Support Vector Machines - II

Prof. Dan A. Simovici

UMB
Recall that the optimization problem for SVMs was

\[
\begin{align*}
\text{minimize} \quad & \frac{1}{2} \| \mathbf{w} \|^2 \\
\text{subject to} \quad & y_i (\mathbf{w}' \mathbf{x} + b) \geq 1 \quad \text{for} \quad 1 \leq i \leq m
\end{align*}
\]

Equivalently, the constraints are

\[
1 - y_i (\mathbf{w}' \mathbf{x} + b) \leq 0
\]

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

The Lagrangean is

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

\[
= \frac{1}{2} \| \mathbf{w} \|^2 + \sum_{i=1}^{m} a_i - \sum_{i=1}^{m} a_i y_i \mathbf{w}' \mathbf{x}_i - b \sum_{i=1}^{m} a_i y_i.
\]
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

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

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

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

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 $y_i(w'\mathbf{x}_i + b) = \pm 1$.

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

Since

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

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

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

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

$$= \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j y_i y_j x'_j x_i.$$

Therefore,

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

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

\[
\text{maximize } \sum_{i=1}^{m} a_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j y_i y_j x_i' x_j \\
\text{subject to } a_i \geq 0 \text{ for } 1 \leq i \leq m \text{ and } \sum_{i=1}^{m} a_i y_i = 0.
\]

Note that the objective function depends on \( a_1, \ldots, a_m \).
The constraints are affine, so they are qualified and 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 $S \sim D^m$, where $D$ is the distribution of examples. If the algorithm $A$ 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

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

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,

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

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

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

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

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

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

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

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

If \( h_S \) is the hypothesis returned by the SVM algorithm \( \mathcal{A} \) for a sample \( S \), then

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

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, \( h_{S \setminus \{x\}} = h_S \) and \( h_{S \setminus \{x\}} \) correctly classifies \( x \). Thus, if \( h_{S \setminus \{x\}} \) misclassifies \( x \), then \( x \) must be a support vector which implies

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

Taking the expectation of both sides yields the result.
Slack Variables

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

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

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:

\[
\nabla_w L = w - \sum_{i=1}^{m} a_i y_i x_i = 0 \quad \Rightarrow \quad w = \sum_{i=1}^{m} a_i y_i x_i
\]

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

\[
\nabla_{\xi_i} L = C - a_i - b_i = 0 \quad \Rightarrow \quad a_i + b_i = C
\]

and

\[a_i [y_i(w'x_i + b) - 1 + \xi_i] = 0 \text{ for } 1 \leq i \leq m \Rightarrow a_i = 0 \text{ or } y_i(w'x_i + b) = 1 - \xi_i,\]

\[b_i \xi_i = 0 \Rightarrow b_i = 0 \text{ or } \xi_i = 0.\]
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 Lagrangean can be rewritten by substituting $\mathbf{w}$:

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

$$\text{maximize for } a \sum_{i=1}^{m} a_i - \frac{1}{2} a_i a_j y_i y_j x_i^t x_j$$
subject to $0 \leq a_i \leq C$ and $\sum_{i=1}^{m} a_i y_i = 0$
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 \(i \neq j\), \(E(y_iy_j) = E(y_i)E(y_j) = 0\); also, \(E(y_iy_i) = 1\).

Since \(d \leq \Lambda \| \sum_{i=1}^{d} y_i x_i \|\) holds for all \(y \in \{-1, 1\}^d\), it holds over expectations and we have

\[
d \leq \Lambda E_y \left( \left\| \sum_{i=1}^{d} y_i x_i \right\| \right) \leq \Lambda \left( E_y \left( \left\| \sum_{i=1}^{d} y_i x_i \right\|^2 \right) \right)^{1/2}
\]

\[
= \Lambda \left( \sum_{i=1}^{m} \sum_{j=1}^{m} E_y (y_iy_j)(x'_i x'_j) \right)^{1/2}
\]

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

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