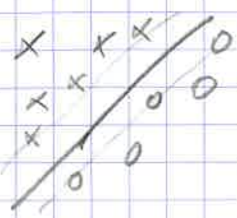
SVM₁ Separable case



First, a remark: • $w\vec{x}+b=0$ defines a plane

• $d_\pm = \dfrac{\vec{w}\vec{x}+b}{\|w\|_2}$ is the distance between the point $\vec{x}$ and the plane $(\vec{w},b)$. (signed distance)

• On one side of the plane, $d_\pm \geqslant 0$, on the other side, we have $d_\pm \leqslant 0$.

• If $\vec{x}_n$ is well classified, $\text{sign}(d_\pm) = \pm 1 = t_n$.

• So, for well-classified points, $d_\pm(\vec{x}_n)\cdot t_n = \dfrac{t_n(w\vec{x}_n+b)}{\|w\|_2} = d$ is the (positive) distance between $\vec{x}_n$ and the plane.

$$d = \frac{t_n(wx_n+b)}{\|w\|} \geqslant 0 \quad \text{is a distance (positive)}$$

• Rescaling: the plane $(\vec{w},b)$ is the same as the plane encoded in $(c\vec{w}, cb)$, $(\forall c \in \mathbb{R}^{+*} \ (c>0))$:

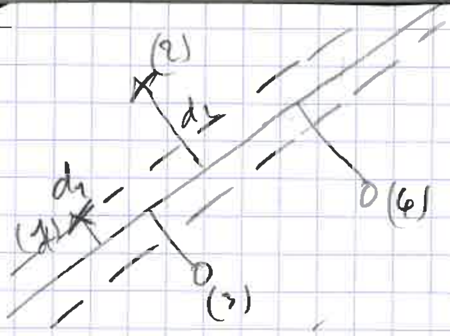
we can check this: $d' = \dfrac{t_n(\vec{w}.\cancel{c}x_n + b\cancel{c})}{\cancel{c}\|w\|} = d$.

• So, we can always choose c (rescale $\|w\|$) such that we have $t_n(\vec{w}\vec{x}_n + b) \geqslant 1, \ \forall n$ (the choice of "1" is arbitrary)

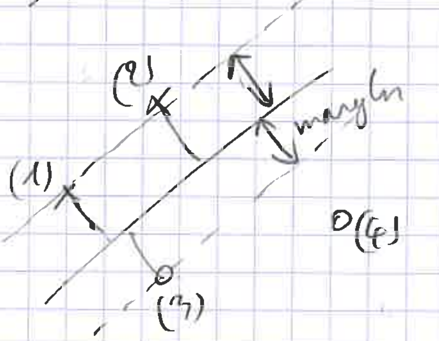
• There are many solutions to correctly classify a linearly separable set of data.

• We call "margin" the minimal distance between the points and the plane:

$$\text{margin} = \min_n (d(\vec{x}_n)) = \min_n\left(\frac{t_n(\vec{w}\vec{x}_n+b)}{\|\vec{w}\|_2}\right)$$



d_1 is minimal, so the margin is $marg = d_1 = d(\vec{x}_1)$ in this example.



here, d_3 is the smallest dist (almost equal to d_1), so margin $= d_3 = d(\vec{x}_3)$.

This $(\vec{w}, b)$ seems better than the first one.

- The core idea of SVM (also re-named "Separators with Vast Margin") is to find the solution with the largest margin, which is expected to work better than those with small margin.

- So, we define the SVM solut° (among all the sol° to the pbm of linear separation) as:

$$w^*_{SVM} = \underbrace{\operatorname{argmax}_{w,b}\left(\underbrace{\min_n\left(\underbrace{\frac{t_n(w\vec{x}_n + b)}{\|w\|_2}}_{\text{distance}}\right)}_{\text{margin}}\right)}_{\text{the largest margin (or its corresp. } \vec{w}\text{, rather)}}$$

From there, it's just maths (optimizat°).

$$b^*, w^*_{SVM} = \underset{w,b}{\operatorname{argmax}} \left(\min_n \left(\frac{t_n(\vec{w}\vec{x}_n + b)}{\|w\|_2} \right) \right)$$

$$= \underset{w,b}{\operatorname{argmax}} \left(\frac{1}{\|w\|_2} \min_n \left(\underbrace{\qquad\qquad}_{} \right) \right)$$

the distances d_n can be set to ≥ 1 by rescaling $\vec{w}, b$ by a constant $c > 0$.

$$= \underset{w,b \text{ such that } t_n(w\vec{x}_n + b) \geq 1, \forall n}{\operatorname{argmax}} \left(\frac{1}{\|w\|_2} \right)$$

because $\frac{1}{2x^2}$ is $\Downarrow$ for all $x > 0$, argmax becomes argmin

$$= \underline{\underline{\operatorname{argmin}}} \qquad \left(\frac{1}{2} \|W\|_2^2 \right)$$

$$= \underset{W,b}{\operatorname{argmin}} \underbrace{\left(\frac{1}{2} \|w\|_2^2 - \sum_n \overset{\geq 0}{\alpha_n} \left(\overbrace{t_n(wx_n + b) - 1}^{\geq 0} \right) \right)}_{\triangleq \mathcal{L}}$$

$$= \operatorname{argmin}_{w,b}(\mathcal{L})$$

The α_n are positive, they are <u>Lagrange multipliers</u>.

We can now solve for $\vec{w}, b$:

$$\vec{\nabla}_w \mathcal{L} = \vec{0} \iff \frac{\cancel{2}}{\cancel{2}} \vec{w} - \sum_n \alpha_n t_n (\vec{x}_n) = \vec{0}$$

$$\iff \vec{w} = \sum_n \alpha_n t_n \vec{x}_n$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \iff \sum_n \alpha_n t_n = 0$$

We now need to find the α_n, so we just moved (re-wrote) the problem. At its $\nabla \mathcal{L} = \vec{0}$ point (extremum), $\mathcal{L}$ can be written: (we replace $\vec{w}$ with its expression in α_n)

$$\mathcal{L} = \frac{1}{2} \|w\|_2^2 - \sum_n \vec{w} \underbrace{\alpha_n t_n x_n}_{= \vec{w}} + \sum_n \alpha_n t_n + \sum_n \alpha_n$$

$$= \underbrace{\frac{1}{2} \vec{w} \cdot \vec{w} - \vec{w} \cdot \vec{w}}_{-1/2\, \vec{w} \cdot \vec{w}} + \underbrace{\sum_n \alpha_n t_n}_{= 0} + \sum_n \alpha_n$$

$$= -\frac{1}{2}\left(\sum_n \alpha_n x_n t_n\right)\left(\sum_m \alpha_m x_m t_m\right) + 0 + \sum_n \alpha_n$$

$$= \sum_n \alpha_n - \frac{1}{2}\sum_n\sum_m \left(\alpha_n \alpha_m t_n t_m \underbrace{\vec{x}_n \vec{x}_m}\right)$$

<u>dual form</u>.

$\vec{x}_n \vec{x}_m$ can be replaced with $K(x_n, x_m)$

This is then the Kernelized SVM

Note that we needed this dual form.

. At prediction time, we would have:

$$y^{pred} = sign(\vec{w}\,\vec{x}_{test}) = sign\left(\sum_n \alpha_n t_n \underbrace{\vec{x}_n \cdot \vec{x}_{test}}\right)$$

we can insert $K(\vec{x}_n, \vec{x}_{test})$ here

. The goal is then to <u>find the best α_n</u>.

. The KKT conditions tell us things about the α_n that do define the $\vec{w}^*_{SVM}$, $\alpha^*_m = \underset{\alpha_1,\dots,\alpha_N}{\text{argmin}}(\mathcal{L})$

$$\text{KKT:}\quad \begin{cases} \alpha_n \geq 0 \\ t_n d'(x_n) - 1 \geq 0 \iff t_n(\vec{w}\,\vec{x}_n + b) \geq 1 \\ \alpha_n(t_n d'(x_n) - 1) = 0 \iff \alpha_n(t_n(wx_n + b) - 1) = 0 \end{cases}$$

$$\left(d'(x_n) = \frac{\vec{w}x + b}{\cancel{\|w\|}}\right)$$

i.e the α_n are Lagrange multipliers, $\alpha_n \geq 0$

i.e each inequality is either <u>saturated</u> $(t_n d(x_n) - 1 = 0)$,

↳ point $\vec{x}_n$ is <u>on the margin</u>

or the α_n is 0 (point ignored)

There are 2 kind of points:

$$\begin{cases} \alpha_n = 0, & t_n d(x_n) - 1 > 0 & : \text{points well. class. outside the margin} \\ \alpha_n > 0, & \text{————} = 0 & : \text{on the margin} \end{cases}$$

These are the <u>support vectors</u>!

Note that the sol° is $w = \sum_n \alpha_n t_n x_n = \sum_{s.v} \cdots$ — a linear combinat° of the support vectors only!

b can be computed: