf{b}^*\} = argmin\ \psi(\mathbf{w}, \mathbf{b}) \quad : \quad \forall i \in \mathbb{D}\;\; y_i(\mathbf{w}^{*T} x_i + \mathbf{b}^*) > 1$$

This gives us the loss function:

$$\{\mathbf{w}^*, \mathbf{b}^*\} = argmin\ ||\mathbf{w}||^2 \quad : \quad \forall i \in \mathbb{D}\;\; y_i(\mathbf{w}^T x_i + \mathbf{b}) > 1$$

We have two different classes in the picture below. There is a maximum margin between the data points of the two classes. Optimal hyperplane nearly lies between the two classes.

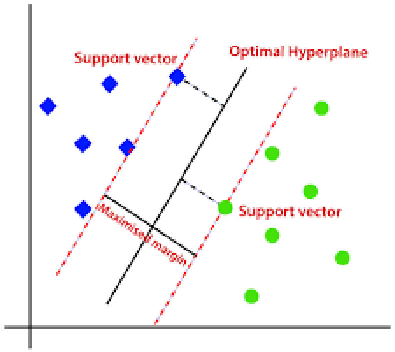

Figure 1: Separable SVM Classification

# 2 Non-Separable SVM

In many real-life cases, the given sample may not be linearly separable and thus we will not have an optimal hyperplane perfectly dividing the hyperplane. There might be an overlap in the convex hull of the negative and the positive points. There does not exist a $\boldsymbol{w}$ such that $y_i(\boldsymbol{w}^T\boldsymbol{x_i} + b) > 1$ satisfies for all the data points of our interest.

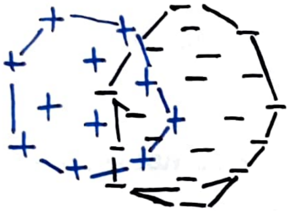

Figure 2: Non-Separable SVM Classification

## 2.1 Removing the outliers

Excluding the *outliers* to reduce the problem to separable-SVM might lead to a significant loss in data. It will hamper classification results and give us a poor model, this is clear when we consider the case shown below.

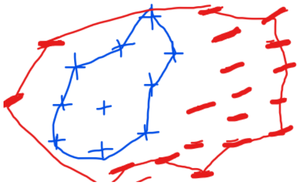

Figure 3: Overlapping SVM classes

### 2.2 Maximise the margin

Another process could be to **maximize the margin**. The Margin is given as

$$\frac{\boldsymbol{w}^T(\boldsymbol{x}_+ - \boldsymbol{x}_-)}{||\boldsymbol{w}||}$$

This margin should be high for all $\boldsymbol{x}_i$. We are worried only about the points where this margin term attains a very low value. We can change our bias such that our $\boldsymbol{w}$ is not affected, so the problem

$$\frac{\boldsymbol{w}^T(\boldsymbol{x}_+ - \boldsymbol{x}_-)}{||\boldsymbol{w}||} \approx \frac{\boldsymbol{\delta}_+ - \boldsymbol{\delta}_-}{||\boldsymbol{w}||}$$

We know that $\boldsymbol{\delta}_+ \approx +1$ while $\boldsymbol{\delta}_- \approx -1$
This gives the target optimization to be

$$max(\frac{2}{||\boldsymbol{w}||})$$

### 2.3 Introduction of Relaxation

The best way to get around such a scenario is to introduce relaxation in our constraints by introducing a parameter $\xi$. Initial assumptions, such as that of the Separable Support Vector Machine problem, will cause trouble here. The parameter is introduced so that every point satisfies the condition:

$$y_i(\boldsymbol{w}^T\boldsymbol{x_i} + b) \geq 1 - \xi_{(x_i,y_i)}$$

Reasoning and mathematics of this scenario are discussed extensively in the sections ahead.

## 3 Mathematical Understanding of Soft-SVM

We know our initial assumptions will cause trouble here. Thus as mentioned above, a slackness parameter $\xi_{x,y}$ is introduced so that all the points satisfy the variable condition:

$$y_i(\boldsymbol{w}^T\boldsymbol{x_i} + b) \geq 1 - \xi_{(x_i,y_i)}$$

where $\xi_{x,y} \geq 0$. To optimize the function, we need to **minimize** $\xi_{x,y}$. This gives us:

$$\min_{\boldsymbol{w},\, b,\, \xi_{(\boldsymbol{x}_i,y_i)}} ||\boldsymbol{w}||^2 + \psi \sum_{i \in D} \xi_{(x_i,y_i)+}$$

If $y_i(\boldsymbol{w}^T\boldsymbol{x_i} + b) > 1$ , the point is correctly classified and $1 - y_i(\boldsymbol{w}^T\boldsymbol{x_i} + b) < 0$, otherwise if $y_i(\boldsymbol{w}^T\boldsymbol{x_i} + b) < 1$, then $1 - y_i(\boldsymbol{w}^T\boldsymbol{x_i} + b) > 0$. So we can use $\xi_{(x_i,y_i)} = max(1 - y_i(\boldsymbol{w}^T\boldsymbol{x_i} + b))$ .

This has been done since we want $\xi_{x_i,y_i} = 0$ for optimal points as they perfectly fall in a class and thus don't come inside other function's convex hull.

Minimizing $\psi \sum_i \xi_{(x_i,y_i)}$ results in the following relation:

$$\min_{\boldsymbol{w},\, b} (||\boldsymbol{w}||^2 + \psi \sum_{i \in \mathcal{D}} ReLU(1 - y_i(\boldsymbol{w}^T \boldsymbol{x_i} + b))$$

Increasing $\psi$ might lead to over-fitting as the classes would be well separated.

# 4 Dual Formulation

## 4.1 Convex Optimization

Some background in convex optimization is needed before we move forward with our new formulation of SVM.

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \text{ such that } g((\theta) \leq \boldsymbol{r}$$

Note that $g(\boldsymbol{\theta})$ can be a vector and then the inequality would be pointwise $(g_i(\theta) \leq r_i)$
We can approximate the above equation as:

$$\max_{\boldsymbol{\lambda} \geq 0} \min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) + \boldsymbol{\lambda}^\top (g(\boldsymbol{\theta}) - \boldsymbol{r}) \tag{1}$$

When $f(\boldsymbol{\theta})$ and $g(\boldsymbol{\theta})$ both are strictly convex, the above two equations are exactly equivalent. This means that either $\boldsymbol{\lambda} = 0$ or $g(\boldsymbol{\theta}) = \boldsymbol{\psi}$. This is known as Slater's condition.

## 4.2 Objective Function

SVM problem at hand:

$$\min_{\boldsymbol{w},b,\xi} \lambda ||\boldsymbol{w}||^2 + \sum_{i \in \mathcal{D}} \xi_i$$
$$\text{Constraints: } y_i(\boldsymbol{w}^\top \boldsymbol{x_i} + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0 \quad \forall i \in \mathcal{D}$$

Rewriting constraints as $1 - \xi_i - y_i(\boldsymbol{w}^\top \boldsymbol{x_i} + b) \leq 0$ and $- \xi_i \leq 0$ to make them of the form $g(w) \leq c$.
Using convex optimisation of (1) we rewrite our formulation as:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \min_{\boldsymbol{w},b,\xi} \lambda ||\boldsymbol{w}||^2 + \sum_{i \in \mathcal{D}} \xi_i + \sum_{i \in \mathcal{D}} \alpha_i (1 - \xi_i - y_i(\boldsymbol{w}^\top \boldsymbol{x_i} + b)) + \sum_{i \in \mathcal{D}} \beta_i(-\xi_i)$$
$$\text{s.t. } \alpha_i \geq 0, \beta_i \geq 0 \;\; \forall i \in \mathcal{D}$$

This is our final optimisation problem in dual space and the function to be maximised is called Lagrangian dual function $g(\boldsymbol{\alpha}, \boldsymbol{\beta})$.
Note : $\lambda$ used here is different from the $\lambda$ used in (1) of convex optimisation.

## 4.3 Optimality Conditions

Differentiating $g(\boldsymbol{\alpha}, \boldsymbol{\beta})$ w.r.t $\boldsymbol{w}, b, \xi_i$ and equating to 0, we get:

$$\begin{aligned}
\frac{\partial g}{\partial \boldsymbol{w}} &= 0 \Rightarrow 2\lambda \boldsymbol{w}^* + \sum_{i \in \mathcal{D}} \alpha_i(-y_i \boldsymbol{x_i}) = 0 \\
&\Rightarrow \boldsymbol{w}^* = \sum_{i \in \mathcal{D}} \frac{\alpha_i y_i \boldsymbol{x_i}}{2\lambda} \\
\frac{\partial g}{\partial b} &= 0 \Rightarrow \sum_{i \in \mathcal{D}} \alpha_i y_i = 0 \\
\frac{\partial g}{\partial \xi_i} &= 0 \Rightarrow 1 - \alpha_i - \beta_i = 0 \\
&\Rightarrow \alpha_i + \beta_i = 1 \quad \forall i \in \mathcal{D}
\end{aligned}$$

We also see that our objective function is quadratic in $\boldsymbol{w}$ and constraints are linear in $\boldsymbol{w}, \xi_i$. Therefore, all functions are strictly convex and we can apply Slater's condition. Therefore,

$$\begin{aligned}
\alpha_i^*(1 - \xi_i^* - y_i(\boldsymbol{w}^{*\top} \boldsymbol{x_i} + b^*)) =& 0 \quad \forall i \in \mathcal{D} \\
\beta_i^* \xi_i^* =& 0 \quad \forall i \in \mathcal{D}
\end{aligned}$$

This gives the following 2 cases:

1. $\xi_i^* > 0 \Rightarrow$ point is incorrectly classified or it is inside the margin of hyperplanes
$\Rightarrow \beta_i^* = 0 \Rightarrow \alpha_i^* = 1$.
$\alpha_i^*$ being high can be seen as penalty being high, which intuitively means that the point is misclassified.

2. $\xi_i = 0 \Rightarrow$ point is on or outside the margin of hyperplanes
If $1 - y_i(\boldsymbol{w}^{*\top} \boldsymbol{x_i} + b^*) > 0 \Rightarrow \alpha_i^* = 0 \Rightarrow \beta_i^* = 1$. This means that the penalty is low and the point is correctly classified.
If $1 - y_i(\boldsymbol{w}^{*\top} \boldsymbol{x_i} + b^*) = 0$, then we cannot comment anything about $\alpha_i^*$.