Support Vector Machines - I

Prof. Dan A. Simovici

UMB

Problem Setting

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

Problem Statement

- the hypothesis space H is $H \subseteq \mathcal{Y}^{\mathcal{X}}$;
- task: find $h \in H$ such that the generalization error

$$L_{\mathcal{D}}(h) = P_{x \sim \mathcal{D}}(h(\mathbf{x}) \neq f(\mathbf{x}))$$

  is small.

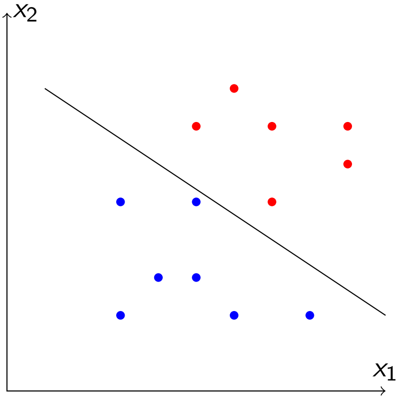
The smaller the VCD(H) the more efficient the process is. One possibility is the class of linear functions from $\mathcal{X}$ to $\mathcal{Y}$:

$$H = \{\mathbf{x} \rightsquigarrow sign(\mathbf{w}'\mathbf{x} + b) \mid \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}\},$$

where

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

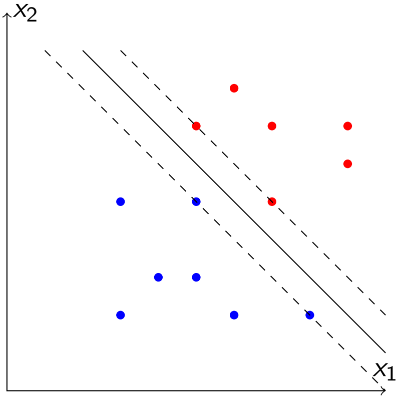
A Fundamental Assumption: Linear Separability of S



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.



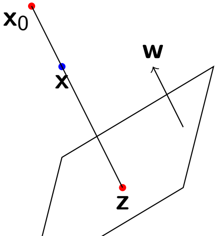
The distance of a point $\mathbf{x}_0$ to a hyperplane $\mathbf{w}'\mathbf{x} + b = 0$

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

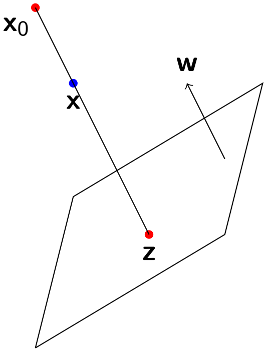
$$\mathbf{x} - \mathbf{x}_0 = t\mathbf{w};$$

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

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



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



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

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

Primal Optimization Problem

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

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

where $\mathbf{w} \in \mathbb{R}^n$ is a vector normal to the hyperplane and $b \in \mathbb{R}$ is a scalar.

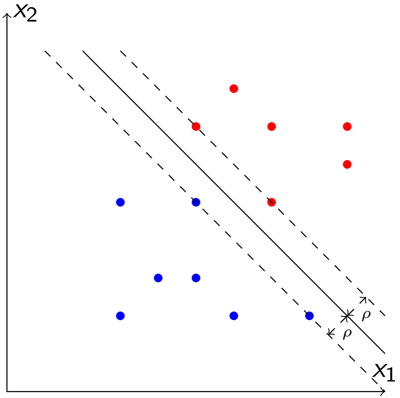
A hyperplane $\mathbf{w}'\mathbf{x} + b = 0$ that does not pass through a point of a set S is in canonical form relative to S if

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

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 $\mathbf{w}$ and b).

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

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



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 $1 \leqslant i \leqslant m$.

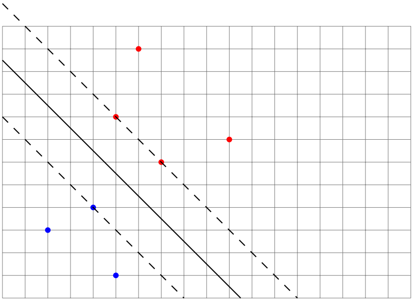
Example

Consider a set S that consists of seven points in $\mathbb{R}^2 \times \{-1, 1\}$:

$$\text{positive examples:} \quad \begin{pmatrix} 5 \\ 8 \end{pmatrix}, \begin{pmatrix} 7 \\ 6 \end{pmatrix}, \begin{pmatrix} 10 \\ 7 \end{pmatrix}, \begin{pmatrix} 6 \\ 11 \end{pmatrix},$$

$$\text{negative examples:} \quad \begin{pmatrix} 4 \\ 4 \end{pmatrix}, \begin{pmatrix} 5 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \end{pmatrix}.$$

Example cont'd



Example cont'd

We seek a hyperplane (in this case, a line in $\mathbb{R}^2$) having the equation

$$w_1 x_1 + w_2 x_2 + b = 0.$$

The support points are

$$\begin{pmatrix} 5 \\ 8 \end{pmatrix}, \begin{pmatrix} 7 \\ 6 \end{pmatrix}, \begin{pmatrix} 4 \\ 4 \end{pmatrix},$$

and we must have

$$5w_1 + 8w_2 + b = 1, 7w_1 + 6w_2 + b = 1.4w_1 + 4w_2 + b = -1.$$

The solution of the above system is:

$$w_1 = \frac{2}{5}, w_2 = \frac{-2}{5}, b = \frac{11}{5}.$$

Since $\| \mathbf{w} \| = \sqrt{0.4^2 + 0.4^2} = 0.4\sqrt{2}$, we have $\rho = \frac{1}{\sqrt{\|w\|}} = \frac{5\sqrt{2}}{4} \sim 1.76.$

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

Note that this objective function,

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

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

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

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

Support Vectors

The Lagrangean of the optimization problem

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

is

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

The Karush-Kuhn-Tucker Optimality Conditions

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

imply

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

Consequences of the KKT Conditions

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

Recall that the optimization problem for SVMs was

$$\begin{aligned}&\textit{minimize } \frac{1}{2} \parallel \mathbf{w} \parallel^2\\ &\qquad \textit{subject to } y_i(\mathbf{w}'\mathbf{x} + b) \geqslant 1 \textit{ for } 1 \leqslant i \leqslant m\end{aligned}$$

Equivalently, the constraints are

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

for $1 \leqslant i \leqslant m$.
The Lagrangean is

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

The Dual Problem

$$\textit{maximize } L(\mathbf{w}, b, \mathbf{a})$$

The KKT conditions are

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

which are equivalent to

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

respectively.

Implications

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

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

Transforming the Lagrangean

Since

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

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

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

Further Transformation of the Lagrangean

Note that

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

Therefore,

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

The Dual Optimization Problem for Separable Sets

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

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 $\mathbf{a}$ of the dual problem can be used directly to determine the hypothesis returned by the SVM as

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

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

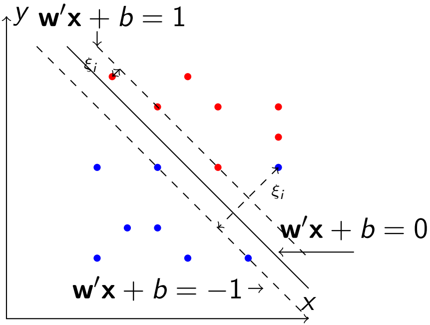
$$b = y_i - \sum_{j=1}^{m} a_j y_j (\mathbf{x}_j'\mathbf{x}).$$

Slack Variables

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

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

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



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

- a vector $\mathbf{x}_i$ with $0 < y_i(\mathbf{w}'\mathbf{x}_i + b) < 1$ is still an outlier even if it is correctly classified by the hyperplane $\mathbf{w}'\mathbf{x} + b = 0$ (see the red point);
- if we omit the outliers the data is correctly separated by the hyperplane $\mathbf{w}'\mathbf{x} + b = 0$ with a soft margin $\rho = \frac{1}{\|\mathbf{w}\|}$;
- 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

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

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:

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

The KKT conditions are:

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

and

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

Consequences of the KKT Conditions

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

The Dual Optimization Problem

The Lagrangean can be rewritten by substituting **w**:

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

- 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:

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

Consequences

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

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

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