GI01/4C55: Supervised Learning

6. Support Vector Machines

November 14, 2005

Massimiliano Pontil
Today’s plan

- Optimal separating hyperplane
- Soft margin separation
- Support vector machines
- Connection to regularization

**Bibliography:** These lecture notes are available at:
http://www.cs.ucl.ac.uk/staff/M.Pontil/courses/index-SL05.htm
Separating hyperplane

Let \( S = \{(x_i, y_i)\}_{i=1}^m \in \mathbb{R}^d \times \{-1, 1\} \) be a training set.

By a hyperplane we mean a set \( H_{w,b} = \{x \in \mathbb{R}^d : w^\top x + b = 0\} \) (affine linear space) parameterized by \( w \in \mathbb{R}^d \) and \( b \in \mathbb{R} \).

We assume that the data are linearly separable, that is, there exist \( w \in \mathbb{R}^d \) and \( b \in \mathbb{R} \) such that

\[ y_i(w^\top x_i + b) > 0, \quad i = 1, \ldots, m \]  

(1)

in which case we call \( H_{w,b} \) a separating hyperplane.

Note that we require the inequality in eq.(1) to be strict (we do not admit that the data lie on a hyperplane).
Separating hyperplane (cont.)

The distance $\rho_x(w, b)$ of a point $x$ from a hyperplane $H_{w,b}$ is

$$\rho_x(w, b) := \frac{|w^\top x + b|}{||w||}$$

If $H_{w,b}$ separates the training set $S$ we define its margin as

$$\rho_S(w, b) := \min_{i=1}^{m} \rho_{x_i}(w, b)$$

If $H_{w,b}$ is a hyperplane (separating or not) we also define the margin of a point $x$ as $w^\top x + b$ (note that this can be positive or negative)
Optimal separating hyperplane (OSH)

This is the separating hyperplane with maximum margin. It solves the optimization problem

$$\rho(S) := \max_{w,b} \min_i \left\{ \frac{y_i (w^\top x_i + b)}{\|w\|} : y_i (w^\top x_i + b) \geq 0, \ i = 1, \ldots, m \right\} > 0$$
Choosing a parameterization

A separating hyperplane is parameterized by \((w, b)\), but this choice is not unique (rescaling with a positive constant gives the same separating hyperplane). Two possible ways to fix the parameterization:

- **Normalized hyperplane:** set \(\|w\| = 1\), in which case \(\rho_x(w, b) = |w^\top x + b|\) and \(\rho_S(w, b) = \min_{i=1}^{m} y_i(w^\top x_i + b)\)

- **Canonical hyperplane:** choose \(\|w\|\) such that \(\rho_S(w, b) = \frac{1}{\|w\|}\), i.e. we require that \(\min_{i=1}^{m} y_i(w^\top x_i + b) = 1\) (a data-dependent parameterization)

We will mainly work with the second parameterization
Optimal separating hyperplane

- If we work with normalized hyperplanes we have
  \[ \rho(S) = \max_{w,b} \min_i \left\{ y_i(w^\top x_i + b) : y_i(w^\top x_i + b) \geq 0, ||w|| = 1 \right\} \]

- If we work with canonical hyperplanes, instead, we have
  \[
  \rho(S) = \max_{w,b} \left\{ \frac{1}{||w||} : \min_i \{y_i(w^\top x_i + b)\} = 1, y_i(w^\top x_i + b) \geq 0 \right\} \\
  = \max_{w,b} \left\{ \frac{1}{||w||} : y_i(w^\top x_i + b) \geq 1 \right\} \\
  = \frac{1}{\min_{w,b} \{||w|| : y_i(w^\top x_i + b) \geq 1\}}
  \]
Optimal separating hyperplane (cont.)

We choose to work with canonical hyperplanes and, so, look at the optimization problem

<table>
<thead>
<tr>
<th>Problem P1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimize $\frac{1}{2}w^T w$</td>
</tr>
<tr>
<td>subject to $y_i(w^T x_i + b) \geq 1$, $i = 1, \ldots, m$</td>
</tr>
</tbody>
</table>

The quantity $1/\|w\|$ is the **margin** of the OSH
Saddle point

The solution of problem $P1$ is equivalent to determine the **saddle point** of the Lagrangian function

$$L(w, b; \alpha) = \frac{1}{2}w^\top w - \sum_{i=1}^{m} \alpha_i \left\{ y_i(w^\top x_i + b) - 1 \right\}$$  \hspace{1cm} (2)$$

where $\alpha_i \geq 0$ are the Lagrange multipliers

We minimize $L$ over $(w, b)$ and maximize over $\alpha$. Differentiating w.r.t $w$ and $b$ we obtain:

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{m} y_i \alpha_i = 0$$

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{m} \alpha_i y_i x_i$$  \hspace{1cm} (3)
Dual problem

Substituting eq.(3) in eq.(2) leads to the **dual problem**

```
Problem P2

Maximize \( Q(\alpha) := -\frac{1}{2}\alpha^T A\alpha + \sum_i \alpha_i \)

subject to \( \sum_i y_i \alpha_i = 0 \)
\( \alpha_i \geq 0, \quad i = 1, \ldots, m \)
```

where \( A \) is an \( m \times m \) matrix \( A = (y_i y_j x_i^T x_j : i, j = 1, \ldots, m) \)

Note that the complexity of this problem depends on \( m \), not on the number of input components \( d \) (same as ridge regression).
Kuhn-Tucker conditions and support vectors

If $\bar{\alpha}$ is a solution of the dual problem then the solution $(\bar{w}, \bar{b})$ of the primal problem is given by

$$\bar{w} = \sum_{i=1}^{m} \bar{\alpha}_i y_i x_i$$

Note that $\bar{w}$ is a linear combination of only the $x_i$ for which $\bar{\alpha}_i > 0$. These $x_i$ are termed support vectors (SVs).

Parameter $\bar{b}$ can be determined by looking at the Kuhn-Tucker conditions

$$\bar{\alpha}_i \left( y_i (\bar{w}^\top x_i + \bar{b}) - 1 \right) = 0$$

Specifically if $x_j$ is a SV we have that

$$\bar{b} = y_j - \bar{w}^\top x_j$$
Some remarks

• The fact that the OSH is determined only by the SVs is most remarkable. Usually, the support vectors are a small subset of the training data

• All the information contained in the data set is summarized by the support vectors: The whole data set could be replaced by only these points and the same hyperplane would be found

• A new point \( x \) is classified as \( \text{sgn} \left( \sum_{i=1}^{m} y_i \alpha_i x_i^T x + b \right) \)
Linearily nonseparable case

If the data is not linearly separable (or one simply ignores whether this is the case) the previous analysis can be generalized by looking at the problem

Problem $\mathbf{P3}$

Minimize $\frac{1}{2}w^T w + C \sum_{i=1}^m \xi_i$

subject to $y_i (w^T x_i + b) \geq 1 - \xi_i$, $\xi_i \geq 0$, $i = 1, \ldots, m$

The idea is to introduce the slack variables $\xi_i$ to relax the separation constraints ($\xi_i > 0 \Rightarrow x_i$ has margin less than 1)
New dual problem

A saddle point analysis (similar to that above) leads to the dual problem

Problem **P4**

Maximize

\[ Q(\alpha) := -\frac{1}{2}\alpha^\top A\alpha + \sum_i \alpha_i \]

subject to

\[ \sum_i y_i \alpha_i = 0 \]
\[ 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, m \]

This is like problem **P2** except that now we have “box constraints” on \( \alpha_i \). If the data is linearly separable, by choosing \( C \) large enough we obtain the OSH
Nonseparable case (cont)

Again we have

$$\tilde{w} = \sum_{i=1}^{m} \tilde{\alpha}_i y_i x_i,$$

while $\tilde{b}$ can be determined from $\tilde{\alpha}$, solution of the problem $\textbf{P4}$, and from the new Kuhn-Tucker conditions

$$\tilde{\alpha}_i \left( y_i (\tilde{w}^T x_i + \tilde{b}) - 1 + \xi_i \right) = 0 \quad (*)$$

$$\left( C - \tilde{\alpha}_i \right) \xi_i = 0 \quad (**)$$

Again, points for which $\tilde{\alpha}_i > 0$ are termed support vectors
A closer look at the KKT conditions

Equation (*) tell us that if

- \( y_i(\mathbf{w}^\top \mathbf{x}_i + \bar{b}) > 1 \Rightarrow \bar{\alpha}_i = 0 \) (not a SV)

- \( y_i(\mathbf{w}^\top \mathbf{x}_i + \bar{b}) < 1 \Rightarrow \bar{\alpha}_i = C \) (a SV with positive slack \( \xi_i \))

- \( y_i(\mathbf{w}^\top \mathbf{x}_i + \bar{b}) = 1 \Rightarrow \bar{\alpha}_i \in [0, C] \) (if \( \bar{\alpha}_i > 0 \) a SV “on the margin”)

Remark: Conversely, from eqs.(*),(**) if \( \bar{\alpha}_i = 0 \) then \( y_i(\mathbf{w}^\top \mathbf{x}_i + \bar{b}) \geq 1, \xi_i = 0 \); if \( \bar{\alpha}_i \in (0, C) \) then \( y_i(\mathbf{w}^\top \mathbf{x}_i + \bar{b}) = 1, \xi_i = 0 \); if \( \bar{\alpha}_i = C \) then \( y_i(\mathbf{w}^\top \mathbf{x}_i + \bar{b}) \leq 1, \xi_i \geq 0 \)
The role of the parameter $C$

Optimal separating hyperplane for four increasing values of $C$. Both the margin and the training error are non-increasing functions of $C$. 
The role of the parameter $C$ (cont.)

The parameter $C$ controls the trade-off between $\|w\|^2$ and the training error $\sum_{i=1}^{m} \xi_i$

It can be shown that the optimal value of the Lagrange multipliers $\bar{\alpha}_i$ (and, so, $\bar{w}, \bar{b}$) are piecewise continuous functions of $C$. This helps re-computing the solution when varying $C$.

$C$ is often selected by minimizing the leave-one-out cross validation error
Support Vector Machines (SVMs)

The above analysis holds true if we work with a feature map \( \phi : \mathcal{X} \to \mathcal{W} \). We simply replace \( x \) by \( \phi(x) \) and \( x^\top t \) by \( \langle \phi(x), \phi(t) \rangle = K(x, t) \)

An SVM with kernel \( K \) is the function

\[
f(x) = \sum_{i=1}^{m} y_i \alpha_i K(x_i, x) + b, \quad x \in \mathcal{X}
\]

where the parameters \( \alpha_i \) solve problem \( P4 \) with \( A = (y_i y_j K(x_i, x_j) : i, j = 1, \ldots, m) \) and \( b \) is obtained as discussed above

A new point \( x \in \mathcal{X} \) is classified as \( \text{sgn}(f(x)) \)
Connection to regularization

The SVM formulation above is equivalent to the problem

\[ E_\lambda(w, b) = \sum_{i=1}^{m} \max(1 - y_i(\langle w, \phi(x_i) \rangle + b), 0) + \lambda \|w\|^2 \]

with \( \lambda = \frac{1}{2C} \)

In fact, we have

\[
\begin{align*}
\min_{w,b,\xi} & \left\{ C \sum_{i=1}^{m} \xi_i + \frac{1}{2} \|w\|^2 : y_i(\langle w, \phi(x_i) \rangle + b) \geq 1 - \xi_i, \xi_i \geq 0 \right\} = \\
\min_{w,b} & \left\{ \min_{\xi} \left\{ C \sum_{i=1}^{m} \xi_i + \frac{1}{2} \|w\|^2 : \xi_i \geq 1 - y_i(\langle w, \phi(x_i) \rangle + b), \xi_i \geq 0 \right\} \right\} = \\
\min_{w,b} & \left\{ C \sum_{i=1}^{m} \max(1 - y_i(\langle w, \phi(x_i) \rangle + b), 0) + \frac{1}{2} \|w\|^2 \right\} = CE_{\frac{1}{2C}}(w, b)
\end{align*}
\]
SVM regression

SVM’s can be developed for regression as well. Here we choose the loss \( |y - f(x)|_\epsilon = \max(|y - f(x)| - \epsilon, 0) \)

\[
\begin{align*}
\text{Minimize} & \quad \frac{1}{2} w^T w + C \sum_{i=1}^{m} (\xi_i + \xi_i^*) \\
\text{subject to} & \quad w^T x_i + b - y_i \leq \epsilon + \xi_i, \\
& \quad y_i - w^T x_i - b \leq \epsilon + \xi_i^*, \\
& \quad \xi_i, \xi_i^* \geq 0, \quad i = 1, \ldots, m
\end{align*}
\]

SVMs loss functions (both for classification and regression) are scale sensitive: errors below a certain resolution do not count. This leads to sparse solutions!
Solution methods

The above optimization problems are Quadratic Programming (QP) problems. Several methods (eg, interior point methods) from convex optimization exist for solving QP problems.

If we work with a non-linear kernel, the number of underlying features, \( N \), is typically much larger (or infinite) than the number of examples. Thus, we need to solve the dual problem.

However, if \( m \gg N \) it is more efficient to solve the primal problem.
Decomposition of the dual problem

For large datasets (say $m > 10^5$) it is practically impossible to solve the dual problem with standard optimization techniques (matrix $A$ is dense!)

A typical approach is to iteratively optimize wrt. an “active set” $\mathcal{A}$ of variables. Set $\alpha = 0$, choose $q \leq m$ and a subset $\mathcal{A}$ of $q$ variables, $\mathcal{A} = \{\alpha_{i_1}, \ldots, \alpha_{i_q}\}$. We repeat until convergence:

- Optimize $Q(\alpha)$ wrt. the variables in $\mathcal{A}$
- Remove one variable from $\mathcal{A}$ which satisfies the KKT conditions and add one variable, if any, which violates the KKT conditions. If no such variable exists stop

One can show that after each iteration $Q$ increases
Removing $b$

If we are looking for a hyperplane which passes through the origin we do not need to optimize wrt. $b$ (set $b \equiv 0$)

In this case we have the simplified dual problem

Problem $P4'$

Maximize \[-\frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j) + \sum \alpha_i\]

subject to \[0 \leq \alpha_i \leq C, \quad i = 1, \ldots, m\]

(the constraint $\sum_{i=1}^{m} y_i \alpha_i = 0$ disappears)
b or not b?

In general, however, it is important to learn b as well

A simple way is to introduce a small regularization on b, i.e. to set \( K'(x,t) = K(x,t) + \lambda_0 \) and solve problem \( \mathbf{P4'} \) using this new kernel

In the limit \( \lambda_0 \to \infty \), the regularization on b is removed and we get the additional constraint \( \sum_{i=1}^{m} \alpha_i y_i = 0 \) (see why?), so, we are back to problem \( \mathbf{P4} \)
Conditionally positive semidefinite kernels

Note that in order to have a solution to problem $P4$ the kernel only needs to be *conditionally positive semidefinite* (cpsd)

Definition: A kernel $K$ is cpsd if for every $m > 0$ and $x_1, \ldots, x_m \in \mathbb{R}^d$, $\sum_{i,j=1}^{m} c_i c_j K(x_i, x_j) \geq 0$ if $\sum_{i=1}^{m} c_i = 0$

Example: The kernel $K(x, t) = -\|x - t\|^2$ is conditionally positive semidefinite but not positive semidefinite

In fact, for every $x_i \in \mathbb{R}^d, c_i \in \mathbb{R}, i = 1, \ldots, m$ such that $\sum_{i} c_i = 0$ we have that

$$- \sum_{i,j} c_i c_j \|x_i - x_j\|^2 = -2 \sum_i c_i \sum_j c_j \|x_j\|^2 + 2 \sum_{i,j} c_i c_j x_i^\top x_j$$

$$= 2 \sum_{i,j} c_i c_j x_i^\top x_j = 2 \left\| \sum_i c_i x_i \right\|^2 \geq 0$$

(however, this is not psd because if all $c_i$ are non-negative we get “≤”)