Support Vector Machines - II

Prof. Dan A. Simovici

UMB

Recall that the optimization problem for SVMs was

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

Equivalently, the constraints are

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

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

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

The Dual Problem

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

The KKT conditions are

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

which are equivalent to

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

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

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

Therefore,

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

The Dual Optimization Problem for Separable Sets

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

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

- 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 $S = \{(\mathbf{x}_j, y_j) \mid 1 \leqslant j \leqslant m\}$.
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 $\mathbf{x}_i$, that is, is trained on $S - \{\mathbf{x}_i\}$ the hypothesis returned is $h_{S-\{\mathbf{x}_i\}}$ and the error is

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

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 $m \geqslant 2$ is an unbiased estimate of the average generalization error for sample of size $m-1$, that is,

$$E_{S\sim\mathcal{D}^m}\left(ERM_{LOO}(\mathcal{A})\right) = E_{S'\sim\mathcal{D}^{m-1}}\left(R(h_{S'}\right).$$

Proof

$$
\begin{aligned}
&E_{S\sim\mathcal{D}^m}\left(\mathrm{ERM}_{LOO}(\mathcal{A})\right)\\
&= \frac{1}{m}\sum_{i=1}^{m} E_{S\sim\mathcal{D}^m}\left(h_{S-\{\mathbf{x}_i\}}(\mathbf{x}_i)\neq y_i\right)\\
&= E_{S\sim\mathcal{D}^m}\left(h_{S-\{\mathbf{x}_1\}}(\mathbf{x}_1)\neq y_1\right)\\
&\quad \text{(since all points of } S \text{ are drawn at random and are equally distributed)}\\
&= E_{S'\sim\mathcal{D}^{m-1},x_1\sim\mathcal{D}}\left(h_{S'}(\mathbf{x}_1)\neq y_1\right)\\
&= E_{S'\sim\mathcal{D}^{m-1}}\left(E_{x_1\sim\mathcal{D}}\left(h_{S'}(\mathbf{x}_1)\neq y_1\right)\right)\\
&= E_{S'\sim\mathcal{D}^{m-1}}\left(R(h_{S'})\right).
\end{aligned}
$$

Theorem

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

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

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

$$\mathrm{ERM}_{LOO}(\mathcal{A}) \leqslant \frac{N_{SV}(S)}{m+1}.$$

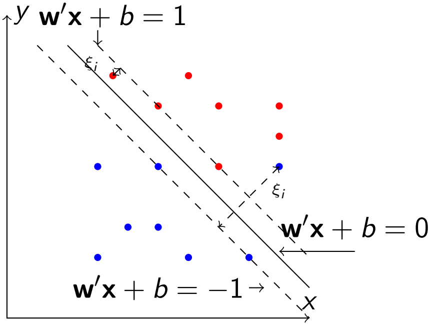
Taking the expectation of both sides yields the result.

Slack Variables

If data is not separable the conditions $y_i(\mathbf{w}'\mathbf{x}_i + b) \geqslant 1$ 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

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

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:

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

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

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

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 $\mathbf{w}$:

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

- 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

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

- 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.

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

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

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

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

Therefore, we obtain

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

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, $S \subseteq \{\mathbf{x} \mid \parallel \mathbf{x} \parallel \leqslant r\}$. The VC dimension of the set of canonical hyperplanes of the form

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

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

Proof

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

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

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 $d \leqslant \Lambda \left\| \sum_{i=1}^{d} y_i \mathbf{x}_i \right\|$ holds for all $\mathbf{y} \in \{-1, 1\}^d$, it holds over expectations and we have

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

Thus,

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

- recall that when the data is linearly separable the margin ρ is given by:

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

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

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