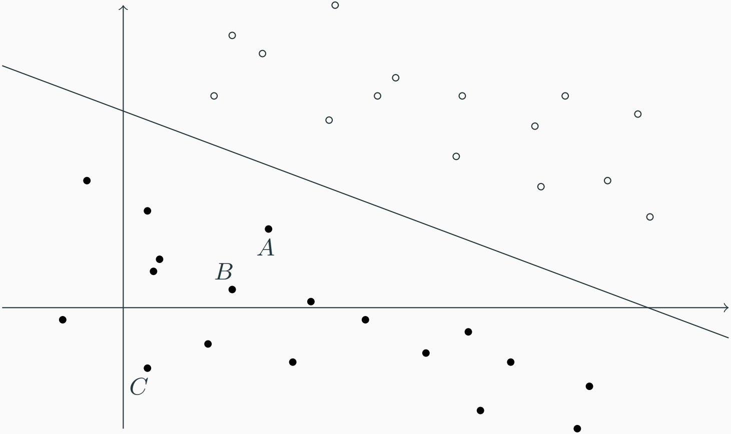
Support vector machines

Course of Machine Learning
Master Degree in Computer Science
University of Rome “Tor Vergata”

Giorgio Gambosi

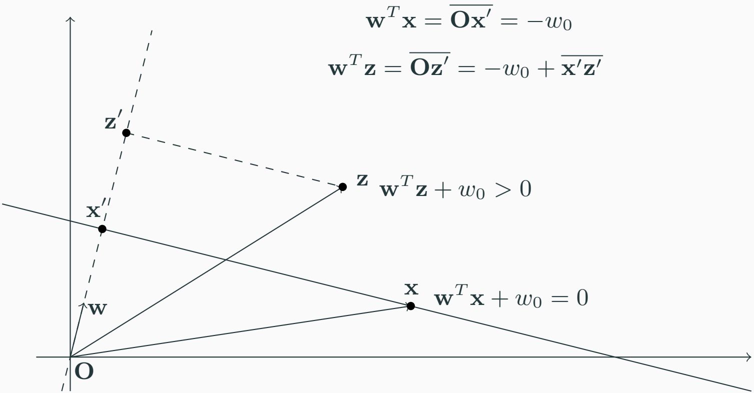
a.a. 2017-2018

Margins



A can be assigned to $\mathcal{C}_1$ with greater confidence than B and even greater confidence than C.

Some geometry



$$\mathbf{w}^T\mathbf{x} = \overline{\mathbf{Ox'}} = -w_0$$

$$\mathbf{w}^T\mathbf{z} = \overline{\mathbf{Oz'}} = -w_0 + \overline{\mathbf{x'z'}}$$

Binary classifiers

Consider a binary classifier which, for any element $\mathbf{x}$, returns a value $y \in \{-1, 1\}$, where we assume that $\mathbf{x}$ is assigned to $\mathcal{C}_0$ if $y = -1$ and to $\mathcal{C}_1$ if $y = 1$.

Moreover, we consider linear classifier such as

$$h(\mathbf{x}) = g(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) + w_0)$$

where $g(z) = 1$ if $z \geq 0$ and $g(z) = -1$ if $z < 0$. The prediction on the class of $\mathbf{x}$ is then provided by deriving a value in $\{-1, 1\}$ just as in the case of a perceptron, that is with no estimation of the probabilities $p(\mathcal{C}_i|\mathbf{x})$ that $\mathbf{x}$ belongs to each class.

Margins

For any training set item $(\mathbf{x}_i, t_i)$, the functional margin of $(\mathbf{w}, w_0)$ wrt such item is defined as

$$\overline{\gamma}_i = t_i(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) + w_0)$$

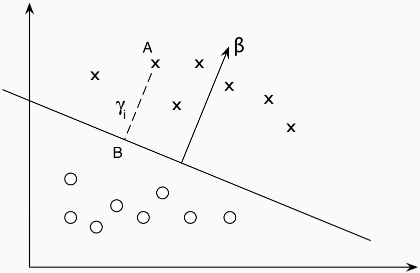
Observe that the resulting prediction is correct iff $\overline{\gamma}_i > 0$. Moreover, larger values of $\overline{\gamma}_i$ denote greater confidence on the prediction.

Given a training set $\mathcal{T} = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_n, t_n)\}$ the functional margin of $(\mathbf{w}, w_0)$ wrt $\mathcal{T}$ is the minimum functional margin for all items in $\mathcal{T}$

$$\overline{\gamma} = \min_i \overline{\gamma}_i$$

Margins

The geometric margin γ_i of a training set item $\mathbf{x}_i, t_i$ is defined as the product of t_i and the distance from $\mathbf{x}_i$ to the boundary hyperplane, that is as the length of the line segment from $\mathbf{x}_i$ to its projection on the boundary hyperplane



Margins

Since, in general, the distance of a point $\overline{\mathbf{x}}$ from a hyperplane $\mathbf{w}^T\mathbf{x} = 0$ is $\frac{\mathbf{w}^T\overline{\mathbf{x}}}{||\mathbf{w}||}$, it results

$$\gamma_i = t_i \left(\frac{\mathbf{w}^T}{||\mathbf{w}||} \boldsymbol{\phi}(\mathbf{x}_i) + \frac{w_0}{||\mathbf{w}||} \right) = \frac{\overline{\gamma}_i}{||\mathbf{w}||}$$

So, differently from$\overline{\gamma}_i$, the geometric margin γ_i is invariant wrt parameter scaling. In fact, by substituting $c\mathbf{w}$ to $\mathbf{w}$ and cw_0 to w_0, we get

$$\overline{\gamma}_i = t_i(c\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) + cw_0) = ct_i(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) + w_0)$$

$$\gamma_i = t_i \left(\frac{c\mathbf{w}^T}{||c\mathbf{w}||} \boldsymbol{\phi}(\mathbf{x}_i) + \frac{cw_0}{||c\mathbf{w}||} \right) = t_i \left(\frac{\mathbf{w}^T}{||\mathbf{w}||} \boldsymbol{\phi}(\mathbf{x}_i) + \frac{w_0}{||\mathbf{w}||} \right)$$

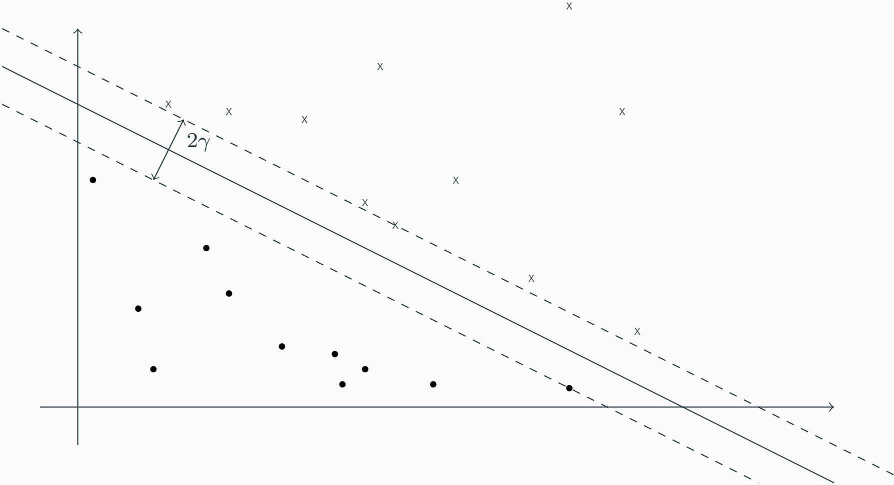
Margins

- The geometric margin wrt the training set $\mathcal{T} = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_n, t_n)\}$ is then defined as the smallest geometric margin for all items $(\mathbf{x}_i, t_i)$

$$\gamma = \min_i \gamma_i$$

- a useful interpretation of γ is as half the width of the largest strip, centered on the hyperplane $\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) + w_0 = 0$, containing none of the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$
- the hyperplanes on the boundary of such strip, each at distance γ from the hyperplane and passing (at least one of them) through some point $\mathbf{x}_i$ are said maximum margin hyperplanes.

Margins



Optimal margin classifiers

Given a training set $\mathcal{T}$, we wish to find the hyperplanes which separates the two classes (if one does exist) and has maximum γ: by making the distance between the hyperplanes and the set of points corresponding to elements as large as possible, the confidence on the provided classification increases.

Assume classes are linearly separable in the training set: hence, there exists a hyperplane (an infinity of them, indeed) separating elements in C_1 from elements in C_2. In order to find the one among those hyperplanes which maximizes γ, we have to solve the following optimization problem

$$\max_{\mathbf{w},w_0} \gamma = \max_{\mathbf{w},w_0} \min_i \gamma_i = \max_{\mathbf{w},w_0} \left(\frac{1}{||\mathbf{w}||} \min_i t_i(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) + w_0) \right)$$

This seems quite difficult to deal with, but we may simplify the formulation

Optimal margin classifiers

Observe that if all parameters are scaled by any constant c, all distances γ_i between elements and hyperplane are unchanged: we may then exploit this freedom to introduce the constraint

$$\gamma = \min_i t_i(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) + w_0) = 1$$

that is, to choose a specific constant c.

As a consequence, for each element $\mathbf{x}_i, t_i$,

$$\gamma_i = t_i(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) + w_0) \geq 1$$

An element (point) is said active if the equality holds, that is if

$$t_i(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) + w_0) = 1$$

and inactive if this does not hold. Observe that, by definition, there must exists at least one active point.

Optimal margin classifiers

For any element $\mathbf{x}, t$,

1. $t(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}) + w_0) > 1$ if $\boldsymbol{\phi}(\mathbf{x})$ is in the region corresponding to its class, outside the margin strip
2. $t(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}) + w_0) = 1$ if $\boldsymbol{\phi}(\mathbf{x})$ is in the region corresponding to its class, on the maximum margin hyperplane
3. $0 < t(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}) + w_0) < 1$ if $\boldsymbol{\phi}(\mathbf{x})$ is in the region corresponding to its class, inside the margin strip
4. $t(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}) + w_0) = 0$ if $\boldsymbol{\phi}(\mathbf{x})$ is on the separating hyperplane
5. $-1 < t(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}) + w_0) < 0$ if $\boldsymbol{\phi}(\mathbf{x})$ is in the region corresponding to the other class, inside the margin strip
6. $t(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}) + w_0) = -1$ if $\boldsymbol{\phi}(\mathbf{x})$ is in the region corresponding to the other class, on the maximum margin hyperplane
7. $t(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}) + w_0) < -1$ if $\boldsymbol{\phi}(\mathbf{x})$ is in the region corresponding to the other class, outside the margin strip

Optimal margin classifiers

The optimization problem, is then transformed into

$$\max_{\mathbf{w},w_0} ||\mathbf{w}||^{-1}$$

$$\text{where } t_i(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)+w_0)\geq 1 \qquad i=1,\ldots,n$$

Maximizing $||\mathbf{w}||^{-1}$ is equivalent to minimizing $||\mathbf{w}||^2$: hence we may formulate the problem as

$$\min_{\mathbf{w},w_0} \frac{1}{2}||\mathbf{w}||^2$$

$$\text{where } t_i(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)+w_0)\geq 1 \qquad i=1,\ldots,n$$

This is a convex quadratic optimization problem. The function to be minimized is in fact convex and the set of points satisfying the constraint is a convex polyhedron (intersection of half-spaces).

Duality

Form optimization theory it derives that, given the problem structure (linear constraints + convexity):

- there exists a dual formulation of the problem
- the optimum of the dual problem is the same the the original (primal) problem

Dual problem

Consider the optimization problem

$$\min_{\mathbf{x}\in\Omega} f(\mathbf{x})$$

$$g_i(\mathbf{x}) \geq 0 \qquad i = 1, \dots, k$$

$$h_j(\mathbf{x}) = 0 \qquad i = 1, \dots, k'$$

where $f(\mathbf{x})$, $g_i(\mathbf{x})$, $h_j(\mathbf{x})$ are convex functions and Ω is a convex set.

Define the Lagrangian

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = h(\mathbf{x}) + \sum_{i=1}^{k} \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^{k'} \mu_j h_j(\mathbf{x})$$

and the minimum

$$\theta(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$$

Then, the solution of the original problem is the same as the solution of

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu}} \theta(\boldsymbol{\lambda}, \boldsymbol{\mu})$$

$$\lambda_i \geq 0 \qquad i = 1, \dots, k$$

Dual problem

In our case,

- $f(\mathbf{x})$ corresponds to $\frac{1}{2}||\mathbf{w}||^2$
- $g_i(x)$ corresponds to $t_i(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)+b)-1\geq 0$
- there is $h_j(\mathbf{x})$
- Ω is the intersection of a set of hyperplanes, that is a polyhedron, hence convex.

Under these conditions, the solution is the same as the solution of

$$\begin{aligned}\min_{\mathbf{w},b}\max_{\boldsymbol{\lambda}} L(\mathbf{w},w_0,\boldsymbol{\lambda}) &= \min_{\mathbf{w},b}\max_{\boldsymbol{\lambda}}\left(\frac{1}{2}\mathbf{w}^T\mathbf{w}-\sum_{i=1}^{n}\lambda_i\left(t_i(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)+w_0)-1\right)\right)\\ &= \min_{\mathbf{w},b}\max_{\boldsymbol{\lambda}}\left(\frac{1}{2}\mathbf{w}^T\mathbf{w}-\sum_{i=1}^{n}\lambda_i t_i(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)+w_0)+\sum_{i=1}^{n}\lambda_i\right)\end{aligned}$$

under the constraints

$$\begin{aligned} t_i(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)+b)-1 &\geq 0 && i=1,\ldots,k\\ \lambda_i &\geq 0 && i=1,\ldots,k\end{aligned}$$

Karush-Kuhn-Tucker conditions

The following necessary and sufficient conditions hold, in this case, for the existence of an optimum $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$.

$$
\begin{aligned}
\frac{\partial L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \mathbf{x}}\bigg|_{\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*} &= \mathbf{0} && \\
\frac{\partial L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \lambda_i}\bigg|_{\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*} &= g_i(\mathbf{x}^*) \geq 0 && i = 1, \ldots, k \\
\frac{\partial L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \mu_j}\bigg|_{\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*} &= h_j(\mathbf{x}^*) = 0 && i = j, \ldots, k' \\
\lambda_i^* &\geq 0 && i = 1, \ldots, k \\
\lambda_i^* g_i(\mathbf{x}^*) &= 0 && i = 1, \ldots, k
\end{aligned}
$$

Note: the last condition states that a Lagrangian multiplier λ_i^* can be non-zero only if $g_i(\mathbf{x}^*) = 0$, that is of $\mathbf{x}^*$ is“at the limit” for the constraint $g_i(\mathbf{x})$. In this case, the constraint is said active.

Applying the KKT conditions

Since the KKT conditions hold for the maximum point, it must be, at that point:

$$
\begin{aligned}
\frac{\partial L(\mathbf{w}, w_0, \boldsymbol{\lambda})}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^{n} \lambda_i t_i \boldsymbol{\phi}(\mathbf{x}_i) = \mathbf{0} \\
\frac{\partial L(\mathbf{w}, w_0, \boldsymbol{\lambda})}{\partial w_0} &= \sum_{i=1}^{n} \lambda_i t_i = 0 \\
t_i(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) + w_0) - 1 &\geq 0 \qquad i = 1, \ldots, n \\
\lambda_i &\geq 0 \qquad i = 1, \ldots, n \\
\lambda_i \left(t_i(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) + w_0) - 1 \right) &= 0 \qquad i = 1, \ldots, n
\end{aligned}
$$

Lagrange method: dual problem

Eliminating $\mathbf{w}$ and b from both $L(\mathbf{w}, b, \lambda)$ and the constraints using the above relations, we get a new dual formulation of the problem

$$\max_{\boldsymbol{\lambda}} \overline{L}(\boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda}} \left(\sum_{i=1}^{n} \lambda_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j t_i t_j \boldsymbol{\phi}(\mathbf{x}_i) \boldsymbol{\phi}(\mathbf{x}_j) \right)$$

where

$$\begin{aligned} \lambda_i &\geq 0 \qquad\qquad i = \ \ldots, n \\ \sum_{i=1}^{n} \lambda_i t_i &= 0 \end{aligned}$$

Dual problem and kernel function

By defining the kernel function

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\phi}(\mathbf{x}_i)^T \boldsymbol{\phi}(\mathbf{x}_j)$$

the dual problem's formulation can be given as

$$\begin{aligned}
\max_{\boldsymbol{\lambda}} \tilde{L}(\boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda}} & \left(\sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j t_i t_j \kappa(\mathbf{x}_i, \mathbf{x}_j) \right) \\
\lambda_i \geq 0 & \qquad i = 1, \dots, n \\
\sum_{i=1}^{n} \lambda_i t_i = 0 &
\end{aligned}$$

Passing from primal to dual

Disadvantage The number variables increases from m to n (in particualar, if $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}$, from d to n).

Advantage The number of variables to be considered, which are relevant for classification, turns out to be quite smaller than n.

Deriving coefficients

By solving the dual problem, the optimal values of Langrangian multipliers $\boldsymbol{\lambda}^*$ are obtained.

The optimal values of parameters $\mathbf{w}^*$ are then derived through the relations

$$w_i^* = \sum_{j=1}^{n} \lambda_j^* t_j \phi_i(\mathbf{x}_j) \qquad\qquad i = 1, \ldots, m$$

The value of b^* can be obatined by observing that, for any support vector $\mathbf{x}_k$, characterized by the condition $\lambda_k \geq 0$ it must be

$$\begin{aligned} 1 &= t_k \left(\boldsymbol{\phi}(\mathbf{x}_i)^T \mathbf{w}^* + b^* \right) \\ &= t_k \left(\sum_{j=1}^{n} \lambda_j^* t_j \boldsymbol{\phi}(\mathbf{x}_j)^T \boldsymbol{\phi}(\mathbf{x}_k) + b^* \right) \\ &= t_k \left(\sum_{j=1}^{n} \lambda_j^* t_j \kappa(\mathbf{x}_j, \mathbf{x}_k) + b^* \right) \\ &= t_k \left(\sum_{j \in \mathcal{S}} \lambda_j^* t_j \kappa(\mathbf{x}_j, \mathbf{x}_k) + b^* \right) \end{aligned}$$

where $\mathcal{S}$ is the set of indices of support vectors.

Deriving coefficients

As a consequence, since $t_k = \pm 1$

$$t_k = \sum_{j \in \mathcal{S}} \lambda_j^* t_j \kappa(\mathbf{x}_j, \mathbf{x}_k) + b^*$$

and

$$b^* = t_k - \sum_{j \in \mathcal{S}} \lambda_j^* t_j \kappa(\mathbf{x}_j, \mathbf{x}_k)$$

A more precise solution can be obtained as the mean value obtained considering all support vectors

$$b^* = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left(t_i - \sum_{j \in \mathcal{S}} \lambda_j^* t_j \kappa(\mathbf{x}_j, \mathbf{x}_i) \right)$$

Classification through SVM

A new element $\mathbf{x}$ can be classified, given a set of base functions $\boldsymbol{\phi}$ or a kernel function κ, by checking the sign of

$$y(\mathbf{x}) = \sum_{i=1}^{m} w_i^* \phi_i(\mathbf{x}) + b^* = \sum_{j=1}^{n} \lambda_j^* t_j \kappa(\mathbf{x}_j, \mathbf{x}) + b^*$$

As noticed, if $\mathbf{x}_i$ is not a support vector, then it must be $\lambda_i^* = 0$. Thus, the above sum can be written as

$$y(\mathbf{x}) = \sum_{j \in \mathcal{S}} \lambda_j^* t_j \kappa(\mathbf{x}_j, \mathbf{x}) + b^*$$

The classification performed through the dual formulation, using the kernel function, does not take into account all training set items, but only support vectors, usually a quite small subset of the training set.

Non separability in the training set

- The linear separability hypothesis for the classes is quite restrictive
- In general, a suitable set of base functions $\boldsymbol{\phi}$, or a suitable kernel function $\kappa(\mathbf{x}_1, \mathbf{x}_2)$, may map all training set elements onto a larger-dimensional feature space where classes turn out to be (at least approximately) linearly separable.

Non separability in the training set

- The approach described before, when applied to non linearly separable sets, does not provide acceptable solutions: it is in fact impossibile to satisfy all constraints

$$t_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 \qquad i = 1, \dots, n$$

- These constraints must then be relaxed in order to allow them to not hold, at the cost of some increase in the objective function to be minimized
- A slack varaible ξ_i is introduced for each constraint, to provide a measure of how much the constraint is not verified

Non separability in the training set

- This can be formalized as

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i$$

$$t_i(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) + b) \geq 1 - \xi_i \qquad i = 1, \ldots, n$$

$$\xi_i \geq 0 \qquad i = 1, \ldots, n$$

  where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$

- By introducing suitable multipliers, the following Lagrangian can be obtained

$$\begin{aligned} &L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) \\ &\quad= \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\lambda_i(y_i(\mathbf{w}^T\boldsymbol{\phi}(x_i) + b) - 1 + \xi_i) - \sum_{i=1}^{n}\alpha_i\xi_i \\ &\quad= \frac{1}{2}\sum_{i=1}^{n}w_i^2 + \sum_{i=1}^{n}(C - \alpha_i)\xi_i - \sum_{i=1}^{n}\lambda_i(t_i(\sum_{j=1}^{m}w_j\phi_j(\mathbf{x}_i)) + b) - 1 + \xi_i) \\ &\quad= \frac{1}{2}\sum_{i=1}^{n}w_i^2 + \sum_{i=1}^{n}(C - \alpha_i - \lambda_i)\xi_i - \sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_i t_i w_j\phi_j(\mathbf{x}_i) + b\sum_{i=1}^{n}\lambda_i t_i + \sum_{i=1}^{n}\lambda_i \end{aligned}$$

  where $\alpha_i \geq 0$ and $\lambda_i \geq 0$, for $i = 1\ldots, n$.

KKT conditions

The Karush-Kuhn-Tucker conditions are now:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) &= \mathbf{0} && && \text{null gradient} \\
\frac{\partial}{\partial b} L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) &= 0 && && \text{null gradient} \\
\frac{\partial}{\partial \boldsymbol{\xi}} L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) &= \mathbf{0} && && \text{null gradient} \\
t_i(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) + b) - 1 + \xi_i &\geq 0 && i = 1, \ldots, n && \text{constraints} \\
\xi_i &\geq 0 && i = 1, \ldots, n && \text{constraints} \\
\lambda_i &\geq 0 && i = 1, \ldots, n && \text{multipliers} \\
\alpha_i &\geq 0 && i = 1, \ldots, n && \text{multipliers} \\
\lambda_i \left(t_i(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) + b) - 1 + \xi_i \right) &= 0 && i = 1, \ldots, n && \text{complementary slackness} \\
\alpha_i \xi_i &= 0 && i = 1, \ldots, n && \text{complementary slackness}
\end{aligned}$$

Deriving a dual formulation

From the null gradient conditions wrt w_i, b, ξ_j it derives

$$
\begin{aligned}
w_i &= \sum_{j=1}^{n} \lambda_j t_j \phi_i(\mathbf{x}_j) \qquad i = 1, \ldots, m \\
0 &= \sum_{i=1}^{n} \lambda_i t_i \\
\lambda_i &= C - \alpha_i \qquad i = 1, \ldots, n
\end{aligned}
$$

Deriving a dual formulation

By plugging the above relations into $L(\mathbf{w}, w_0, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\alpha})$, the dual problem results

$$\max_{\boldsymbol{\lambda}} \tilde{L}(\boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda}} \left(\sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j t_i t_j \kappa(\mathbf{x}_i, \mathbf{x}_j) \right)$$

$$0 \leq \lambda_i \leq C \qquad i = 1, \ldots, n$$

$$\sum_{i=1}^{n} \lambda_i y_i = 0$$

Observe that the only difference wrt the linearly separable case is given by constraints $0 \leq \lambda_i$ transformed into in $0 \leq \lambda_i \leq C$

Item characterization

Given a solution of the above problem, the elements of the training set can be partitioned into three subsets:

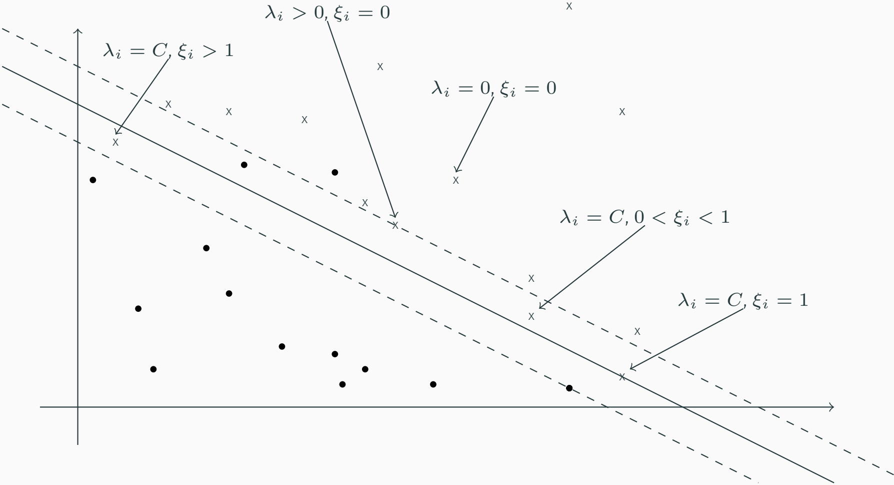
- elements correctly classified and not relevant, the ones such that $\lambda_i = 0$ and $\xi_i = 0$: such elements are in the correct halfspace, in terms of classification, and do not lie on the maximum margin hyperplanes (they are not support vectors)
- elements correctly classified and not relevant, the ones such that $\lambda_i > 0$ and $\xi_i = 0$: such elements are in the correct halfspace, in terms of classification, on the maximum margin hyperplanes (they are support vectors). The se are the only elements to affect classification
- elements correctly classified, the ones with $\xi_i > 0$, hence with $\alpha_i = 0$ and $\lambda_i = C$: such elements are in the wrong halfspace.

Item characterization

Let $\mathbf{x}_i$ be a training set element, then one of the following conditions holds:

1. $\xi_i = 0, \lambda_i = 0$ if $\boldsymbol{\phi}(\mathbf{x}_i)$ is in the correct halfspace, outside the margin strip
2. $\xi_i = 0, 0 < \lambda_i < C$ if $\boldsymbol{\phi}(\mathbf{x}_i)$ is in the correct halfspace, on the maximum margin hyperplane
3. $0 < \xi_i < 1, \lambda_i = C$ if $\boldsymbol{\phi}(\mathbf{x}_i)$ is in the correct halfspace, within the margin strip
4. $\xi_i = 1, \lambda_i = C$ if $\boldsymbol{\phi}(\mathbf{x}_i)$ is on the seprating hyperplane
5. $\xi_i > 1, \lambda_i = C$ if $\boldsymbol{\phi}(\mathbf{x}_i)$ is in the wrong halfspace

Item characterization



$\lambda_i > 0, \xi_i = 0$

$\lambda_i = C, \xi_i > 1$

$\lambda_i = 0, \xi_i = 0$

$\lambda_i = C, 0 < \xi_i < 1$

$\lambda_i = C, \xi_i = 1$

Classification

From the optimal solution $\boldsymbol{\lambda}^*$ of the dual problem, the coefficients $\mathbf{w}^*$ and b^* can be derived just as done in the linearly separable case.

A new element $\mathbf{x}$ can then be classified, again, through the sign of

$$y(\mathbf{x}) = \sum_{i=1}^{m} w_i^* \phi_i(\mathbf{x}) + b^*$$

or, equivalently, of

$$y(\mathbf{x}) = \sum_{i \in \mathcal{S}} \lambda_j^* t_j \kappa(\mathbf{x}_i, \mathbf{x}_j) + b^*$$