# Support Vector Machines - II
# Slide Set 14

Prof. Dan A. Simovici

UMB

# Problem Setting

- the input space is $\mathcal{X} \subseteq \mathbb{R}^n$;
- the output space is $\mathcal{Y} = \{-1, 1\}$;
- concept sought: a function $f : \mathcal{X} \longrightarrow \mathcal{Y} = \{-1, 1\}$;
- sample: a sequence $S = ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$ extracted from a distribution $\mathcal{D}$.

# Problem Statement

We are exploring a hypothesis space $H$ that consists of functions of the form $h : \mathcal{X} \longrightarrow \{-1, 1\}$ such that

$$h(\boldsymbol{x}) = sign\ (\boldsymbol{w}'\boldsymbol{x} + b),$$

where

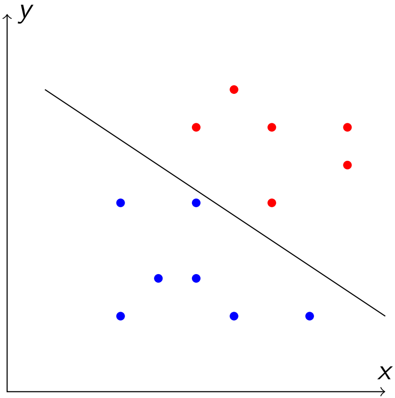
$$sign\ (a) = \begin{cases} 1 & \text{if } a \geq 0, \\ -1 & \text{if } a < 0. \end{cases}$$

such that the quantity

$$L(h) = P_{x \sim \mathcal{D}}(h(\boldsymbol{x}) \neq f(\boldsymbol{x}))$$

is small. This is the generalization error of $h$.

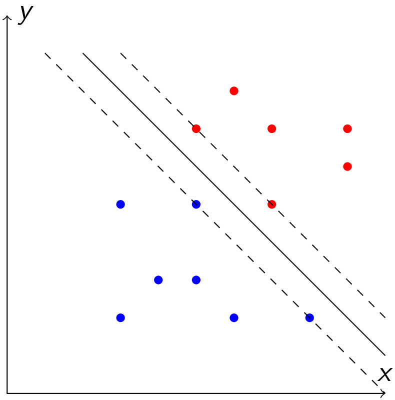
# A Fundamental Assumption: Linear Separability of $S$




If $S$ is linearly separable there are, in general, infinitely many hyperplanes that can do the separation.

# Solution returned by SVMs

SVMs seek the hyperplane with the maximum separation margin.



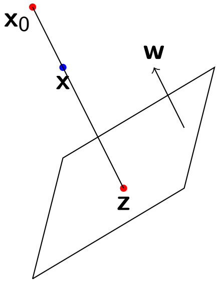
# The distance of a point $\boldsymbol{x}_0$ to a hyperplane $\boldsymbol{w}'\boldsymbol{x} + b = 0$

Equation of the line passing through $\boldsymbol{x}_0$ and perpendicular on the hyperplane is

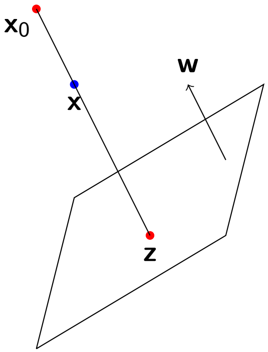
$$\boldsymbol{x} - \boldsymbol{x}_0 = t\boldsymbol{w};$$

Since $\boldsymbol{z}$ is a point on this line that belongs to the hyperplane, to find the value of $t$ that corresponds to $\boldsymbol{z}$ we must have $\boldsymbol{w}'(\boldsymbol{x}_0 + t\boldsymbol{w}) + b = 0$, that is,

$$t = -\frac{\boldsymbol{w}'\boldsymbol{x}_0 + b}{\parallel \boldsymbol{w} \parallel^2}$$



# The distance of a point $\boldsymbol{x}_0$ to a hyperplane $\boldsymbol{w}'\boldsymbol{x} + b = 0$




Thus, $\boldsymbol{z} = \boldsymbol{x}_0 - \frac{\boldsymbol{w}'\boldsymbol{x}_0 + b}{\|\boldsymbol{w}\|^2}\boldsymbol{w}$, hence the distance from $\boldsymbol{x}_0$ to the hyperplane is

$$\| \boldsymbol{x}_0 - \boldsymbol{z} \| = \frac{|\boldsymbol{w}'\boldsymbol{x}_0 + b|}{\| \boldsymbol{w} \|}.$$

# Primal Optimization Problem

We seek a hyperplane in $\mathbb{R}^n$ having the equation

$$\boldsymbol{w}'\boldsymbol{x} + b = 0,$$

where $\boldsymbol{w} \in \mathbb{R}^n$ is a vector normal to the hyperplane and $b \in \mathbb{R}$ is a scalar. A hyperplane $\boldsymbol{w}'\boldsymbol{x} + b = 0$ that does not pass through a point of $S$ is in canonical form relative to a sample $S$ if

$$\min_{(\boldsymbol{x},y)\in S} |\boldsymbol{w}'\boldsymbol{x} + b| = 1.$$

Note that we may always assume that the separating hyperplane are in canonical form relative by $S$ by rescaling the coefficients of the equation that define the hyperplane (the components of $\boldsymbol{w}$ and $b$).

## Example

Consider the points:

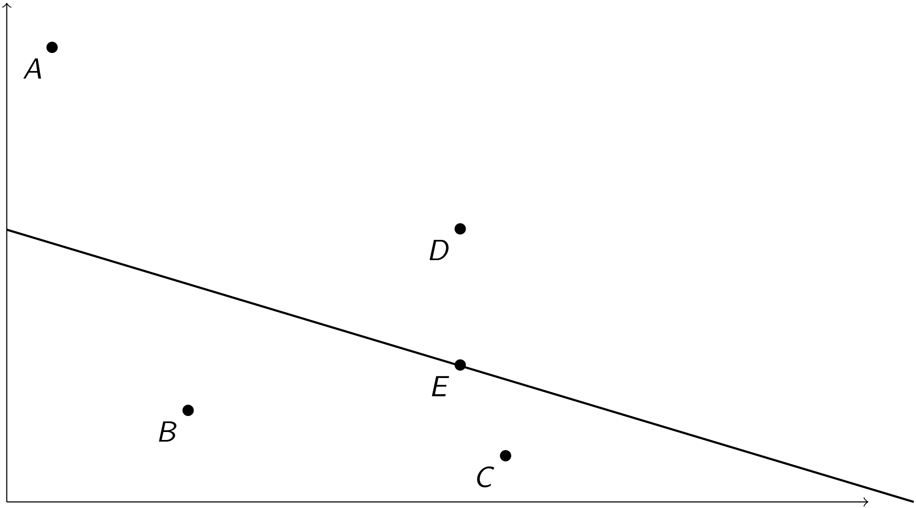
$$A = \begin{pmatrix} 1 \\ 9 \end{pmatrix}, B = \begin{pmatrix} 4 \\ 2 \end{pmatrix}, C = \begin{pmatrix} 11 \\ 1 \end{pmatrix} D = \begin{pmatrix} 10 \\ 6 \end{pmatrix}, D = \begin{pmatrix} 10 \\ 3 \end{pmatrix},$$

in $\mathbb{R}^2$ and the hyperplane $P$ (in this case, a line)

$$3x_1 + 10x_2 - 60 = 0.$$

For this hyperplane we have $\boldsymbol{w} = \begin{pmatrix} 3 \\ 10 \end{pmatrix}$ and $b = -60$. Also, $\| \boldsymbol{w} \| = \sqrt{109}$.

Note that $A, B, C, D$ do not belong to he hyperplane (e.g. $3 \cdot 1 + 10 \cdot 9 = 93 \neq 60$), except $E$ for which we have $3 \cdot 10 + 10 \cdot 3 - 60 = 0$. We say that $E$ is a support point for $P$.




$$A = \begin{pmatrix} 1 \\ 9 \end{pmatrix}, B = \begin{pmatrix} 4 \\ 2 \end{pmatrix}, C = \begin{pmatrix} 11 \\ 1 \end{pmatrix} D = \begin{pmatrix} 10 \\ 6 \end{pmatrix}, D = \begin{pmatrix} 10 \\ 3 \end{pmatrix},$$

# Example (cont'd)

## Example

Distances from the points to the hyperplane are:

$$\begin{aligned}
d(A, P) &= \frac{|3 + 90 - 60|}{\sqrt{109}} = \frac{33}{\sqrt{109}}, \\
d(B, P) &= \frac{|12 + 20 - 60|}{\sqrt{109}} = \frac{28}{\sqrt{109}}, \\
d(C, P) &= \frac{33 + 10 - 60}{\sqrt{109}} = \frac{17}{\sqrt{109}}, \\
d(D, P) &= \frac{30 + 60 - 60}{\sqrt{109}} = \frac{30}{\sqrt{109}}, \\
d(E, P) &= \frac{30 + 30 - 60}{\sqrt{109}} = 0.
\end{aligned}$$

## Example

The closest point to $P$ is $C$ (except $E$), which means that we can rescale the coefficients of the hyperplane by dividing them by 17 Thus, the equation of the hyperplane in canonical form becomes

$$\frac{3}{17}x_1 + \frac{10}{17}x_2 - \frac{60}{17} = 0.$$

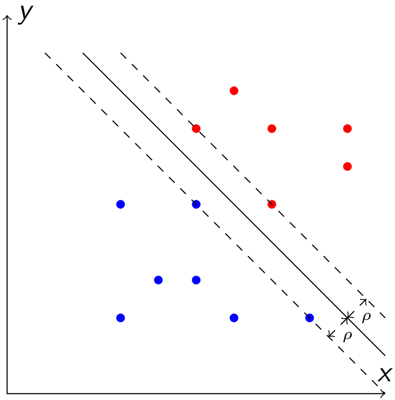
## Example

The minimum distance from one of the external points to the hyperplane is

$$d(C,P) = \frac{\left|\frac{3}{17}x_1 + \frac{10}{17}x_2 - \frac{60}{17}\right|}{\sqrt{109}} = \frac{\left|\frac{33}{17} + \frac{10}{17} - \frac{60}{17}\right|}{\sqrt{109}} = \frac{1}{\sqrt{109}}.$$

If the hyperplane $\boldsymbol{w}'\boldsymbol{x} + b = 0$ is in canonical form relative to the sample $S$, then the distance to the hyperplane to the closest points in $S$ (the margin of the hyperplane) is the same, namely,

$$\rho = \min_{(\boldsymbol{x},y)\in S} \frac{|\boldsymbol{w}'\boldsymbol{x} + b|}{\parallel \boldsymbol{w} \parallel} = \frac{1}{\parallel \boldsymbol{w} \parallel}.$$



# Canonical Separating Hyperplane

For a canonical separating hyperplane we have

$$|\mathbf{w}'\mathbf{x} + b| \geqslant 1$$

for any point $(\mathbf{x}, y)$ of the sample and

$$|\mathbf{w}'\mathbf{x} + b| = 1$$

for every support point. The point $(\mathbf{x}_i, y_i)$ is classified correctly if $y_i$ has the same sign as $\mathbf{w}'\mathbf{x}_i + b$, that is, $y_i(\mathbf{w}'\mathbf{x}_i + b) \geqslant 1$.
Maximizing the margin is equivalent to minimizing $\parallel \mathbf{w} \parallel$ or, equivalently, to minimizing $\frac{1}{2} \parallel \mathbf{w} \parallel^2$. Thus, in the separable case the SVM problem is equivalent to the following convex optimization problem:

- minimize $\frac{1}{2} \parallel \mathbf{w} \parallel^2$;
- subjected to $y_i(\mathbf{w}'\mathbf{x}_i + b) \geqslant 1$ for $1 \leqslant i \leqslant m$.

# Why $\frac{1}{2} \parallel \boldsymbol{w} \parallel^2$?

Note that this objective function,

$$\frac{1}{2} \parallel \boldsymbol{w} \parallel^2 = \frac{1}{2}(w_1^2 + \cdots + w_n^2)$$

is differentiable!

We have $\nabla \left(\frac{1}{2} \parallel \boldsymbol{w} \parallel^2\right) = \boldsymbol{w}$ and that

$$H_{\frac{1}{2}\|\boldsymbol{w}\|^2} = \boldsymbol{I}_n,$$

which shows that $\frac{1}{2} \parallel \boldsymbol{w} \parallel^2$ is a convex function of $\boldsymbol{w}$.

# Support Vectors

The Lagrangean of the optimization problem

- minimize $\frac{1}{2} \parallel \boldsymbol{w} \parallel^2$;
- subjected to $y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) \geqslant 1$ for $1 \leqslant i \leqslant m$.

is

$$L(\boldsymbol{w}, b, \boldsymbol{a}) = \frac{1}{2} \parallel \boldsymbol{w} \parallel^2 - \sum_{i=1}^{m} a_i \left( y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) - 1 \right).$$

# The Karush-Kuhn-Tucker Optimality Conditions

$$\begin{aligned}
\nabla_{\boldsymbol{w}} L &= \boldsymbol{w} - \sum_{i=1}^{m} a_i y_i \boldsymbol{x}_i = 0, \\
\nabla_b L &= -\sum_{i=1}^{m} a_i y_i = 0, \\
a_i(y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) - 1) &= 0 \text{ for all } i
\end{aligned}$$

imply

$$\begin{aligned}
\boldsymbol{w} &= \sum_{i=1}^{m} a_i y_i \boldsymbol{x}_i = 0, \\
\sum_{i=1}^{m} a_i y_i &= 0, \\
a_i &= 0 \text{ or } y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) = 1 \text{ for } 1 \leqslant i \leqslant m.
\end{aligned}$$

# Consequences of the KKT Conditions

- the weight vector is a linear combination of the training vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$, where $\boldsymbol{x}_i$ appears in this combination only if $a_i \neq 0$ (support vectors);
- since $a_i = 0$ or $y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) = 1$ for all $i$, if $a_i \neq 0$, then $y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) = 1$ for the support vectors; thus, all these vectors lie on the marginal hyperplanes $\boldsymbol{w}'\boldsymbol{x} + b = 1$ or $\boldsymbol{w}'\boldsymbol{x} + b = -1$;
- if non-support vector are removed the solution remains the same;
- while the solution of the problem $\boldsymbol{w}$ remains the same different choices may be possible for the support vectors.

Recall that the optimization problem for SVMs was

*minimize* $\frac{1}{2} \parallel \boldsymbol{w} \parallel^2$

*subject to* $y_i(\boldsymbol{w}'\boldsymbol{x} + b) \geqslant 1$ *for* $1 \leqslant i \leqslant m$

Equivalently, the constraints are

$$1 - y_i(\boldsymbol{w}'\boldsymbol{x} + b) \leqslant 0$$

for $1 \leqslant i \leqslant m$.

The Lagrangean is

$$\begin{aligned} & L(\boldsymbol{w}, b, \boldsymbol{a}) \\ = \; & \frac{1}{2} \parallel \boldsymbol{w} \parallel^2 + \sum_{i=1}^{m} a_i(1 - y_i(\boldsymbol{w}'\boldsymbol{x}_i + b)) \\ = \; & \frac{1}{2} \parallel \boldsymbol{w} \parallel^2 + \sum_{i=1}^{m} a_i - \sum_{i=1}^{m} a_i y_i \boldsymbol{w}'\boldsymbol{x}_i - b \sum_{i=1}^{m} a_i y_i. \end{aligned}$$

# The Dual Problem

*maximize* $L(\boldsymbol{w}, b, \boldsymbol{a})$

The KKT conditions are

$$\begin{aligned}
(\nabla_{\boldsymbol{w}} L) &= \boldsymbol{w} - \sum_{i=1}^{m} a_i y_i \boldsymbol{x}_i = \boldsymbol{0}, \\
(\nabla_b L) &= -\sum_{i=1}^{m} a_i y_i = 0, \\
& \quad a_i(1 - y_i(\boldsymbol{w}'\boldsymbol{x}_i + b)) = 0,
\end{aligned}$$

which are equivalent to

$$\begin{array}{rcl}
\boldsymbol{w} & = & \sum_{i=1}^{m} a_i y_i \boldsymbol{x}_i, \\
\sum_{i=1}^{m} a_i y_i & = & 0, \\
a_i = 0 & \text{or} & y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) = 1,
\end{array}$$

respectively.

# Implications

- the weight vector $\boldsymbol{w}$ is a linear combination of the training vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$;
- a vector $\boldsymbol{x}_i$ appears in $\boldsymbol{w}$ if and only if $a_i \neq 0$ (such vectors are called support vectors);
- if $a_i \neq 0$, then $y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) = \pm 1$.

Note that support vectors define the maximum margin hyperplane, or the SVM solution.

# Transforming the Lagrangean

Since

$$L(\boldsymbol{w}, b, \boldsymbol{a}) = \frac{1}{2} \parallel \boldsymbol{w} \parallel^2 + \sum_{i=1}^{m} a_i - \sum_{i=1}^{m} a_i y_i \boldsymbol{w}' \boldsymbol{x}_i - b \sum_{i=1}^{m} a_i y_i,$$

$\boldsymbol{w} = \sum_{j=1}^{m} a_j y_j \boldsymbol{x}_j$ (note that we changed the summation index from $i$ to $j$), and $\sum_{i=1}^{m} a_i y_i = 0$, we have

$$L(\boldsymbol{w}, b, \boldsymbol{a}) = \frac{1}{2} \parallel \boldsymbol{w} \parallel^2 + \sum_{i=1}^{m} a_i - \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j y_i y_j \boldsymbol{x}_j' \boldsymbol{x}_i.$$

# Further Transformation of the Lagrangean

Note that

$$\begin{aligned}
\| \boldsymbol{w} \|^2 &= \boldsymbol{w}'\boldsymbol{w} = \left(\sum_{j=1}^{m} a_j y_j \boldsymbol{x}_j'\right)\left(\sum_{i=1}^{m} a_i y_i \boldsymbol{x}_i\right), \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j y_i y_j \boldsymbol{x}_j' \boldsymbol{x}_i.
\end{aligned}$$

Therefore,

$$\begin{aligned}
L(\boldsymbol{w}, b, \boldsymbol{a}) &= \frac{1}{2} \| \boldsymbol{w} \|^2 + \sum_{i=1}^{m} a_i - \sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j y_i y_j \boldsymbol{x}_j' \boldsymbol{x}_i \\
&= \sum_{i=1}^{m} a_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j y_i y_j \boldsymbol{x}_j' \boldsymbol{x}_i.
\end{aligned}$$

# The Dual Optimization Problem for Separable Sets

$$\begin{aligned} &\textit{maximize } \sum_{i=1}^{m} a_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j y_i y_j \mathbf{x}_i' \mathbf{x}_j \\ &\qquad \textit{subject to } a_i \geqslant 0 \textit{ for } 1 \leqslant i \leqslant m \textit{ and } \sum_{i=1}^{m} a_i y_i = 0. \end{aligned}$$

Note that the objective function depends on $a_1, \ldots, a_m$.

- in this case the strong duality holds; therefore, the primal and the dual problems are equivalent;
- the solution $\boldsymbol{a}$ of the dual problem can be used directly to determine the hypothesis returned by the SVM as

$$h(\boldsymbol{x}) = sign\,(\boldsymbol{w}'\boldsymbol{x} + b) = sign\,\left(\sum_{i=1}^{m} a_i y_i (\boldsymbol{x}_i'\boldsymbol{x}) + b\right);$$

- since support vectors lie on the marginal hyperplanes, for every support vector $\boldsymbol{x}_i$ we have $\boldsymbol{w}'\boldsymbol{x}_i + b = y_i$, so

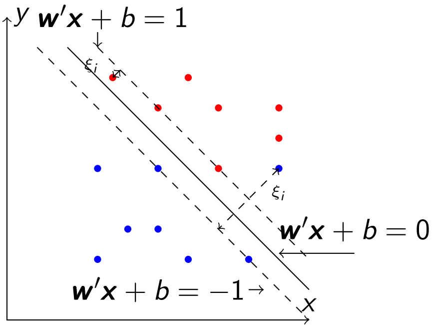
$$b = y_i - \sum_{j=1}^{m} a_j y_j (\boldsymbol{x}_j'\boldsymbol{x}).$$

# Slack Variables

If data is not separable the conditions $y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) \geqslant 1$ cannot all hold (for $1 \leqslant i \leqslant m$). Instead, we impose a relaxed version, namely

$$y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) \geqslant 1 - \xi_i,$$

where $\xi_i$ are new variables known as slack variables.
A slack variable $\xi_i$ measures the distance by which $\boldsymbol{x}_i$ violates the desired inequality $y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) \geqslant 1$.




A vector $\boldsymbol{x}_i$ is an outlier if $\boldsymbol{x}_i$ is not positioned correctly on the side of the appropriate hyperplane.

- a vector $\boldsymbol{x}_i$ with $0 < y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) < 1$ is still an outlier even if it is correctly classified by the hyperplane $\boldsymbol{w}'\boldsymbol{x} + b = 0$ (see the red point);
- if we omit the outliers the data is correctly separated by the hyperplane $\boldsymbol{w}'\boldsymbol{x} + b = 0$ with a soft margin $\rho = \frac{1}{\|\boldsymbol{w}\|}$;
- we wish to limit the amount of slack due to outliers $(\sum_{i=1}^{m} \xi_i)$, but we also seek a hyperplane with a large margin (even though this may lead to more outliers).

# Optimization for Non-Separable Data

$$\begin{aligned} &\textit{minimize } \frac{1}{2} \parallel \mathbf{w} \parallel^2 + C \sum_{i=1}^{m} \xi_i^p \\ &\textit{subject to } y_i(\mathbf{w}'\mathbf{x}_i + b) \geqslant 1 - \xi_i \textit{ and } \xi_i \geqslant 0 \textit{ for } 1 \leqslant i \leqslant m. \end{aligned}$$

The parameter $C$ is determined in the process of cross-validation.
This is a convex optimization problem with affine constraints.

# Support Vectors

As in the separable case:

- constraints are affine and thus, qualified;
- the objective function and the affine constraints are convex and differentiable;
- thus, the KKT conditions apply.

# Variables

- $a_i \geqslant 0$ for $1 \leqslant i \leqslant m$ are variables associated with $m$ constraints;
- $b_i \geqslant 0$ for $1 \leqslant i \leqslant m$ are variables associated with the non-negativity constraints of the slack variables.

The Lagrangean is defined as:

$$
\begin{array}{rcl}
L(\boldsymbol{w}, b, \xi_1, \ldots, \xi_m, \boldsymbol{a}, \boldsymbol{b}) & = & \frac{1}{2} \parallel \boldsymbol{w} \parallel^2 + C \sum_{i=1}^{m} \xi_i \\
& & - \sum_{i=1}^{m} a_i [y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^{n} b_i \xi_i .
\end{array}
$$

The KKT conditions are:

$$
\begin{array}{rclcl}
\nabla_{\boldsymbol{w}} L & = & \boldsymbol{w} - \sum_{i=1}^{m} a_i y_i \boldsymbol{x}_i = 0 & \Rightarrow & \boldsymbol{w} = \sum_{i=1}^{m} a_i y_i \boldsymbol{x}_i \\
\nabla_b L & = & - \sum_{i=1}^{m} a_i y_i = 0 & \Rightarrow & \sum_{i=1}^{m} a_i y_i = 0 \\
\nabla_{\xi_i} L & = & C - a_i - b_i = 0 & \Rightarrow & a_i + b_i = C
\end{array}
$$

and

$$
\begin{array}{l}
a_i [y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) - 1 + \xi_i] = 0 \text{ for } 1 \leqslant i \leqslant m \Rightarrow a_i = 0 \text{ or} \\
y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) = 1 - \xi_i, \\
b_i \xi_i = 0 \Rightarrow b_i = 0 \text{ or } \xi_i = 0.
\end{array}
$$

# Consequences of the KKT Conditions

- $\boldsymbol{w}$ is a linear combination of the training vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$, where $\boldsymbol{x}_i$ appears in the combination only if $a_i \neq 0$;
- if $a_i \neq 0$, then $y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) = 1 - \xi_i$;
- if $\xi_i = 0$, then $y_i(\boldsymbol{w}'\boldsymbol{x}_i + b) = 1$ and $\boldsymbol{x}_i$ lies on marginal hyperplane as in the separable case; otherwise, $\boldsymbol{x}_i$ is an outlier;
- if $\boldsymbol{x}_i$ is an outlier, $b_i = 0$ and $a_i = C$ or $\boldsymbol{x}_i$ is located on the marginal hyperplane.
- $\boldsymbol{w}$ is unique; the support vectors are not.

# The Dual Optimization Problem

The Lagrangean can be rewritten by substituting $\boldsymbol{w}$:

$$
\begin{aligned}
L &= \frac{1}{2}\left\| \sum_{i=1}^{m} a_i y_i \boldsymbol{x}_i \right\|^2 - \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j y_i y_j \boldsymbol{x}'_i \boldsymbol{x}_j \\
&\quad - \sum_{i=1}^{m} a_i y_i b + \sum_{i=1}^{m} a_i \\
&= \sum_{i=1}^{m} a_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j y_i y_j \boldsymbol{x}'_i \boldsymbol{x}_j,
\end{aligned}
$$

- the Lagrangean has exactly the same form as in the separable case;
- we need $a_i \geqslant 0$ and, in addition $b_i \geqslant 0$, which is equivalent to $a_i \leqslant C$ (because $a_i + b_i = C$);

The dual optimization problem for the non-separable case becomes:

*maximize for* $\boldsymbol{a}$ $\sum_{i=1}^{m} a_i - \frac{1}{2} a_i a_j y_i y_j \boldsymbol{x}_i' \boldsymbol{x}_j$
*subject to* $0 \leqslant a_i \leqslant C$ *and* $\sum_{i=1}^{m} a_i y_i = 0$
*for* $1 \leqslant i \leqslant m$.

# Consequences

- the objective function is concave and differentiable;
- the solution can be used to determine the hypothesis

$$h(\boldsymbol{x}) = sign\,(\boldsymbol{w}'\boldsymbol{x} + b);$$

- for any support vector $b_i$ we have $b = y_i - \sum_{j=1}^{m} a_j y_j \boldsymbol{x}_i' \boldsymbol{x}_j$.
- the hypothesis returned depends only on the inner products between the vectors and not directly on the vectors themselves.

## Definition

The geometric margin relative to a linear classifier $h(\mathbf{x}) = \mathbf{w}'\mathbf{x} + b$ is its distance to the hyperplane $\mathbf{w}'\mathbf{x} + b = 0$:

$$\rho(\mathbf{x}) = \frac{y(\mathbf{w}'\mathbf{x} + b)}{\parallel \mathbf{w} \parallel}.$$

The margin for a linear classifier $h$ for a sample $S = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ is

$$\rho = \min_{1 \leqslant i \leqslant m} \frac{y_i(\mathbf{w}'\mathbf{x} + b)}{\parallel \mathbf{w} \parallel}$$

### Theorem

*Let $S$ be a sample included in a sphere of radius $r$, $S \subseteq \{\boldsymbol{x} \mid \parallel \boldsymbol{x} \parallel \leqslant r\}$. The VC dimension of the set of canonical hyperplanes of the form*

$$h(\boldsymbol{x}) = sign\ (\boldsymbol{w}'\boldsymbol{x}), \min_{\boldsymbol{x} \in S} |\boldsymbol{w}'\boldsymbol{x}| = 1 \ and \ \parallel \boldsymbol{w} \parallel \leqslant \Lambda,$$

*verifies $d \leqslant r^2\Lambda^2$.*

# Proof

Suppose that $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d\}$ is a set that can be fully shattered. Then, for all $\boldsymbol{y} = (y_1, \ldots, y_d) \in \{-1, 1\}^d$ there exists $\boldsymbol{w}$ such that $1 \leqslant y_i(\boldsymbol{w}'\boldsymbol{x})$ for $1 \leqslant i \leqslant d$.
Summing up these inequalities yields:

$$d \leqslant \boldsymbol{w}' \sum_{i=1}^{d} y_i \boldsymbol{x}_i \leqslant \parallel \boldsymbol{w} \parallel \cdot \left\| \sum_{i=1}^{d} y_i \boldsymbol{x}_i \right\| \leqslant \Lambda \left\| \sum_{i=1}^{d} y_i \boldsymbol{x}_i \right\|.$$

# Proof (cont'd)

Since $y_1, \ldots, y_d$ are independent, if $i \neq j$, $E(y_i y_j) = E(y_i)E(y_j) = 0$; also, $E(y_i y_i) = 1$.
Since $d \leqslant \Lambda \left\| \sum_{i=1}^{d} y_i \boldsymbol{x}_i \right\|$ holds for all $\boldsymbol{y} \in \{-1, 1\}^d$, it holds over expectations and we have

$$
\begin{aligned}
d &\leqslant \Lambda E_{\boldsymbol{y}} \left( \left\| \sum_{i=1}^{d} y_i \boldsymbol{x}_i \right\| \right) \leqslant \Lambda \left( E_{\boldsymbol{y}} \left( \left\| \sum_{i=1}^{d} y_i \boldsymbol{x}_i \right\|^2 \right) \right)^{1/2} \\
&= \Lambda \left( \sum_{i=1}^{m} \sum_{j=1}^{m} E_y(y_i y_j)(\boldsymbol{x}_i' \boldsymbol{x}_j) \right)^{1/2} \\
&= \Lambda \left( \sum_{i=1}^{d} \boldsymbol{x}_i' \boldsymbol{x}_i \right)^{1/2} \leqslant \Lambda (d r^2)^{1/2} = \Lambda r \sqrt{d}.
\end{aligned}
$$

Thus,

$$d \leqslant \Lambda^2 r^2$$

- recall that when the data is linearly separable the margin $\rho$ is given by:

$$\rho = \min_{(\boldsymbol{x},y)\in S} \frac{|\boldsymbol{w}'\boldsymbol{x} + b|}{\| \boldsymbol{w} \|} = \frac{1}{\| \boldsymbol{w} \|};$$

- if we restrict the sample $S$ such that the resulting $\boldsymbol{w}$ is such that $\| \boldsymbol{w} \| = \frac{1}{\rho} = \Lambda$, it follows that

$$d \leqslant \frac{r^2}{\rho^2}.$$