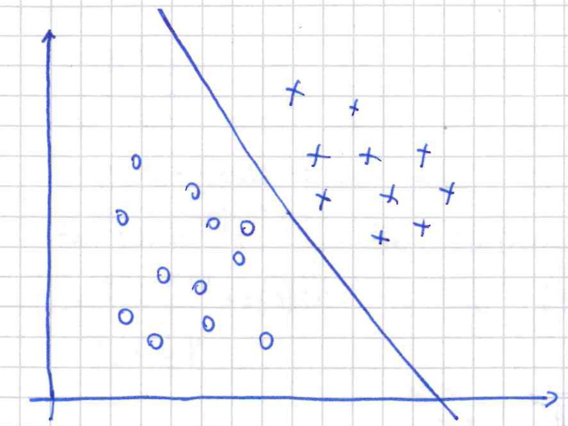
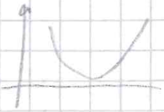
# SVM



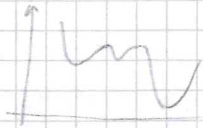
CLASSIFICATION TASK

$y : \{x, o\}$

Ribadire che le funzioni convesse sono solo quelle



e quelle



non sono convesse

(se proprio vogliamo possiamo dire che sono convesse su intervalli)

SGD ? spiegare bene

$$y = ax + q = \theta_0 + \theta_1 x$$

$$q = \theta_0$$

$$a = \theta_1$$

In general we want to find an hyperplane $H = \{ x \in \bar{x}^T \bar{\theta} = \emptyset \}$

In SVM we want to find a couple of hyperplanes such that

1) they are parallel and they separate the two classes
2) the distance between the two hyperplanes is the maximum distance

- DATA SET LINEARLY SEPARABLE

$$\bar{x} \in \mathbb{R}^d, \quad y \in \{+1, -1\}$$

Let's suppose that exists an hyperplane $\bar{w}^T \bar{x} + \theta > 0$ if $y = 1$ and $\bar{w}^T \bar{x} + \theta < 0$ if $y = -1$ $\Rightarrow$

(The data set is linearly separable)

It is possible to proove that the previous condition is satisfied if and only if

$$\hat{w}^T \bar{x} + \hat{\theta} \geq 1 \quad , y = 1$$

and

$$\hat{w}^T \bar{x} + \hat{\theta} \leq -1 \quad , y = -1$$

$\rightarrow$ N.B.: we will use $w$ instead of $\hat{w}$ and $\theta$ instead of $\hat{\theta}$

USING A SET OF OUR POINT (TRAINING SET) we want to determine $w$ and $\theta$ for any point $\bar{x}$

(The dimension of the data set is $N$) such that

$$y_i \left( \bar{w}^T \bar{x}^i + \theta \right) \geq 1 \qquad \text{for all } \bar{x}^i \qquad i = 1, \dots, N$$

Moreover the distance between the two hyperplanes has to be the maximum distance

So we consider $H$, $H_+$ and $H_-$ as three parallel hyperplanes

$$H = \{\bar{x} \in \mathbb{R}^d : \bar{w}^T\bar{x} + \theta = 0\}$$

$$H_+ = \{\bar{x} \in \mathbb{R}^d : \bar{w}^T\bar{x} + \theta = 1\}$$

$$H_- = \{\bar{x} \in \mathbb{R}^d : \bar{w}^T\bar{x} + \theta = -1\}$$

$r: ax + by + c = 0$
$P(x_P, y_P)$

$$d(P, r) = \frac{|ax_P + by_P + c|}{\sqrt{a^2 + b^2}}$$

if $\bar{x} \in H_+$

$$\text{dist}(x, H) = \frac{|\bar{w}^T\bar{x} + \theta|}{\|\bar{w}\|} = \frac{1}{\|w\|}$$

if $\bar{x} \in H_-$

$$\text{dist}(x, H) = \frac{|\bar{w}^T\bar{x} + \theta|}{\|\bar{w}\|} = + \frac{1}{\|w\|}$$

So $\text{dist}(H_+, H_-) = \dfrac{2}{\|w\|}$

In the TRAINING phase we need to solve the following problem

$$\min_{(w,\theta)} \frac{1}{2}\bar{w}^T\bar{w} \qquad (*)$$

$$\text{subject to} \quad y_i(\bar{x}^T\bar{x}^i + \theta) \geq 1, \quad i = 1, \ldots, N$$

Obtained the optimal $w^*$ and $\theta^*$ we can write the CLASSIFIER as

$$h(x) = \text{sign}(f(x)) = \begin{cases} 1 & \text{, if } f(x) > 0 \\ -1 & \text{, if } f(x) < 0 \end{cases}$$

$$f(x) = \bar{w}^T\bar{x} + \theta$$

- DATA SET NOT LINEARLY SEPARABLE

Actually, in real problems data are not linearly separable and in this case there is no solution for the problem (*)
For this reason we need to modify that problem, introducing the possibility for the training data to be missclassified
The new problem is the following

$$\min_{w,\sigma,\xi} \quad \frac{1}{2}\bar{w}^T w + C \sum_{i=1}^{N} \xi_i$$

$$\text{subject to} \quad y_i\left(\bar{w}^T \bar{x}^i + \sigma\right) \geq 1 - \xi_i \quad , \quad i = 1, \dots, N$$

$$\xi_i \geq \emptyset \quad , \quad i = 1, \dots, N$$

- $\frac{1}{2}\bar{w}^T\bar{w}$ controls the "ability" to learn of the system, maximizing the distance $dist(H_+, H_-)$
- $\bar{e}^T\bar{\xi}$ controls the missclassification, C penalizes the missclassified data.

$\bar{e}^T = [1, 1, \dots, 1] \in \mathbb{R}^N$

Actually instead of solving the previous minimization problem we solve the, so called, DUAL PROBLEM

$$\min_{\alpha} \quad \frac{1}{2}\bar{\alpha}^T Q \bar{\alpha} - \bar{e}^T\bar{\alpha}$$

$$\text{subject to} \quad \bar{y}^T\bar{\alpha} = \emptyset$$

$$0 \leq \alpha \leq C$$

$$\text{where} \quad Q_{ij} = y_i y_j x^{iT} x^j$$

once we obtain $\alpha \Rightarrow$ $w = \sum_{i=1}^{N} \alpha_i y_i x^i$

$$f(x) = \bar{w}^T x + \sigma = \sum_{i=1}^{N} \alpha_i y_i x^{iT} x + \sigma$$

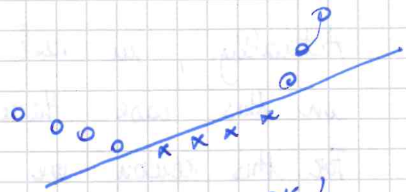
Vectors $\bar{x}^i$ such that $0 < \alpha_i < C$ are the support vector

• NON LINEAR MODEL (NON LINEARLY SEPARABLE)

ARE NOT LINEARLY SEPARABLE

ORIGINAL SPACE

$$\phi : x \in \mathbb{R} \to \begin{bmatrix} x \\ x^2 \end{bmatrix} \in \mathbb{R}^2$$



OK!

FEATURE SPACE

$$\min_{w,\theta,\xi} \quad \frac{1}{2}\langle w, w \rangle + C \sum_{i=1}^{N} \xi_i$$

subject to $\quad y_i\left(\langle w, \phi(x^i) \rangle + \theta\right) \geq 1 - \xi_i \quad , \quad i = 1, \dots, N$

$\qquad \xi_i \geq 0 \quad , \quad$ "

DUAL

$$\min_{\alpha} \quad \frac{1}{2}\bar{\alpha}^T Q \bar{\alpha} - \bar{e}^T \alpha$$

s.t. $\quad \bar{y}^T \bar{\alpha} = \emptyset$

$\qquad 0 < \alpha < C$

$$w^* = \sum_{i=1}^{N} \bar{\alpha}_i^* y_i \, \phi(x^i)$$

where $\quad Q_{ij} = y_i y_j K_{ij} \qquad K_{ij} = K(x^i, x^j) := \langle \phi(x^i), \phi(x^j) \rangle$

$K : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ is the Kernel function

→ IN THE DUAL PROBLEM IT IS NOT NECESSARY TO KNOW exactly the functio $\phi(\cdot)$ but we can only KNOW

$$K(\bar{x}, \hat{x}) = \langle \phi(\bar{x}), \phi(\hat{x}) \rangle = \phi(\bar{x})^T \phi(\hat{x})$$

$$h(x) = \text{sign}(f(x))$$

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i \langle \phi(x^i), \phi(x) \rangle + \theta = \sum_{i=1}^{N} \alpha_i y_i \, k(x^i, x) + \theta$$