# 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
Lecture notes are in part based on: Pontil & Verri. Properties of support vector machines. *Neural Comp.*, 10: 955–974, 1998. See also: Shawe-Taylor and Cristianini, Chapter 7.2; Hastie, Tibshirani & Friedman, Chapter 12.2-3

# Separating hyperplane

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

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

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

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 0, \quad i = 1, \ldots, m \tag{1}$$

in which case we call $H_{\mathbf{w},b}$ 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 $\rho_{\mathbf{x}}(\mathbf{w}, b)$ of a point $\mathbf{x}$ from a hyperplane $H_{\mathbf{w},b}$ is

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

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

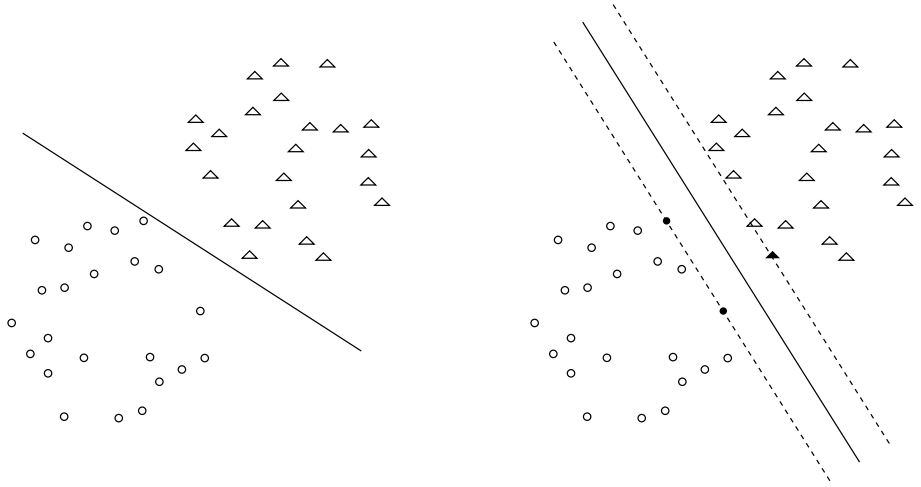
$$\rho_S(\mathbf{w}, b) := \min_{i=1}^{m} \rho_{\mathbf{x}_i}(\mathbf{w}, b)$$

If $H_{\mathbf{w},b}$ is a hyperplane (separating or not) we also define the *margin of a point* $\mathbf{x}$ as $\mathbf{w}^\top \mathbf{x} + b$ (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

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



(a) (b)

# Choosing a parameterization

A separating hyperplane is parameterized by $(\mathbf{w}, b)$, 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 $\|\mathbf{w}\| = 1$, in which case $\rho_{\mathbf{x}}(\mathbf{w}, b) = |\mathbf{w}^\top \mathbf{x} + b|$ and $\rho_S(\mathbf{w}, b) = \min_{i=1}^m y_i(\mathbf{w}^\top \mathbf{x}_i + b)$

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

We will mainly work with the second parameterization

# Optimal separating hyperplane

- If we work with normalized hyperplanes we have

$$\rho(S) = \max_{\mathbf{w},b} \min_i \left\{ y_i(\mathbf{w}^\top \mathbf{x}_i + b) \ : \ y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 0, \|\mathbf{w}\| = 1 \right\}$$

- If we work with canonical hyperplanes, instead, we have

$$\begin{aligned}
\rho(S) &= \max_{\mathbf{w},b} \left\{ \frac{1}{\|\mathbf{w}\|} : \min_i \{ y_i(\mathbf{w}^\top \mathbf{x}_i + b) \} = 1, y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 0 \right\} \\
&= \max_{\mathbf{w},b} \left\{ \frac{1}{\|\mathbf{w}\|} : y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \right\} \\
&= \frac{1}{\min_{\mathbf{w},b} \{ \|\mathbf{w}\| : y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \}}
\end{aligned}$$

## Optimal separating hyperplane (cont.)

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

| | |
|---|---|
| Problem **P1** | |
| Minimize | $\frac{1}{2}\mathbf{w}^\top\mathbf{w}$ |
| subject to | $y_i(\mathbf{w}^\top\mathbf{x}_i + b) \geq 1,\ i = 1, \ldots, m$ |

The quantity $1/\|\mathbf{w}\|$ is the **margin** of the OSH

# Saddle point

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

$$L(\mathbf{w}, b; \alpha) = \frac{1}{2}\mathbf{w}^\top \mathbf{w} - \sum_{i=1}^{m} \alpha_i \left\{ y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1 \right\} \qquad (2)$$

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

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

$$\begin{aligned} \frac{\partial L}{\partial b} &= -\sum_{i=1}^{m} y_i \alpha_i = 0 \\ \frac{\partial L}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i = 0 \;\Rightarrow \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \end{aligned} \qquad (3)$$

# Dual problem

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

Problem **P2**

Maximize $\quad Q(\alpha) := -\frac{1}{2}\boldsymbol{\alpha}^\top A\boldsymbol{\alpha} + \sum_i \alpha_i$

subject to $\quad \sum_i y_i\alpha_i = 0$

$\alpha_i \geq 0, \qquad i = 1, \ldots, m$

where $\mathbf{A}$ is an $m \times m$ matrix $\mathbf{A} = (y_i y_j \mathbf{x}_i^\top \mathbf{x}_j : i, j = 1, \ldots, m)$

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 $(\bar{\mathbf{w}}, \bar{b})$ of the primal problem is given by

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

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

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

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

Specifically if $\mathbf{x}_j$ is a SV we have that

$$\bar{b} = y_j - \bar{\mathbf{w}}^\top \mathbf{x}_j$$

## Some remarks

- The fact that 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 $\mathbf{x}$ is classified as $\operatorname{sgn}\left(\sum_{i=1}^{m} y_i \bar{\alpha}_i \mathbf{x}_i^\top \mathbf{x} + \bar{b}\right)$

## Linearly 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 **P3**

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

subject to $y_i(\mathbf{w}^\top\mathbf{x}_i + b) \geq 1 - \xi_i,$
$\xi_i \geq 0, \qquad i = 1, \ldots, m$

The idea is to introduce the slack variables $\xi_i$ to relax the separation constraints ($\xi_i > 0 \Rightarrow \mathbf{x}_i$ has margin less than 1)

# New dual problem

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

Problem **P4**

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

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

$$\bar{\mathbf{w}} = \sum_{i=1}^{m} \bar{\alpha}_i y_i \mathbf{x}_i,$$

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

$$\begin{aligned} \bar{\alpha}_i \left( y_i(\bar{\mathbf{w}}^\top \mathbf{x}_i + \bar{b}) - 1 + \bar{\xi}_i \right) &= 0 \quad (*) \\ (C - \bar{\alpha}_i)\bar{\xi}_i &= 0 \quad (**) \end{aligned}$$

Again, points for which $\bar{\alpha}_i > 0$ are termed **support vectors**

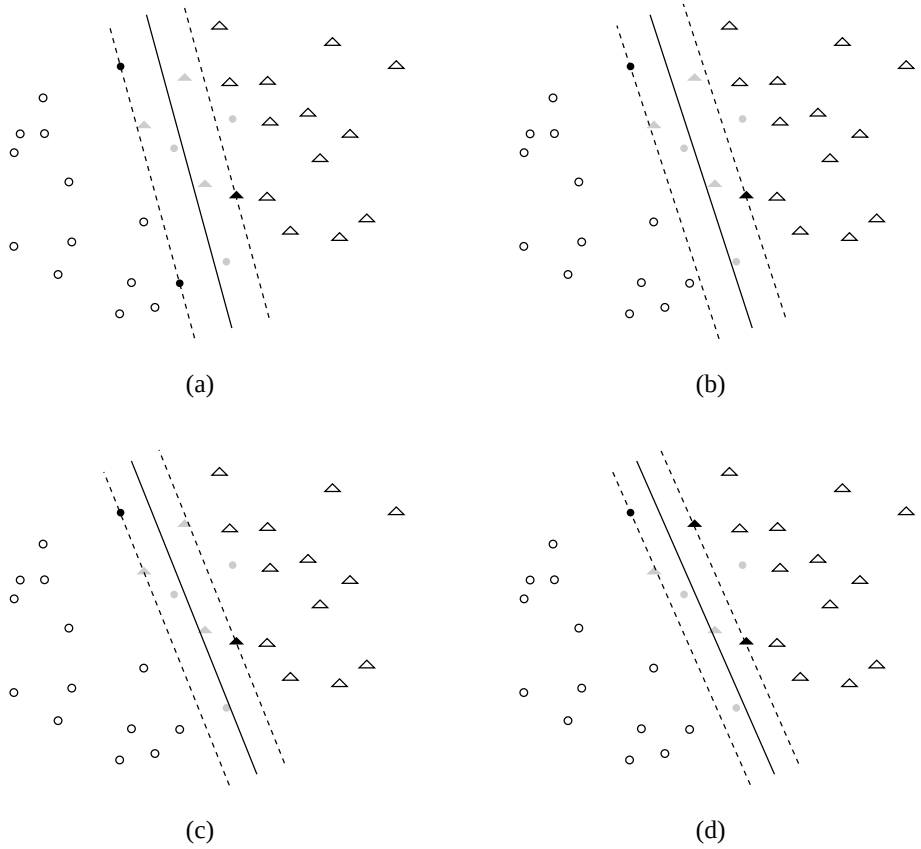
# A closer look at the KKT conditions

Equation (*) tell us that if

- $y_i(\bar{\mathbf{w}}^\top \mathbf{x}_i + \bar{b}) > 1 \Rightarrow \bar{\alpha}_i = 0$ (not a SV)
- $y_i(\bar{\mathbf{w}}^\top \mathbf{x}_i + \bar{b}) < 1 \Rightarrow \bar{\alpha}_i = C$ (a SV with positive slack $\bar{\xi}_i$)
- $y_i(\bar{\mathbf{w}}^\top \mathbf{x}_i + \bar{b}) = 1 \Rightarrow \bar{\alpha}_i \in [0, C]$ (if $\bar{\alpha}_i > 0$ a SV "on the margin")

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

# 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 $\|\mathbf{w}\|^2$ 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, $\bar{\mathbf{w}}, \bar{b}$) 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 $\mathbf{x}$ by $\phi(\mathbf{x})$ and $\mathbf{x}^\top \mathbf{t}$ by $\langle \phi(\mathbf{x}), \phi(\mathbf{t}) \rangle = K(\mathbf{x}, \mathbf{t})$

An SVM with kernel $K$ is the function

$$f(x) = \sum_{i=1}^{m} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b, \quad \mathbf{x} \in \mathcal{X}$$

where the parameters $\alpha_i$ solve problem **P4** with $\mathbf{A} = (y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) : i, j = 1, \ldots, m)$ and $b$ is obtained as discussed above

A new point $\mathbf{x} \in \mathcal{X}$ is classified as $\text{sgn}(f(\mathbf{x}))$

## Connection to regularization

The SVM formulation above is equivalent to the problem

$$E_\lambda(\mathbf{w}, b) = \sum_{i=1}^{m} \max(1 - y_i(\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle + b), 0) + \lambda \|\mathbf{w}\|^2$$

with $\lambda = \frac{1}{2C}$

In fact, we have

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \left\{ C \sum_{i=1}^{m} \xi_i + \frac{1}{2}\|\mathbf{w}\|^2 : y_i(\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle + b) \geq 1 - \xi_i, \xi_i \geq 0 \right\} =$$

$$\min_{\mathbf{w},b} \left\{ \min_{\boldsymbol{\xi}} \left\{ C \sum_{i=1}^{m} \xi_i + \frac{1}{2}\|\mathbf{w}\|^2 : \xi_i \geq 1 - y_i(\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle + b), \xi_i \geq 0 \right\} \right\} =$$

$$\min_{\mathbf{w},b} \left\{ C \sum_{i=1}^{m} \max\left(1 - y_i(\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle + b), 0\right) + \frac{1}{2}\|\mathbf{w}\|^2 \right\} = C E_{\frac{1}{2C}}(\mathbf{w}, b)$$

# SVM regression

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

| | |
|---|---|
| Minimize | $\frac{1}{2}\mathbf{w}^\top\mathbf{w} + C\sum_{i=1}^{m}(\xi_i + \xi_i^*)$ |
| subject to | $\mathbf{w}^\top\mathbf{x}_i + b - y_i \leq \epsilon + \xi_i,$ |
| | $y_i - \mathbf{w}^\top\mathbf{x}_i - b \leq \epsilon + \xi_i^*,$ |
| | $\xi_i, \xi_i^* \geq 0, \quad i = 1, \ldots, m$ |

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 $\mathbf{A}$ is dense!)

A typical approach is to iteratively optimize wrt. an "active set" $\mathcal{A}$ of variables. Set $\boldsymbol{\alpha} = 0$, 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 $\quad -\frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) + \sum_i \alpha_i$

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

(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 $K'(\mathbf{x}, \mathbf{t}) = K(\mathbf{x}, \mathbf{t}) + \lambda_0$ and solve problem **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 **P4**

## Conditionally positive semidefinite kernels

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

**Definition:** A kernel $K$ is cpsd if for every $m > 0$ and $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^d$, $\sum_{i,j=1}^m c_i c_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$ if $\sum_{i=1}^m c_i = 0$

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

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

$$\begin{aligned} -\sum_{i,j} c_i c_j \|\mathbf{x}_i - \mathbf{x}_j\|^2 &= -2\sum_i c_i \sum_j c_j \|\mathbf{x}_j\|^2 + 2\sum_{i,j} c_i c_j \mathbf{x}_i^\top \mathbf{x}_j \\ &= 2\sum_{i,j} c_i c_j \mathbf{x}_i^\top \mathbf{x}_j = 2\left\|\sum_i c_i \mathbf{x}_i\right\|^2 \geq 0 \end{aligned}$$

(however, this is not psd because if all $c_i$ are non-negative we get "$\leq$")