Support Vector Machines Models

Laura Palagi[1]

[1]Department of Computer, Control, and Management Engineering Antonio Ruberti, Sapienza University of Rome

1st MINOA PhD school
Mixed-Integer Nonlinear Optimization meets Data Science
Ischia (Italy) - June 25, 2019

MINOA
MIXED-INTEGER NON-LINEAR OPTIMISATION: ALGORITHMS AND APPLICATIONS

SAPIENZA
UNIVERSITÀ DI ROMA

SVM formulation

1. Hyperplanes
2. Wolfe duality theory
3. Unconstrained formulation

SAPIENZA UNIVERSITÀ DI ROMA

Recap

Supervised binary classification

$$\mathcal{T} = \{(x^i, y^i) : x_i \in \mathbb{R}^n,\ y^i \in \{-1, 1\},\ i = 1, \ldots, l\},$$

Manage the trade-off

- Empirical risk $\rightarrow 0$
- Complexity h of $\mathcal{F}$

$$\min R_{emp}(\omega) + C_{VC}(h, l)$$

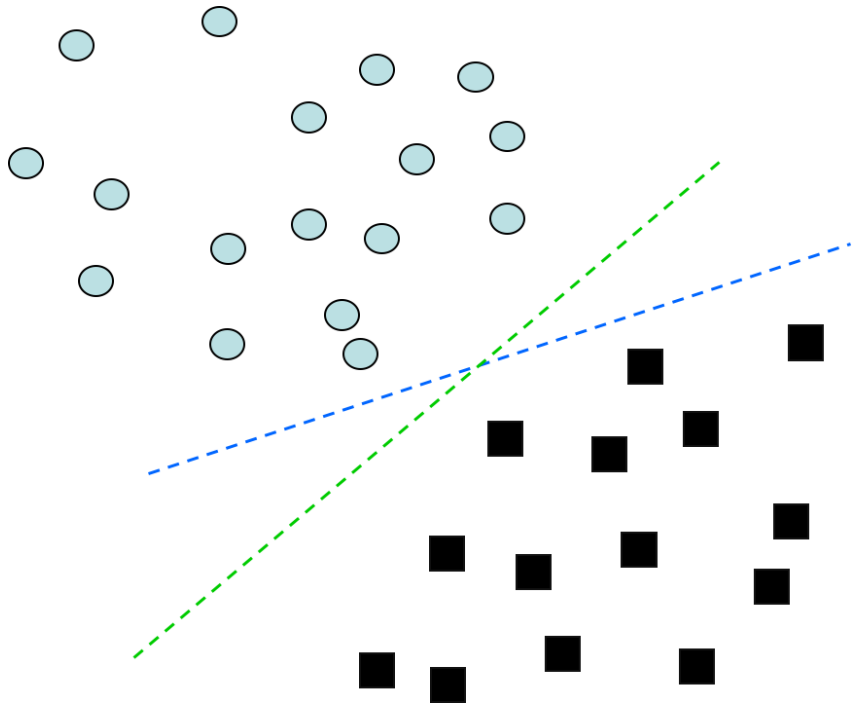
- we need to compute h for a given class of functions
- NOT an easy task in general
- h is the maximum number of points that can be classified for all possible assigned labels ± 1

SAPIENZA
UNIVERSITÀ DI ROMA

Hyperplane: VC dimension

Consider the linear classifiers

$$f_{w,b}(x) = \mathrm{sgn}(w^T x + b)$$



Both hyperplanes have $R_{emp} = 0$, and $h = 3$

$$R(\omega) \leq 0 + C_{VC}(3, l)$$

Hyperplane VC dimension

Theorem: VC dimension of linear classifiers

The VC dimension of $f(x; w, b) = w^T x + b$ in $\mathbb{R}^n$ is $n + 1$

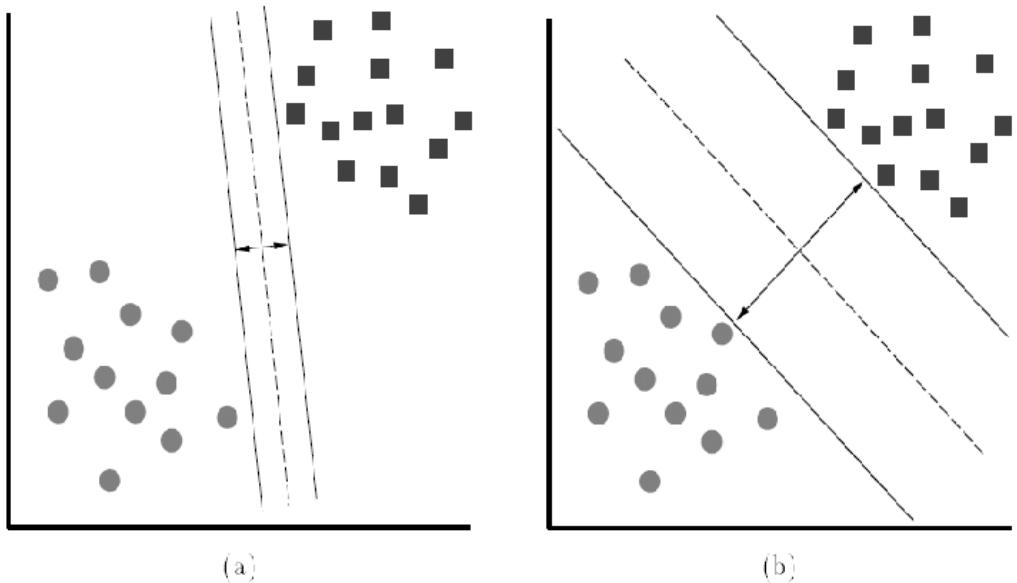
This follows from the fact that p points $x^1, \ldots, x^p$ can be shattered by the class of linear functions if and only if the p points are affinely independent (i.e. $x^2 - x^1, \ldots, x^p - x^1$ are linearly independent)

Since in $\mathbb{R}^n$ the maximum number of linearly independent vectors is n, we get $h = n + 1$

Any separating hyperplane gives the same upper bound

SAPIENZA UNIVERSITÀ DI ROMA

Margin: Intuition



Both hyperplanes have zero empirical risk, but what about the generalization capability.

Which hyperplane is better?

A hyperplane too close to the training examples will be sensitive to noise and less likely to generalize well for new data

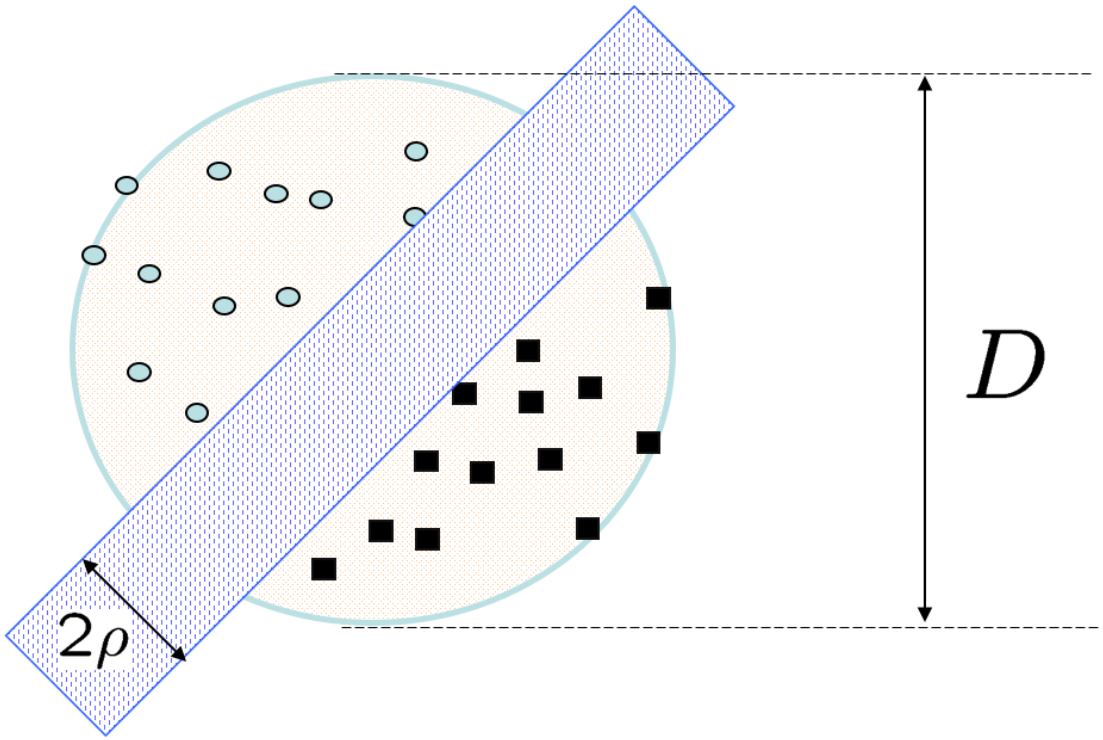
A hyperplane that is far from all the training examples would have better generalization capabilities

Hyperplan with margin

Idea

restrict the choice in the class of linear classifier by restricting to hyperplanes with tolerance gap

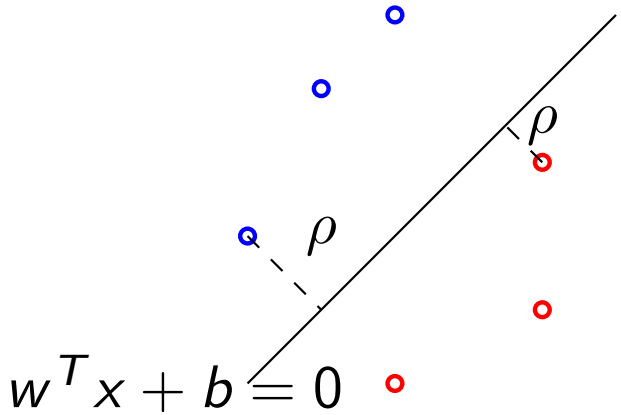
Assume the data belong to a sphere with diameter D.



Evaluate the margin

The distance of a point x^i from an hyperplane is obtained as the solution of the *projection problem* on the hyperplane $w^T z + b = 0$

$$\begin{aligned} \min_{z \in \mathbb{R}^n} \quad & \|x^i - z\| \\ & w^T z + b = 0 \end{aligned}$$

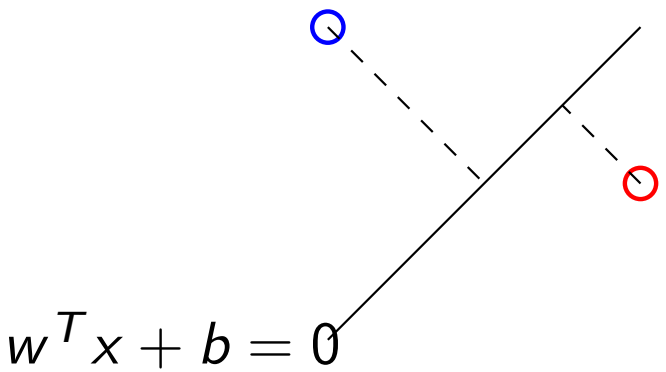


Applying Lagrangian optimality condition we get

$$d(x^i; w, b) = \frac{|w^T x^i + b|}{\|w\|}$$

Hyperplanes with margin

Consider the hyperplane $w^T x + b = 0$



Let

$$\varrho(x^i, \mathcal{H}(w, b)) = \frac{w^T x^i + b}{\|w\|}$$

The distance between x^i and the hyperplane is
$d(x^i; w, b) = |\varrho(x^i; w, b)|$

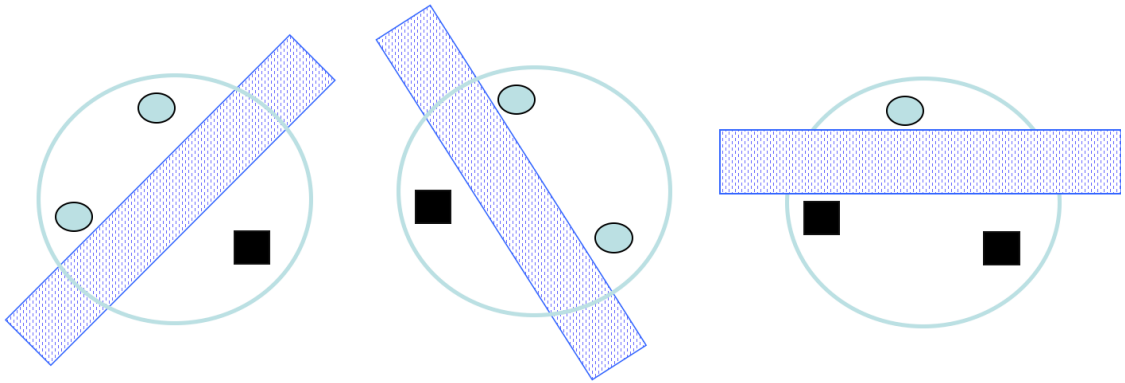
Consider the classifier

$$f(x^i) = \begin{cases} 1 & \text{if } \varrho(x^i; w, b) \geq \rho \\ -1 & \text{if } \varrho(x^i; w, b) \leq -\rho \end{cases}$$

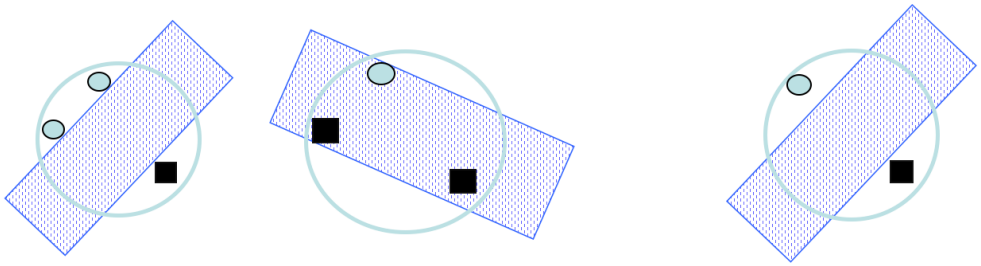
SAPIENZA UNIVERSITÀ DI ROMA

Hyperplanes with margin

Assuming D given, the VC dimension can change with ρ



For small values of ρ it is still possible to shatter 3 points



No more possible to shatter 3 points

still possible to shatter 2 points

SAPIENZA UNIVERSITÀ DI ROMA

Margin and VC confidence

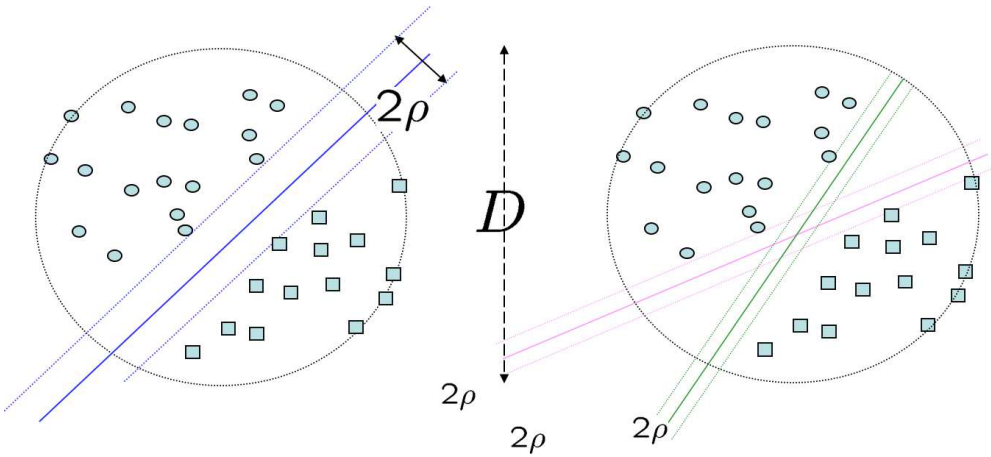
Theorem VC dimension of linear classifiers with margin

Consider the class of functions

$$\mathcal{F} = \{f : \Re^n \to \mathbb{R} : f(x) = w^T x + b, d(x; w, b) \geq \rho\}$$

then

$$h \leq \min\left\{\left\lceil \frac{D^2}{\rho^2} \right\rceil, n\right\} + 1 \leq n + 1$$



VC term decreases when the margin ρ increases

Linear SVM - Geometric viewpoint

Hyperplane with margin has a value of VC dimension h not less than the one of simple hyperplane $h = n + 1$ thus leading to a better value of the VC confidence term.

The idea in Linear Hard SVM is to "fix" the value of the empirical error to zero by imposing the separating condition

$$\begin{aligned} w^T x^i + b > 0 \quad & i : \ y^i = 1 \\ w^T x^j + b < 0 \quad & j : \ y^j = -1 \end{aligned}$$

and by maximizing the margin.

The separating conditions can be rewritten in compact form as

$$y^i (w^T x^i + b) > 0 \quad i = 1, \ldots, l$$

SAPIENZA UNIVERSITÀ DI ROMA

Margin

$$\mathcal{T} = \{(x^i, y^i) : x_i \in \mathbb{R}^n,\ y^i \in \{-1, 1\},\ i = 1, \ldots, l\},$$

The distance of a point x^i from a given hyperlane $w^t x + b = 0$ is

$$d(x^i; w, b) = \frac{|w^T x^i + b|}{\|w\|}$$

The margin is

$$\min_{i=1,\ldots,l} \left\{ \frac{|w^T x^i + b|}{\|w\|} \right\}$$

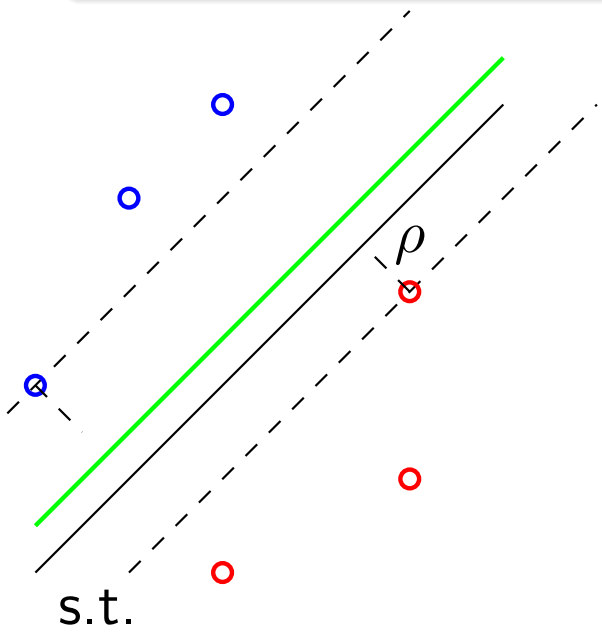
SAPIENZA UNIVERSITÀ DI ROMA

The maximum Margin pb

The maximum margin problem

$$\max_{w \in \Re^n, b \in \mathbb{R}} \quad \min_{i=1,\ldots,\ell} \left\{ \frac{|w^T x^i + b|}{\|w\|} \right\}$$

$$\text{s.t.} \quad y^i \left[w^T x^i + b \right] \geq 1 \quad i = 1, \ldots, \ell.$$



for each separating (w, b), there exists a $(\hat{w}, \hat{b})$ and points x^+ e x^- s.t.

$$\bar{w}^T x^+ + \bar{b} = 1$$
$$\bar{w}^T x^- + \bar{b} = -1$$

$$\rho(w, b) \leq \rho(\bar{w}, \bar{b}) = \frac{1}{\|\bar{w}\|}.$$

SAPIENZA UNIVERSITÀ DI ROMA

The maximum Margin pb

Further

- for each separating hyperplane (w, b) it holds hat

$$\rho(w, b) = \min_{i=1,\dots,l} \left\{ \frac{|w^T x^i + b|}{\|w\|} \right\} \geq \frac{1}{\|w\|}.$$

- $$\min \quad \frac{1}{\|w\|}$$

$$y^p \left[w^T x^p + b \right] - 1 \geq 0, \quad p = 1, \dots, l \ ,$$

Hard - Linear SVM

$$\min_{w \in \mathbb{R}^n, b \in \mathbb{R}} \quad \frac{1}{2}\|w\|^2$$

$$\text{s.t.} \quad y^i \left[w^T x^i + b \right] \geq 1, i = 1, \dots, l.$$

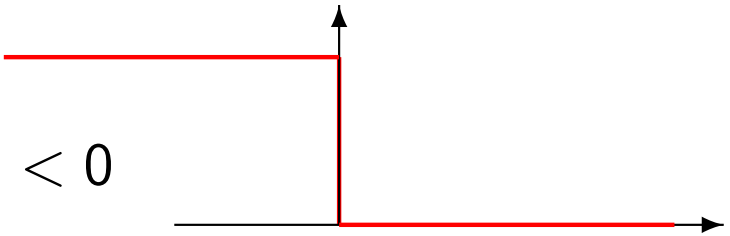
Hard Linear SVM optimization

$$\min_{w\in\mathbb{R}^n, b\in\mathbb{R}} \quad \frac{1}{2}\|w\|^2$$

$$\text{s.t.} \quad y^i\left[w^T x^i + b\right] - 1 \geq 0, i = 1, \ldots, l.$$

- it is a (loosely) convex quadratic problem
- it admits a **unique** solution (w^*, b^*)
- the classifier is given by $f(x^i; w, b) = \text{sgn}(w^T x^i + b)$
- we use the $0-1$ loss which counts the misclassifed samples

$$\ell(x; w^T x+b) = \begin{cases} 1 & \text{if } y^i(w^T x^i + b) < 0 \\ 0 & \text{otherwise} \end{cases}$$



What happens if the two sets are not linearly separable ?

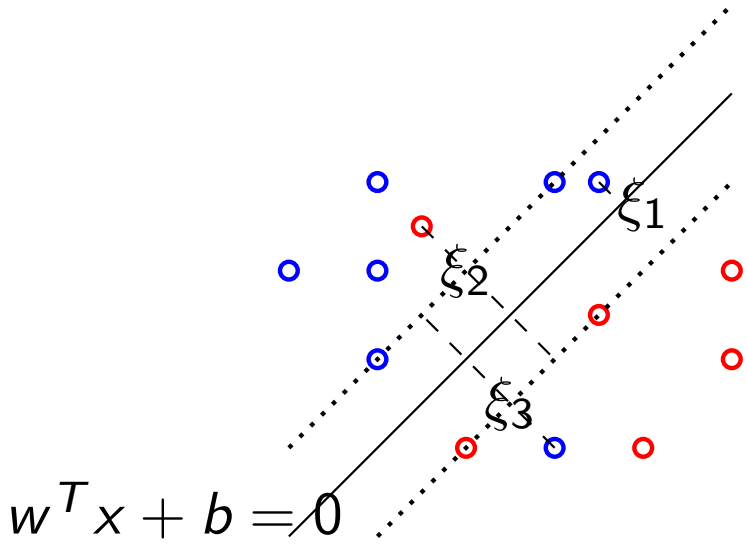
The feasible set is empty !

Allow violating constraints

Assume that the points are NOT linearly separable, but still we use linear classifiers

We add slack variables to allow violation of the constraints
$i = 1, \ldots, l$

$$y^i \left[w^T x^i + b \right] \geq 1 - \xi_i$$
$$\xi_i \geq 0$$



ξ_i measures how much a constraint is violated.

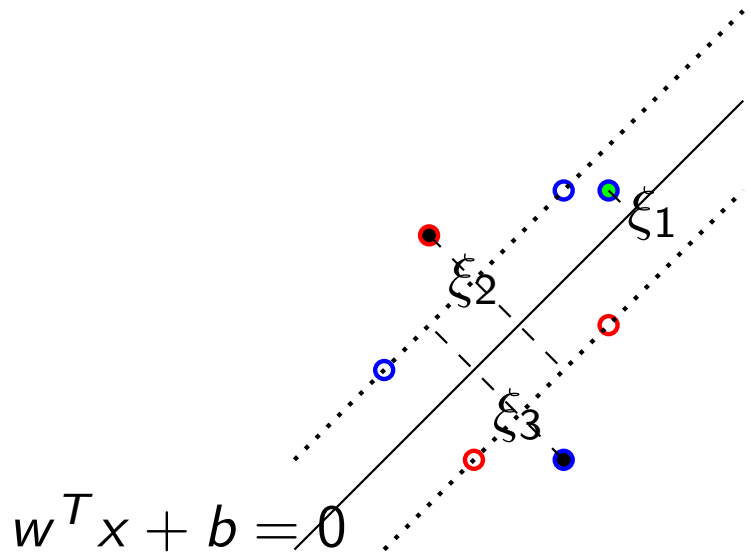
A feasible solution (w, b, ξ) always exist: choose large ξ !

SAPIENZA UNIVERSITÀ DI ROMA

Linear Soft-SVM

We also want to control the error on the training set (i.e. control R_{emp}).

- $0 \leq \xi_i < 1$ the point is correctly classified although it violates the constraint
- $\xi_i \geq 1$ the point is misclassified



$$\boxed{\text{sgn}(y^i(w^T x^i + b)) < 0 \text{ implies } \xi_i \geq 1.}$$

Therefore, we have an upper bound on the number of errors

$$\#errors \leq \sum_{i=1}^{l} \xi_i^p$$

SAPIENZA UNIVERSITÀ DI ROMA

Linear Soft-SVM

A penalization on the upper bound of misclassified points is added

$$\min_{w\in\mathbb{R}^n, b\in\mathbb{R}, \xi\in\mathbb{R}^l} \underbrace{\frac{1}{2}\|w\|^2}_{\text{max margin}} + \underbrace{\sum_{i=1}^{l} C_i \xi_i^p}_{\text{min empirical error}}$$

$$\text{s.t.} \quad y^i \left[w^T x^i + b\right] \geq 1 - \xi_i, i = 1, \ldots, l$$
$$\xi_i \geq 0$$

where $C_i > 0$ are user-dependent **hyper parameters** to assess the training error.

- it is a (loosely) convex problem in (w, b, ξ)
- depending on value of $p \in [1, \infty)$ it can admits a **unique** classifiers (w^*, b^*)
- the classifier is given by $f(x^i; w, b) = \text{sgn}(w^{*T} x^i + b^*)$

Special cases

$p = 1$ and $C_i = C$: C-SVM or L_1-SVM

$$\min_{w \in \mathbb{R}^n, b \in \mathbb{R}, \xi \in \mathbb{R}^l} \quad \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{l} \xi_i$$

$$\text{s.t.} \quad y^i \left[w^T x^i + b \right] \geq 1 - \xi_i, i = 1, \ldots, l$$
$$\xi_i \geq 0$$

$p = 2$ and $C_i = C$: L_2-SVM

$$\min_{w \in \mathbb{R}^n, b \in \mathbb{R}, \xi \in \mathbb{R}^l} \quad \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{l} \xi_i^2$$

$$\text{s.t.} \quad y^i \left[w^T x^i + b \right] \geq 1 - \xi_i, i = 1, \ldots, l$$
$$\xi_i \geq 0$$

UNIVERSITÀ DI ROMA

- it is quadratic (loosely) convex problem

The homogeneous case - UNBIASED SVM

- If (# of features) is small, b may be important, otherwise not.
- Adding fictitious feature so that $x \to (x, 1)$ and $w \to (w,\ b)$, the constraints reads as $w^T x \geq 1 - \xi$ BUT the objective function is biased $\|(w,\ b)\|^2 = \|w\|^2 + b^2$
- posing $b = 0$ is not equivalent to homogenization of the constraints
- Simplification in the model and hence in algorithms

Unbiased SVM $b = 0$

$$\min_{w \in \Re^n, \xi \in \Re^l} \quad \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{l} \xi_i$$

$$\text{s.t.} \quad y^i w^T x^i \geq 1 - \xi_i, i = 1, \ldots, l$$
$$\xi_i \geq 0$$

Optimization recap

It is a constrained problem

$$\begin{aligned} \min \quad & f(w,\xi) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{l} C_i \xi_i^p \\ & (y^i x^i) w^T + y^i b + \xi_i \geq 1, i = 1, \ldots, l \\ & \xi \geq 0 \end{aligned}$$

- $f(w,\xi)$ is continuously differentiable $p \geq 1$ and convex
- linear constraints (feasible set polyhedron)

$$A \cdot (w, b, \xi) \geq g$$

- KKT are necessary optimality conditions for global optimality

$$\nabla_{w,b,\xi} f - A^T \alpha = 0 \qquad \text{stationarity}$$
$$\alpha \geq 0, \ \alpha^T (g - A \cdot (w, b, \xi)) = 0 \qquad \text{non negat + complementarity}$$

KKT conditions

Define the Lagrangian

$$L = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{I}\left\{C_i\xi_i^p + \lambda_i\left[1 - \xi_i - y^i\left(w^T x^i + b\right)\right] - \mu_i\xi_i\right\}$$

the global solution (w^*, b^*, ξ^*) satisfies with some $\alpha^* = (\lambda^*, \mu^*)$

$$\nabla_w L(w, b, \xi; \lambda) = w^\star - \sum_{i=1}^{I}\lambda_i^\star y^i x^i = 0$$

$$\nabla_b L(w, b, \xi; \lambda) = -\sum_{i=1}^{I}\lambda_i^\star y^i = 0$$

$$\nabla_\xi L(w, b, \xi; \lambda) = 0$$

$$\mu_i \geq 0, \qquad \xi_i\mu_i = 0,$$

$$\lambda_i \geq 0, \qquad \lambda_i\left[1 - \xi_i - y^i\left(w^T x^i + b\right)\right] = 0$$

SAPIENZA UNIVERSITÀ DI ROMA

Find the w^*

Whatever the penalization used $p \in [1, \infty)$ the expression of the optimal w^* is obtained as

$$w^\star = \sum_{i=1}^{l} \lambda_i^\star y^i x^i,$$

being $\lambda_i^* \geq 0$.
The classifier is

$$f(x; w, b) = \sum_{i=1}^{l} \lambda_i^\star y^i x^T x^i + b^*,$$

Those sample i such that $\lambda_i^* > 0$ are called *Support Vectors*.

SAPIENZA UNIVERSITÀ DI ROMA

C-SVM or L_1-SVM

Consider $p = 1$ then

$$\nabla_\xi L(w, b, \xi; \lambda) = 0 \quad \rightarrow \quad C - \lambda = \mu$$

$$\mu_i \geq 0, \qquad \xi_i \mu_i = 0,$$

$$\lambda_i \geq 0,$$

reads as

$$0 \leq \lambda \leq C$$

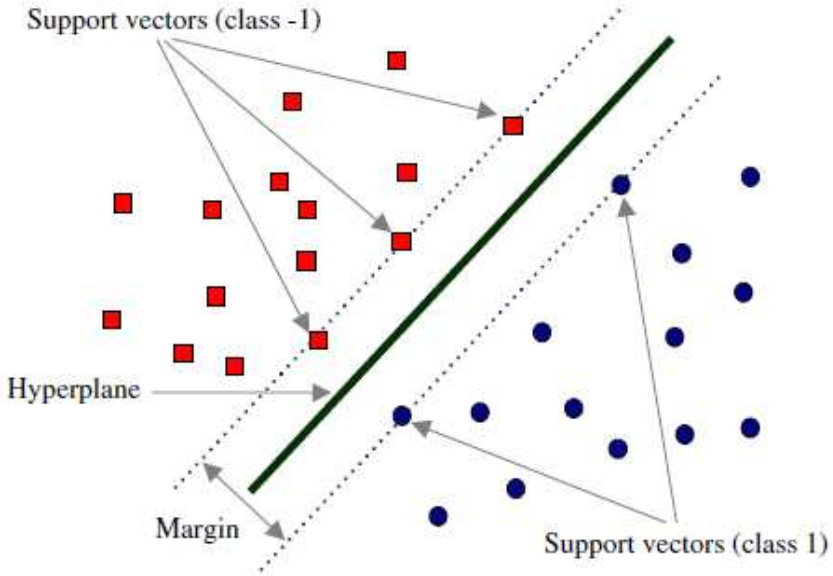
Further when $0 < \lambda_i < C$ we get $\xi_i = 0$ and

$$1 - y^i \left(w^T x^i + b \right) = 0$$

Support Vectors & active constraints

From the KKT conditions (complementarity ones) we have that $0 < \lambda_i < C$ corresponds to *active constraints* and $\xi_i = 0$ so that

$$y^i \left(w^T x^i + b \right) = 1$$



Remind $f(x; w, b) = \sum_{i=1}^{l} \lambda_i^\star y^i x^T x^i + b^*$, so $\lambda_i > 0$ (Support Vectors) includes $\lambda_i = C$ too (Bounded Support Vectors).
If we knew the SV (aka we knew the active constraints at the optimum) we could eliminate the useless samples and reduce the number of constraints

The C-SVM optimization problem

Quadratic primal problem

$$\min_{z \in \mathbb{R}^d} \quad \frac{1}{2} z^T Q z + q^T z$$
$$Az \geq g$$

with $z = (w, b, \xi)$, Q symmetric positive semidefinite matrix; A representing the matrix of the linear constraints $m \times d$

- (loosely) convex quadratic objective function
- KKT conditions characterize the global solutions
- Strong duality holds

SAPIENZA UNIVERSITÀ DI ROMA

Wolfe dual

Let us consider the convex programming problem

$$\begin{aligned} \min_{z} \quad & f(z) \\ & Az \geq g, \end{aligned} \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex, continuously differentiable function
Introducing the Lagrangian function

$$L(z, \lambda) = f(z) + \lambda^T (g - Az)$$

Wolfe's dual of (1) is

$$\begin{aligned} \max_{z,\lambda} \quad & L(z, \lambda) \\ & \nabla_z L(z, \lambda) = 0 \\ & \lambda \geq 0. \end{aligned} \tag{2}$$

Theorem

If problem (1) admits a solution $z^\star$, then there exists a vector of Lagrange multipliers $\lambda^\star$ such that $(z^\star, \lambda^\star)$ is a solution of (2).

Duality

Primal pb

$$\min_{z} f(z)$$
$$Az \geq g,$$

Wolfe dual pb

$$\max_{z,\lambda} \quad f(z) + \lambda^T(g - Az)$$
$$\nabla f(z) - A^T\lambda = 0$$
$$\lambda \geq 0.$$

Optimality result under convexity of f

Primal opt

z^* optimal solution primal
λ^* KKT multiplier

$\Longrightarrow$

Dual Opt

z^*, λ^* optimal for the Wolfe dual

Duality: Quadratic case

In the quadratic case stronger results hold

Primal pb

$$\min_{z} \quad \frac{1}{2} z^T Q z + q^T z \qquad Az \geq g$$

Wolfe dual pb

$$\max_{x,\lambda} \quad f(z) + \lambda^T (g - Az) \qquad Qz + q - A^T \lambda = 0 \qquad \lambda \geq 0.$$

By simple manipulation we get

$$\min_{x,\lambda} \quad \frac{1}{2} z^T Q z - \lambda^T g \qquad Qz + q - A^T \lambda = 0 \qquad \lambda \geq 0.$$

It is again quadratic and convex

SAPIENZA UNIVERSITÀ DI ROMA

Duality QP case

Strong duality

Assume $Q \succeq 0$. Let $(\bar{z}, \bar{\lambda})$ be a solution of Wolfe's dual. Then, there exists a vector $z^\star$ (not necessarily equal to $\bar{z}$) such that

(i) $Q(z^* - \bar{z}) = 0$;

(ii) z^* is a global minimum of primal QP with associated multipliers $\bar{\lambda}$.

Primal opt

z^* optimal solution primal
λ^* KKT multiplier

$\Longleftrightarrow$

Dual Opt

$\bar{z}, \lambda^*$ optimal for the Wolfe dual

SAPIENZA UNIVERSITÀ DI ROMA

Dual of C-SVM QP problem

$$\min_{w\in\Re^n, b\in\Re, \xi\in\Re^l} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i$$

$$\text{s.t.} \quad y^i\left[w^T x^i + b\right] \geq 1-\xi_i, i = 1,\ldots,l$$
$$\xi_i \geq 0$$

The primal variables are $z = (w, b, \xi)$, and the condition $\nabla_w L(z,\lambda) = 0$, $\nabla_b L(z,\lambda) = 0$ gives the constraints

$$w = \sum_{i=1}^{l}\lambda_i y^i x^i \qquad \sum_{i=1}^{l}\lambda_i y^i = 0.$$

SAPIENZA UNIVERSITÀ DI ROMA

Dual of SVM training

With some manipulations variables ξ can be eliminated and the Wolfe's dual is

$$\begin{aligned} \min_{\lambda} \quad & \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y^i y^j (x^i)^T x^j \lambda_i \lambda_j - \sum_{i=1}^{l} \lambda_i \\ s.t. \quad & \sum_{i=1}^{l} \lambda_i y^i = 0 \\ & 0 \leq \lambda_i \leq C, \quad i = 1, \ldots, l \end{aligned}$$

The objective function is of the form $\lambda^T Q \lambda$ with $Q = X^T X \succeq 0$ and $X = \left[y^1 x^1, \ldots, y^l x^l\right]$

- $Q = \{q_{ij}\}$ with $q_{ij} = y^i y^j x^{i\,T} x^j$
- It is a **convex quadratic programming problem**
- one linear constraints + box constraints
- dense $l \times l$ hessian matrix Q

SAPIENZA UNIVERSITÀ DI ROMA

Homogeneous case

If $b = 0$, then the condition $\nabla_b L = y^T \lambda = 0$ disappears and the dual simplifies in

$$\begin{aligned} \min_{\lambda} \quad & \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y^i y^j (x^i)^T x^j \lambda_i \lambda_j - \sum_{i=1}^{l} \lambda_i \\ s.t. \quad & 0 \leq \lambda_i \leq C, \quad i = 1, \ldots, l \end{aligned}$$

- It is a **convex quadratic programming problem**
- only box constraints
- dense $l \times l$ hessian matrix Q

SAPIENZA UNIVERSITÀ DI ROMA

Dual solution of C-SVM

By strong duality we can solve the dual and find λ^*

Thus, by considering any support vector x^i such that $0 < \lambda_i^\star < C$, we can get b^* from

$$y^i \left({w^\star}^T x^i + b^\star \right) - 1 = 0, \tag{3}$$

The decision function of a linear SVM is

$$f(x) = \operatorname{sgn}\left({w^*}^T x + b^\star \right) = \operatorname{sgn}\left(\sum_{i=1}^{I} \lambda_i^\star y^i x^T x^i + b^\star \right).$$

Recap: formulations

Primal L_1-SVM

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \\ & y^i\left[w^T x^i + b\right] \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned}$$

Dual L_1-SVM

$$\begin{aligned} \min_{\lambda\in\mathbb{R}^l} \quad & \frac{1}{2}\lambda^T Q\lambda - e^T\lambda \\ \text{s.t.} \quad & \lambda^T y = 0 \\ & 0 \leq \lambda \leq C \end{aligned}$$

Primal L_1- unbiased SVM

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \\ & y^i w^T x^i \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned}$$

Dual L_1- unbiased SVM

$$\begin{aligned} \min_{\lambda\in\mathbb{R}^l} \quad & \frac{1}{2}\lambda^T Q\lambda - e^T\lambda \\ \text{s.t.} \quad & 0 \leq \lambda \leq C \end{aligned}$$

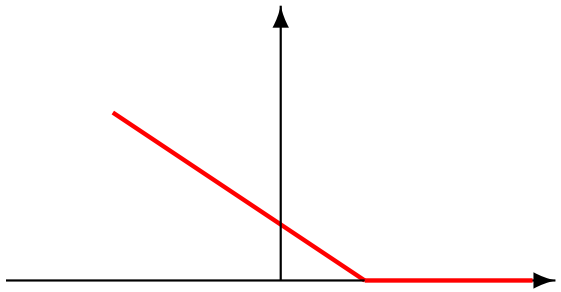
Hinge Loss formulation

The constraints $y^i \left(w^T x^i + b\right) \geq 1 - \xi_i \qquad \xi_i \geq 0$ are equivalently written as $\xi_i \geq \max\left\{0, 1 - y^i \left(w^T x^i + b\right)\right\}$

The L1-SVM unconstrained problem

hinge loss

$$\ell(x; w^T x + b) = \max\{0, 1 - y^i (w^T x^i + b)\}$$



$$\min_{w \in \mathbb{R}^n, b \in \mathbb{R}} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{l} \max\left\{0, 1 - y^i \left(w^T x^i + b\right)\right\}$$

$$\min_{w \in \mathbb{R}^n} \frac{1}{2C}\|w\|^2 + \sum_{i=1}^{l} \max\left\{0, 1 - y^i w^T x^i\right\}$$

SAPIENZA UNIVERSITÀ DI ROMA

Unconstrained L_2-SVM (quadratic loss)

$$\min_{w\in\mathbb{R}^n, b\in\mathbb{R}, \xi\in\mathbb{R}^l} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i^2$$

$$\text{s.t.} \quad y^i\left[w^T x^i + b\right] \geq 1 - \xi_i, i = 1, \ldots, l$$
$$\xi_i \geq 0$$

The L_2-SVM unconstrained problem

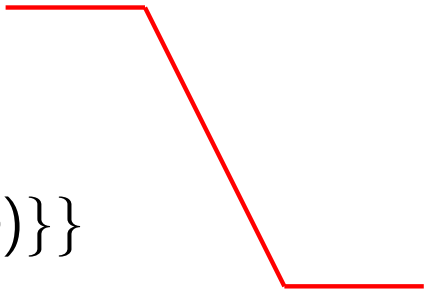
$$\min_{w\in\mathbb{R}^n, b\in\mathbb{R}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\max\left\{0, 1 - y^i\left(w^T x^i + b\right)\right\}^2$$

- convex smooth but not twice continuously diffentiable

Ramp Loss formulation

It wants to account the fact that points within the margin are correctly classified.

ramp loss



$$\ell(x; w^T x + b) = \min\{1, \max\{0, 1 - y^i(w^T x^i + b)\}\}$$

Consider an integer formulation with additional $z_i \in \{0, 1\}$ used to model the constraints

$$y^i \left[w^T x^i + b \right] \geq 1 - \xi_i, \qquad \text{if } z_i = 0$$

$$\begin{aligned} \min_{w \in \mathbb{R}^n, b \in \mathbb{R}, \xi \in \mathbb{R}^l} \quad & \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{l} (\xi_i + 2z_i) \\ \text{s.t.} \quad & y^i \left[w^T x^i + b \right] \geq 1 - \xi_i, \quad \text{if } z_i = 0 \qquad i = 1, \ldots, l \\ & 0 \leq \xi_i \leq 2 \end{aligned}$$

References

See the surveys and references therein


E. Carrizosa and D. R. Morales.
Supervised classification and mathematical optimization.
Computers & Operations Research, 40(1):150–165, 2013.


V. Piccialli and M. Sciandrone.
Nonlinear optimization and support vector machines.
4OR, 16(2):111–149, 2018.

SAPIENZA UNIVERSITÀ DI ROMA