The distance of a training point $\underline{x_i}$ to a hyperplane $(\underline{w}, b)$ is

$$\frac{|\langle \underline{w}, \underline{x_i} \rangle + b|}{\|\underline{w}\|}$$

Assuming that $\|\underline{w}\| = 1$, the distance of the closest point to the hyperplane is $\min_{i \in [m]} |\langle \underline{w}, \underline{x_i} \rangle + b|$

The hard SVM learning rule says that pick up the hyperplane which maximizes the margin.

abs value.

$$\underset{\substack{(\underline{w}, b) \\ \|\underline{w}\| = 1}}{\arg\max} \underbrace{\left[\underbrace{\min_{i \in [m]}}_{\text{min}} \underbrace{|\langle \underline{w}, \underline{x_i} \rangle + b|}_{\text{distance}} \right]}_{\text{margin}}$$

such that $\forall i \quad \boxed{y_i(\langle \underline{w}, \underline{x_i} \rangle + b) > 0}$

for the separable case when the hyperplane can correctly classify all the examples.

For the separable case we are sure that

$$y_i(\langle \underline{w}, \underline{x_i} \rangle + b) > 0$$

ground truth

prediction.

1st formulation $\therefore$ $\underline{|\langle \underline{w}, \underline{x_i} \rangle + b|} = \underline{y_i(\langle \underline{w}, \underline{x_i} \rangle + b)}$

$\therefore$ the equivalent problem is $\underset{\substack{(\underline{w}, b) \\ \|\underline{w}\| = 1}}{\arg\max} \underbrace{\min_{i \in [m]} y_i(\langle \underline{w}, \underline{x_i} \rangle + b)}_{\text{margin}}$

Hard SVM learning rule.

In another formulation of Hard SVM, we assume that $\|w\| \neq 1$

Therefore the distance of a point x_i to the hyperplane is

$$\frac{|\langle w, x_i\rangle + b|}{\|w\|}$$

The margin for a given training set is therefore

$$\min_{i\in[m]} \left[\frac{|\langle w, x_i\rangle + b|}{\|w\|} \right]$$

We now assume that the smallest value for

achieved through scaling of w and b.

$$\min_i |\langle w, x_i\rangle + b| = \min_i y_i(\langle w, x_i\rangle + b) = 1.$$

$$\text{i.e. } \forall i \quad y_i(\langle w, x_i\rangle + b) \geq 1$$

Again consider

$$\arg\max_{(w,b)} \left[\min_{i\in[m]} \left(\frac{y_i(\langle w, x_i\rangle + b)}{\|w\|} \right) \right]$$

$$= \arg\max_{(w,b)} \left(\frac{1}{\|w\|}\right) \underbrace{\min_{i\in[m]} y_i(\langle w, x_i\rangle + b)}_{=1}$$

$$= \arg\max_{(w,b)} \left(\frac{1}{\|w\|}\right)$$

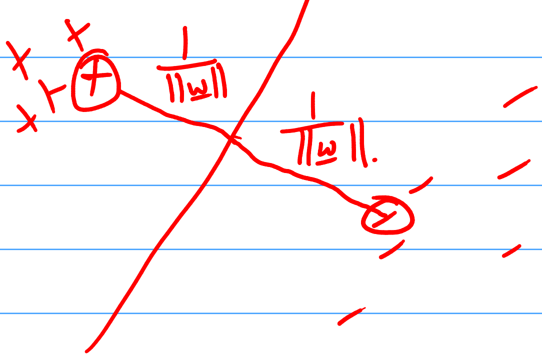
$\therefore$ the 2nd formulation of Hard margin learning rule

$$(w_0, b) = \arg\min_{(w,b)} \|w\|^2 \quad \text{such that } \forall i \ \left(y_i(\langle w, x_i\rangle + b)\right) \geq 1$$

Not normalized. Quadratic objective.

$1\left(\frac{1}{\|\underline{w}\|}\right)$: Margin.

$\frac{1}{\|\underline{w}\|}$: unit to measure the margin.



$\underline{w} \in \mathbb{R}^d$

$\underline{x_i} \in \mathbb{R}^d$.

In many applications d is very large.

The 2^{nd} formulation is called as the quadratic formulation of the Hard Margin SVM problem.

Final output $(\hat{\underline{w}} \quad \hat{b})$

$$\hat{\underline{w}} = \frac{\underline{w}_0}{\|\underline{w}_0\|} \qquad \hat{b} = \frac{b_0}{\|\underline{w}_0\|}$$

Here we have enforced that the margin is 1.

Let's verify that the solution to the first formulation will also be a valid ~~formulation~~ solution to the 2^{nd} formulation and vice versa.

Let $\underline{w}^*, b^*$ be a solution of the first formulation

$\|\underline{w}^*\| = 1$. Since $\forall i \quad y_i(\langle \underline{w}^*, \underline{x}_i \rangle + b^*) > 0$, $\Rightarrow$ correct classification

Let $\gamma^* = \min_i y_i(\langle \underline{w}^*, \underline{x}_i \rangle + b^*)$

$\therefore \forall i$ we have $y_i(\langle \underline{w}^*, \underline{x}_i \rangle + b^*) \geq \gamma^*$

Normalizing both sides by γ^* we have

$$\underbrace{y_i\left(\left\langle \frac{\underline{w}^*}{\gamma^*}, \underline{x}_i \right\rangle + \frac{b^*}{\gamma^*}\right) \geq 1.}_{\text{These are the constraints of the } 2^{nd} \text{ formulation.}}$$

Candidate

$\therefore \left(\frac{\underline{w}^*}{\gamma^*}, \frac{b^*}{\gamma^*}\right)$ satisfies all the constraints of the 2^{nd} formulation

but we are not sure that whether this leads to the min value of the objective fn of the 2^{nd} formulation

$\hat{w}, \hat{b}$: normalized.

The unnormalized answer given by the 2^{nd} formulation is $\underline{w_0}$.

$\therefore$ $\underline{w_0}$ has the minimum norm among all the $\underline{w}$ vectors that satisfy the constraints of the 2^{nd} formulation.

$$\therefore \underbrace{\|\underline{w_0}\|}_{\text{min}} \leq \underbrace{\left\| \frac{\underline{w}^*}{\gamma^*} \right\|}_{\text{Candidate}} = \frac{1}{\gamma^*} \qquad \therefore \boxed{\|\underline{w}^*\| = 1}$$

Now consider the final answer of the 2^{nd} formulation

Normalized $\hat{\underline{w}}$ and $\hat{b}$ $\|\hat{\underline{w}}\| = 1$.

We should check whether it satisfies the constraints of the first formulation

2nd formulation

$$y_i\left(\langle \hat{\underline{w}}, \underline{x_i} \rangle + \hat{b}\right) = \underbrace{\frac{1}{\|\underline{w_0}\|}}\ \underbrace{y_i\left(\langle \underline{w_0}, \underline{x_i} \rangle + b_0\right)}_{\geq 1}$$

As per the 2^{nd} formulation,

$$y_i\left(\langle \underline{w_0}, \underline{x_i} \rangle + b_0\right) \geq 1 \quad \forall i$$

$\gamma^* > 0$ positive.

$$\therefore \underbrace{\frac{1}{\|\underline{w_0}\|}}\ \underbrace{y_i\left(\langle \underline{w_0}, \underline{x_i} \rangle + b_0\right)}_{\geq 1} \geq \frac{1}{\|\underline{w_0}\|} \geq \gamma^* \qquad \therefore \|\underline{w_0}\| \leq \frac{1}{\gamma}$$

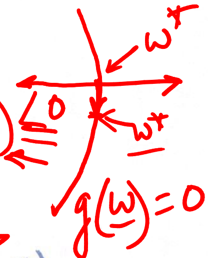
$\therefore$ $(\hat{\underline{w}}, \hat{b})$ satisfy the constraints of the 1^{st} formulation.

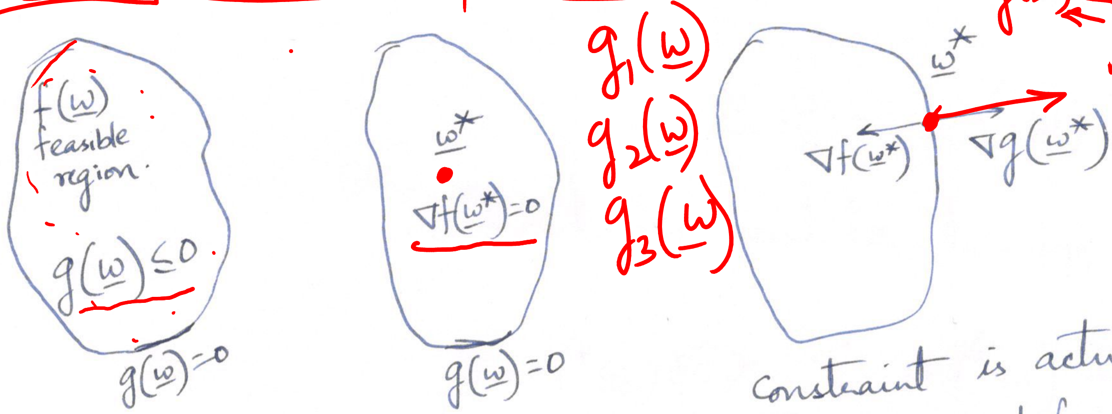
Further, $\|\hat{\underline{w}}\| = 1$

$\Rightarrow$ $(\hat{\underline{w}}, \hat{b})$ is an optimal solution to the first formulation.

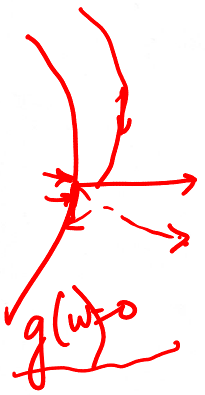
Dual Formulation of SVM

min $f(\underline{w})$ subject to $g(\underline{w}) \le 0$ $\qquad h(w) \ge \gamma$.





Constraint is active because $\underline{w}^*$ satisfies $g(\underline{w}^*) = 0$



α : scalar

$$\nabla f(\underline{w}^*) = -\alpha \nabla g(\underline{w}^*)$$

Hold at $\underline{w}^*$ $\rightarrow$ $\boxed{\nabla f(\underline{w}^*) + \alpha \nabla g(\underline{w}^*) = 0}$

When there are several constraints $g_i(\underline{w}) \le 0$, the feasible region is the intersection of all the regions $g_i(\underline{w}) \le 0$. $\forall i$

Assuming that f and g_i are differentiable,

$$\nabla f(\underline{w}^*) = -\sum_{i \in I} \alpha_i g_i(\underline{w}^*)$$

where I is the set of constraints which are active at $\underline{w}^*$

$$\nabla f(\underline{w}^*) + \sum_{i \in I} \alpha_i g_i(\underline{w}^*) = \underline{0}$$

The Hard SVM learning rule specifies

$f(\underline{w}) = \frac{1}{2}\|\underline{w}\|^2$ and $g_i(\underline{w}) \le 0$ $\quad g_i = 1 - y_i \langle w, x_i \rangle$