Support Vector Machines - I

Prof. Dan A. Simovici

UMB
1. Linear Classification
2. SVM - The Separable Case
3. Leave-One Out (LOO) Analysis in the Separable Case
4. SVM - The Non-Separable Case
5. Margins
Problem Setting

- the input space is $\mathcal{X} \subseteq \mathbb{R}^n$;
- the output space is $\mathcal{Y} = \{-1, 1\}$;
- concept sought: a function $f : \mathcal{X} \rightarrow \mathcal{Y}$;
- sample: a sequence $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$ extracted from a distribution $\mathcal{D}$. 
Problem Statement

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

$$L_D(h) = P_{x \sim D}(h(x) \neq f(x))$$

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 $\mathcal{Y}$:

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

where

$$\text{sign}(a) = \begin{cases} 1 & \text{if } a \geq 0, \\ -1 & \text{if } a < 0. \end{cases}$$
A Fundamental Assumption: Linear Separability of $S$

If $S$ is linearly separable there are, in general, infinitely many hyperplanes that can do the separation.
SVMs seek the hyperplane with the maximum separation margin.
The distance of a point $x_0$ to a hyperplane $w'x + b = 0$

Equation of the line passing through $x_0$ and perpendicular on the hyperplane is

$$x - x_0 = tw;$$

Since $z$ is a point on this line that belongs to the hyperplane, to find the value of $t$ that corresponds to $z$ we must have $w'(x_0 + tw) + b = 0$, that is,

$$t = -\frac{w'x_0 + b}{\|w\|^2}$$
The distance of a point $x_0$ to a hyperplane $w'x + b = 0$

Thus, $z = x_0 - \frac{w'x_0 + b}{\|w\|^2}w$, hence the distance from $x_0$ to the hyperplane is

$$\| x_0 - z \| = \frac{|w'x_0 + b|}{\|w\|}.$$
We seek a hyperplane in $\mathbb{R}^n$ having the equation

$$w^T x + b = 0,$$

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

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

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 $w$ and $b$).
If the hyperplane $w'x + b = 0$ 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,

$$\rho = \min_{(x,y) \in S} \frac{|w'x + b|}{\|w\|} = \frac{1}{\|w\|}.$$
For a canonical separating hyperplane we have

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

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, \(y_i(\mathbf{w}'\mathbf{x}_i + b) \geq 1\).

Maximizing the margin is equivalent to minimizing \(\|\mathbf{w}\|\) or, equivalently, to minimizing \(\frac{1}{2} \|\mathbf{w}\|^2\). Thus, in the separable case the SVM problem is equivalent to the following convex optimization problem:

- minimize \(\frac{1}{2} \|\mathbf{w}\|^2\);
- subjected to \(y_i(\mathbf{w}'\mathbf{x}_i + b) \geq 1\) for \(1 \leq i \leq m\).
Linear Classification

Why $\frac{1}{2} \| w \|^2$?

Note that this objective function,

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

is differentiable!

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

$$H_{\frac{1}{2}\|w\|^2} = I_n,$$

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

The Lagrangean of the optimization problem

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

is

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2} \| \mathbf{w} \|^2 - \sum_{i=1}^{m} a_i \left( y_i (\mathbf{w}' \mathbf{x}_i + b) - 1 \right).$$
The Karush-Kuhn-Tucker Optimality Conditions

\[ \nabla_w L = \mathbf{w} - \sum_{i=1}^{m} a_i y_i \mathbf{x}_i = 0, \]

\[ \nabla_b L = - \sum_{i=1}^{m} a_i y_i = 0, \]

\[ a_i(y_i(\mathbf{w}'\mathbf{x}_i + b) - 1) = 0 \text{ for all } i \]

imply

\[ \mathbf{w} = \sum_{i=1}^{m} a_i y_i \mathbf{x}_i = 0, \]

\[ \sum_{i=1}^{m} a_i y_i = 0, \]

\[ a_i = 0 \text{ or } y_i(\mathbf{w}'\mathbf{x}_i + b) = 1 \text{ for } 1 \leq i \leq m. \]
Consequences of the KKT Conditions

- The weight vector is a linear combination of the training vectors $x_1, \ldots, x_m$, where $x_i$ appears in this combination only if $a_i \neq 0$ (support vectors);
- Since $a_i = 0$ or $y_i(w'x_i + b) = 1$ for all $i$, if $a_i \neq 0$, then $y_i(w'x_i + b) = 1$ for the support vectors; thus, all these vectors lie on the marginal hyperplanes $w'x + b = 1$ or $w'x + b = -1$;
- If non-support vector are removed the solution remains the same;
- While the solution of the problem $w$ remains the same different choices may be possible for the support vectors.
Recall that the optimization problem for SVMs was

$$\text{minimize } \frac{1}{2} \| w \|^2$$

subject to \( y_i (w' x + b) \geq 1 \) for \( 1 \leq i \leq m \)

Equivalently, the constraints are

$$1 - y_i (w' x + b) \leq 0$$

for \( 1 \leq i \leq m \).

The Lagrangean is

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

$$= \frac{1}{2} \| w \|^2 + \sum_{i=1}^{m} a_i - \sum_{i=1}^{m} a_i y_i w' x_i - b \sum_{i=1}^{m} a_i y_i.$$
The Dual Problem

\[ \text{maximize } L(w, b, a) \]

The KKT conditions are

\[
(\nabla_w L) = w - \sum_{i=1}^{m} a_i y_i x_i = 0, \\
(\nabla_b L) = - \sum_{i=1}^{m} a_i y_i = 0, \\
a_i (1 - y_i (w' x_i + b)) = 0,
\]

which are equivalent to

\[
w = \sum_{i=1}^{m} a_i y_i x_i, \\
\sum_{i=1}^{m} a_i y_i = 0, \\
a_i = 0 \text{ or } y_i (w' x_i + b) = 1,
\]

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

\[ L(w, b, a) = \frac{1}{2} \| w \|^2 + \sum_{i=1}^{m} a_i - \sum_{i=1}^{m} a_i y_i w' x_i - b \sum_{i=1}^{m} a_i y_i, \]

\[ w = \sum_{j=1}^{m} a_j y_j x_j \] (note that we changed the summation index from \( i \) to \( j \)),

and \( \sum_{i=1}^{m} a_i y_i = 0 \), we have

\[ L(w, b, a) = \frac{1}{2} \| w \|^2 + \sum_{i=1}^{m} a_i - \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j y_i y_j x_j' x_i. \]
Further Transformation of the Lagrangean

Note that

\[ \| \mathbf{w} \|^2 = \mathbf{w}' \mathbf{w} = \left( \sum_{j=1}^{m} a_j y_j \mathbf{x}_j' \right) \left( \sum_{i=1}^{m} a_i y_i \mathbf{x}_i \right), \]

\[ = \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j y_i y_j \mathbf{x}_j' \mathbf{x}_i. \]

Therefore,

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

\[ = \sum_{i=1}^{m} a_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j y_i y_j \mathbf{x}_j' \mathbf{x}_i. \]
The Dual Optimization Problem for Separable Sets

\[
\begin{align*}
\text{maximize} & \quad \sum_{i=1}^{m} a_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j y_i y_j x'_i x_j \\
\text{subject to} & \quad a_i \geq 0 \text{ for } 1 \leq i \leq m \text{ and } \sum_{i=1}^{m} a_i y_i = 0.
\end{align*}
\]

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

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

• 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

\[
b = y_i - \sum_{j=1}^{m} a_j y_j (\mathbf{x}_j'\mathbf{x}).
\]
Let $N_{SV}$ the number of support vectors that define the hypothesis $h_S$ returned for a sample $S$ in the separable case, where $S = \{(x_j, y_j) \mid 1 \leq j \leq m\}$.

Suppose the sample $S$ is $S \sim \mathcal{D}^m$, where $\mathcal{D}$ is the distribution of examples. If the algorithm $\mathcal{A}$ is trained on all points of $S$ with the exception of $x_i$, that is, is trained on $S - \{x_i\}$ the hypothesis returned is $h_S - \{x_i\}$ and the error is

$$\hat{R} <_{LOO} (\mathcal{A}) = \frac{1}{m} \sum_{i=1}^{m} (h_{S-\{x_i\}}(x_i) \neq y_i).$$

The leave-one error is the average of the errors obtained by leaving one example out.
**Lemma**

The average leave-one-out error for sample of size $m \geq 2$ is an unbiased estimate of the average generalization error for sample of size $m - 1$, that is,

$$E_{S \sim \mathcal{D}^m} (\text{ERM}_{\text{LOO}}(A)) = E_{S' \sim \mathcal{D}^{m-1}} (R(h_{S'})).$$
Proof

\[ E_{S \sim D^m} \left( \text{ERM}_\text{LOO}(A) \right) \]

\[ = \frac{1}{m} \sum_{i=1}^{m} E_{S \sim D^m} \left( h_{S \setminus \{x_i\}}(x_i) \neq y_i \right) \]

\[ = E_{S \sim D^m} \left( h_{S \setminus \{x_1\}}(x_1) \neq y_1 \right) \]

(since all points of \( S \) are drawn at random and are equally distributed)

\[ = E_{S' \sim D^{m-1}, X_1 \sim D} \left( h_{S'}(x_1) \neq y_1 \right) \]

\[ = E_{S' \sim D^{m-1}} \left( E_{X_1 \sim D} \left( h_{S'}(x_1) \neq y_1 \right) \right) \]

\[ = E_{S' \sim D^{m-1}} \left( R(h_{S'}) \right). \]
Theorem

If $h_S$ is the hypothesis returned by the SVM algorithm $A$ for a sample $S$, then

$$E\left(\text{ERM}(h_S)\right) \leq E_{S \sim \mathcal{D}^{m+1}} \left(\frac{N_{SV}(S)}{m+1}\right).$$

Proof: Let $S$ be a linearly separable sample of size $m + 1$. If $x$ is not a support vector of $h_S$, removing it does not change the solution. Thus, $h_{S-\{x\}} = h_S$ and $h_{S-\{x\}}$ correctly classifies $x$. Thus, if $h_{S-\{x\}}$ misclassifies $x$, then $x$ must be a support vector which implies

$$\text{ERM}_{LOO}(A) \leq \frac{N_{SV}(S)}{m+1}.$$ 

Taking the expectation of both sides yields the result.
If data is not separable the conditions $y_i(w'x_i + b) \geq 1$ cannot all hold (for $1 \leq i \leq m$). Instead, we impose a relaxed version, namely

$$y_i(w'x_i + b) \geq 1 - \xi_i,$$

where $\xi_i$ are new variables known as slack variables. A slack variable $\xi_i$ measures the distance by which $x_i$ violates the desired inequality $y_i(w'x_i + b) \geq 1$. 
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 $x_i$ with $0 < y_i(w'x_i + b) < 1$ is still an outlier even if it is correctly classified by the hyperplane $w'x + b = 0$ (see the red point);

if we omit the outliers the data is correctly separated by the hyperplane $w'x + b = 0$ with a soft margin $\rho = \frac{1}{||w||}$;

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

\[
\text{minimize } \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{m} \xi_i
\]
\[
\text{subject to } y_i (\mathbf{w}' \mathbf{x}_i + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0 \text{ for } 1 \leq i \leq m.
\]

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

- $a_i \geq 0$ for $1 \leq i \leq m$ are variables associated with $m$ constraints;
- $b_i \geq 0$ for $1 \leq i \leq m$ are variables associated with the non-negativity constraints of the slack variables.
The Lagrangean is defined as:

\[ L(w, b, \xi_1, \ldots, \xi_m, a, b) = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} a_i[y_i(w'x_i + b) - 1 + \xi_i] - \sum_{i=1}^{n} b_i \xi_i. \]

The KKT conditions are:

\[ \nabla_w L = w - \sum_{i=1}^{m} a_i y_i x_i = 0 \quad \text{with} \quad w = \sum_{i=1}^{m} a_i y_i x_i \]
\[ \nabla_b L = -\sum_{i=1}^{m} a_i y_i = 0 \quad \text{with} \quad \sum_{i=1}^{m} a_i y_i = 0 \]
\[ \nabla_{\xi_i} L = C - a_i - b_i = 0 \quad \text{with} \quad a_i + b_i = C \]

and

\[ a_i[y_i(w'x_i + b) - 1 + \xi_i] = 0 \text{ for } 1 \leq i \leq m \]
\[ a_i = 0 \]
\[ b_i \xi_i = 0 \text{ for } \xi_i = 0. \]
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 $w$:

$$
L = \frac{1}{2} \left\| \sum_{i=1}^{m} a_i y_i x_i \right\|^2 - \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j y_i y_j x'_i x_j \\
- \sum_{i=1}^{m} a_i y_i b + \sum_{i=1}^{m} a_i \\
= \sum_{i=1}^{m} a_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j y_i y_j x'_i x_j,
$$
• the Lagrangean has exactly the same form as in the separable case;
• we need \(a_i \geq 0\) and, in addition \(b_i \geq 0\), which is equivalent to \(a_i \leq C\) (because \(a_i + b_i = C\));

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

\[
\text{maximize for } \mathbf{a} \quad \sum_{i=1}^{m} a_i - \frac{1}{2} \sum_{i,j} a_i a_j y_i y_j \mathbf{x}_i^{'} \mathbf{x}_j \\
\text{subject to } 0 \leq a_i \leq C \text{ and } \sum_{i=1}^{m} a_i y_i = 0 \\
\text{for } 1 \leq i \leq m.
\]
Consequences

- the objective function is concave and differentiable;
- the solution can be used to determine the hypothesis
  \[ h(x) = \text{sign}(w'x + b); \]
- for any support vector \( b_i \) we have
  \[ b = y_i - \sum_{j=1}^{m} a_j y_j x'_i x_j. \]
- 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 $h(x) = w'x + b$ is its distance to the hyperplane $w'x + b = 0$:

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

The margin for a linear classifier $h$ for a sample $S = (x_1, \ldots, x_m)$ is

$$\rho = \min_{1 \leq i \leq m} \frac{y_i(w'x + b)}{\|w\|}.$$
The VCD of the family of hyperplanes in $\mathbb{R}^n$ is $n + 1$. By the application of the VCD bound we have that for any $\delta > 0$, with probability at least $1 - \epsilon$ we have

$$R(h) \leq \text{ERM}(h) + \sqrt{\frac{2d \log \frac{\epsilon m}{d}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$ 

Therefore, we obtain

$$R(h) \leq \text{ERM}(h) + \sqrt{\frac{2(N + 1) \log \frac{\epsilon m}{N+1}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$ 

When $N$ is large compared to $m$ the bound is not helpful.
Theorem

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

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

verifies $d \leq r^2 \Lambda^2$. 
Proof

Suppose that \( \{x_1, \ldots, x_d\} \) is a set that can be fully shattered. Then, for all \( y = (y_1, \ldots, y_d) \in \{-1, 1\}^d \) there exists \( w \) such that \( 1 \leq y_i(w'x) \) for \( 1 \leq i \leq d \).

Summing up these inequalities yields:

\[
d \leq w' \sum_{i=1}^{d} y_i x_i \leq \|w\| \cdot \left\| \sum_{i=1}^{d} y_i x_i \right\| \leq \Lambda \left\| \sum_{i=1}^{d} y_i x_i \right\|.
\]
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 $d \leq \Lambda \left\| \sum_{i=1}^{d} y_i x_i \right\|$ holds for all $\mathbf{y} \in \{-1, 1\}^d$, it holds over expectations and we have

$$d \leq \Lambda E_{\mathbf{y}} \left( \left\| \sum_{i=1}^{d} y_i x_i \right\| \right) \leq \Lambda \left( E_{\mathbf{y}} \left( \left\| \sum_{i=1}^{d} y_i x_i \right\|^2 \right) \right)^{1/2}$$

$$= \Lambda \left( \sum_{i=1}^{m} \sum_{j=1}^{m} E_{\mathbf{y}}(y_i y_j)(x'_i x_j) \right)^{1/2}$$

$$= \Lambda \left( \sum_{i=1}^{d} x'_i x_i \right)^{1/2} \leq \Lambda (dr^2)^{1/2} = \Lambda r \sqrt{d}.$$
Thus,

\[ d \leq \Lambda^2 r^2 \]

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

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

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

\[ d \leq \frac{r^2}{\rho^2}. \]