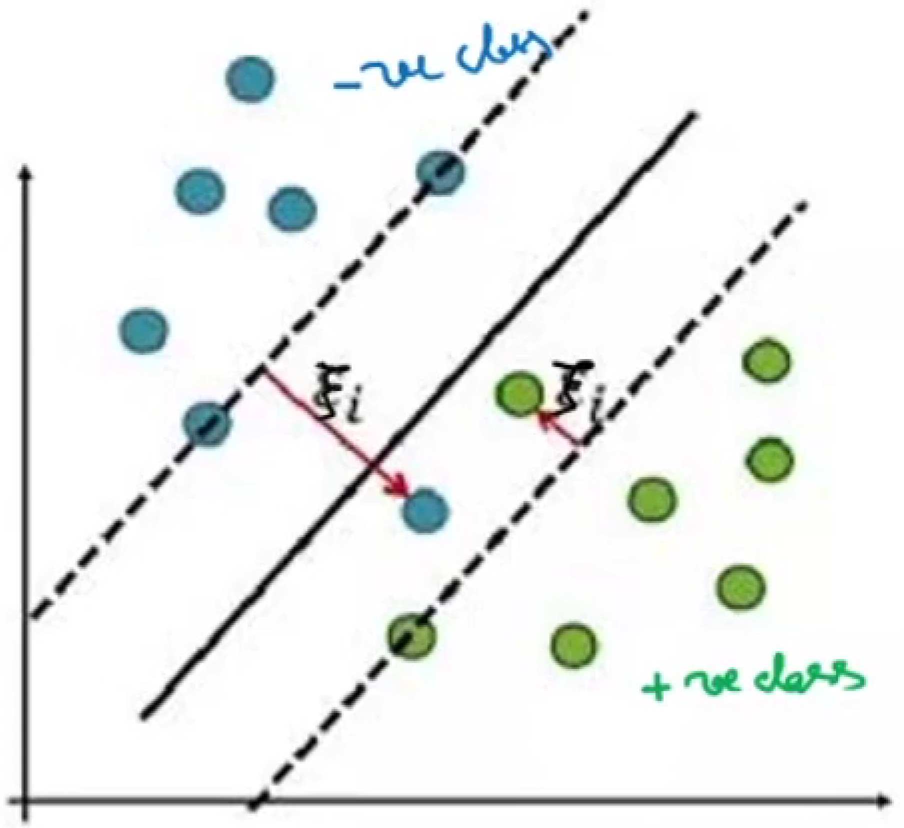
Regularization in SVM



Soft Margin SVM.

Let us take the Primal Prob:

$$\min_{\bar{w}, \xi_i, b} \quad \frac{1}{2} \|\bar{w}\|^2 + C \sum_{i=1}^{n} \xi_i$$

$$\text{s.t.} \quad Y^{(i)}(\bar{w} \cdot \bar{x}^{(i)} + b) \geq 1 - \xi_i \qquad (8)$$

$$\xi_i \geq 0, \quad i = 1, 2, \dots n.$$

Regularization of SVM

Recap: Primal problem with regularizer

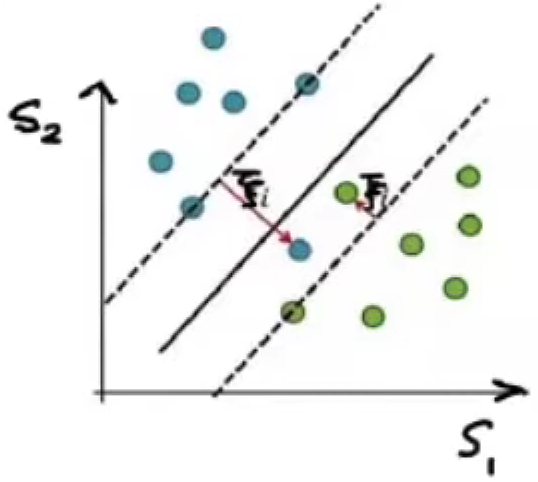
$$\min_{\bar{w}, b, \xi} \; \frac{1}{2}\|\bar{w}\|^2 + C\sum_{i=1}^{m} \xi_i$$

, role of C here are
- To make $\|\bar{w}\|^2$ small
- To ensure that most examples have functional margin at least 1.

Subject to: $Y^{(i)}(\bar{w}\cdot\bar{s}^{(i)} + b) + \xi_i \geqslant 1$, $i = 1, 2, \dots m$

$$Y^{(i)}(\bar{w}\cdot\bar{s}^{(i)} + b) \geqslant 1 - \xi_i$$

$$\xi_i \geqslant 0, \quad i = 1, 2, \dots m. \qquad (1)$$



$$L(\bar{w}, b, \xi, \alpha, \beta) = \underbrace{\frac{1}{2}\|\bar{w}\|^2 + C\sum_{i=1}^{m}\xi_i}\; - \sum_{i=1}^{m}\alpha_i\left\{ Y^{(i)}(\bar{w}\cdot\bar{s}^{(i)} + b) - 1 + \xi_i \right\} - \sum_{i=1}^{m}\beta_i\xi_i \qquad (2)$$

where α_i's and β_i's are Lagrangian multiplier $\geqslant 0$.

Like earlier setting derivaties w.r.t $\bar{w}$ and b to zero, substituting them back in and simplifying we should obtain the Dual problem:

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m} Y^{(i)} Y^{(j)} \alpha_i \alpha_j \langle \bar{s}^{(i)}, \bar{s}^{(j)} \rangle. \qquad (3)$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq C, \quad i = 1, 2, \dots m$$

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

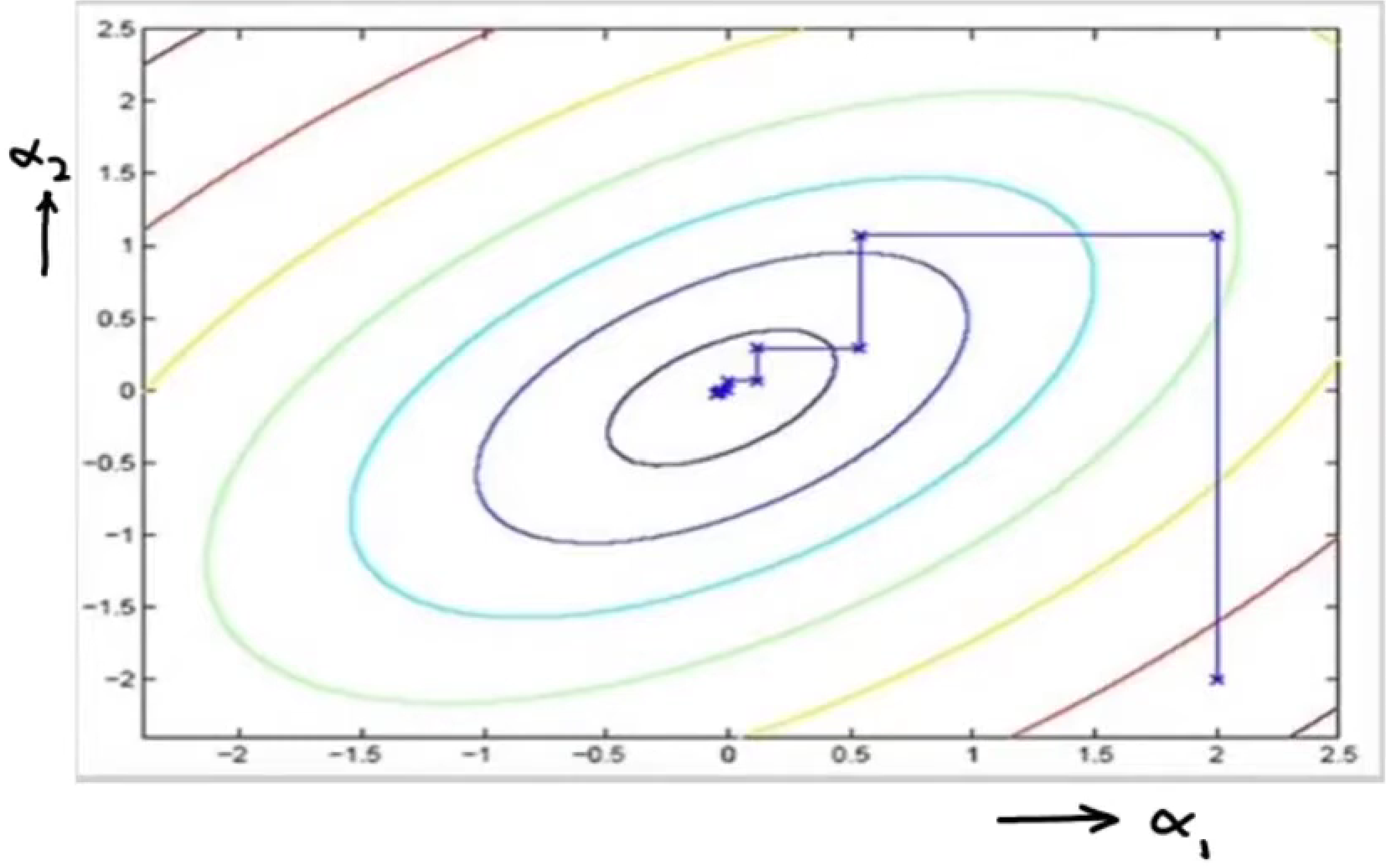
Prediction:

After solving the dual problem we can use the same decision rule:

$\bar{w}\cdot\bar{s}^{(i)} + b \geqslant 0$, $Y^{(i)} = 1$, $\bar{s}^{(i)}$ in +ve class.

< 0, $Y^{(i)} = -1$, $\bar{s}^{(i)}$ in -ve class.

Co-ordinate Ascent Algorithm



Consider solving an unconstrained optimization problem:

$$\max_{\alpha} W(\alpha_1, \alpha_2 \ldots\ldots\ldots\ldots \alpha_m)$$

Repeat till Convergence: {

For $i = 1, 2, \ldots m$ {

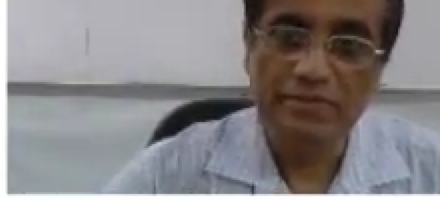
$$\alpha_i := \arg\max_{\hat{\alpha}_i} W(\alpha_1, \alpha_2, \ldots \alpha_{i-1}, \hat{\alpha}_i, \alpha_{i+1}, \ldots \alpha_m)$$

}

}

When the function W is such that "arg max" in the inner loop can be performed efficiently, then this algorithm will be an efficient algorithm.

Sequential Minimal Optimization (SMO)Algorithm



John Platt : Sequential Minimal Optimization : a Fast Algorithm for training SVM, 1998, Microsoft Research , Technical report.
The algorithm is very useful to train a SVM.

Recap: KKT dual –complementarity conditions will now be useful to test the convergence of the SMO algorithm.

$$\alpha_i^* \underbrace{\left[Y^{(i)} (\bar{w}^* \cdot \bar{s}^{(i)} + b) - 1 \right]}_{g_i(\bar{w}^*)} = 0 \qquad L = \frac{1}{2}\|\bar{w}\|^2 - \alpha_i \cdot g_i(\bar{w}^*)$$

$$\alpha_i = 0 \Rightarrow Y^{(i)} (\bar{w} \cdot \bar{s}^{(i)} + b) \geq 1 \quad \text{—— (4)}$$

$$\alpha_i = C \Rightarrow Y^{(i)} (\bar{w} \cdot \bar{s}^{(i)} + b) \leq 1 \quad \text{—— (5)}$$

$$0 < \alpha_i < C \Rightarrow Y^{(i)} (\bar{w} \cdot \bar{s}^{(i)} + b) = 1 \quad \text{—— (6)}$$

SMO Algorithm

Dual Problem to be solved :

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m} Y^{(i)} Y^{(j)} \alpha_i \alpha_j \langle \bar{s}^{(i)} \cdot \bar{s}^{(j)} \rangle \quad \text{——— (7)} \quad \Rightarrow a\alpha^2 + b\alpha + c = W(\alpha)$$

$$\text{st}: \quad 0 \le \alpha_i \le C, \; i = 1, 2, \dots m \quad \text{——— (8)}$$

$$\sum_{i=1}^{m} \alpha_i Y^{(i)} = 0 \quad \text{——— (9)}$$

> Say, now we have a set of $\alpha_i's$ that satisfy the constraints (8) & (9).

> Next say, we want to vary only α_1, keeping $\alpha_2, \alpha_3, \dots \alpha_m$ fixed.

> Take a co-ordinate ascent step and reoptimize the objective w.r.t α_1

> Can the algo make a progress ? No $\quad \alpha_1 Y^{(1)} + \alpha_2 Y^{(2)} + \dots \alpha_m Y^{(m)} = 0 \Rightarrow \alpha_1 Y^{(1)} + \sum_{i=2}^{m} \alpha_i Y^{(i)} = 0$

Why ? — Look on constraint (9) $\Rightarrow \quad \alpha_1 Y^{(1)} = -\sum_{i=2}^{m} \alpha_i Y^{(i)} \Rightarrow$ Multiply both sides by $Y^{(1)}$

$$\alpha_1 \underbrace{Y^{(1)} \cdot Y^{(1)}}_{=1} = -Y^{(1)} \sum_{i=2}^{m} \alpha_i Y^{(i)} \qquad \left[\begin{array}{l} \text{since } Y^{(1)} \in \{-1, 1\} \\ (Y^{(1)})^2 = 1 \end{array} \right]$$

$$\alpha_1 = -Y^{(1)} \sum_{i=2}^{m} \alpha_i Y^{(i)}$$

So α_1 is exactly determined by other $\alpha_i's$

So the algorithm cannot make any progress.

SMO Algorithm

So we have seen that without violating constraint (9), Co-ordinate ascent algorithm can't make any progress.

This motivates SMO algorithm to take at least two α_i's simultaneously for updating satisfying the constraints.

Repeat till convergence {

1. Select some pair α_i, α_j using some heuristic that tries to pick the two which would make biggest progress towards global maximum.

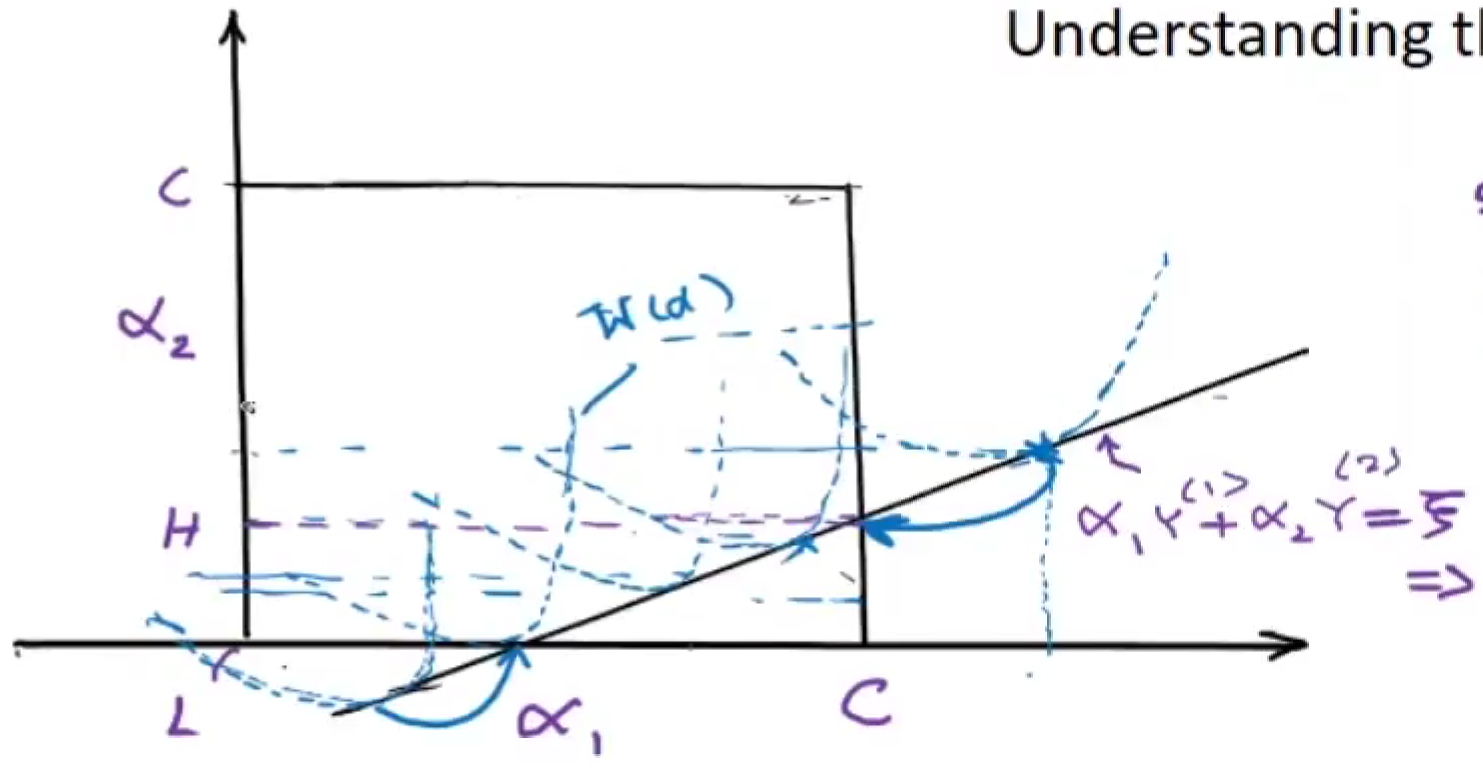
2. Reoptimize $W(\alpha)$ w.r.t α_i and α_j, holding all other α_k's $(k \neq i, j)$ fixed.

}

You can test the convergence of the algorithm by checking the KKT conditions as given in equations (4-6).

SMO Algorithm

Understanding the main ideas of SMO Algorithm



Suppose we decide to optimize $W(\alpha)$ by varying α_1 and α_2, keeping all other fixed. From constraint eqn (9) $\Rightarrow$

$$\alpha_1 Y^{(1)} + \alpha_2 Y^{(2)} + \sum_{i=3}^{m} \alpha_i Y^{(i)} = 0$$

$$\Rightarrow \alpha_1 Y^{(1)} + \alpha_2 Y^{(2)} = -\sum_{i=3}^{m} \alpha_i Y^{(i)} = \xi \text{ (say)} \quad \text{———} \quad (10)$$

From constraint (8), we know that α_1 and α_2 must lie within the box $[0, C] \times [0, C]$

Also to obey the box constraint and straight line simultaneously, $L \leq \alpha_2 \leq H$.

Using equation (10), we can write α_1 as a function of $\alpha_2 \Rightarrow \alpha_1 = (\xi - \alpha_2 Y^{(2)}) \cdot Y^{(1)}$ ——— (11)

Hence the objective function $W(\alpha_i) = W((\xi - \alpha_2 Y^{(2)}) Y^{(1)}, \alpha_2, \alpha_3 \cdots \alpha_m)$, —

This is a quadratic function in α_2 of the form $\boxed{a\alpha_2^2 + b\alpha_2 + c}$. Now without box constraint or $L \leq \alpha_2 \leq H$ we can easily find α_2 by calculus method. Let it be α_2^{actual}. But if we want to maximize W subject to box constraint?

SMO Algorithm

Let $\alpha_2^{new, unclipped}$ denote the resulting value of α_2. If now you want to maximize $W(\alpha)$ subject to the box constraint, then simply we can use the value of $\alpha_2^{new, unclipped}$ and clipping it to lie in the [L, H] interval to get

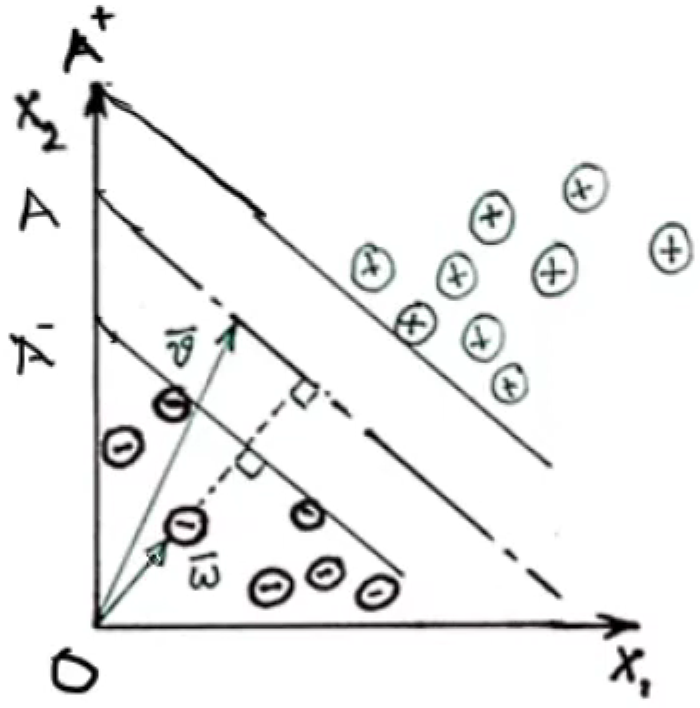
$$\alpha_2^{new} = \begin{cases} H & \text{if } \alpha_2^{new, unclipped} > H \\ \alpha_2^{new, unclipped} & \text{if } \alpha_2^{new, unclipped} \le H \\ L & \text{if } \alpha_2^{new, unclipped} < L \end{cases} \qquad \text{—— (12)}$$

After getting all α^*'s, W and b can be found and the SVM is assumed to be trained. It can now be used for prediction.

> $\bar{W}^* \Rightarrow$ From the solution of the Primal Prob $= \sum_{i=1}^{m} Y^{(i)} \alpha_i \bar{S}^{(i)}$

> Knowing $\bar{W}$, b can be found

How to find : b?



Recall Primal Problem Constraint:

$$Y^{(i)}\left(\bar{W}\cdot\bar{S}^{(i)}+b\right)-1=0 \text{, for samples on the gutter.}$$

$$\Rightarrow \left(Y^{(i)}\right)^2\left(\bar{W}\cdot\bar{S}^{(i)}+b\right)=1\cdot Y^{(i)}$$

$$\Rightarrow b = Y^{(i)}-\bar{W}\cdot\bar{S}^{(i)}$$

$$OA^- = b^- = \max_{i:Y^{(i)}=-1}\left(Y^{(i)}-\bar{W}\cdot\bar{S}_-^{(i)}\right)$$

$$OA^+ = b^+ = \min_{i:Y^{(i)}=1}\left(Y^{(i)}-\bar{W}\cdot\bar{S}_+^{(i)}\right)$$

$$b^* = OA = \frac{OA^+-OA^-}{2} = \frac{b^+-b^-}{2} = \frac{\min_{i:Y^{(i)}=1}\left(Y^{(i)}-\bar{W}\cdot\bar{S}_+^{(i)}\right)-\max_{i:Y^{(i)}=-1}\left(Y^{(i)}-\bar{W}\cdot\bar{S}_-\right)}{2} \quad \text{--- (13)}$$

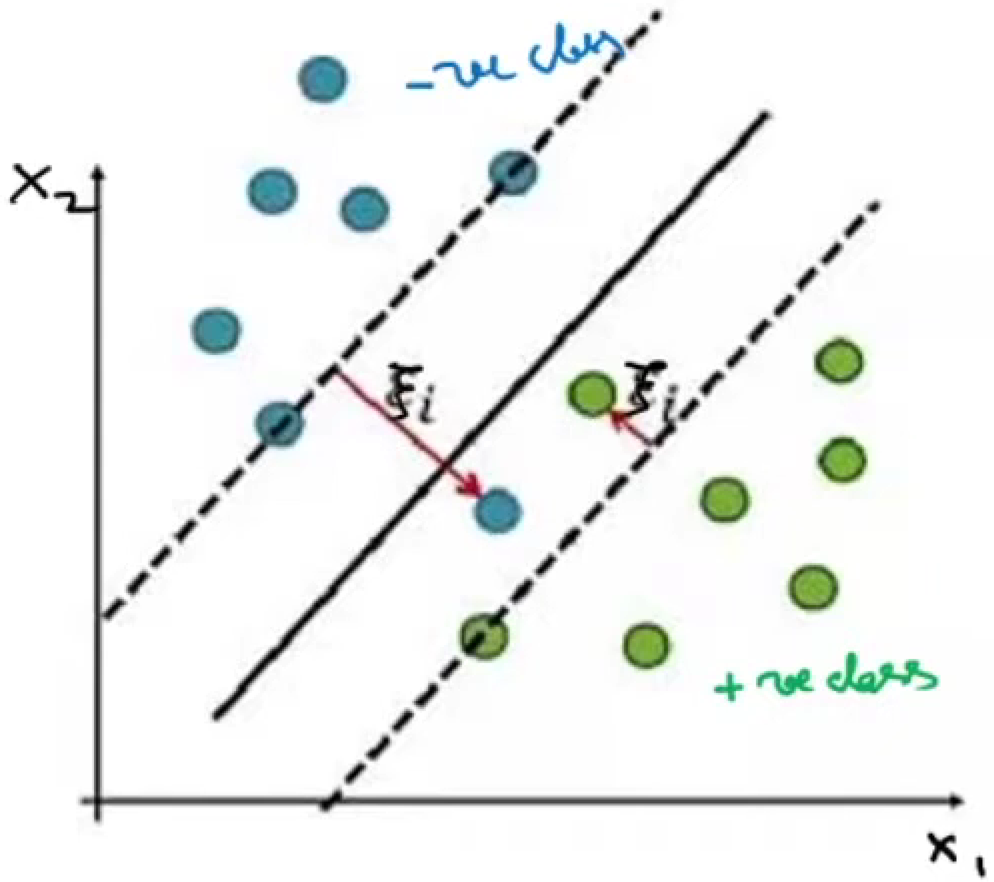
How to Predict?

From decision rule $\bar{W}\,\bar{S}^{(i)}+b=\sum_{i=1}^{m}\alpha_i Y^{(i)}\left\langle \underbrace{\bar{S}^{(i)}}_{\text{support vector}}\cdot\underbrace{\bar{S}^{(j)}}_{\text{any sample}}\right\rangle+b$ --- (14)

If expression (14) >0, sample belongs to +ve class

<0, sample belongs to −ve class.

Regularization in SVM



Soft Margin SVM.

Let us take the Primal Prob:

$$\min_{\bar{w}, \xi_i, b} \ \frac{1}{2}\|\bar{w}\|^2 + C\sum_{i=1}^{n} \xi_i$$

$$\text{s.t.} \quad \underline{Y^{(i)}(\bar{w}\cdot\bar{s}^{(i)} + b) \geq 1 - \xi_i} \qquad (8)$$

$$\underbrace{\xi_i \geq 0}, \quad i = 1, 2, \ldots n.$$

$$L_p(\bar{w}, b, \alpha, \xi, \beta) = \frac{1}{2}\|\bar{w}\|^2 + C\sum \xi_i - \sum_{i=1}^{m} \alpha_i \left(Y^{(i)}(\bar{w}\cdot\bar{s}^{(i)} + b) - 1 + \xi \right) - \sum \beta_i \xi_i \qquad (9)$$

Dual problem:

$$\max_{\alpha} W(\alpha) = \sum \alpha_i - \frac{1}{2}\sum_i\sum_j Y^{(i)} Y^{(j)} \alpha_i \alpha_j \langle \bar{s}^{(i)} \cdot \bar{s}^{(j)} \rangle$$